Published in final edited form as:

Phys Biol. ; 14(2): 025002. doi:10.1088/1478-3975/aa64a4.

An evolution-based strategy for engineering allosteric regulation

David Pincus^{1,*,#}, Orna Resnekov^{2,*,#}, and Kimberly A. Reynolds^{3,4,*,#}

¹Whitehead Institute for Biomedical Research, Cambridge, MA 02142

²1255 La Canada Road, Hillsborough, CA 94010

³The Green Center for Systems Biology, University of Texas Southwestern Medical Center, Dallas, TX 75390

⁴Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX 75390

Abstract

Allosteric regulation provides a way to control protein activity at the time scale of milliseconds to seconds inside the cell. An ability to engineer synthetic allosteric systems would be of practical utility for the development of novel biosensors, creation of synthetic cell signaling pathways, and design of small molecule pharmaceuticals with regulatory impact. To this end, we outline a general approach – termed Rational Engineering of Allostery at Conserved Hotspots (REACH) – to introduce novel regulation into a protein of interest by exploiting latent allostery that has been hard-wired by evolution into its structure. REACH entails the use of statistical coupling analysis (SCA) to identify “allosteric hotspots” on protein surfaces, the development and implementation of experimental assays to test hotspots for functionality, and a toolkit of allosteric modulators to impinge on endogenous cellular circuitry. REACH can be broadly applied to rewire cellular processes to respond to novel inputs.

Motivation

Allosteric regulation enables natural proteins to integrate one or more external signals to determine a functional response (Cui and Karplus, 2008). In natural systems, allosteric control is critical for regulating metabolism, maintaining homeostasis, and signaling in response to environmental cues. Because it is often desirable to rewire or modulate the function of existing cellular systems, a capacity to rationally engineer custom regulation would be broadly useful. Allosteric regulation has several properties that are attractive from this viewpoint: it is reversible, fast, and can be mediated by a variety of effector signals including small molecules, light, voltage, peptides, and nucleic acids. Practically, a capacity to engineer synthetic allostery would provide a route to design allosteric drugs with improved specificity, create reagents for probing cell signaling (e.g. optogenetic tools), construct environmental biosensors, understand mutations associated with disease, and test mechanisms of host-pathogen interactions.

[#]Correspondence: pincus@wi.mit.edu; orna.resnekov@gmail.com; kimberly.reynolds@utsouthwestern.edu.

* All authors contributed equally and are listed alphabetically.

An emerging principle for the engineering of synthetic regulation comes from the finding of allosteric “hotspots” – protein surface patches that participate in a network of physically contiguous, co-evolving amino acids coupled to protein function (Reynolds et al., 2011). The idea is that these positions are privileged sites for the engineering and evolution of new allosteric control. In this tutorial, we describe a general approach for installing regulation at these positions termed *Rational Engineering of Allostery at Conserved Hotspots (REACH)*. We begin by briefly summarizing the main findings underlying this approach, and then outline a roadmap for engineering allostery in a focused and directed manner. Our goal is to provide a springboard for further usage, testing, and development of this approach by the scientific community.

Allosteric Regulation – a short introduction

The word allostery is derived from two Greek roots: *allos* meaning “other”, and *stereos* which means “structure” or “solid”. The phrase “allosteric control” (introduced by Monod, Changeux and Jacob in 1963; (Monod et al., 1963)) is intended to convey the reversible conformational change a protein can undergo in response to specific input signals, called effectors. The assertions of Monod and colleagues were that: 1) the effector binds in a site distinct from the active site, 2) the “input” effector signal has no chemical relationship to the output activity, and 3) that allostery is mediated by changes in protein conformational state (Monod et al., 1963). The main concept is that effector binding modulates the conformational equilibrium of the protein and thus changes protein activity. As a result, information can be transmitted from a distal allosteric site to the active site (Figure 1A, B). While the original work of Monod and colleagues focused on small molecule metabolites as effectors, regulation can more generally be mediated by a diversity of signals including post-translational modification and interaction with other protein domains.

A large body of prior work has investigated the physical mechanism of allosteric regulation, with a central finding that allostery involves the cooperative action of multiple, spatially distributed amino acids (Luque et al., 2002). Because an allosteric surface does not provide a useful (and thus evolutionarily selectable) function until it is coupled to protein activity, allostery would seem difficult to evolve through a process of stepwise variation and selection. Nevertheless, experimental studies have shown that allostery is present in many different protein families, that allosteric control of protein activity can be manifested in many different ways (effector molecules), and that homologous proteins can exhibit different allosteric control mechanisms (Kuriyan and Eisenberg, 2007). Further, several groups have described the construction of new allosteric switches using empirical screening – in many cases, a random process of domain insertion and screening for regulation is sufficient to arrive at new allosteric activity (see (Nadler et al., 2016; Ostermeier, 2005) for examples). Considered together, both the ubiquity of allostery in natural proteins and the success of these naïve engineering approaches suggest that proteins have the intrinsic capacity for regulation at a diversity of sites.

Recent work from Wendell Lim and colleagues supports this idea. In budding yeast (*S. cerevisiae*) two allosteric activators (effectors) within the scaffold protein Ste5 regulate the MAP kinase Fus3 (Coyle et al., 2013). Surprisingly, when tested *in vitro*, these activators

also regulate evolutionarily distant MAP kinases – kinases present in fungal species that diverged prior to the evolution of a Ste5 scaffold protein. This study illustrates how specific effectors may be used to reveal latent allosteric control potential in a protein family, and is consistent with the idea that proteins contain a hard-wired allosteric conduit. These results intriguingly suggest that we should be able to leverage the natural design of proteins to facilitate the rational engineering of allosteric regulation. If one could identify the interactions between amino acid residues responsible for latent allostery in a protein – the “paths” that connect functional sites to distal surfaces – they could be exploited to engineer artificial input sensors at specific sites.

Evolution-based design of regulation

One approach for mapping the pattern of interactions between residues is the Statistical Coupling Analysis, or SCA, developed by Rama Ranganathan and colleagues. The basic premise behind SCA is that functionally interacting amino acid positions should experience a joint evolutionary constraint. As a result, a pair of functionally coupled residues should show correlated changes in amino acid identity across homologous sequences. In essence, what SCA does is compute the evolutionary correlation (i.e., co-evolution) between all possible pairs of amino acids in a protein. By analyzing these correlations, one can make a model for the pattern of interactions between amino acid residues. Notably, SCA is one of several approaches for analyzing the pattern of amino acid sequence coevolution in proteins, some of which differ substantially in overall goal and technical implementation (de Juan et al., 2013). Our purpose is not to provide a review or comparison of these methods, but rather to provide a tutorial using one method for coevolutionary analysis that was experimentally shown to identify allosteric hotspots.

The central finding from SCA is that many protein families contain groups of co-evolving amino acids termed *sectors* (Halabi et al., 2009; Lockless and Ranganathan, 1999). Empirically, the sectors studied to date show several properties. They are sparse (comprising only 20–30% of the amino acid positions), physically contiguous in the tertiary structure, and connect the active site to specific, distant *sector-associated surface sites* on the protein (Figure 1C). In a number of protein families, sectors have been observed to connect known allosteric sites to the active site (Ferguson et al., 2007; Hatley et al., 2003; Suel et al., 2003). In Figure 2A–C, we show several examples from recent work. The sector can connect small molecule regulatory sites to the active site, provide the basis for regulation between domains, and link protein-protein interfaces. In all cases, the salient feature is that the sector provides a cooperative network linking functional surfaces.

The approach delineated below, Rational Engineering of Allostery at Conserved Hotspots (REACH), uses sector-connected surface residues for the installation of allosteric control in a given protein (Figure 1C) and has been successfully implemented in a limited number of cases in the literature. Using domain insertion scanning, Reynolds et al. showed that the metabolic enzyme dihydrofolate reductase can be engineered to become regulated *in vivo* via a light-sensitive signaling module and that sector-associated surface sites are the statistically preferred locations for the regulation to occur (Figure 2D) (Lee et al., 2008; Reynolds et al., 2011). A separate study used sector-associated surface sites to introduce

regulation by a small molecule. Small molecule allosteric control modifiers may also be a useful alternative to active site inhibitors, which have sometimes exhibited problems with toxicity and/or off-target effects. Cathepsin K is a key target for the treatment of osteoporosis – Novinec et al., used SCA to predict sectors and surface sector-connected residues, AutoLigand to predict potential binding sites on the surface of the protein for compound libraries, and through screening, successfully identified the first low molecular weight allosteric modifier of cathepsin K (Figure 2E) (Novinec et al., 2014).

In this tutorial we illustrate: (1) how to use the SCA to predict candidate allosteric sites (Figure 3) and (2) how to introduce and test new regulation at these putative allosteric sites in a protein design and testing pipeline. We look forward to testing and implementing REACH further, and our goal is to make REACH accessible to the broader scientific community.

Roadmap

In the body of this tutorial we present practical steps for implementing REACH in the system of your choice. As stated above, REACH is an emerging strategy for installing allosteric regulation at specific sector-associated surface residues. As such, we provide general practical guidelines for implementing this approach rather than a procedurally detailed protocol. The most critical steps are: 1) conducting SCA with an adequately large and diverse sequence alignment and 2) choice of an appropriate functional assay for measuring allosteric activity. Here is an overview of REACH:

1. Perform SCA on your protein family of interest.
2. Identify putative “allosteric hotspots”: sector-connected surface residues on the protein surface.
3. Choose or develop a functional assay for regulation given a specific member of your protein family of interest.
4. Screen sector-connected surface residues for functionality. Do mutations in putative allosteric sites constitutively alter activity of the protein? The goal of REACH is to focus the search for putative allosteric sites down to a subset of sector-connected surface residues.
5. Introduce modulators at functional surfaces to regulate activity of your protein. In many cases, step 4 (screening) and step 5 (engineering of regulation) might be naturally combined.
6. Optimize regulation
7. Connect the allosteric switch to endogenous cellular circuitry.

1 & 2. Performing the Statistical Coupling Analysis (SCA) and Identification of Sector Connected Surface Sites

The mathematics of the SCA calculation are described in their entirety elsewhere (Halabi et al., 2009; Reynolds et al., 2013; Rivoire et al., 2016); here we briefly review the principles

behind the approach and the main steps to execute the analysis. Because we are interested not only in defining the sector, but in choosing specific sites for engineering allosteric regulation, we also describe the steps needed to identify sector-connected surface sites. Figure 3 provides a summary of the process.

The central challenge behind any statistical analysis of amino acid correlations is how to distinguish functionally significant correlations from those caused by limited sequence sampling (*statistical noise*) or historical relationships between the sequences in the alignment (*phylogenetic noise*). The defining feature of SCA is that it addresses these problems by computing a *weighted* covariance matrix, in which the correlations between amino acids are weighted by their evolutionary conservation. The origin of the weighting factor is explained and derived elsewhere (Rivoire et al., 2016), but the intuition is that correlations between conserved positions are more likely to be relevant to the function of the protein. The resulting SCA matrix is a square, symmetric matrix that describes the statistical coupling between all pairs of amino acid positions in the protein. Once the conservation-weighted correlation matrix is constructed, standard tools from linear algebra (principal component analysis and independent component analysis) are used to identify groups of evolutionarily correlated amino acid positions. These groups of positions provide the basis for defining sectors.

An open-source python toolbox for computing the SCA (PySCA) is freely available through the online project hosting website github (<https://github.com/reynoldsk/PySCA>). The code implementing the calculations is contained in a python module (`scaTools.py`), and is executed by running a series of three scripts (`scaProcessMSA.py`, `scaCore.py`, and `scaSectorID.py`). After the calculations are complete, the results can be loaded and analyzed using a jupyter notebook, a web-based tool that lets you interactively plot and analyze your data (<http://jupyter.org/index.html>). These notebooks are designed to be easily customizable and readily shared with your colleagues and collaborators. Importantly, the PySCA distribution comes with several interactive tutorials that walk through the analysis for three different protein families: 1) the Ras-like small G-proteins 2) the metabolic enzyme DHFR and 3) the antibiotic resistance enzyme β -lactamase. The G-protein example is particularly relevant for our purposes; it provides a nice illustration of how PySCA can be used to identify residues that are known to be important for allosteric regulation (http://reynoldsk.github.io/pySCA/SCA_G.html).

- **Prerequisites and preliminary steps**

1. To run SCA, you'll need some familiarity with python and the UNIX command line. A web-based app for non-coders is under development, but the current version requires some basic programming knowledge.
2. A protein family with many available homologous sequences (hundreds or more) and a high resolution crystal structure (PDB file) for at least one representative family member.
3. Download and install the toolbox. Installation and usage instructions may be found here: <http://reynoldsk.github.io/pySCA/>.

4. Become familiar with PySCA by running the tutorials (<http://reynoldsk.github.io/pySCA/tutorials.html>). Modification of the tutorials is also a useful starting point for developing your own code.

- **Alignment Construction and Annotation**

1. Collect homologous sequences for your protein of interest. In general, more is better. One good starting place for obtaining sequences is the PFAM database (<http://pfam.xfam.org/>) (Finn et al., 2016). Another option is to identify sequences using several iterations of PSI-BLAST (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) (Altschul et al., 1997). In some cases, bioinformaticians studying your protein family may have previously published curated alignments that can be downloaded (in which case skip to step 4 in this subsection).
2. Once sequences have been obtained, they should be filtered for length. This step removes partial sequences and/or sequences that would potentially introduce a large number of gaps in the alignment. Typically we retain sequences that are within 50 amino acids of the average sequence length.
3. Align the sequences to produce a fasta format alignment. We often use the tools MUSCLE (<http://www.ebi.ac.uk/Tools/msa/muscle/>) and/or Promals3D (<http://prodata.swmed.edu/promals3d/promals3d.php>) for this step (Edgar, 2004; Pei and Grishin, 2014).
4. If desired, the sequences can be annotated with phylogenetic information using the script `annotateMSA.py` (distributed as part of the PySCA toolbox). Though not essential for the determination of sector positions, this step is sometimes useful for interpreting how the sequence of sector positions diverges between different subfamilies or clades.

- **Evaluation of alignment statistics**

Once the alignment is constructed, you will need to decide if it is suitable for analysis. An alignment that is too small, or does not contain enough sequence diversity will not give informative results. There are three main indicators calculated by the script `scaProcessMSA.py`.

1. The effective number of sequences (M_{eff}). This is the number of sequences retained after filtering the alignment for highly gapped sequences and applying sequence weights. The sequence weights are chosen to down-weight high identity sequences in the alignment, and act as a correction for biased sampling. For SCA, you will need $M_{eff} > 100$.
2. The variable N_{pos} is the number of amino acid positions in the alignment after filtering out highly gapped positions. As a rough estimate, N_{pos} should be at least 70% the size of the protein of interest.

N_{pos} much less than the size of the protein indicates that the starting alignment is overly gapped, or potentially missing a domain.

3. The distribution of pairwise sequence identities, as calculated between all pairs of sequences in the alignment. This provides a good indicator of diversity in the alignment. Typical alignments for SCA have a distribution of pairwise sequence identities with a peak near 20–40%.

- **Calculation of conservation and coevolution (scaCore.py)**

Given a suitable sequence alignment, the next step is computing the conservation of each amino acid position, and the co-evolution between all pairs of positions. This is the heart of SCA, and understanding this step is essential. Thorough descriptions of the procedure are required reading (Halabi et al., 2009; Reynolds et al., 2013; Rivoire et al., 2016). Computing the co-evolution of all pairs of amino acids is accomplished by the script `scaCore.py`, and the end result of this step is the SCA matrix. For all alignments analyzed to date, the SCA matrix is sparse – roughly 20–30% of positions are highly coupled, while the remaining positions evolve relatively independently.

- **Analysis of the SCA matrix and sector identification**

The SCA matrix is a model for the pattern of coupling between all pairs of positions in the protein. This pattern can now be analyzed to identify groups of amino acids that are significantly correlated to one another. This step is accomplished by the scripts `scaCore.py` and `scaSectorID.py`. The main idea is to use a combination of principal component analysis (PCA) and independent component analysis (ICA) to find groups of co-evolving residues. These positions are combined to define a sector. When mapped to the protein tertiary structure, the expectation is that the sector positions will form a physically contiguous network. Sector positions that appear scattered throughout the tertiary structure and lack physical contiguity can be one indication of an underlying problem with the alignment.

- **Defining the protein surface**

In this step, we identify all of the solvent accessible positions on the protein, with the rationale that these are suitable sites for introducing regulation; the allosteric hotspots occur where the sector intersects with the surface. Several high-quality tools exist for calculating solvent accessible surface area given a protein structure. In past work, we have used the Maximal Speed Molecular Surface (MSMS) package with a probe size of 1.4 Å, and a relative solvent accessibility (RSA) cutoff of 0.25 (this is the fraction of the amino acid side chain surface area that is solvent exposed) (Sanner et al., 1996).

- **Computing sector connected surface sites**

In the last step, the list of solvent accessible surface positions (obtained in the previous step) is compared to the set of sector positions to define a list of sector-connected surface sites. Whether or not two positions are connected depends on

their location in the protein tertiary structure, not the primary amino acid sequence. A sector contact is defined whenever any atom of the sector residue occurs within 4 Å of the backbone atoms (N, C α , C, or O) of the surface accessible position. This calculation can readily be made using python code in the PySCA jupyter notebook, or in the PyMOL molecular graphics package (Schrodinger, 2015).

These sector connected surface sites now become the set of candidate positions to screen for allosteric function—Screening of at least a few sites is a necessary step because we do not anticipate that every identified sector-connected site will result in experimentally demonstrated allosteric control in a particular protein. Remember, SCA uses the entire alignment of >100 effective sequences to define the sectors, and as such sectors are an ensemble property of the family of proteins that can be more or less penetrant in any given family member. The prediction is that sector-connected surface sites have a higher probability or likelihood of being regulatable than other surface sites. A general question that deserves further exploration is the extent to which particular allosteric sites are conserved (or diverge) among homologous domains (Tullman et al., 2016).

In our limited experience about 40% of the identified sector-connected sites prove to be functional, regulatable sites in any individual protein. This proportion will depend not only on the protein family, but the nature of the regulation being introduced. Transmitting an allosteric signal depends on the efficiency of the perturbation at the surface site. For example, in the case of engineering allostery by domain insertion, the length or the amino acid composition of the inter-domain linker might influence how well conformational change is transmitted between domains.

3. Establish an Experimental System

Determining which of the sector-connected surface residues identified by SCA are coupled to protein activity requires robust, quantitative assays that either directly measure or reliably report on protein function. There are three key considerations for this step. First, the throughput of the assay is important – one would ideally like to screen multiple sites, and thus relatively high-throughput assays are more desirable. Second, the allosteric activity obtained in the initial round of engineering may be low, so sensitivity to small changes in protein activity is critical. For example, in the case of engineering DHFR regulation by insertion of a LOV2 domain, the largest allosteric effects obtained were approximately two-fold changes in DHFR activity – an effect size that is barely outside the measurement error of many *in vitro* biochemical enzyme assays (Reynolds et al., 2011). Lastly, because SCA is based on sequence analysis (and does not provide mechanistic insight into *how* the regulation is accomplished), it cannot predict the directionality of the allosteric effect. Thus, you may wish to design a screen that is sensitive to both up- and down- regulation of protein activity.

- First, choose a representative of your protein family that is experimentally tractable. The best models will have a crystal structure of your specific protein to map the sector(s) precisely, but at a minimum a crystal structure of a homolog is

needed for REACH. The structure is necessary to define which sites on your protein are predicted to be sector-connected and surface accessible (see above).

- Next, choose an assay. Assays broadly fall into two categories: biochemical assays and cell-based assays.
- Biochemical assays (Table 1) allow direct measurement of the activity of isolated proteins and can provide valuable insight into structure-function relationships. However, biochemical assays tend to be low throughput since they require laborious protein purification.
- Cell-based assays (Table 2) provide a functional readout of a protein's activity, allow high-throughput mutant screening and can establish platforms for subsequent rewiring applications. In many cases, it is advantageous – or only possible – to measure a protein's activity in its native context in the cell.
- Cell-based assays can be performed in any organism or cell type, but the choice of model system is critical to success. General requirements include that the cells perform the biology you are interested in and are genetically tractable enough to allow you to express and assay mutants.
- In many cases, you will want to perform an initial high-throughput cell-based assay and subsequently follow up with more mechanistic cell-based or biochemical studies.
- For proof-of-principle experiments, budding yeast is an excellent model system due to the breadth of eukaryotic biological processes it performs and the facility of genetic manipulations. The bacteria *E. coli* and *B. subtilis* are two alternative model systems that are genetically tractable, suitable for both biochemical and cell-based assays and have fast doubling times (a useful feature for directed evolution optimization steps).

Here are the pros and cons of common biochemical and cell-based assays:

4. Screen putative allosteric positions for functionality

After establishing an assay, an intermediate step is to generate point mutants of your protein of interest at the putative allosteric sites (Figure 4). Whether performing biochemical or cell-based assays, the first step is validation: do mutations in putative allosteric sites alter activity of the protein? It is also possible to skip this step, and directly modify the sector connected sites with some allosteric modulator (step 5). Including this step (rather than moving directly to the addition of an allosteric modulator) allows one to distinguish between negative outcomes. There are two ways for a site to not provide allosteric regulation: 1) it is not coupled to protein function or 2) the allosteric modulator introduced does not efficiently transmit a signal (for example, if domain insertion linker length requires optimization). For biochemical assays, you will be limited in the number of mutants you can screen. For cell-based assays, your initial screening approach will depend on whether you are interested in comprehensively identifying functional sites or focused on finding a single functional site at which you can engineer novel regulation. For comprehensive screening, one possibility is

saturation mutagenesis at all sector connected sites (McLaughlin et al., 2012). Our strategy is outlined below:

- Perform site-directed mutagenesis of the desired number of putative functional residues to alanine (or glycine if already alanine).
- Express mutants as the only copy of the protein in cells. Ensure you have wild type and loss-of-function controls.
- Assay wild type and mutants. Each sample should be performed in replicate and ideally at multiple time points or inducer concentrations.
- Identify mutants with altered functionality (gain-, loss-, or partial-loss-of-function). In pooled strategies, this can mean sorting a distribution of high or low reporter activity or applying selection and deep sequencing.
- Explore the scope of the functional plasticity of any identified sites by mutating those positions to all other amino acids and repeat the assay. In this way, you may generate gain-of-function or neomorphic mutants.

5 & 6. Harness latent allostery to engineer novel regulation and optimize regulation, if necessary

“Hacking” into a protein by harnessing latent allostery requires perturbation of allosteric hotspots (Figure 4). The simplest perturbations are mutations to the hotspots themselves, some of which were described above to validate the hotspot. These mutations can disrupt, enhance or alter intraand inter-molecular interactions that regulate protein activity. More complex perturbations, such as introducing a ligand binding domain, a post-translational modification site or a conditional protein-protein interaction can be engineered to generate inducible and dynamic systems. One interesting direction could be the creation of logic gates and signals, through the introduction of multiple modulators. Here we describe a toolkit of perturbations that can be used as allosteric modulators (Table 3, Figure 1C, Figure 4). Introduce these perturbations into your protein of interest and adapt your assay to determine if you have engineered novel regulation.

In many cases, we anticipate that the regulation initially obtained will be modest. In that instance, optimization may be a necessary next step. This might include optimizing linker length and sequence for domain fusions, or taking other steps to amplify signaling (the steps to consider will depend on the system that you are working in). These optimizations could be rationally designed, or perhaps more powerfully, explored by directed evolution or screening.

7. Connect the allosteric switch to endogenous cellular circuitry

Successful implementation of engineered allosteric regulation enables construction of novel cellular circuits and provides orthogonal control to explore endogenous cellular circuitry. Linking regulation of a protein of interest to other synthetic or endogenous systems results in the creation of novel cellular networks that can be used to generate desired behaviors or cell-based computation. The major benefit of building such circuits by allosteric protein regulation is that signaling and other processes that occur at the level of protein interactions

occur at much more rapid time scales than transcription-based circuits. More fundamentally, engineering novel regulation into proteins provides researchers with unique opportunities to gain insight into natural systems via synthetic genetic probes.

Conclusions

In this tutorial we present REACH, an evolution-based strategy to rationally engineer allosteric control in the protein of your choice. We anticipate that the strategies described in the tutorial can be broadly applied to rewire cellular processes to respond to novel inputs (effectors). We hope that this tutorial inspires creative usage and further development of this approach, and that it provides a basis for exploring fundamental processes in disease, cell communication, protein biophysics and evolution through engineered allostery. By following the roadmap presented here, engineered allosteric control is within your REACH.

Acknowledgments

We thank William Russ and Rama Ranganathan for the Caspase-7 SCA analysis. This work was funded by an Early Independence Award from the NIH Office of the Director (DP5 OD017941-01 to D.P.) and by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative (GBMF4557 to K.R.).

References

- ** Reference that we highlight for the purpose of this tutorial (reason for highlighting)
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*. 1997; 25:3389–3402. [PubMed: 9254694]
- ** Coyle SM, Flores J, Lim WA. Exploitation of latent allostery enables the evolution of new modes of MAP kinase regulation. *Cell*. 2013; 154:875–887. (example of latent allostery). [PubMed: 23953117]
- Cui Q, Karplus M. Allostery and cooperativity revisited. *Protein Sci*. 2008; 17:1295–1307. [PubMed: 18560010]
- de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nature reviews Genetics*. 2013; 14:249–261.
- Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*. 2004; 5:113. [PubMed: 15318951]
- Ferguson AD, Amezcua CA, Halabi NM, Chelliah Y, Rosen MK, Ranganathan R, Deisenhofer J. Signal transduction pathway of TonB-dependent transporters. *Proc Natl Acad Sci U S A*. 2007; 104:513–518. [PubMed: 17197416]
- Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic acids research*. 2016; 44:D279–285. [PubMed: 26673716]
- Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: evolutionary units of three-dimensional structure. *Cell*. 2009; 138:774–786. [PubMed: 19703402]
- Hardy JA, Lam J, Nguyen JT, O'Brien T, Wells JA. Discovery of an allosteric site in the caspases. *Proc Natl Acad Sci U S A*. 2004; 101:12461–12466. [PubMed: 15314233]
- Hatley ME, Lockless SW, Gibson SK, Gilman AG, Ranganathan R. Allosteric determinants in guanine nucleotide-binding proteins. *Proc Natl Acad Sci U S A*. 2003; 100:14445–14450. [PubMed: 14623969]
- Kuriyan J, Eisenberg D. The origin of protein interactions and allostery in colocalization. *Nature*. 2007; 450:983–990. [PubMed: 18075577]

- ** . Lee J, Natarajan M, Nashine VC, Socolich M, Vo T, Russ WP, Benkovic SJ, Ranganathan R. Surface sites for engineering allosteric control in proteins. *Science*. 2008; 322:438–442. (original paper using sector connection to establish allosteric regulation between domains). [PubMed: 18927392]
- Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*. 1999; 286:295–299. [PubMed: 10514373]
- Luque I, Leavitt SA, Freire E. The linkage between protein folding and functional cooperativity: two sides of the same coin? *Annual review of biophysics and biomolecular structure*. 2002; 31:235–256.
- McLaughlin RN Jr, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. *Nature*. 2012; 491:138–142. [PubMed: 23041932]
- ** . Monod J, Changeux JP, Jacob F. Allosteric proteins and cellular control systems. *J Mol Biol*. 1963; 6:306–329. (key assertions regarding how small molecule metabolites act as allosteric effectors). [PubMed: 13936070]
- Nadler DC, Morgan SA, Flamholz A, Kortright KE, Savage DF. Rapid construction of metabolite biosensors using domain-insertion profiling. *Nature communications*. 2016; 7:12266.
- ** . Novinec M, Korenc M, Caflisch A, Ranganathan R, Lenarcic B, Baici A. A novel allosteric mechanism in the cysteine peptidase cathepsin K discovered by computational methods. *Nature communications*. 2014; 5:3287. (sector-based design of an allosteric drug).
- Ostermeier M. Engineering allosteric protein switches by domain insertion. *Protein Eng Des Sel*. 2005; 18:359–364. [PubMed: 16043448]
- Pei J, Grishin NV. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. *Methods in molecular biology*. 2014; 1079:263–271. [PubMed: 24170408]
- Peterson FC, Penkert RR, Volkman BF, Prehoda KE. Cdc42 regulates the Par-6 PDZ domain through an allosteric CRIB-PDZ transition. *Mol Cell*. 2004; 13:665–676. [PubMed: 15023337]
- ** . Reynolds KA, McLaughlin RN, Ranganathan R. Hotspots for allosteric regulation on protein surfaces. *Cell*. 2011; 147:1564–1575. (experimental demonstration of using allosteric hotspots to implement allosteric control). [PubMed: 22196731]
- Reynolds KA, Russ WP, Socolich M, Ranganathan R. Evolution-based design of proteins. *Methods in enzymology*. 2013; 523:213–235. [PubMed: 23422432]
- ** . Rivoire O, Reynolds KA, Ranganathan R. Evolution-Based Functional Decomposition of Proteins. *PLoS computational biology*. 2016; 12:e1004817. (describes the SCA approach and mathematics). [PubMed: 27254668]
- Sanner MF, Olson AJ, Spehner JC. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*. 1996; 38:305–320. [PubMed: 8906967]
- Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. 2015
- Smock RG, Rivoire O, Russ WP, Swain JF, Leibler S, Ranganathan R, Gierasch LM. An interdomain sector mediating allostery in Hsp70 molecular chaperones. *Mol Syst Biol*. 2010; 6:414. [PubMed: 20865007]
- Suel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol*. 2003; 10:59–69. [PubMed: 12483203]
- Tullman J, Nicholes N, Dumont MR, Ribeiro LF, Ostermeier M. Enzymatic protein switches built from paralogous input domains. *Biotechnology and bioengineering*. 2016; 113:852–858. [PubMed: 26461040]

Figure 1. Allosteric Regulation in Proteins

A generic protein can exist in a low activity and high activity state. In the absence of allostery (top panel - A), the transition between the two states depends entirely on the local remodeling of the active site due to stochastic conformational fluctuations. In the presence of allostery (middle panel - B), binding of another protein to a distal surface drives the transition of the active site from the inactive to the active state. Allosteric hotspots – many of which are likely to be latent in any given protein under most conditions – occur where protein sectors revealed by the statistical coupling analysis (SCA) intersect with the protein surface (lower panel - C). These hotspots can be harnessed to engineer novel regulation by a

variety of orthogonal regulatory modules that, for instance, respond to light ($\hbar\nu$) or phosphorylation (PO_4).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 2. Hot spots for Allosteric Regulation in Natural and Engineered Systems

In all panels, the protein backbone is shown in cartoon, the sector in dark blue spheres, the active and allosteric site residues in red spheres, and interface residues in green spheres. **A**, The guanine nucleotide binding protein Cdc42 (G protein, grey cartoon) regulates ligand binding affinity of the human Par6 PDZ domain. A sector within the G protein connects the nucleotide binding site of Cdc42 to the interface with Par6 PDZ, and a sector within the PDZ domain connects the interface to the ligand binding site (Lee et al., 2008; Peterson et al., 2004). **B**, In the molecular chaperone Hsp70, nucleotide binding in the N-terminal ATPase regulates ligand affinity in the C-terminal substrate-binding domain. A cross-domain sector links the two sites (Smock et al., 2010). **C**, An allosteric site on caspase-7 was identified by Wells and co-workers using small-molecule tethering (Hardy et al., 2004). A sector links the small molecule regulatory site to the catalytic site (personal communication, William Russ and Rama Ranganathan). **D**, A comprehensive domain insertion screen on DHFR identified fourteen sites capable of allosteric control. These positions are highlighted

in red; all are sector connected (Reynolds et al., 2011). **E**, Baici and co-workers used a combination of SCA and computational ligand docking to identify an allosteric small molecule inhibitor of the collagenolytic cysteine peptidase cathepsin K (Novinec et al., 2014).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 3. Using SCA to identify allosteric hotspots

SCA begins with the construction of a sequence alignment. In step 1, the lines represent individual sequences, and the small circles are amino acids (with color indicating amino acid identity). In step 2, The SCA matrix is computed from this alignment. The diagonal pixels are related to the degree of conservation at a single amino acid position, and the off-diagonal pixels indicate co-evolution between each pair of positions. *Conservation* measures the degree to which amino acid frequency at a particular site differs from what one would expect randomly. Position nine of the cartoon alignment is absolutely conserved (the frequency of the purple amino acid is 100%). *Co-evolution* quantifies the correlations in amino acid identity between positions. For example, every time position one of the alignment has a purple amino acid, position two is green. Once the SCA matrix is generated, it is analyzed to identify groups of evolutionarily correlated amino acids – these form the sector (step 3). The sector is shown in dark blue spheres, and sector connected surface sites are indicated in red spheres.

Figure 4. A toolbox for engineering allosteric control

Allosteric hotspots can be modulated by a variety of genetically encoded alterations that lead to altered protein activity. Possible modifications include constitutive gain- or loss-of-function mutations or inducible switches that can be inserted such as light- and hormone-activated domains and phosphorylation sites that match the consensus motif of effector kinases.

Table 1**Biochemical Assays**

Biochemical Assays	Requirements	Pros/Cons
Binding assays (protein-protein interactions (PPI), peptide binding, small molecule binding)	<ul style="list-style-type: none"> • Purified wild type protein of interest and selected sector mutants thereof • Purified interaction partner (other protein, peptide, small molecule, etc.) • Binding interaction readout (co-IP, anisotropy, SPR, AlphaLISA, etc.) 	<ul style="list-style-type: none"> • Direct, quantitative measurement of interaction of interest • Allows deep structure-function analysis of sector sites • Requires purified proteins and challenging assays • Can have low sensitivity/poor signal-to-noise
Enzymatic assay (ATPase/GTPase assay, kinase assay, metabolic reactions)	<ul style="list-style-type: none"> • Purified wild type enzyme of interest and selected sector mutants thereof • Substrate and other reagents • Readout (kinetic, end point, colorimetric, chromatographic, etc.) 	<ul style="list-style-type: none"> • Direct, quantitative measurement of reaction of interest • Allows deep structure-function analysis of sector sites • Requires purified proteins and challenging experiments • Can have low sensitivity/poor signal-to-noise

Table 2

Cell-Based Assays

Cell-Based Assays	Requirements	Pros/Cons
Fitness/viability (growth rate, competitive fitness, drug suppression, cell titer-glo)	<ul style="list-style-type: none"> Protein of interest must be involved in a process with clear functional impact Wild type and sector mutant cells (can be arrayed or pooled and barcoded) Growth/viability monitoring 	<ul style="list-style-type: none"> Quantitatively determines phenotypic consequence of mutations Can be performed in medium- and high-throughput Allows detection of small (or at least physiologically relevant) effects Requires obvious functional readout for protein
Morphology and subcellular localization (differentiation, yeast mating projection, live-cell or immuno-fluorescence)	<ul style="list-style-type: none"> Wild type and sector mutant cells (separated) Imaging by microscopy 	<ul style="list-style-type: none"> Visually-convincing phenotype Only applicable to certain proteins Low throughput – each mutant must be independently assayed
Transcriptional reporter (fluorescent reporter of a kinase pathway or stress response)	<ul style="list-style-type: none"> Wild type and sector mutant cells expressing the reporter (can be arrayed or pooled and barcoded) Flow cytometry or other readout for reporter activity 	<ul style="list-style-type: none"> Quantitatively determines consequence of mutations Can be performed in medium- and high-throughput Requires reporters for specific signaling pathways
Phosphorylation (western blot, IF, flow cytometry)	<ul style="list-style-type: none"> Wild type and sector mutant cells (separated) Antibodies to monitor phosphorylated state of protein of interest 	<ul style="list-style-type: none"> Direct readout of signaling pathway activity Only applicable to certain proteins Low throughput – each mutant must be independently assayed Can have low sensitivity/poor signal-to-noise

Table 3

Toolkit of Modulators

Modulator	Implementation	Utility
Alanine mutation	Mutate sector-connected surface residues to alanine	Scan for functional sites
Saturation mutagenesis	Mutate sector-connected surface residues to all amino acids	Screen for gain-of-function and neomorphic activity
Light-activated domain (e.g. LOV, CRY or Phy domains)	Introduce LOV or other domains that undergo conformational changes upon exposure to specific wavelengths of light at sector-connected surface residues	Conditional regulation by an orthogonal signal
Hormone-binding domain (e.g. estradiol, progesterone or glucocorticoid binding domains)	Introduce ligand binding domains from nuclear hormone receptors that change conformation upon binding to cognate hormones at sector-connected surface residues; can restrict localization of target protein to the cytoplasm	Conditional regulation by an orthogonal signal
Periplasmic solute-binding protein domains	Introduce maltose binding protein domain	Conditional regulation by an orthogonal signal
Nucleic acids	Introduce nucleic acid binding domain	Nucleic acid libraries can be synthesized
Phosphorylation sites	Introduce consensus motifs from kinases of interest at sector-connected surface residues	Conditional regulation by orthogonal or endogenous signaling pathways
Protein-protein interaction domain	Introduce protein-protein interaction domain for a protein of interest at sector-connected surface residues	Conditional regulation by interaction with signaling proteins; can be used as second wave signaling in complex circuits