SynPHARM and the Guide to Pharmacology database: a toolset for conferring drug control on engineered proteins

SynPHARM and the Guide to Pharmacology database:

a toolset for conferring drug control on engineered proteins

Jamie A. Davies
Synthsys Centre for Systems and Synthetic Biology,
and Deanery of Biomedical Science.
University of Edinburgh,
Scotland, UK.

Abstract: (250wd)

Optimizing synthetic biological systems, for example novel metabolic pathways, becomes more complicated with more protein components. One method of taming the complexity and allowing more rapid optimization is engineering external control into components. Pharmacology is essentially the science of controlling proteins using (mainly) small molecules, and a great deal of information, spread between different databases, is known about structural interactions between these ligands and their target proteins. In principle, protein engineers can use an inverse pharmacological approach to include drug response in their design, by identifying ligand-binding domains from natural proteins that are amenable to being included in a designed protein. In this context, 'amenable' means that the ligand-binding domain is in a relatively self-contained subsequence of the parent protein, structurally independent of the rest of the molecule so that its function should be retained in another context. The Synpharm database is a tool, built on to the Guide to PHARMACOLOGY database and connected to various structural databases, to help protein engineers identify ligand-binding domains suitable for transfer. This article describes the tool, and illustrates its use in seeking candidate domains for transfer. It also briefly describes already-published proof-of-concept studies in which the CRISPR effectors Cas9 and Cpf1 were placed separately under the control of tamoxifen and mefipristone, by including ligand-binding domains of the Estrogen Receptor and Progesterone Receptor in modified versions of Cas9 and Cpf1. The advantages of drug control or the rival protein-control technology of optogenetics, for different purposes and in different situations, are also briefly discussed.

Keywords:

protein structure, protein binding, drug, pharmacological control, biological engineering, protein engineering, CRISPR, gene editing.

Lay audience statement:

When we design novel proteins with useful properties, it would be useful if we could also engineer in external control. Drugs control natural proteins by interacting with them in a specific way. The Synpharm database is a software tool to help protein engineers identify drug-interacting elements in normal proteins, that they can copy into their engineered proteins to make these, too, respond to the same drug.

Introduction

Broadly, the challenge of optimizing engineered biological systems increases exponentially with the number of components involved [1 -3]. This challenge does not lie primarily in physical construction of DNA sequences etc., because any difficulties here increase only linearly with system size. Rather, it arises because each extra active molecule in a system adds further dimensions to the parameter space in which the finished system will operate. To illustrate this with an example, first consider a very simple system involving only one engineered protein, the activity of which might be chosen to be any one of ten possible values (according, for example, to the amino acid sequence of an active site). Clearly, with these restrictions, optimizing the system just requires comparing the performance of ten possible versions. A slightly more complicated system, with two such proteins, offers 100 possible combinations of parameters. A system with three such proteins offers 1000, and so on (Fig 1). By the time systems reach even a dozen components, the parameter space is massive and finding parameters for optimal performance is not trivial.

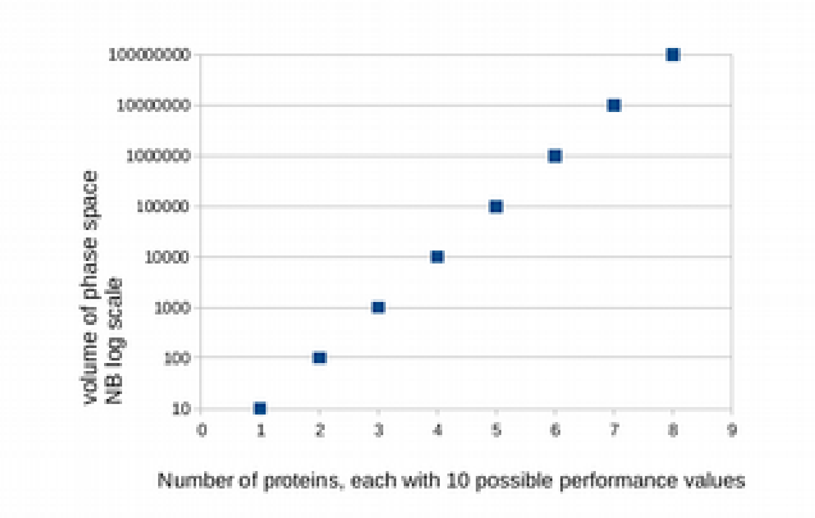


Figure 1: The exponential rise of the volume of phase space to be explored during the optimization of systems containing more and more engineered proteins. For simplicity, each protein is idealized to have only ten possible performance values; the reality may well be a great deal worse.

There are several common approaches to meet this challenge. One is mathematical modelling, which tends to be most useful when the system and its interactions with a host cell are reasonably well understood [4-6]. Models that are mathematically tractable can be solved algebraically, while less tractable ones can be explored by computer, if necessary by the brute-force exploration of parameter space, something that can be done much more rapidly and economically in silicon than in culture plates. Another approach, which can be used either with or as an alternative to computer modelling, is an evolutionary one of constructing many different genetic mutations of a system and selecting that with the best performance [7, 8]. The genetic mutations need not be restricted to exogenous proteins; host genes may also be mutated for this 'directed evolution' approach. Directed evolution is especially straightforward if performance can be linked to fitness, so that cells harbouring the best performing version of the system have a selective advantage and come to dominate the culture. Without this, though, identification and selection of cells harbouring the best system, amongst vast numbers of other ones, becomes very difficult.

A third approach is to parameterize the system itself. Instead of being constructed with fixed components, at least parts of the system are made externally controllable, so that one physical version of the system can be used to explore different volumes of parameter space. This saves building many different versions. Most methods for doing this have centred on controlling the concentration of the protein in question by controlling the transcription of the gene that encodes it, using systems such as Tetracycline-inducible operators (the 'Tet' system; [9]) or more advanced systems [10]. These work well, and have been used for parameter exploration [10], but their use is restricted to controlling how much of a protein is made, not the specific activity of that protein.

An alternative approach is to engineer external regulation into the proteins themselves. Here, we describe an open-to-all database and tool-set to facilitate this type of protein engineering, and describe some examples of proteins in which drug-mediated control has been successfully engineering into DNA editing proteins.

Inverse pharmacology: a backwards approach to drugs and targets

In conventional pharmacology, a researcher begins with a target protein and attempts to find or design a small molecule that will modulate its activity in some desired way.

Molecules with this property are then tested for useful kinetic properties, such as half-life and ability to pass from vessels to tissues, and tested for safety and for efficacy. Those that pass all the tests may go on to become clinically registered drugs. The whole process is difficult, expensive, and time-consuming; clearly it would not be an appropriate way of attaching control to new synthetic biological systems.

The conventional approach is forced on pharmacologists because they have to work with the proteins that are naturally present in the body. Designers of synthetic systems tend to engineer proteins anyway, to adapt or alter binding, enzymatic rates etc. This opens up the possibility of inverting the normal order of pharmacology, to begin with a drug and to design responsiveness to that drug to be a property of a new engineered protein. This approach has considerable advantages; the pharmacokinetic and safety properties of human and veterinary drugs are already well-known and thousands are licensed for clinical use. This is the approach that the tool we describe in this article is intended to facilitate.

Designing the tool: what do we know, and how can we use it?

Many decades of progress in molecular pharmacology have resulted in a great deal of knowledge about which of the approximately 10,000 drugs and similar molecules used in clinical medicine and research bind to which of the approximately 3000 human proteins that are drug targets. This knowledge is summarized in open resources, such as the IUPHAR/ BPS Guide to PHARMACOLOGY database [11]. In addition, for a significant proportion of these drugs, there is high resolution (generally crystallographic) information about precisely how the drug interacts with its target; which protein residues are involved, and how primary, secondary, and tertiary protein structures are involved in making the required protein residues available to the ligand [12-15]. In principle, these datasets might be used to identify drug-binding motifs from natural proteins that can be included in engineered proteins (by encoding them as part of the coding sequence of a transgene in the usual way)

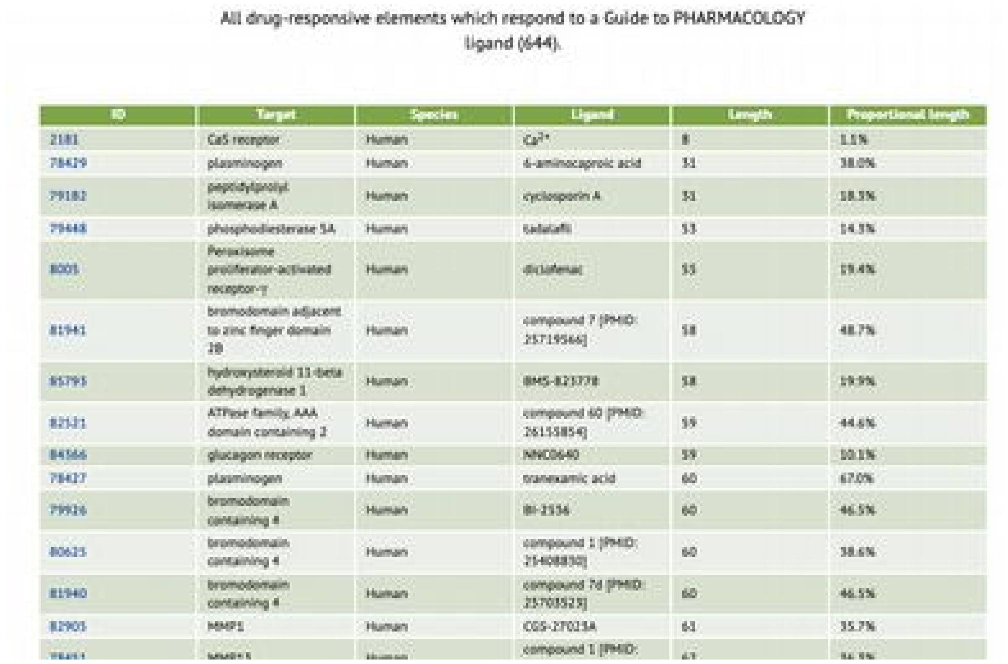
There are, however, problems with this idea. The greatest comes from the nature of many drug-binding sites. Proteins are complex, folded, three-dimensional structures, and amino acids that are close to one another in space are not necessarily close to one another in the primary sequence of the peptide chain. If a drug-binding site is formed by the spatial apposition of amino acids from many different parts of the peptide, brought to that location by the structure of the rest of the protein, it would be very difficult to re-create the drug binding site in an engineered protein with different overall shapes and properties. The suitability of a known drug-binding site for use in novel proteins therefore depends on the extent to which it is formed from a relatively self-contained run of amino acids, forming a structure relatively independent of the rest of the protein. This 'ligand-binding module' can itself be highly folded, as long as it is relatively self-contained.

Scanning drug-protein binding structures manually to find promising examples is possible but laborious. Given that we already curate the IUPHAR/ BPS Guide to PHARMACOLOGY database, which has rich links to structural databases, we decided to build a tool to make identification of promising ‘drugability modules’ easier.

The SynPharm tool

The tool we have constructed, SynPharm (from ‘synthetic biology’ and ‘pharmacological control’), is open to anyone to use at https://synpharm.guidetopharmacology.org/. Its home-page presents simple statistics about the number of ligands (drugs and drug-like molecules) and natural protein targets about which it holds binding information (at the time of writing, 515 ligands and 644 targets), search boxes, and links to tutorials and other information. It contains no new information that is not in other databases; rather, it provides new ways to interact with that information. The manner in which it was constructed and populated has been described elsewhere [17] and this information will not be repeated here.

There are several possible search strategies but the most useful to the focus of this article – identifying modules to confer drugability on engineered proteins – is simply to click ‘search sequences that interact with a ligand’. Doing so results in an ordered table (Fig 2), listing targets, species, ligand, length of peptide over which binding residues are scattered, and what proportion of the whole protein length this is. Clicking on any table heading will cause ordering of the table by that criterion (a second click reverses the order). Clicking on the ‘length’ column header, as has been done in Fig 2, will order the list by the length of the binding domain, low-to-high. Proteins for which drug binding is a property of amino acids clustered on a short length of the peptide chain are the most likely to be useful as a source of drugability modules, but length is not the only consideration; the ‘independence’ of that domain from the structures elsewhere in the protein is highly relevant. This independence is extremely hard to define computationally, so judgement is left to human users, who are assumed to be reasonably expert at protein engineering.

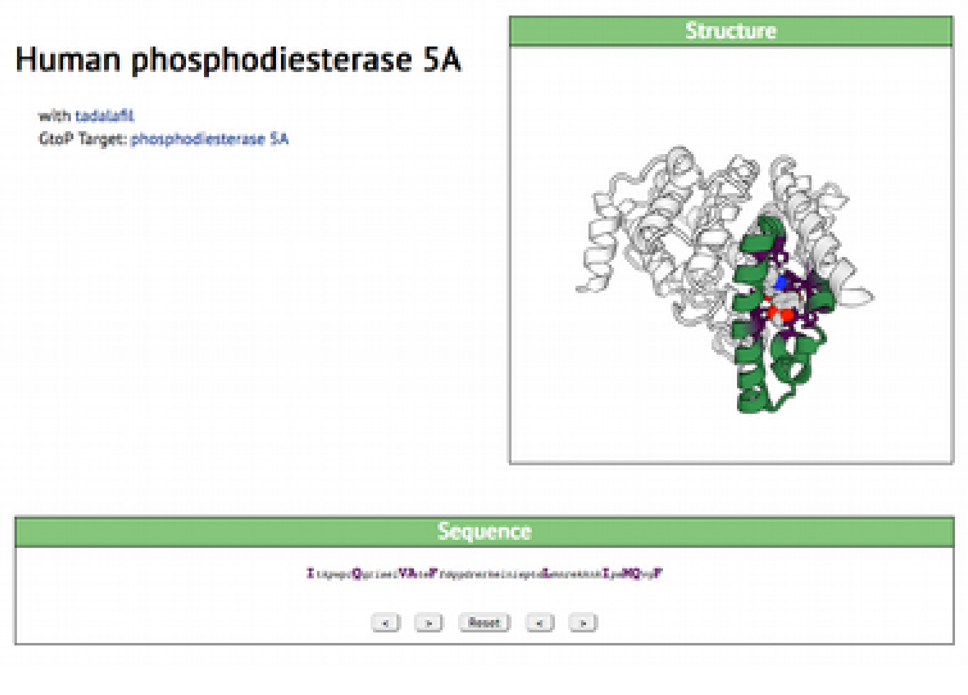


All drug-responsive elements which respond to a Guide to PHARMACOLOGY ligand (644).

ID	Target	Species	Ligand	Length	Proportional length
2181	CaS receptor	Human	Ca^{2+}	8	1.1%
78429	plasminogen	Human	6-aminocaproic acid	31	38.0%
79182	peptidylprolyl isomerase A	Human	cyclosporin A	31	18.3%
79448	phosphodiesterase 5A	Human	tadalafil	53	14.3%
8005	Peroxisome proliferator-activated receptor-γ	Human	diclofenac	55	19.4%
81941	bromodomain adjacent to zinc finger domain 2B	Human	compound 7 [PMID: 25719566]	58	48.7%
85793	hydroxysteroid 11-beta dehydrogenase 1	Human	BMS-823778	58	19.9%
82521	ATPase family, AAA domain containing 2	Human	compound 60 [PMID: 26155854]	59	44.6%
84566	glucagon receptor	Human	NNC0640	59	10.1%
78427	plasminogen	Human	tranexamic acid	60	67.0%
79926	bromodomain containing 4	Human	BI-2536	60	46.5%
80625	bromodomain containing 4	Human	compound 1 [PMID: 25408830]	60	38.6%
81940	bromodomain containing 4	Human	compound 7d [PMID: 25703523]	60	46.5%
82905	MMP1	Human	CGS-27023A	61	35.7%
[illegible]	MMP13	Human	compound 1 [PMID:	[illegible]	[illegible]

Figure 2: the top part of a Synpharm summary table of proteins, for which the database has structural binding data, that interact with ligands. The output has been ordered by length of ligand-binding domain, low to high, by clicking on the 'length' column header.

Clicking on the index number of any protein brings up a more detailed page. The page provides a simple display of the relevant amino acid sequence, with the ligand-interacting amino acids highlighted to show how they cluster (or do not). It also provides a rotatable 3D model of the protein with its bound ligand, which can be used to estimate how independent of the rest of the protein is the structure of the ligand-binding site. The interaction of the human phosphodiesterase 5A with tadalafil, for example, involves a run of only 53 amino acids but the binding structure lies at the interface of two alpha helices, the angles of which would be likely to depend strongly on details of the rest of the protein (Fig 3). This would suggest that, for ligand binding to be transferred to an engineered protein, a substantial fraction of the natural protein would need to be included (and linked to the rest of the engineered protein in a way that did not interfere with its folding). From a protein engineering point of view, this would not seem promising.

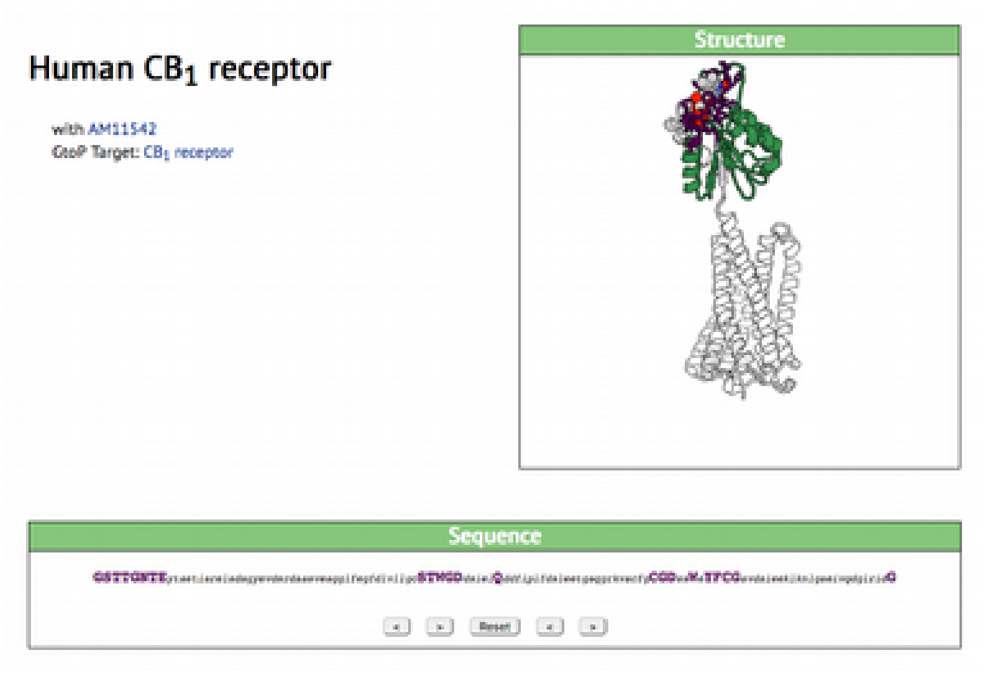


Human phosphodiesterase 5A

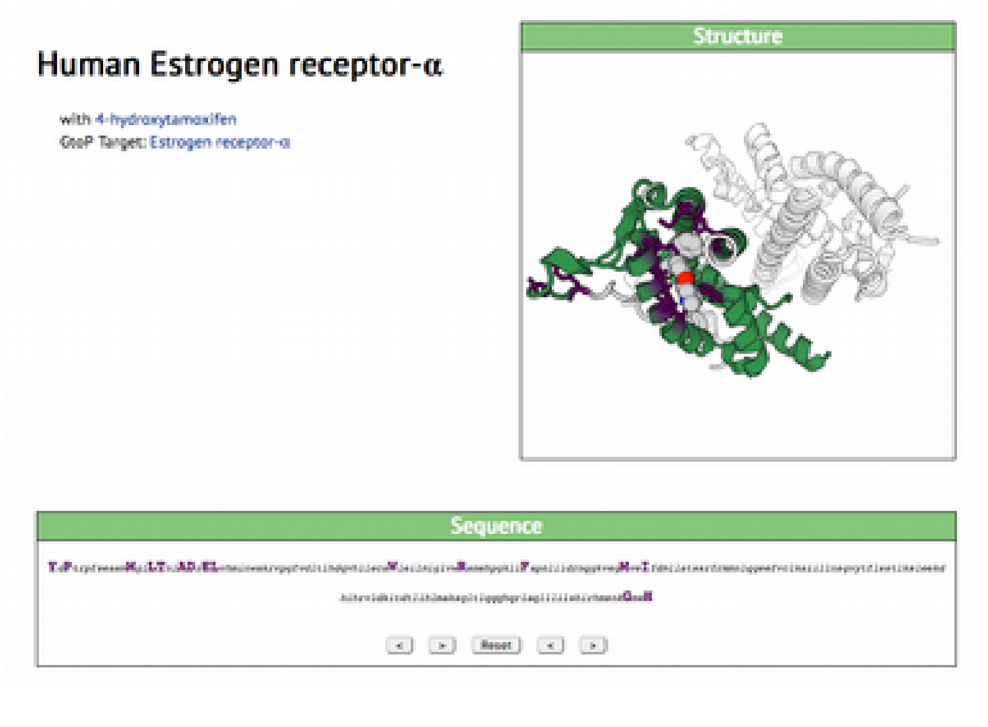
with tadalafil
GtoP Target: phosphodiesterase 5A

Figure 3: The interaction of phosphodiesterase 5A with tadalafil, as displayed on a Synpharm page. In the rotatable molecular model, the green shading represents the whole ligand-binding sequence, also shown in single-letter codes below. The brown, and the large amino acids in the sequence, represent amino acides directly involved in the interaction. The ligand is shown as a multi-colour chemical structure. See main text for comments on the probable (non-) utility of this interaction for protein engineering.

A similar view of the interaction of human CB1 receptor with the ligand AM11542, on the other hand, shows a binding site dominated by 23 amino acids widely spaced along a long sequence, but one that forms a compact and self-contained domain of the protein that is coupled relatively flexibly to the rest (Fig 4). This suggests that this section of the natural protein might be included in an engineered protein to bring in the ligand-binding domain.

	 Human CB_1 receptor with AM11542 GtoP Target: CB_1 receptor 	
	Figure 4: Synpharm display of the interaction of the CB1 receptor with compound AM11542. Here the interacting domain, though fairly long, is relatively independent of the rest of the molecule, linkage between the two parts of the receptor that contain many alpha helices being via a relatively flexible chain. Colours etc. are as in Fig 3.	

A famous example of ligand-interaction domains that can be ported to other, engineered proteins is provided by nuclear hormone receptors. The ligand-binding site of ESR1 (estrogen receptor α), shown in Fig 5, has been connected (in a slightly mutant form) to Cre recombinase and confers tamoxifen-dependency on that recombinase [18]. The Synpharm tool identifies a ligand-binding domain of 193 amino-acids, but in fact the domain actually used to confer tamoxifen control on other molecules is around 300 amino acids long (ESR1 is 595 amino acids long in all). This acts as a warning that, though the algorithms behind Synpharm offer a useful sketch, human judgement is again needed to ensure adequate environment and 'spacing' for the transferred domain in its new context.



Human Estrogen receptor-α

with 4-hydroxytamoxifen
GtoP Target: Estrogen receptor-α

Figure 5: Synpharm display of interaction of Estrogen receptor α with hydroxytamoxifen, colours etc. as in Fig 3. See main text for commentary.

An example of engineering drug control into effector proteins

One of the most important technologies to emerge in molecular biology has been gene editing, using the effectors of the bacterial CRISPR system (reviewed in [19]). These effectors can be targeted to specific genes using guide RNAs (gRNA), and can introduce either random indel-type mutations or, in combination with templates, introduce targeted insertions or replacements [20]. Modified versions of the effectors can also be used as transcription activators [21].

The bacterial-derived effectors, such as Cas9 and Cpf1, are constitutively active provided they have gRNA. For gene editing in simple 2-dimensional cultures, this is not a problem because the reagents can be introduced to cells only when editing is needed. There is, though, increasing interest in performing gene editing in solid 3-dimensional culture systems, such as organoids made from human pluripotential cells, for example to mimic the effects of loss-of-heterozygosity at a locus connected with a congenital disease. In these 3-D systems, access to cells for transfection with Cas9 and gRNA is highly restricted and usually only the outer layer can be reached, but transfection of cells before the organoid is made would result in gene editing happening before the developmental stage at which it is needed to mimic the disease. For these cases, having the gRNA and a drug-controllable version of the Cas9 or Cpf1 effectors expressed in the cells all the time would allow organoids to be built in the absence of the drug, gene editing to be induced by a small molecule that can diffuse well even through an organoid.

We therefore engineered ligand-binding domains from either the estrogen receptor (in the ERT2 mutant form), or the progesterone receptor, into both Cas9 and Cpf1 (Fig 6) [22]. Estrogen, progesterone and their pharmacological analogues diffuse very well in tissues because they can cross membranes. In a simple 2-dimension proof-of-concept study, in which gene editing destroyed a transcriptional repressor and thus freed production of a fluorescent signal from repression, the activity of these engineered CRISPR effectors was found to be highly dependent on presence of their ligands. For the tamoxifen-inducible Cas9, for example, there was a 49-fold difference in reporter fluorescence between dishes treated with and without 1μM hydroxytamoxifen [22].

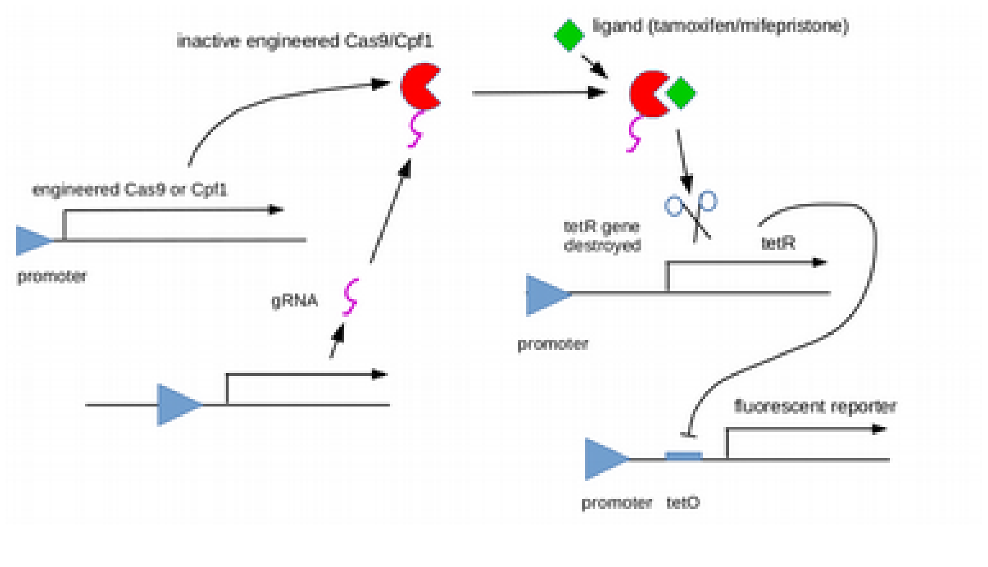


Figure 6: a schematic of the system used in [22] to demonstrate drug control engineered separately into the Cas9 and the Cof1 CRISPR effectors. In the absence of ligand (tamoxifen or mifepristone, depending on the ligand-binding domain engineered into the CRISPR effector), the effector is inactive the gene controlling TetR is intact, allowing this transcriptional repressor to act on its operator upstream of a reporter gene, to hold that gene 'off'. When the ligand is present the CRISPR effector is active, the gene controlling TetR is mutated by the effector, and once residual TetR has cleared the reporter switches 'on'.

Concluding remarks

The utility of the Synpharm resource will depend on several things. One is the popularity of engineering ligand control into proteins at all. At around the same time that Synpharm was being developed, alternative technologies for protein control were maturing quickly. Optogenetic technologies, in particular, have made great advances both in control of protein expression and control of protein function, particularly that of channels [23]. Light-mediated control offers very high spatial and temporal resolutions offer more flexibility for control than drugs can. Indeed, this lab has begin to move towards optogenetics for these reasons [24-25]. But light has limited penetrance in deep tissues, which continues limits its utility in animal models despite recent work that has extended the depth to which light can still be used [26, 27]. For engineered proteins used in this context, drug control still seems to be the best option.

The second influence on the utility of the Synpharm resource is the extent to which pre-packaged 'kit' approaches to introducing ligand control replace the need for individual protein engineers to do their own research and make their own decisions. At present, as far as the author knows, no such kits exist but if one module is developed that will work in a large range of protein hosts, then it may well dominate the field as, for example, the Tet-operator system has come do dominate transcriptional control [9].

References

1. Naseri G, Koffas MAG (2020) Application of combinatorial optimization strategies in synthetic biology. *Nat Commun*. 2020; 11: 2446. doi: 10.1038/s41467-020-16175-y

2. Volk MJ, Lourentzou I, Mishra S, Vo LT, Zhai C, Zhao H (2020) Biosystems Design by Machine Learning. *ACS Synth Biol* 17;9(7):1514-1533. doi: 10.1021/acssynbio.0c00129.

3. Shen L, Kohlhaas M, Enoki J, Meier R, Schönenberger B, Wohlgemuth R , Kourist R, Niemeyer F, van Niekerk D, Bräsen C, Niemeyer J, Snoep J, Siebers B (2020) A combined experimental and modelling approach for the Weimberg pathway optimisation. *Nat Commun* 11(1):1098. doi: 10.1038/s41467-020-14830-y.

4. de Arroyo Garcia L, Jones PR (2020) In silico co-factor balance estimation using constraint-based modelling informs metabolic engineering in Escherichia coli. *PLoS Comput Biol* 16(8):e1008125. doi: 10.1371/journal.pcbi.1008125.

5. Küken A, Nikoloski Z. (2019) Computational Approaches to Design and Test Plant Synthetic Metabolic Pathways. *Plant Physiol.* 179(3):894-906. doi: 10.1104/pp.18.01273.

6. Schneider P, von Kamp A, Klamt S (2020) An extended and generalized framework for the calculation of metabolic intervention strategies based on minimal cut sets. *PLoS Comput Biol*
16(7):e1008110. doi: 10.1371/journal.pcbi.1008110.

7. Ye L, Yang C, Yu H (2018) From molecular engineering to process engineering: development of high-throughput screening methods in enzyme directed evolution. *Appl Microbiol Biotechnol*
102(2):559-567. doi: 10.1007/s00253-017-8568-y.

8. Wu J, Chen W, Zhang Y, Zhang X, Jin J-M, Tang S-Y (2020) Metabolic Engineering for Improved Curcumin Biosynthesis in Escherichia coli . *J Agric Food Chem* 2020 Sep 11. doi: 10.1021/acs.jafc.0c04276.

9. Ramos, J. L.; Martínez-Bueno, M; Molina-Henares, A.J.; Terán, W.; Watanabe, K.; Zhang, X.; Gallegos, M.T.; Brennan, R.; Tobes, R. (2005) The TetR family of transcriptional repressors. Microbiol. *Mol Biol. Rev.* 69: 326-356.

10. Naseri G, Behrend J, Rieper L, Mueller-Roeber B (2019) COMPASS for rapid combinatorial optimization of biochemical pathways based on artificial transcription factors. *Nat Commun 2019* Jun 13;10(1):2615. doi: 10.1038/s41467-019-10224-x.

11. Armstrong JF, Faccenda E, Harding SD, Pawson AJ, Southan C, Sharman JL, Campo B, Cavanagh DR, Alexander SPH, Davenport AP, Spedding M, Davies JA (2020) The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY. *Nucleic Acids Res.* 2020 Jan 8;48(D1):D1006-D1021. doi: 10.1093/nar/gkz951.

12. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J (2016) BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* 2016, 44, D1045-53

13. Gaulton A, Hersey A, Nowotka, M, Bento AP, Chambers J, Mendez, D, Mutowo P, Atkinson,F.; Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach A. (2017) The ChEMBL database in 2017. *Nucleic Acids Res*. 2017, 45, D945-D954.

14. Berman HM, Burley SK, Kleywegt GJ, Markley JL, Nakamura H, Velankar S. (2016) The archiving and dissemination of biological structure data. *Curr. Opin. Struct. Biol*. 2016, 40, 17-22.

16. Desaphy J, Bret G, Rognan D, Kellenberger E (2015) sc-PDB: a 3D-database of ligandable binding sites -10 years on. *Nucleic Acids Res.* 2015, 43, D399-404.

16. Liu Z, Li Y, Han L, Li J, Liu J, Zhao Z, Nie W, Liu Y, Wang R. (2015) PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*. 2015, 31, 405-412.

17. Ireland SM, Southan C, Dominguez-Monedero A, Harding SD, Sharman JL, Davies JA (2018) . SynPharm: A Guide to PHARMACOLOGY Database Tool for Designing Drug Control into Engineered Proteins. *ACS Omega*. 2018 Jul 31;3(7):7993-8002. doi: 10.1021/acsomega.8b00659.

18. Feil R, Brocard,J, Mascrez B, LeMeur M, Metzger D, Chambon P. (1996) Ligand-activated site-specific recombination in mice. *Proc. Natl. Acad. Sci. USA*, 1996, 93, 10887–10890

19. Makarova, K.S., Zhang, F., Koonin, E.V. (2017) SnapShot: Class 2 CRISPR-Cas Systems. *Cell*. 168, 328-328.

20. Ceasar SA, Rajan V, Prykhozhij SV, Berman JN, Ignacimuthu S. (2016) Insert, remove or replace: A highly advanced genome editing system using CRISPR/Cas9. *Biochim Biophys Acta*. 2016 Sep;1863(9):2333-44. doi: 10.1016/j.bbamcr.2016.06.009.

21. Tak YE, Kleinstiver BP, Nuñez JK, Hsu JY, Horng JE, Gong J, Weissman JS, Joung JK (2017) Inducible and multiplex gene regulation using CRISPR-Cpf1-based transcription factors. *Nat Methods*. 14, 1163-1166.

22. Dominguez-Monedero A, Davies JA (2018) Tamoxifen- and Mifepristone-Inducible Versions of CRISPR Effectors, Cas9 and Cpf1. *ACS Synth Biol* 2018 Sep 21;7(9):2160-2169.
doi: 10.1021/acssynbio.8b00145. Epub 2018 Sep 4.

23. Paoletti P, Ellis-Davies GCR, Mourot A. (2019) Optical control of neuronal ion channels and receptors. *Nat Rev Neurosci*. 2019 Sep;20(9):514-532. doi: 10.1038/s41583-019-0197-2.

24. Baaske J, Gonschorek P, Engesser R, Dominguez-Monedero A, Raute K, Fischbach P, Müller K, Cachat E, Schamel WWA, Minguet S, Davies JA, Timmer J, Weber W, Zurbriggen MD. (2018) Dual-controlled optogenetic system for the rapid down-regulation of protein levels in mammalian cells. *Sci Rep*. 2018 Oct 9;8(1):15024. doi: 10.1038/s41598-018-32929-7.

25. Fischbach P, Gonschorek P, Baaske J, Davies JA, Weber W, Zurbriggen MD. (2020)

Optogenetic Downregulation of Protein Levels to Control Programmed Cell Death in Mammalian Cells with a Dual Blue-Light Switch. *Methods Mol Biol*. 2020;2173:159-170. doi: 10.1007/978-1-0716-0755-8_11.