20.320

Section 2

Modeling and Manipulating Biomolecular Interactions

Goals:

- **To understand the biophysics of molecular interactions**
- **To computationally model the energetics of interactions**
- **To predict protein structures**
- **To predict the effects of mutations**
- **To solve computationally intractable problems**
- **To design improved molecules**

Overview:

Diverse problems ranging from fundamental questions of molecular biology to drug development and synthetic biology can be analyzed in terms of the interactions between specific biomolecules, including protein-DNA and kinase-substrate interactions. Techniques for modifying these interactions are an essential part of the biological engineer's toolkit. This section of 20.320 will focus on methods for modeling biomolecular interactions to understand and manipulate biology.

Understanding Biology

Low cost DNA-sequencing has recently made it possible to examine the genomes of unprecedented numbers of organisms. These sequences have revealed fascinating aspects of evolutionary history, including human history and identified many genetic variations associated with disease. However, it has also created a huge number of new unanswered questions that the techniques you are learning may help to solve.

"What part of the genome do you not understand?"

Galperin and Koonin pose this amusing question in a recent review [5]. They estimate that there are millions of sequences of unknown function. Inspired perhaps by Donald Rumsfeld, they split these into known unknowns (proteins that can be assigned some general function, but whose specifics remain unknown) and unknown unknowns (truly mysterious). Some of these unknowns are found only in a handful of organisms, but others are widespread.

Ultimately, the molecular function of these proteins must be determined by their sequences, and these in turn determines their physical properties. So, there is great

hope that we can predict structure from sequence and function from structure. We will examine methods for this in this unit.

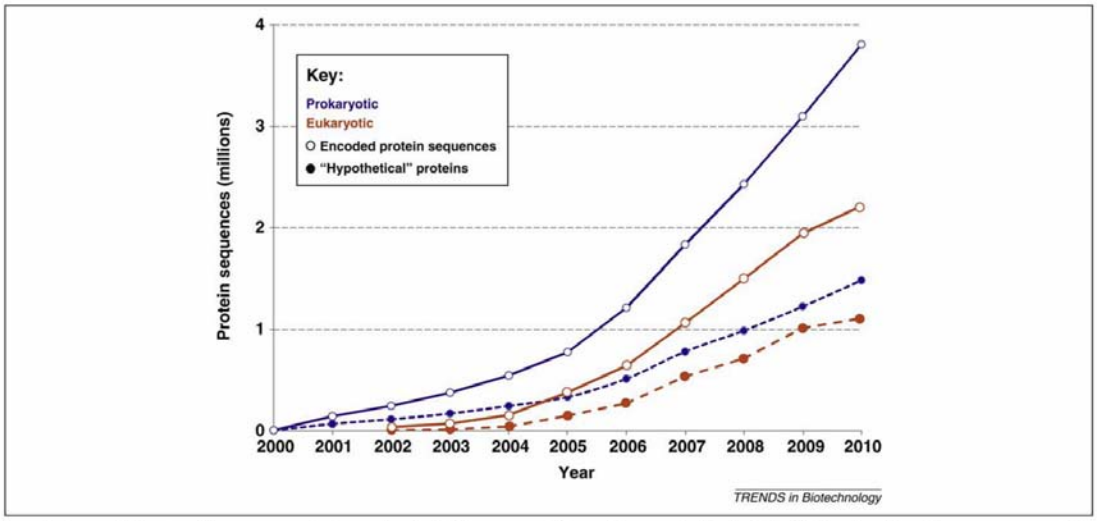


Figure 1. Accumulation of protein sequences of unknown function in the genome databases. Open symbols indicate the total number of protein sequences encoded in prokaryotic (blue) and eukaryotic (red) genomes; filled symbols indicate the number of 'hypothetical' or 'uncharacterized' proteins. The data are taken from the NCBI RefSeq database [68]; the numbers for 2010 are extrapolated from the first 4 months.

From Galperin and Koonin, 2010 [5].

Genetic basis of disease.

Sequencing-based efforts have now revealed the extent of genomic variation in humans. Some of this variation seems to have no impact on phenotype. Other variants are of greater consequence. Some of you may be familiar with diseases that arise from mutations in single proteins. These include Huntington's disease, cystic fibrosis and hemophilia. The recent sequencing efforts have been able to discover more complex cases in which there is no single genetic change that causes a disease, but rather a set of regions in which variations occur associated with increased risk. For some recent results see Hirschhorn and Gajdos [6]. Computational methods for understanding the functional consequences of observed genetic variations could be of tremendous importance in using the genetic observations to develop new therapies.

Manipulating Biology

The tools for protein engineering are growing rapidly and have used for a fascinating range of applications. We will examine how proteins have been engineered to :

1. Edit genomes by predictably altering the DNA sequence at specific sites
2. Design peptides that bind and inactive specific proteins
3. Design sensors to aid in neurobiology research
4. Design light-activated proteins that can control cell morphology
5. Design entirely new enzymes catalyzing reactions that (apparently) do not occur in nature.
6. Design better "biologics" to treat disease

Some of these engineered proteins were created using traditional methods of molecular biology. But others were designed using the computational techniques that we will teach you, and could never have been developed without these tools. In your design project, you will have an opportunity to use the same software, Rosetta, that has been so successful in this area.

1. Tools for editing the genome: The mostly widely used tool for editing the genome relies on rare homologous recombination events in mammalian cells. DNA with high homology to a gene and a selectable marker is introduced into embryonic stem cells, and the rare cells with a recombination event are selected.

An alternative and much more versatile approach is being developed based on protein engineering, reviewed in [3, 7]. Zinc-fingers are a particularly versatile DNA-binding protein that we will explore later in this course. They are remarkable for the fact that the same protein scaffold can be used to recognize many distinct sequences in a relatively predictable way. That means that one can create zinc finger proteins to bind to almost any place in the genome. As far back as 1994, they have been shown to be useful for engineering, when a protein was designed to bind specifically to an oncogenic mutation[8]. The designed protein was shown to repress the oncogene in cell culture.

More recently, efforts have focused on fusing proteins with a desired specificity to a nuclease. The DNA-binding proteins bring two halves of the nuclease to the same region of the genome. When the two halves of the enzyme come together, they cut the DNA. In principle, these proteins will produce double-stranded breaks at exactly one site in the genome. Depending on the DNA-repair mechanism, you can get deletions of various sizes or insertions of specific DNA. The figure below shows some of the types of genomic editing that have been tried. This process could be much more efficient that typical "gene targeting," which relies on rare homologous recombination events and requires selection markers. This approach has now been extended to also take advantage of another class of DNA binding proteins called TAL effectors[9-11].

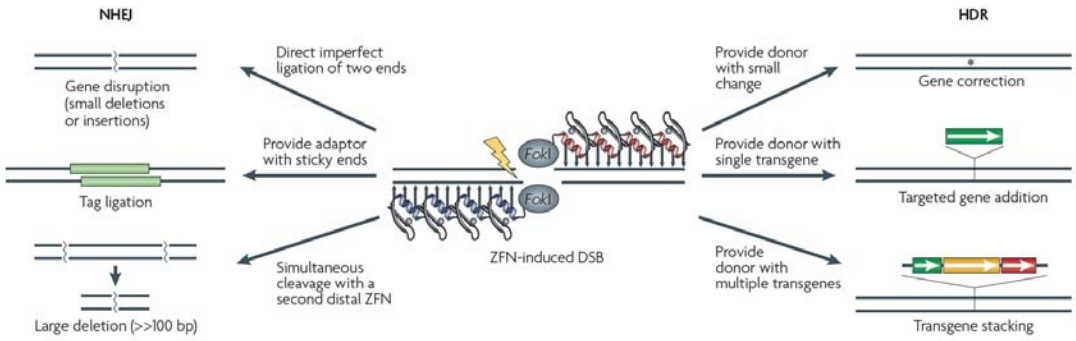


Figure and legend from [3]. A zinc finger nuclease (ZFN)-induced double-strand break (DSB) allows a range of alleles to be generated at endogenous loci, as specified by the investigator. The diagram shows the different outcomes that can result from the introduction of a site-specific DNA break. A ZFN pair is shown bound to a genomic target site (the two different DNA binding domains are shown in red and blue). The DSB generated by ZFN cleavage induces DNA repair processes that may be influenced by the addition of an investigator-designed donor DNA. As shown on the left, if the break is resolved via non-homologous end joining (NHEJ) (which will occur in the absence of donor DNA), this can lead to the following outcomes (from top to bottom): gene disruption — the two ends can be ligated back together, frequently with loss or gain of genetic information at the site of the break, resulting in small insertions or deletions; tag ligation — if a double-stranded oligonucleotide is provided with overhangs complementary to those left by the ZFNs (an adaptor), it will be ligated into the chromosome, thus producing, for example, a tagged allele; large deletion — two simultaneous DSBs made on the same chromosome can lead to a deletion of the entire intervening stretch. As shown in the panel on the right, if the break is resolved via homology-directed repair (HDR) (which will occur in the presence of donor DNA), this can result in (from top to bottom): gene correction — if the donor specifies solely a single-base-pair change (for example, a restriction fragment length polymorphism encoding a novel allele), this will result in 'gene correction' that subtly edits the endogenous allele; targeted gene addition — if a donor is provided that carries an ORF or a transgene at the position corresponding to the site of the break, its sequence will be transferred to the chromosome via the synthesis-dependent strand annealing pathway; transgene stacking — if the donor carries multiple linked transgenes between the homology arms, they will be transferred into the chromosome via the synthesis-dependent strand annealing pathway, producing essentially 'a stacked trait'.

Source: Figure 3 in Urnov, Fyodor D., Edward J. Rebar, et al. Gregory. "Genome Editing with Engineered Zinc Finger Nucleases." *Nature Reviews Genetics* 11, no. 9 (2010): 636-46.

Safety: While this approach sounds very good, there are still some big risks. In particular, even a low level of off-target activity could have very serious consequences by creating undesired “edits” somewhere else in the genome. Some of the strategies used to minimize these risks are (1) designing proteins with very long recognition sequences that will occur only once in the genome and (2) using a nuclease (FokI) that is only functional as a dimer.

A recurring theme in this unit will be understanding and designing specificity.

The controllable specificity of zinc finger proteins for DNA makes them attractive for genome editing. Similarly, the ability to redesign the protein-protein specificity of the FokI nuclease that allows one to design an enzyme that works only when two different versions of the protein, each recognizing half the desired target sequence, bind together.

Clinical trials for genome editing: Despite these risks, this approach has already been used in three clinical trials. In the first trial (NCT01082926), a T-cell line that is being used as a potential therapy for glioma has been engineered be resistant to glucocorticoids, which need to be administered as immunosuppressants. Two other trials are underway for HIV NCT00842634 and NCT01044654. In these studies, the CCR5 gene is deleted from the patients’ T-cells. If effective, these cells would be resistant to HIV and would protect the patient. Even further in the future, there is the potential make the patient’s stem cells CCR5-negative and repopulate the entire

immune system. This approach seems promising because a cancer patient who had HIV was recently apparently cured of HIV by a heterologous bone marrow transplant from a donor who naturally had a mutation in CCR5 that protects from HIV [12].

2. Inhibitors and activators of specific proteins. The ability to design protein-protein interactions is rapidly improving (reviewed in [13]). Recent successes include modifying a peptide to bind and inhibit an enzyme responsible for antibiotic resistance [14] and work by the Keating lab who designed peptides that could specifically bind to different members of set of structurally similar proteins (the bZIP family) [15].

In addition to looking at the specificity of interactions between macromolecules such as protein and DNA, we will spend considerable time looking at the specificity of proteins for small molecules. There is obviously great interest in the design of pharmaceutical compounds, such as kinase inhibitors, that bind with high affinity to proteins. We will examine the techniques for drug discovery in some depth.

3. There is also increasing interest in designing proteins to bind to small molecules to be biological sensors. Such sensors can have exquisite specificity for particular small molecules, providing an unparalleled tool for *in vivo* analysis. Alan Jasanoff's group [16] showed a very interesting example of this, although it was not done using structure-based design. Their goal was to produce a molecule that could respond to neurotransmitter release and provide a signal that could be detected by MRI. This would revolutionize our understanding of neurobiology, because we could study animals as they respond to stimuli and we could start to map out brain function at the cellular level. They started with a bacterial enzyme that contains a paramagnetic iron and binds arachidonic acid at the heme. They mutated it to bind dopamine with reasonable specificity and affinity and used it to monitor dopamine release *in vivo*.

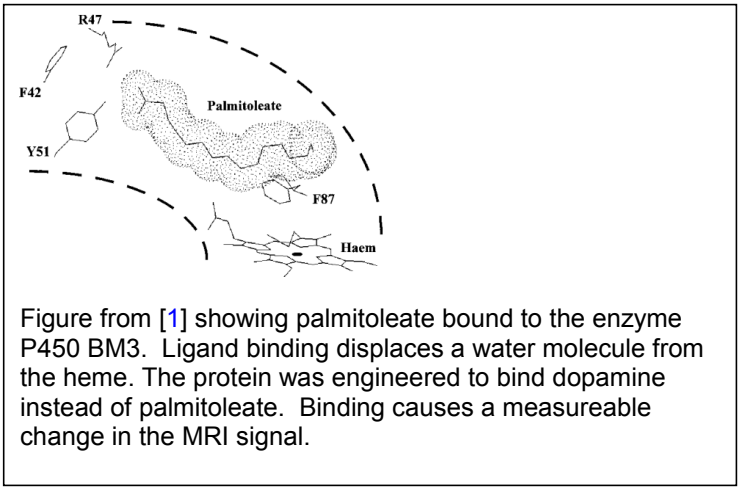


Figure from [1] showing palmitoleate bound to the enzyme P450 BM3. Ligand binding displaces a water molecule from the heme. The protein was engineered to bind dopamine instead of palmitoleate. Binding causes a measureable change in the MRI signal.

Source: Figure 1 in Noble, M., C. Miles, et al. "Roles of Key Active-site Residues in Flavocytochrome P450 BM3." *Biochem. J* 339 (1999): 371-9.

4. Sensors have also been designed to react to light. A number of domains have been identified that organisms use to sense light (reviewed in [17]). One of these is the PAS domain, which occurs in proteins that respond to light, chemical ligands and redox potentials. Because the PAS domain is connected to many different effector domains, it is presumed to function in a modular way that could allow it to be attached to other proteins to engineer new sensors [18]. The LOV domain [19], which is a type of PAS domain, has been used by several groups to make photo-responsive proteins. LOV domains are used by plants to control phototropic bending. They form a reversible covalent bond with a flavin mononucleotide in response to exposure to light.

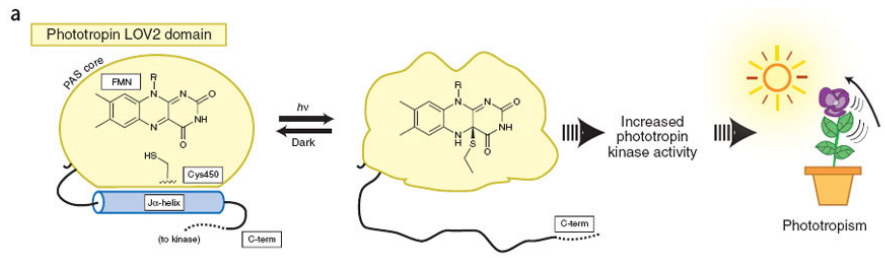


Figure from [4] showing the light-inducible conformational change in an LOV domain. Light causes the flavin cofactor to make a covalent bond with a cysteine residue in the protein, and this causes a conformational change.

Wu et al.[20] fused an LOV domain to the G-protein Rac1. After some mutations in the linker, they were able to get a light responsive change in conformation that allowed Rac1 to interact with the PAK effector only in the illuminated state. Upon illumination, the cell began to form lamellipodia at the site of the light. This construct allowed them to control the direction in which cells moved.

How general is approach? Wu et al.[20] were also able to fuse the domain with a different G-protein, Cdc42. While the initial construct did not work, modeling helped them identify mutations that produced a functional protein. Some parts of the modeling were done with Rosetta, which will be used in this course.

Related work by Yazawa et al. [21] used this same type of domain to create light-inducible protein-protein interactions. In addition to making a light-inducible signaling change that created lamellipodia, as in the previous example, they also created a light-inducible transcription factor. Light-repsonsive proteins have also been used to control neurons, an approach which is called optogenetics: This work has been pioneered by the Boyden lab[22] and the Deisseroth lab [23].

5. Enzyme design is another area with great potential. A number of approaches have been successful. In some cases, existing enzymes have been redesigned to catalyze new reactions. For example, Savile et al [24] were able to create an enzyme to improve the process for synthesizing the drug sitagliptin, an anti-

diabetic. The normal approach requires high-pressure hydrogen, toxic transition metals and results in poor stereoselectivity. They modified a transaminase ultimately obtaining an enzyme that carried out the desired reaction under practical conditions. The final enzyme had 10% better yield, 53% increase in productivity (kg/l per day) and 19% reduction in waste.

A more radical approach is to design a completely new enzyme. Rothlisberger et al. [25] designed an enzyme from scratch to catalyze the Kemp elimination, for which there is no known biological catalyst. The Kemp elimination is a ring-opening reaction that involves abstraction of a proton from carbon, which has a high activation barrier. They began by designing the ideal active site and then found a protein scaffold on which they could build these proteins. The modeling was done with Rosetta, and resulted in an enzyme with a modest rate enhancement.

Siegel, et al. [2] designed an enzyme for an even more challenging problem. Their synthetic enzyme catalyzes the Diels-Alder reaction: this reaction produces a cyclohexane ring, which is useful for many organic syntheses. The reaction is not known to be catalyzed by any biological enzyme. The previous reaction involved breaking a bond. This reaction requires carefully position two substrates to form a bond (See the figure below from their paper).

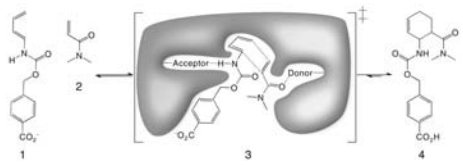
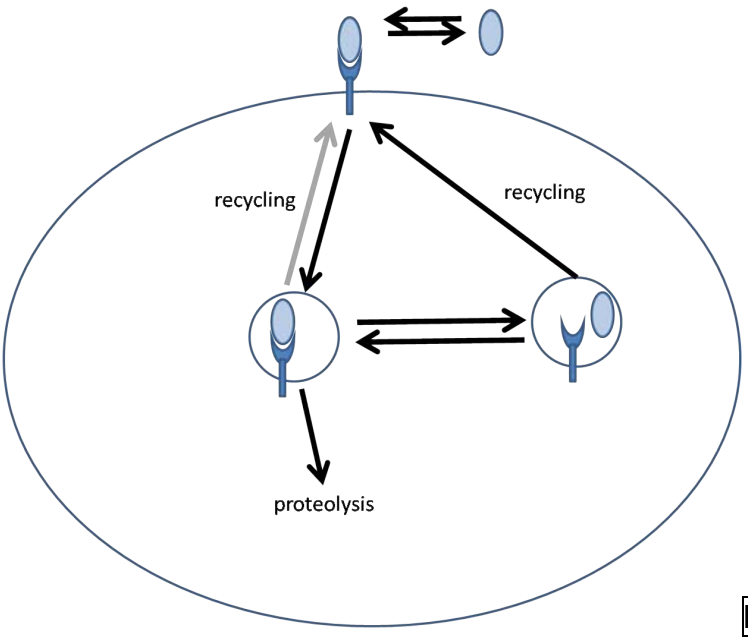


Fig. 1 from Siegel, et al. [2]. The Diels-Alder reaction. Diene (1) and dienophile (2) undergo a pericyclic [4 + 2] cycloaddition (3) to form a chiral cyclohexene ring (4). Also shown in (3) is a schematic of the design target active site, with hydrogen bond acceptor and donor groups activating the diene and dienophile and a complementary binding pocket holding the two substrates in an orientation optimal for catalysis.

6. Designing therapeutic proteins. A number of naturally occurring proteins are currently used to treat disease. The most common of these is probably insulin. In cases where the normal version of the protein is not ideal for therapeutic purposes, there is the potential to redesign the protein. An important example of this is GCSF, which is a protein used to stimulate bone marrow precursor cells in neutropenic patients.

GSF binds to a receptor, GCSFR, on the surface of these cells. If you want to make a more potent version of this drug you might expect that your best bet would be to make a version of GCSF that binds even tighter to the receptor than the wild-type. That turns out to be (1) hard and (2) counterproductive. It's hard because the wild-type binding has a K_d that is measured in pMoles! But it is also counterproductive. As you increase the affinity, you actually get less potency!

Differential equation-based modeling of the type you learned about in the first part of this course reveals why. In brief: the GCSF receptor doesn't just occur on the bone marrow precursors. It also occurs on the surface of **neutrophils**. These cells will internalize the protein and do one of two things: they either recycle the ligand and receptor to the cell surface or proteolytically degrade it.

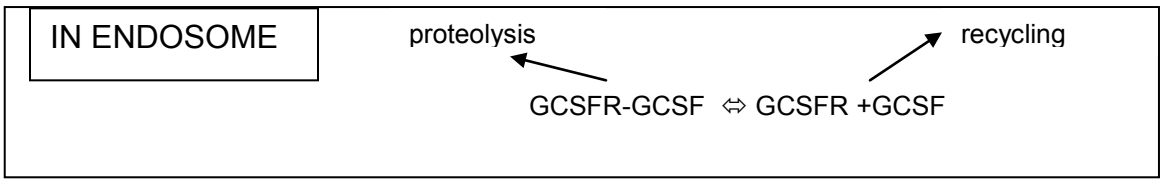


NEUTROPHIL

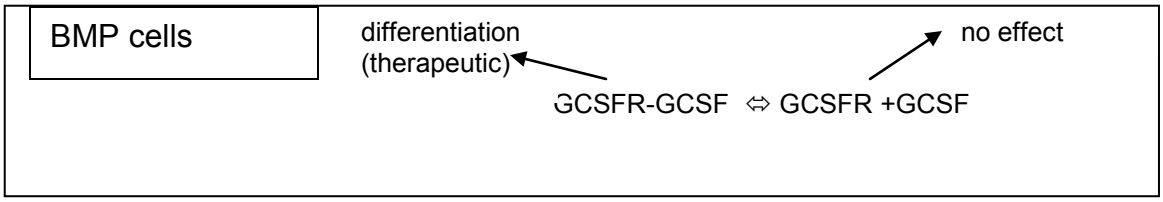
Let's look at the consequences. If the neutrophil proteolyses GCSF, the concentration of GCSF in circulation decreases and the drug's efficacy is lowered. If, instead, GCSF goes through the recycling pathway, it gets back to the circulation.

What control's this? It turns out that if the complex dissociates in the endosome, the GCSF gets recycled.

Let's focus on the endosomal binding:



At the same time, on the bone marrow precursor we have:



So we want to achieve two contradictory aims. In the endosome, we need to drive the equilibrium toward dissociation. On the surface we need to drive it toward association. The key to solving this problem is recognizing that there is a difference in pH between the outside of the cell and the endosome. It's about 7 on the surface and between 5 and 6 in the endosomes. So we want to find a mutation that leaves the affinity relatively constant at pH 7, but alters it at ~pH 5**. In this section of the course, you will learn how to design such mutations.**

Necessary Background. This unit will build off your existing knowledge of python and will assume a basic, but solid knowledge of probability. If you feel in need of a refresher course on probability and statistics, please look at the material posted on the course website.

References

1. Noble, M.A., et al., *Roles of key active-site residues in flavocytochrome P450 BM3.* Biochem J, 1999. **339 (Pt 2)**: p. 371-9.
2. Siegel, J.B., et al., *Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction.* Science, 2010. **329**(5989): p. 309-13.
3. Urnov, F.D., et al., *Genome editing with engineered zinc finger nucleases.* Nat Rev Genet, 2010. **11**(9): p. 636-46.
4. Ko, W.H., A.I. Nash, and K.H. Gardner, *A LOVely view of blue light photosensing.* Nat Chem Biol, 2007. **3**(7): p. 372-4.
5. Galperin, M.Y. and E.V. Koonin, *From complete genome sequence to 'complete' understanding?* Trends Biotechnol, 2010. **28**(8): p. 398-406.
6. Hirschhorn, J.N. and Z.K. Gajdos, *Genome-wide association studies: results from the first few years and potential implications for clinical medicine.* Annu Rev Med, 2011. **62**: p. 11-24.
7. Klug, A., *The discovery of zinc fingers and their applications in gene regulation and genome manipulation.* Annu Rev Biochem, 2010. **79**: p. 213-31.
8. Choo, Y., I. Sanchez-Garcia, and A. Klug, *In vivo repression by a site-specific DNA-binding protein designed against an oncogenic sequence.* Nature, 1994. **372**(6507): p. 642-5.
9. Sun, N., et al., *Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease.* Mol Biosyst, 2012. **8**(4): p. 1255-63.
10. Li, T., et al., *Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes.* Nucleic Acids Res, 2011. **39**(14): p. 6315-25.

11. Christian, M., et al., *Targeting DNA double-strand breaks with TAL effector nucleases.* Genetics, 2010. **186**(2): p. 757-61.
12. Hutter, G., et al., *Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation.* N Engl J Med, 2009. **360**(7): p. 692-8.
13. Mandell, D.J. and T. Kortemme, *Computer-aided design of functional protein interactions.* Nat Chem Biol, 2009. **5**(11): p. 797-807.
14. Reynolds, K.A., et al., *Computational redesign of the SHV-1 beta-lactamase/beta-lactamase inhibitor protein interface.* J Mol Biol, 2008. **382**(5): p. 1265-75.
15. Grigoryan, G., A.W. Reinke, and A.E. Keating, *Design of protein-interaction specificity gives selective bZIP-binding peptides.* Nature, 2009. **458**(7240): p. 859-64.
16. Shapiro, M.G., et al., *Directed evolution of a magnetic resonance imaging contrast agent for noninvasive imaging of dopamine.* Nat Biotechnol, 2010. **28**(3): p. 264-70.
17. Moglich, A. and K. Moffat, *Engineered photoreceptors as novel optogenetic tools.* Photochem Photobiol Sci, 2010. **9**(10): p. 1286-300.
18. Moglich, A., R.A. Ayers, and K. Moffat, *Design and signaling mechanism of light-regulated histidine kinases.* J Mol Biol, 2009. **385**(5): p. 1433-44.
19. Crosson, S., S. Rajagopal, and K. Moffat, *The LOV domain family: photoresponsive signaling modules coupled to diverse output domains.* Biochemistry, 2003. **42**(1): p. 2-10.
20. Wu, Y.I., et al., *A genetically encoded photoactivatable Rac controls the motility of living cells.* Nature, 2009. **461**(7260): p. 104-8.
21. Yazawa, M., et al., *Induction of protein-protein interactions in live cells using light.* Nat Biotechnol, 2009. **27**(10): p. 941-5.
22. Chow, B.Y., et al., *High-performance genetically targetable optical neural silencing by light-driven proton pumps.* Nature, 2010. **463**(7277): p. 98-102.
23. Gradinaru, V., et al., *Molecular and cellular approaches for diversifying and extending optogenetics.* Cell, 2010. **141**(1): p. 154-65.
24. Savile, C.K., et al., *Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture.* Science, 2010. **329**(5989): p. 305-9.
25. Rothlisberger, D., et al., *Kemp elimination catalysts by computational enzyme design.* Nature, 2008. **453**(7192): p. 190-5.

20.320 Analysis of Biomolecular and Cellular Systems
Fall 2012