Patrick Caveney

May 5, 2013

Final Report

High Throughput Method for Biosensor Substrate Screening

Protein sensors are the third largest class of proteins encoded in bacterial genomes after metabolic enzymes (~80%) and transport proteins (~10%) [Zhulin 2013]. Even though there are many sensors, an estimated 25,000-40,000 in already sequenced genomes, only about 40 have been characterized. This means that synthetically modified cells are currently limited to sensing only 0.1% of substrates they could be sensing. This leaves a hole in the cycle of self regulation of cells and gives the cells no knowledge of their own performance in the same way that a cook without taste buds could not judge his own work. Fusion proteins would enable a high throughput way of classifying new sensing domains that could open up new possibilities to synthetic biologists.

Proteins are chains of amino acid residues. Proteins can be classified into three divisions, regions, domains, and motifs. A region is a short segment of continuous amino acids. A domain is a segment of continuous amino acids that has an independent fold and thus independent function. Motifs are amino acids that interact in three dimensions. The residues involved may or may not be continuous in the chain. Of the three, the most relevant to this essay are domains. A single protein may have one or more domains which are connected by linker regions of residues. Because the domains have independent form and function they can be shuffled around and connected in novel ways. This is a proposed method of evolution, by

randomly shuffling domains new proteins can be evolved.

This bottleneck in sensors is wonderfully demonstrated in Jarred Callura and Collins et al.'s paper Genetic Switchboard for Synthetic Biology Applications [Collins 2013]. The orthogonal system they design is very elegant in its application and design. However, even given the large number of orthogonal systems that can be build with their technique they are limited by the number of sensors that control each system. Each path is controlled by only two sensors. Thus, no more than n ^ 2 paths can be orthogonally created, where n is the number of sensors available. This is a large number, even with only 40 known biosensors (1600 total orthogonal paths), but it is not as large as the orthogonality bestowed by the RNA switches. Each switch has 20 base pairs of homology which would be 4 ^ 20 unique switches (10^12 orthogonal switches).

Protein sensors are important to the function of the cell and its ability to find resources and avoid toxins. Escherichia coli has five chemotaxis sensors, Tar, Trg, Tsr, Tap, Aer for sensing, aspartate and maltose, ribose and galactose, serine, dipeptides and pyrimidines, and oxygen, respectively [Parkinson Lab 2013]. The standard sensors have a sensing domain that can bind the substrate the protein senses. This usually causes a conformational change in the protein which activates a kinase domain. The kinase phosphorylates, sometimes with the help of ATP [Cozzone 1993], which in turn phosphorylates a transcription factor. The transcription factor binds a promoter or repressor domain and activates or represses a gene or set of genes.

Fusion Proteins

Fusion proteins or chimera proteins are created using the knowledge that domains are independent of the rest of the protein they belong to. Fusion proteins can be created by merging the ORF of two proteins and removing the stop codon between them. This is easily done by gibson assembly or overlap PCR with the respective DNA sequences. The technique of fusing proteins has been used to purify proteins to understand their function. An unknown protein can be fused with a domain that exhibits a known affinity to isolate the fusion protein in an affinity chromatography column.

There are two broad classes of protein sensors, transmembrane and cytoplasmic. Transmembrane sensors tell the cell about the environment just outside the cell membrane. A common example is Tar in E. coli. The periplasmic domain contains the sensing domain, then there is a transmembrane region, then the cytoplasmic kinase domain that relays the signal [Nishiyama 2010]. Transmembrane sensors are convenient because the transmembrane domain provides a location to merge the sensor and the kinase. The transmembrane region can be take from either protein but the point is there are only two options to test.

Cytoplasmic sensors are more difficult to test because their sensing and kinase domains are connected by linker chains of amino acids and it unclear which amino acids are important to the function of the sensor.

The mechanism by which kinases transmit signals is not fully understood. In histidine kinases it is believed that when the substrate binds a conformational change exposes an ATP binding site. ATP then binds and phosphorylates a conserved histidine residue [Marina 2005**].** It is important to understand the mechanism of different families of kinases and sensor domains because they

may not work as well or at all if the sensing and kinase domains from different families are crossed.

A popular example of fusion proteins is Tar-EnvZ [Yoshida 2007]. This is a fusion of the Tar receptor to the EnvZ kinase domain, both of which are native to E. coli. The key development of this fusion is the fact that Tar is a chemoreceptor sensing aspartate and EnvZ senses to osmolarity. Thus, a cell with a Tar-EnvZ fusion protein will respond to aspartate as it would respond to an osmotic change. The fusion protein was transmembrane so the transmembrane region from Tar or from EnvZ could be tested.

A similar technique was used by John Dueber and Jay Keasling et al. to create scaffolds that link metabolic enzymes for higher throughput of reactants and intermediaries. [Keasling 2008]. They added what are essentially linker regions between the proteins This technique and application are very new. The result is enzymes which are sequentially utilized in metabolic pathways can be positioned near each other to increase the titer. Dueber et al. achieved a 77-fold increase in mevalonate titer with this method.

Proposed Process

Fusion proteins (chimeric proteins) can be assembled with a random chimerization. The process is called RACHITT [Pelletier 2001]. The first step creates ssDNA with an exonuclease. The ssDNA is then cut with DNase 1, creating short strands. The pieces can be size fractionated to favor small segments, but considering that most of the final gene (the sensing and kinase domains) will match the scaffold this step may not be necessary. RACHITT is

typically used to recombine numerous homologous genes, but it can be adapted to create a random fusion between two domains. The main difference is in the scaffold. In RACHITT, the scaffold is a length of ssRNA with similarity to the genes (it is key that the scaffold is not identical to any of the genes). In the proposed, modified scaffold would be composed of the identical sensing domain at the N terminus and the identical kinase domain at the C terminus. The linker region would be composed of a mix of amino acids from each protein. To create the scaffold, the two regions can be assembled with gibson assembly. The linker region sequence can be specified with the primers. The linker region is often 20 residues, 60 base pairs and the current primer limit, so longer primers will need to be used than are currently available. After gibson assembly the sensor-kinase sequences can be amplified with PCR, substituting uracil for thymine to facilitate future steps. If the conditions of hybridization during RACHITT are favorable to lower stringency, such as low temperature and higher salt concentration, various pieces of each protein should hybridize with the linker region even though they have low homology. Once the hybridization occurs the ssRNA is removed with an exoribonuclease and the pieces are amplified with PCR. RACHITT produces an average of 14 recombination events. This is not an exorbitant amount, but it is more than could be achieved with different primers and faster and more random than point mutations. From there the mutants can be isolated and tested.

It would be ideal to test both the substrates and the chimerization, so the mutants should be isolated and tested individually. Single copies of the gene can be attached to beads. The beads can then be placed in wells and PCR can amplify each mutant individually, like in emPCR [Ziebolz 2007]. The resulting copies can be divided into new wells. Now each substrate can be tested against each mutant. Crude lysate should be added to the wells. Then the wells should be doped with the substrates to be tested. Positive results can be easily isolated because of the

fluorescent reporter gene. Wells with positive results can then be sequenced and further studied, and its response to substrates can be tested against the natural sensor.

The proposed final process is below.

1. Isolate the nucleotide sequence coding for the sensing domain of the protein.
2. Randomly create chimeras with the linker region between the sensing domain and the standard kinase domain with RACHITT
3. Isolate mutants of the chimeric sensors in wells.
4. PCR each well.
5. Divide the resulting copies into new individual wells.
6. Add crude lysate to the wells.
7. Dope one well of each mutant with each substrate thought to be important to the original organism.
8. Observe the fluorescence of each well.
9. Note the wells with large positive results.
10. Screen the small population of wells with positive results and check the associated substrates with the original sensor through traditional methods.

Challenges

There are six main challenges in the proposed method for rapidly discovering the substrates of sensors. First, where should the sensing and kinase domains be connected? Second, how can the response of the chimera be ensured to respond like the natural sensor? Third, what is the

standard module that the sensing domains will be tested on? Fourth, the cost of cell free systems. Fifth, a large number of substrates need to be tested against each mutant. Sixth, the sensor reaction to the substrate needs to be reproducible.

First, where should the two domains be joined? Without the high levels of conservation observed in other genes and proteins, such as 16S rRNA, it is not possible to know which amino acid residues are important to the function of either the sensing domain or the kinase domain. Compounding the lack of conservation is a lack of knowledge about signal transduction. With only 40 sensors categorized it is hard to determine the conserved residues. Because of this, I would recommend pursuing random chimeragenesis. This way, the ideal merge between the two domains can be elucidated while the sensors are being tested for their substrates.

Second, if the fusion proteins are created by random chimeragenesis how can we be sure the fusion protein responds to the same substrate in the same way as the original sensing protein? Such was the case with FixL [Kumita 2003]. In this study Hideyuki Kumita et al. made five different chimeric proteins with FixL (an O_2 sensing domain) with the kinase from a hyperthermophile, *Thermotoga maritima*. The "UV–visible, resonance Raman, and circular dichroism spectral characteristics" of the five chimeras were determined to be identical to the original; however, "it was likely that the autophosphorylation activity of the histidine kinase domain could not be regulated by the O_2 association/dissociation in the sensor domain in the chimeric proteins." Even though they tried fusion at five locations and had an identical measurements, the chimeric proteins did not respond to O_2 like the original protein. One way of ensuring the two sensors (natural and synthetic) work in the same manner is to screen a large number of substrates with the fusion protein, because it is easier to see working with its fluorescent

reporter gene, and then only screen the positives on the original protein. In this way, the proposed process would be a quick and rough survey of substrates to narrow the focus of researchers to a smaller population of substrates.

Third, is the standard module that each sensing domain is attached to. This domain should be two different modules, one for transmembrane sensors and the other for cytoplasmic sensors. The kinases chosen as the standard module should be well characterized. The reporter gene should be fluorescent to aid in the characterization because flow cytometry is a quick reliable way to search through a large number of samples. EnvZ has been widely studied [UniProt 2013] and the 3D structure has been resolved. Thus the EnvZ / OmpR structure is proposed as the chassis for transmembrane sensors. Ideally, the sensing domain being tested will be fused with the EnvZ kinase domain. When the sensor is activated by its substrate it will cause a conformational change in EnvZ which phosphorylated OmpR. OmpR, a transcriptional regulator, can then bind to the OmpC promoter region and activate transcription of GFP. In the same manner, TodS and its associated TodT transcriptional regulator and TodX promoter can be used as a chassis for cytoplasmic sensors.

Fourth, cell free biology is more expensive than traditional biology because the proper transcriptional and translational machinery must be added in addition to free nucleotides and amino acids. This can be offset by the speed of cell free systems. Kim et al. estimate cell free biology accomplish in hours what traditional methods accomplish in days or weeks [Kim 2011]. In vitro, cell free, results and in vivo results tend to differ, as seen in medical trials or toxicity tests, but they can still provide a general idea of a system.

Fifth, in the current procedure a large number of mutant chimeras need to be tested against a large number of substrates. A few things can be used to cut down on this burden. First, this process can be automated with processes similar emPCR like in current DNA sequencing technologies [Ziebolz 2007]. Second, the environment from which the sensor originated can be sampled with GCMS to determine the substrates present. This will narrow the number of substrates that need to be tested. Third, substrate specificity is not exact in protein sensors [Goldstein 2006]. “Remarkably, cholesterol is only a weak inducer of this degradation. The cholesterol precursor, lanosterol, is much more potent” Thus, a number of unique substrates can be chosen that have varying structure that cover a range of structural families.

Sixth, ensuring the chimera construct is correct and at the same time testing for the substrate. This is the biggest challenge and one that leads to the large number of samples needing to be tested. Because it is not known how to create working fusion proteins de novo, each mutant needs to be tested against each substrate. This problem can be slightly mitigated by using microfluidics and small testing chambers, but the ultimate goal would be to effectively predict where to join the sending domain and the kinase domain. This problem could possible be solved with bioinformatics, but the information needed could be provided by the proposed procedure. If we have a library of natural or synthetic sensors with the associated information about the kinase used and the substrate this knowledge can be used to predict conservation and important residues in any protein sensors.

Future Applications

The possible applications of such a screening technique are broad and varied. The most

important would be categorizing new sensors with high throughput testing. The benefits of such a system are that uses high throughput screening to narrow the possible substrate candidates of a sensing domain. The same technique can be used to create novel synthetic sensors for synthetic biology applications. Knowledge about natural and synthetic sensors will contribute to knowledge about conserved residues and thus the process of signal transduction.

Even if the chimeric sensor responds to different substrates than the chimera can still be used as a biosensor with a known response to a specific substrate. This substrate - sensor combination can then be used in synthetic applications, but it cannot help predict the environment of the original organism.

Bioinformatics is a powerful tool to predict new genes and proteins from existing ones. However, bioinformatics can only compare new genes and proteins to ones that are currently known. Thus, by characterizing a large portion of sensors with unknown responses, bioinformatics can possibly start predicting the rest of the unknown biosensors by homology and conservation.

Another possible application is understanding the minimal environments of new bacteria a priori. Many bacteria cannot be cultured because it is unknown what is needed for the bacteria to survive. E. coli can grow on agar and glucose and a few salts, but not every bacteria can. This limits our ability to study such bacteria. If we knew the sensors in the bacteria by reading their genome we could predict the substrates they sense and thus the environment that matters to them. For example, E. coli are attracted to amino acids, sugars, and oxygen, all of which they have sensors for.

The proposed system would help in the creation of a synthetic feedback loop or controller of sorts for cellular processes. With the limited sensors currently available, not every process can be controlled directly. If new sensors are discovered any metabolite or other substrate can be sensed and thus controlled.

The proposed solution has many benefits and while it has many challenges, they can be overcome with current technology. The proposed procedure provides a high throughput method for testing substrates against sensors. It uses a standard chassis with an easy to observe GFP reporter gene for testing each sensor. The procedure also accounts for variation in the linker region joining the sensing domain with the standard kinase domain. The whole system is cell free to improve processing speeds. This procedure will be interesting to research and implement and its potential impact is large.

Resources

Jarred M. Callura, Charles R. Cantor, and James J. Collins. “Genetic switchboard for synthetic biology applications.” *PNAS.* 109 (2012): 5850-5855 accessed May 3, 2013. doi: 10.1073/pnas.1203808109

Alain J. Cozzone “ATP-dependent protein kinases in bacteria.” *Journal of Cellular Biochemistry.* 51 (1993): 7-13 accessed April 30, 2013. doi: 10.1002/jcb.240510103

John E Dueber, Gabriel C Wu, and Jay D Keasling. “Synthetic Protein Scaffolds Provide Modular Control Over Metabolic Flux.” *Nature Biotechnology.* 27 (2008): 753-759 accessed May 3, 2013. doi: 10.1038/nbt.1557

Joseph L. Goldstein, Russell A. DeBose-Boyd, and Michael S. Brown “Protein Sensors for Membrane Sterols.” *Cell*. 124 (2006): 35-46 accessed April 30, 2013. doi: 10.1016/j.cell.2005.12.022

Dr. Tae-Wan Kim, Dr. Harshal A. Chokhawala, and Prof. Douglas S. Clark. “High-Throughput In Vitro Glycoside Hydrolase (HIGH) Screening for Enzyme Discovery.” *Angewandte Chemie International Edition.* 50 (2011): 11215-11218 accessed May 2, 2013. doi: 10.1002/anie.201104685

Hideyuki Kumita, Seiji Yamada, and Yoshitsugu Shiro. “Chimeric sensory kinases containing O2

sensor domain of FixL and histidine kinase domain from thermophile." *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics.* 1646 (2003): 136-144 accessed April 30, 2013. doi: 10.1016/S1570-9639(02)00555-1

Alberto Marina, Carey D Waldburger, and Wayne A Hendrickson. "Structure of the entire cytoplasmic portion of a sensor histidine-kinase protein." *The EMBO Journal.* 24 (2005): 4247-4259 accessed April 30, 2013. doi: 10.1038/sj.emboj.7600886

So-ichiro Nishiyama, Shinji Ohno, and Ikuro Kawagishi. "Thermosensing Function of the Escherichia coli Redox Sensor Aer." *Journal of Bacteriology.* 192 (2010): 1740-1743 accessed May 2, 2013. doi: 10.1128/JB.01219-09

Parkinson Lab. "An overview of E. coli chemotaxis." Department of Biology, University of Utah. Accessed May 2, 2013. http://chemotaxis.biology.utah.edu/Parkinson_Lab/projects/ecolichemotaxis/ecolichemotaxis.html

Joelle N. Pelletier. "A RACHITT for our toolbox." *Nature Biotechnology.* 19 (2001): 314 - 315 accessed May 3, 2013. doi:10.1038/86681

UniProt. "P0AEJ4 (ENVZ_ECOLI) Osmolarity sensor protein EnvZ." Last Modified May 1, 2013. Accessed May 2, 2013. http://www.uniprot.org/uniprot/P0AEJ4

Takeshi Yoshida, Sangita Phadtare, and Masayori Inouye. "The Design and Development of

Tar-EnvZ Chimeric Receptors." *Methods in Enzymology.* 423 (2007): 166-183 accessed April 30, 2013. doi:10.1016/S0076-6879(07)23007-1

Burkhard Ziebolz and Marcus Droege. "Toward a new era in sequencing." *Biotechnology Annual Review.* 13 (2007): 1-26. accessed April 30, 2013. doi: 10.1016/S1387-2656(07)13001-5

Igor Zhulin. "Whole Genome Reconstruction and Analysis." Lecture. University of Tennessee. April 28, 2013.