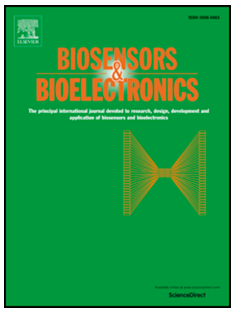
Journal Pre-proof



Machine learning linked evolutionary biosensor array for highly sensitive and specific molecular identification

Haseong Kim, Wonjae Seong, Eugene Rha, Hyewon Lee, Dae-Hee Lee, Seong Keun Kim, Kil Koang Kwon, Kwang-Hyun Park, Seung-Goo Lee

PII: S0956-5663(20)30659-X

DOI: https://doi.org/10.1016/j.bios.2020.112670

Reference: BIOS 112670

To appear in: *Biosensors and Bioelectronics*

Received Date: 16 July 2020

Revised Date: 26 September 2020

Accepted Date: 30 September 2020

Please cite this article as: Kim, H., Seong, W., Rha, E., Lee, H., Lee, D.-H., Kim, S.K., Kwon, K.K., Park, K.-H., Lee, S.-G., Machine learning linked evolutionary biosensor array for highly sensitive and specific molecular identification, *Biosensors and Bioelectronics* (2020), doi: https://doi.org/10.1016/j.bios.2020.112670.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2020 Published by Elsevier B.V.

Journal Pre-proof

CRediT author statement

Haseong Kim: Study design, Biosensor experiments, ML algorithm analysis, Writing-Rewriting, Editing, and Supervision, **Wonjae Seong** and **Eugene Rha**: Mutant library construction, and Writing. **Hyewon Lee**: Array data experiments. **Dae-Hee Lee** and **Seong Keun Kim**: Cell fluorescence data analysis, Revision, Editing. **Kil Koang Kwon**: Cell sorting, Biosensor data analysis. **Kwang-Hyun Park**: DmpR structure analysis. **Seung-Goo Lee**: Study design, Validation, and Supervision

Journal Pre-proof

Journal Pre-proof

Machine learning linked evolutionary biosensor array for highly sensitive and specific molecular identification

Haseong Kim[1, 3*], Wonjae Seong[1, 3], Eugene Rha[1], Hyewon Lee[1], Dae-Hee Lee[1, 3], Seong Keun Kim[1], Kil Koang Kwon[1], Kwang-Hyun Park[2, 3], and Seung-Goo Lee[1, 3*]

[1] Synthetic Biology and Bioengineering Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 34141, Republic of Korea
[2]Disease Target Structure Research Center, KRIBB, Daejeon 34141, Republic of Korea
[3]Department of Biosystems and Bioengineering, KRIBB School of Biotechnology, University of Science and Technology, Daejeon 34113, Republic of Korea

* To whom correspondence should be addressed. Tel: +82-42-860-4372; Fax: +82-42-860-4379; Email: haseong@kribb.re.kr, sglee@kribb.re.kr

ABSTRACT

Bacteria initiate complicated signaling cascades from the detection of intracellular metabolites or exogenous substances by hundreds of transcription factors, which have been widely investigated as genetically-encoded biosensors for molecular recognition. However, the limited number of transcription factors and their broad substrate specificity results in ambiguity in small molecule identification. This study presents a new small molecule fingerprinting technique using evolutionary biosensor arrays with a machine learning technique that can capture highly specific substrate signals. Employing multiple mutant transcription factors derived from a single transcription factor has effectively circumvented the limited availability of transcription factors induced by a small molecule of our interest. This method achieved up to 95.3% true positive rate for identifying small molecules, and the high-resolution protein engineering technique improved the limit of detection 75-fold. The signal trade-offs with background noises caused by the complex cellular biochemistry of mutant transcription factors enable the biosensor arrays to be more informative in terms of statistical variance. The machine-learning technology, coupled with the single transcription factor-driven evolutionary biosensor array, will open new avenues for molecular fingerprinting technologies.

Keywords

Genetically-encoded biosensors; Machine learning; Molecular identification; Protein engineering; Biosensor array; High-throughput screening system

Journal Pre-proof

1. INTRODUCTION

Bacteria initiate complicated signaling cascades for meeting metabolic requirements or identifying exogenous substances using hundreds of transcription factors (TFs). These TFs have been widely investigated as genetically-encoded biosensors over the past decade for molecular recognition (Bucur et al., 2006; Fernandez-López et al., 2015; Zhang and Keasling, 2011) as alternatives to resource-intensive mechanochemical monitoring systems (Beyer and Clausen-Schaumann, 2005; Juribašić et al., 2014; Mandal et al., 2015). Recent advances in synthetic biology enabled the development of a system capable of tightly controlling TF-based biosensors (Meyer et al., 2018; Rogers et al., 2016; Wan et al., 2018). The applications of these biosensor technologies are rapidly increasing that currently includes environmental detection (Chong and Ching, 2016; Jha et al., 2016), high-throughput screening for enzymes (Choi et al., 2014; Kwon et al., 2018), and genetic microbial prototyping (Kim et al., 2016). However, TFs of the biosensors activated by small molecules of interest are usually unavailable. Protein engineering approaches can be applied to alter the substrate specificity (Bucur et al., 2006; Jha et al., 2016; Ray et al., 2016; Yeom et al., 2018), but it could induce biosensor basal noise or false positives, which prevents identification of target molecules, especially in complex and unknown environmental samples. The stochastic and unpredictable molecular interactions in a complex mixture could hinder the fluorescence or luminescence mediated reporting signals. To improve the analytical capabilities of biosensors, it is advisable to use multiple biosensors as in the field of chemical electronics (Dickert et al., 1999; Elad and Belkin, 2013; Psuj, 2018). Cells with different promoters were used as an array for identifying environmentally toxic chemicals (Jin et al., 2005). But TF availability is still problematic, since multiple TFs induced by the specific target molecules are required for the multiple biosensors. In this study, all the engineered TFs derived from a single TF were employed to tackle the limitation of TF availability. This approach takes advantage of the protein engineering problems conversely by generating multiple high-variance signals that enable a machine learning (ML) algorithm to discriminate small molecules effectively.

As proof of concept, the present study demonstrates that coupling ML methods with a set of mutant biosensors significantly improves biosensor performance for identifying phenolic small molecules. There are several inexpensive and easy-to-operate phenol detection methods utilizing colorimetric reagents such as gold nanoparticles and quantum dots (Alkasir et al., 2012; Arciuli et al., 2013; Xue et al., 2018). But regardless of the detection performance of these methods, none of them identify the chemicals triggering the detection mechanism, due to the lack of substrate specificity of the enzyme-based detectors. Only conventional chromatographic methods with mass spectrometry can be used for the detection and identification of chemicals, including phenolic compounds (Moldoveanu and Kiser, 2007). These methods, coupled with appropriate pre-treatment and extraction techniques, have Limit of detection (LOD) of approximately 0.01 ppb for phenolic compounds and linearity between 5 to 10000 ppb. However, these methods require expensive equipment and specialists in operating the devices.

The biosensor array, in this study, uses *Pseudomonas*-derived DmpR as a key TF. DmpR comprises 563 amino acids, and undergoes transcription in response to phenolic compounds at 1–

Journal Pre-proof

100-µM concentrations (Pavel et al., 1994). DmpR has been studied for monitoring hazardous environmental phenolic chemicals (Chong and Ching, 2016; Shin et al., 2005; Xue et al., 2014). However, it could be activated by various analogs, potentially leading to false-positives in biosensor applications (Shingler and Moore, 1994; Wise and Kuske, 2000). Ancillary effectors, leading to basal noise, have been reported while investigating DmpR mutants (Gupta et al., 2012; O'Neill et al., 1999; Wise and Kuske, 2000), even when the mutant residues were not directly involved in the effector binding site. This suggests that improving the sensitivity or the LOD of a TF could be a trade-off with broadening specificity. A good example is the well-studied DmpR E135K mutation related to the ATPase activity of DmpR (Pavel et al., 1994). The mutant displayed a better LOD than wild-type (WT) DmpR and had an improved response to 4-nitrophenol (Choi et al., 2014). It has been proposed that the E135K mutation induces changes in the ATPase-triggering mechanism, which causes 4-nitrophenol to act against the DmpR TF (Kwon et al., 2018). In this study, we developed an evolution-driven biosensor array with 6 TFs (WT and five variants) engineered using high-throughput genetic enzyme screening system (Choi et al., 2014) and next-generation sequencing. We used a random forest algorithm, an ensemble ML method with large-scale simulated response data of the biosensor array, for small molecule identification.

2. MATERIAL AND METHODS

Strains and media

All the strains and plasmids used in this study are displayed in Table S1. Cloning and plasmid propagation was carried out in *Escherichia coli* (*E. coli*) DH5α (fhuA2 Δ(argF-lacZ)U169 phoA glnV44 Φ80 Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17; NEB #2987*), E. coli* JM109 (endA1, recA1, gyrA96, thi, hsdR17 (rk–, mk+), relA1, supE44, Δ(lac-proAB), [F´ *tra*D36, *pro*AB, *laq*IqZΔM15])*, and E. coli* MG1655 *(*K-12 F– λ– ilvG– rfb-50 rph-1*).* DmpR engineering and its activity testing were conducted in *E. coli* HK41 *(*K-12 F– λ– ilvG– rfb-50 rph-1 bglA– pO-eGFP:kmR). The cells were grown in LB medium (10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl) and M9-glucose medium (4% glucose, Na2HPO4 6.78 g/L, KH2PO4 3 g/L, NaCl 1 g/L, 2 mM MgSO4, 0.1 mM CaCl2, and 0.01% thiamine) with appropriate antibiotic supplementation (100 µg/mL ampicillin, 34 µg/mL chloramphenicol, and 25 µg/mL kanamycin), for strain maintenance and plasmid construction in *E. coli* strains.

Materials

All chemical reagents used in this study and *E. coli* MG1655 cells were purchased from Sigma-Aldrich, except for LB broth from Miller BD Difco™ (New Jersey, USA). Restriction endonucleases, Gibson assembly cloning kits, T4 PNK, T4 DNA ligase, pUC19, and *E. coli* DH5α cells were purchased from New England Biolabs (Beverly, MA, USA). PCR Amplification kits were purchased from TaKaRa (Shiga, Japan). TA cloning kits, plasmid DNA isolation, and DNA extraction were performed using plasmid preparation kits from Promega (Madison, WI, USA). Synthesis of oligonucleotides and

Journal Pre-proof

sequencing were carried out by Macrogen (Daejeon, Korea). JM109, Site-Directed Mutagenesis Kit was purchased from Agilent (Santa Clara, USA). The random mutagenesis kit was purchased from Stratagene (La Jolla, CA, USA).

Construction of HK41 reporter strain

An HK41 strain in which only the report module of the genetic enzyme screening system (GESS) (Choi et al., 2014) was inserted into the bglA position of *E. coli* chromosome was prepared by λ Red recombination. The ΔbglA::Km region containing flippase recognition target (FRT) site was amplified by PCR in the ΔbglA::Km strain of the KEIO collection (background strain: BW25113) and cloned by TA cloning as a vector for insertion into the chromosome. To construct the cassette including the ΔbglA::Km region and reporter module, the cloned vector and pGESSv4 were amplified with each primer pair (Rep_BF, Rep_BR for the vector and Rep_IF, Rep_IR for pGESSv4). Note that pGESSv4 plasmids constructed in our previous work(Choi et al., 2014) were built on the backbone of the pUC19 plasmid which contains the mutant pMB1 origin of replication and ampicillin resistance marker. All the primers are listed in Table S2. The amplified sections were ligated by the Gibson assembly cloning kit. After sequence confirmation, the constructed cassette was amplified using primer pairs (bglA_300up and bglA_300dn). Then, the amplified linear DNA was purified and transformed into *E. coli* MG1655 expressing λ-Red recombinase for homologous recombination (Datsenko and Wanner, 2000). After verification of chromosomal integration by kanamycin selection and PCR, colonies were incubated at 42 °C for 24 h to remove the recombinase-containing plasmid with a temperature-sensitive origin of replication (Cherepanov and Wackernagel, 1995).

Construction of the DmpR mutant library

A random mutagenesis library of *dmpR* was constructed by error-prone PCR targeting the 1-630-bp section of the A domain, which is known to include the substrate-binding site. For the random mutagenesis of *dmpR*, mutagenic PCR was performed (EP-Ins-F1 and EP-Ins-R1) using the Genemorph II Random Mutagenesis Kit according to the manufacturer's instructions. The random PCR product was purified and ligated to another PCR product using pSHCE-dmpR plasmid as a template (EP-vec-F1 and EP-vec-R1) via Gibson assembly kits. pSHCE-dmpR and pSHCE-dmpR_M52I plasmids were constructed on the backbone of pSHCE containing a p15A origin of replication and chloramphenicol resistance marker. The DmpR genes were amplified by PCR using pGESSv4 (for the first stage of engineering) and pGESSv4_M52I (for the second stage), and the resulting fragments were ligated using the Gibson assembly kit. To measure library size, Gibson's product was transformed into DH5α strain. Then, all colonies were counted, and a select few were confirmed by Sanger sequencing. After sequencing confirmation, the library size was measured by the mutant ratio (mutated colony/selected colony) × total colony size (colony number / Gibson product (μg)). Then, the plasmid-containing mutant *dmpR* library was transformed into the HK41 to create the 1st DmpR mutant library (library size: 1×10^6) for the first stage of the screening. The 2nd DmpR

mutant library for the second-stage screening was generated using the same protocol with pSHCE-dmpR_M52I, which contained the M52I mutation of DmpR.

High-throughput DmpR mutant screening with FACS

The HK41 cells harboring DmpR mutant library were inoculated at an initial OD_{600} of 0.5~0.8 in 1 mL of M9-glucose and LB supplemented with 25 μg/mL chloramphenicol for the first and second DmpR mutant library, respectively. The cells were incubated at 37 °C for 1 h in a shaking incubator, following the addition of the appropriate concentration of phenol. The cells were incubated at 37 °C for 3 h to initiate the expression of fluorescence. The fluorescence intensity of the induced cells was analyzed with a FACSAria III system (BD Biosciences, Franklin Lakes, NJ, USA), and strains showing the top 1% fluorescence intensities were selected. The flow-cytometry system was set to detect FSC, SSC, and FITC (excitation = 488 nm, emission = 525 nm), and the observed fluorescence data were analyzed using R software. The selected cells were spread onto a solid M9 medium plate supplemented with 25 μg/mL chloramphenicol and then incubated at 37 °C for 24 h. The cells were collected with LB broth (final OD_{600} was approximately 20) and stored at -70 °C after 15% glycerol treatment. For the second round of screening, 0.04 mL of the stored strains was added into 1 mL of liquid minimal M9 or LB medium supplemented with 25 μg/mL chloramphenicol (final OD_{600} of approximately 0.5–0.8) and then incubated at 37 °C for 1 h. An appropriate concentration of phenol was added to the media, followed by incubation at 37 °C for 3 h to initiate the expression of fluorescence. The same steps were repeated for subsequent rounds of selection.

NGS and clustering analysis for high-resolution mutant screening

A MiSeq system (Illumina, San Diego, CA) was used to sequence the four DNA samples extracted from the sorted cells of 2nd, 3rd, and 4th, and 5th rounds in the second-stage screening. Approximately 0.34 million paired-end reads (at a read length of 300 bp) were obtained for each round of DNA samples. The fastq (https://github.com/OpenGene/fastp), seqtk (https://github.com/lh3/seqtk), and bwa (https://github.com/lh3/bwa) script tools were used for quality control, fastq-fasta format conversion, and reference sequence mapping, respectively. BAM files were read using R based package of Rsamtools, and the nucleotide frequencies at each position of the *dmpR* reference sequence were obtained from the mapped reads. A measure of nucleotide frequency ratio at each nucleotide position (FRi) is defined as non-reference nucleotide frequency divided by reference nucleotide frequency.

$$\mathrm{FR}_i = \frac{\sum_{k\in\{A,T,G,C\}\setminus j} N_{ik}}{N_{ij}} \qquad \text{Eq. (1)}$$

where N_{ij} is the j nucleotide ($j\in${A, T, G, C}) frequency (the number of NGS reads) at ith ($i\in${1, 2, …, 630}) position of the reference (original) *dmpR* gene sequence. Let *FR* be 630 by 4 matrix, *FR* = {$FR_{2,i}$,

Journal Pre-proof

$FR_{3,i}$, $FR_{4,i}$, $FR_{5,i}$}, where $FR_{2,i}$, $FR_{3,i}$, $FR_{4,i}$, and $FR_{5,i}$ are the 630 length column vector of FR_i from 2nd, 3rd, and 4th, and 5th round screenings, respectively. In order to find the position *i* where its nucleotide frequencies are monotonically increasing as the screening round progresses, ($FR_{2,i} < FR_{3,i} < FR_{4,i} < FR_{5,i}$), K-means clustering was performed with various numbers of centers, from 0 to 15. We used NbClust R package(Charrad et al., 2014) that provides 30 indices that determine the number of clusters in a dataset. The ideal number of clusters was found to be three, but six was used for a more detailed classification of the patterns (Fig. S3a, b). The ggplot2 (https://ggplot2.tidyverse.org/) package was used to visualize the ratio patterns.

Biosensor array construction

For the construction of biosensor array comprising the five selected DmpR mutants (E135K, M52I, M52C, CK, and IR), the corresponding single/double amino acid mutation was introduced to the *dmpR* gene of the pGESSv4 plasmids using the Site-Directed Mutagenesis kit. Each of the mutant plasmids, pGESSv4_E135K, pGESSv4_M52I, pGESSv4_M52C, and pGESSv4_CK, pGESSv4_IR, was transformed to *E. coli* DH5α and spread on a solid LB medium plate supplemented with 50 μg/mL ampicillin. After incubation at 37 °C for 24 h, the cells were collected with LB broth and stored at -70 °C with 15% glycerol treatment. These six types of cells including pGESSv4 with WT *dmpR* gene were used as a biosensor array for the detection of chemicals.

Chemical detection with biosensor array and data collection

The seed culture was prepared by inoculating the six biosensors (WT, E135K, M52I, M52C, CK, and IR) into 1 mL of LB liquid medium supplemented with 50 μg/mL ampicillin, followed by incubation at 37 °C with shaking. For the assay in a 96-well plate, the overnight cultured seed was inoculated at 1% into 0.2 mL of LB liquid medium supplemented with 50 μg/mL ampicillin in each well of the plate, followed by addition of substrates at final concentrations of 0–1000 μM and incubated at 37 °C for 15 h. Total 11 substrates (2-aminophenol, 2AP; 3-aminophenol, 3AP; 4-aminophenol, 4AP; 2-hydroxybenzoic acid, 2HB; 3-hydroxybenzoic acid, 3HB; 4-hydroxybenzoic acid, 4HB; 2-nitrophenol, 2NP; 3-nitrophenol, 3NP; 4-nitrophenol, 4NP; benzene; phenol) were tested at 11 different concentrations ranging from 0-1000 μM. To speed up and improve reproducibility, a semi-automatic liquid handler (INTEGRA Bioscience, Switzerland) performing serial dilution of the substrate was used. After the 15 h incubation in a shaking incubator at 55 rpm, the fluorescence intensity of the biosensors in each of the 96-well plates was measured using a multi-label reader (PerkinElmer, Waltham, MA, USA). Responses of the six biosensors to a total of 11 substrates were measured, and 11 different concentrations were used for each substrate. This process was repeated 6–12 times, and fluorescence data for a total of 7,324 samples were collected. The measured values were stored separately for GFP fluorescence intensity and OD_{600} with the various conditions of the reactions.

Journal Pre-proof

Machine learning for fingerprinting

All obtained values of GFP fluorescence (FL) and OD_{600} (OD) were pre-processed before applying ML algorithms. Let $Z = (X, Y)$ be a observed data matrix where $X = [x_{i,1}, x_{i,2}, \ldots x_{i,j} \ldots, x_{i,12}]$, j is an index of {ODWT, ODE135K, ODM52I, ODM52C, ODCK, ODIR, FLWT, FLE135K, FLM52I, FLM52C, FLCK, FLIR} and $i = \{1, 2, \ldots, N\}$. The total number of samples, N, includes all cases of the different concentrations for each substrate with 6–12 replicated experiments. Let $Y = [y_{i,1}, y_{i,2}]$ where $y_{i,1}$ is substrate name and $y_{i,2}$ is its concentration. To reduce noise among the batches of experiments, the value of $x_{i,j}$ is obtained from the original observed value divided by the observed value at zero concentration of the same condition. Two random forest models (Breiman, 2001), M1 and M2, were trained with the dataset $Z_1 = (X, y_{\cdot,1})$ for substrate classification and with $Z_2 = (X^*, y_{\cdot,2})$ where $X^* = [X, y_{\cdot,1}]$ for concentration prediction, respectively. The random forest is a collection of a large number of decision trees, resulting in a reduction of variance compared to the single decision trees. We used the R packages of "caret" 6.0-81 and "randomForest" 4.6-14 version with default hyperparameter values, mtry = $\sqrt{12}$ (number of variables randomly sampled at each split) and ntree = 500 (number of trees). Bootstrap samples were randomly drawn from Z_1 to build each tree T_b where $b = \{1, 2, \ldots, B\}$. The Gini impurity index was used to select the best splitting variable from a random subset of the mtry variables in each split step. The final ensemble classifier in the substrate classification problem was constructed by majority voting of the trained trees, $M_1(x) = \text{majority vote}\{\hat{C}_b(x)\}_1^B$ where $\hat{C}_b(x)$ is the class prediction of the T_b. In the regression problem of concentration prediction, $M_2(x) = 1/B \sum_{b=1}^{B} T_b(x)$. To evaluate the performance of the classification problem, we evaluated sensitivity (true positive rate), the probability of predicting a substrate given true state is the same substrate, and specificity (true negative rate), which is a probability of predicting a substrate given true state is not the same substrate.

3. RESULTS

3.1 High-resolution protein engineering for DmpR variants

GESS was employed to screen for DmpR mutants that are highly sensitive to detect molecules of our interest (Fig. 1a). In the first stage of the engineering, random mutations were introduced into the A domain (amino acids from 1 to 210) of DmpR, which includes the substrate-binding site, by error-prone polymerase chain reaction (PCR) to generate a DmpR mutant library (1×10^6). For efficient DNA recovery and re-screening, we built a reporter strain HK41 by introducing the promoter containing the DmpR binding site and eGFP gene to the chromosome of *E. coli* DH5α. The HK41 strain was transformed with the DmpR mutant library and cultured in M9 medium with 10 μM phenol. Cells showing higher fluorescence intensities than those of control cells were selected by fluorescence-activated cell sorting (FACS), and this screening process was repeated for four additional rounds.

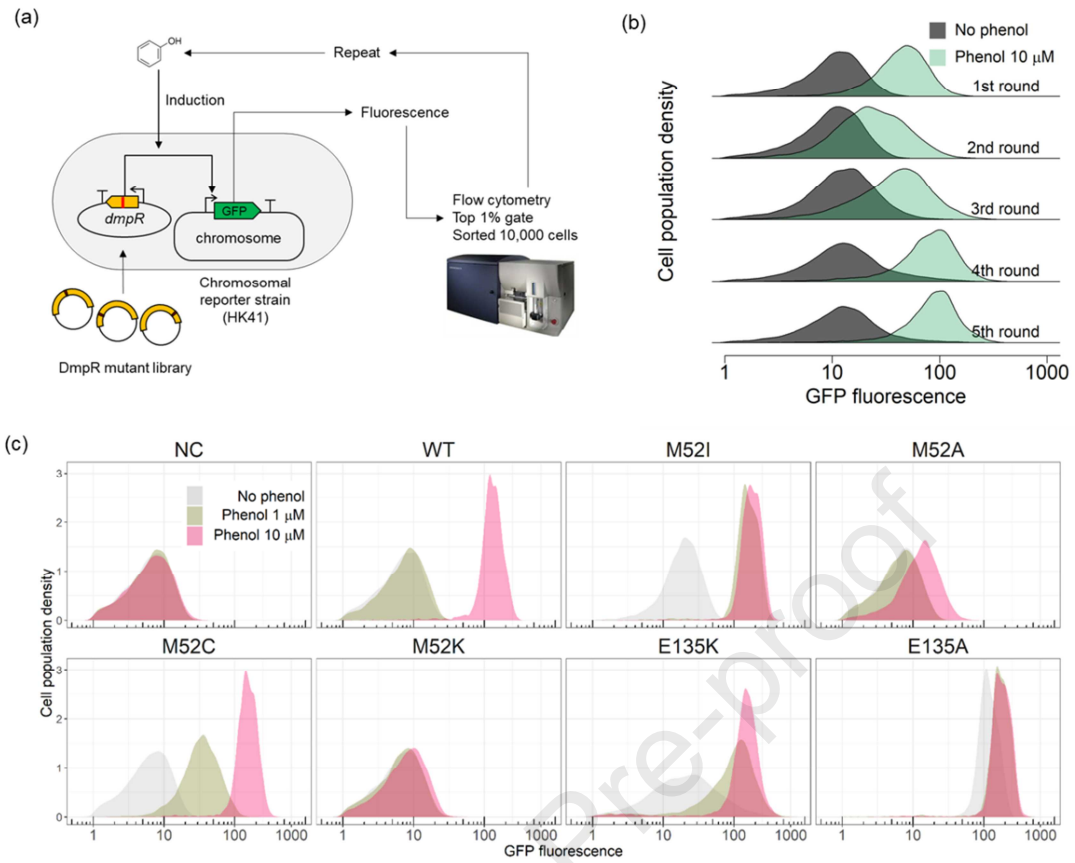


Figure 1. First-stage of DmpR screening (a) Schematic diagram of the repetitive screening process with a reporter strain (HK41) containing the DmpR mutant library by error-prone PCR targeting the 630-bp section of the DmpR A domain. The top 1% of high fluorescence cells were sorted by flow-cytometry after induction of the phenol substrate. (b) Single-cell fluorescence with (green) and without (grey) phenol induction from a total of 5 screening rounds in the first stage of DmpR engineering. Y-axis represents the number of cells, and X-axis represents GFP fluorescence by measuring the fluorescein isothiocyanate (FITC) with an emission wavelength of 525 nm via flow-cytometry. (c) Single-cell fluorescence of six DmpR mutants (M52I, M52A, M52C, M52K, E135K, E135A) and WT in response to phenol. Negative control (NC) is the HK41 strain without the plasmid harboring dmpR. The X-axis represents GFP fluorescence (FITC of the flow-cytometry), and Y-axis shows the cell population density. In each plot of DmpR mutants, the distributions of single-cell fluorescence are depicted in three cases of phenol concentrations with 0, 1, and 10 μM with colors of grey, brown, and pink, respectively.

The difference in fluorescence intensity between phenol-induced (green) and non-phenol-induced (grey) samples increased during subsequent rounds of selection (Fig. 1b). After the colonies grew in the final round of screening, interestingly, most of the selects were identified to share point mutations of methionine to isoleucine at position 52 (M52I) and glutamic acid to lysine at position 135 (E135K) (Fig. S1). M52I is a novel mutant of DmpR, whereas the E135K mutant has been previously reported to exhibit an increased response to phenol (Choi et al., 2014; Pavel et al., 1994). The DmpR structure (Fig. S2) modeled by Modeler software (Webb and Sali, 2014) with the PoxR (Patil et al., 2016) template, strongly suggests that position 52 is located at the interface between two DmpR monomers and is possibly responsible for the dimerization, which will be further described in the Discussion section. To investigate the positional effect of M52I, we constructed additional cysteine (M52C), lysine

(M52K), and alanine (M52A) mutations for their effects on the disulfide bond and residue length of DmpR dimerization (Fig. 1c). The HK41 strain harboring DmpR with the M52K or M52A mutations was not activated by phenol. In contrast, DmpR M52I or M52C exhibited strong fluorescence at sub-μM phenol concentrations, which was not detectable in WT DmpR. Since the maximum sensitivity was shown in the M52I mutant, this was chosen as a new DNA template for the second-stage of DmpR screening.

To generate a mutant library of DmpR M52I, random point mutations were additionally introduced from amino acid position 1 to 210 region of the DmpR A domain. The resulting library (approximately 10^5) was then transformed into the reporter strain of HK41. During the initial selection step, where there no substrate added, high background fluorescence mutants were removed from the library via FACS. Note that Luria-Bertani (LB) medium induces high background fluorescence in some DmpR mutants (Choi et al., 2014). Next, the top 1% of fluorescent cells were rescued by FACS in the presence of 100 nM phenol. The same positive sorting process was repeated for three more rounds, but the fluorescence signal did not markedly increase during the subsequent rounds of selection (Fig. 2a). To investigate the collected samples with a single base-pair resolution, the relevance of screening rounds and concomitant NGS reads were analyzed by a measure of nucleotide frequency ratio (FR),

$$\mathrm{FR}_i = \frac{\sum_{k\in\{A,T,G,C\}\setminus j} N_{ik}}{N_{ij}} \qquad \text{Eq. (2)}$$

where N_{ij} is the *j* nucleotide ($j\in${A, T, G, C}) frequency at *i*th ($i\in${1, 2, …, 563}) position of the reference *dmpR* sequence. N_{ij} is obtained from the number of reads after NGS reads were mapped to the reference (original) *dmpR* sequence.

Journal Pre-proof

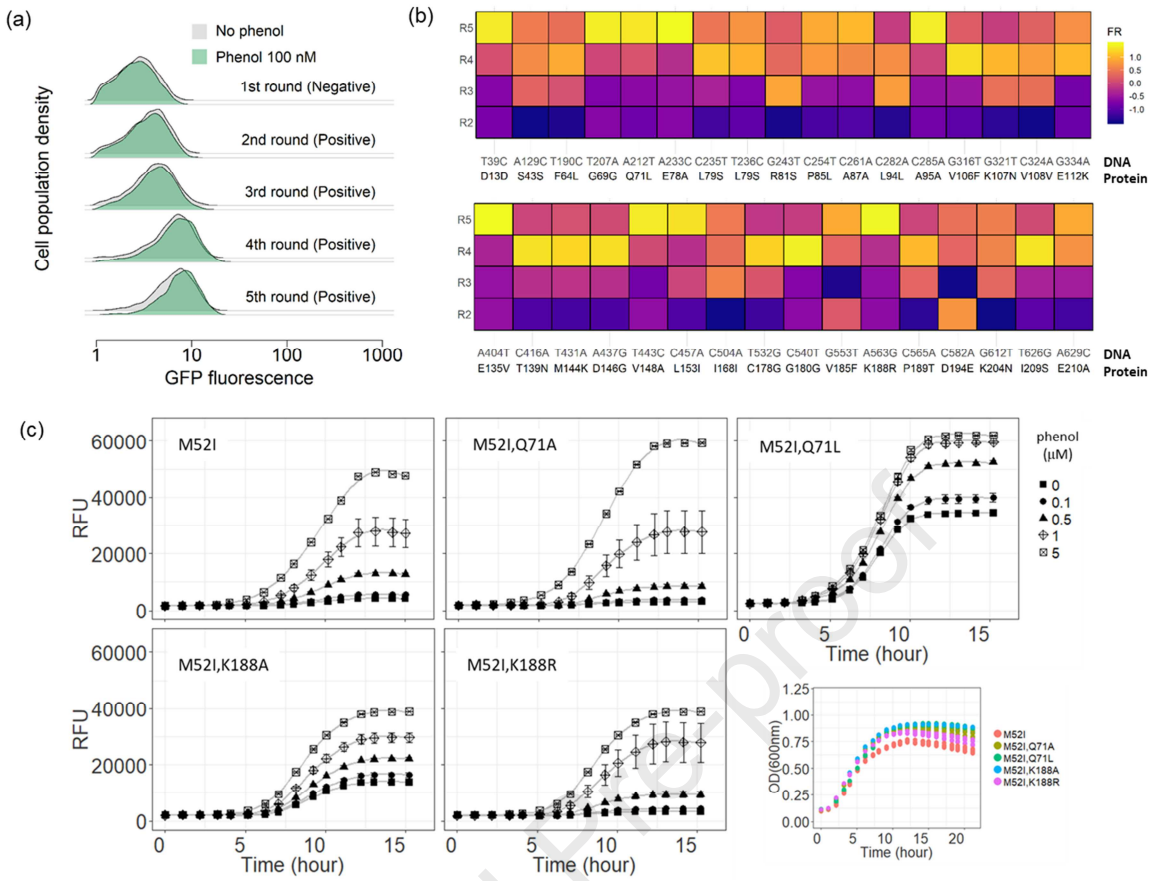


Figure 2. Second-stage of dmpR screening with the template of the DmpR mutant, M52I. (a) Five rounds of FACS-based screening. In each plot, the green distribution represents the dmpR M52I based mutant library cells with 100 nM phenol treatment, while grey shows the same library cells without phenol treatment. Y- and X-axes represent the cell population density and GFP fluorescence (FITC of the flow-cytometry), respectively. (b) 34 dmpR sequence positions grouped in cluster 3 at which nucleotide proportions were monotonically increasing as the screening round progresses. The color in each pixel depicts the nucleotide frequency ratio (FR) at each sequence position, and the screening rounds of R2, R3, R4, and R5 are depicted in the Y-axis. Low FR value indicates that the number of NGS reads with reference nucleotide is larger than the number of NGS reads with non-reference nucleotides, while High FR value in R5 means non-referenced nucleotide portion increases at the final round of screening. The changed codons and amino acids at each position are displayed in the bottom of each panel (c) Time-course fluorescent response of the DmpR M52I mutant and further selected four dmpR double mutants, M52I,Q71A; M52I,Q71L; M52I,K188A; and M52I,K188R. As an indicator of dmpR activity, the fluorescence values were measured for 15 h with five different phenol concentrations (0, 0.1, 0.5, 1, and 5 μM). X- and Y-axes represent second of time and relative fluorescence units (RFU), respectively. For RFU calculation, the measured fluorescence values with excitation 488 nm and emission 515 nm were divided by the value of optical density 600 nm (OD_{600}). On the bottom right, the growth curve of cells harboring the five DmpR mutants were plotted by measuring OD_{600}. The error bars represent the standard deviation of the duplicate samples.

The FR measure effectively quantifies the proportion of a specific nucleotide at the *i*th position with non-reference nucleotide frequency (numerator) divided by reference nucleotide frequency (denominator). $FR_{2,i}$, $FR_{3,i}$, $FR_{4,i}$, and $FR_{5,i}$ are the FR_i values representing the 2nd, 3rd, 4th, and 5th round screenings, respectively. We can then find target position *i* where its non-reference nucleotide

Journal Pre-proof

frequencies are monotonically increasing as screening round progresses, ($FR_{2,i} < FR_{3,i} < FR_{4,i} < FR_{5,i}$). To search for the DmpR mutants whose expression patterns show a monotonic increase, K-means clustering was performed with the number of centers ranging from 0 to 15 (Fig. S3a) and choose six centers for further analysis (see the Methods section for more details). The clustering analysis revealed a monotonic increment in cluster group 3 (Fig. S3b). Figure 2b shows all the positions in cluster group 3 and their altered nucleotides with corresponding amino acids. Note that the final mutant nucleotide at each position was determined by choosing one nucleotide that had the maximum portion among the non-referenced nucleotides. We choose two nucleotides positions at 212 and 563 for further investigation, as these two show high *FR* in the final round of R5, and their values reasonably increase monotonically. The amino acid changes at positions 212 and 563 were glutamine to leucine at position 71 and lysine to arginine at position 188, respectively. To evaluate their response to phenol, the plasmids expressing DmpR with the M52I/Q71A, M52I/Q71L, M52I/K188A, or M52I/K188R mutation were constructed and tested with the HK41 strain (Fig. 2c). With the five different phenol concentrations (0, 0.1 µM, 0.5 µM, 1 µM, 5 µM), we confirmed that mutant biosensors harboring DmpR M52I/Q71A and DmpR M52I/K188R had lower background noise and wider dynamic range compared to the M52I template in the complex medium.

3.2. Characterization and construction of evolutionary biosensor arrays

For the construction of the evolutionary biosensor array, six biosensors with the DmpR mutants, including E135K, M52I, M52C, M52C/E135K (CK), M52I/Q71A (IA), and M52I/K188R (IR) were sub-cloned into the pGESSv4 vector, which has a high copy origin of replication, and quantitatively compared their performance. Each transformant was cultured in LB broth with final phenol concentrations ranging from 50 nM to 500 µM for 15 h with shaking at 37 °C. The fluorescence data were collected using both a multi-label reader (Fig. 3a) and FACS (Fig. S4), that measure fluorescence at the population and single-cell level, respectively. The quantitative characteristics of the six biosensors were compared using the five measures of 1) sensitivity (SENS), 2) Limit of detection (LOD), 3) basal noise (BGN), 4) dynamic range (DYR), and 5) half-maximal effective concentration (EC_{50}), which are shown in Fig 3b. These measurements were obtained by fitting the data to the following sigmoid equation,

$$y = BGN + \frac{MAX - BGN}{1 + (EC_{50}/x)^n} \quad \text{Eq. (3)}$$

where *x* and *y* are the log value of substrate concentration and specific fluorescence, respectively; MAX is the maximum response value of the fitted model; *n* is the Hill coefficient which represents ultra-sensitivity of the sensor. Nonlinear least square function ("cobyla" function of the "nloptr" package in R) was used to fit the data to the sigmoid curve (Fig. 3b and c).

LOD is the minimum concentration point where the difference of predicted *y* and BGN is the basal fluorescence intensity without any substrate induction. For the calculation of SENS, we fitted a simple linear regression model with ten data points of *x* values around EC_{50} and corresponding predicted *y* values. Each of the mutant-based biosensors (referred to as E135K, M52I, M52C, CK, IR,

and IA biosensors) exhibited improved sensitivities than that of WT DmpR biosensor (WT biosensor) (Table S3). The LODs of the IR and IA biosensors were approximately 35–75-fold improved than that of the WT biosensor, while the BGN values of IR and IA biosensors are 3-fold lower than that of M52I as intended. The M52C mutation was most interesting; with a BGN value close to WT, while showing a 2.2-fold lower LOD. Normally, one ideal biosensor is chosen for further study based on its performance measures. However, the combination of all mutant biosensors, including the WT biosensor, is expected to have much wider coverage than any single biosensor, as shown in Figure 3c. Therefore, the use of ensemble biosensor is expected to have the benefit of capturing delicate and discriminative information in response to small molecules as compared to a single biosensor.

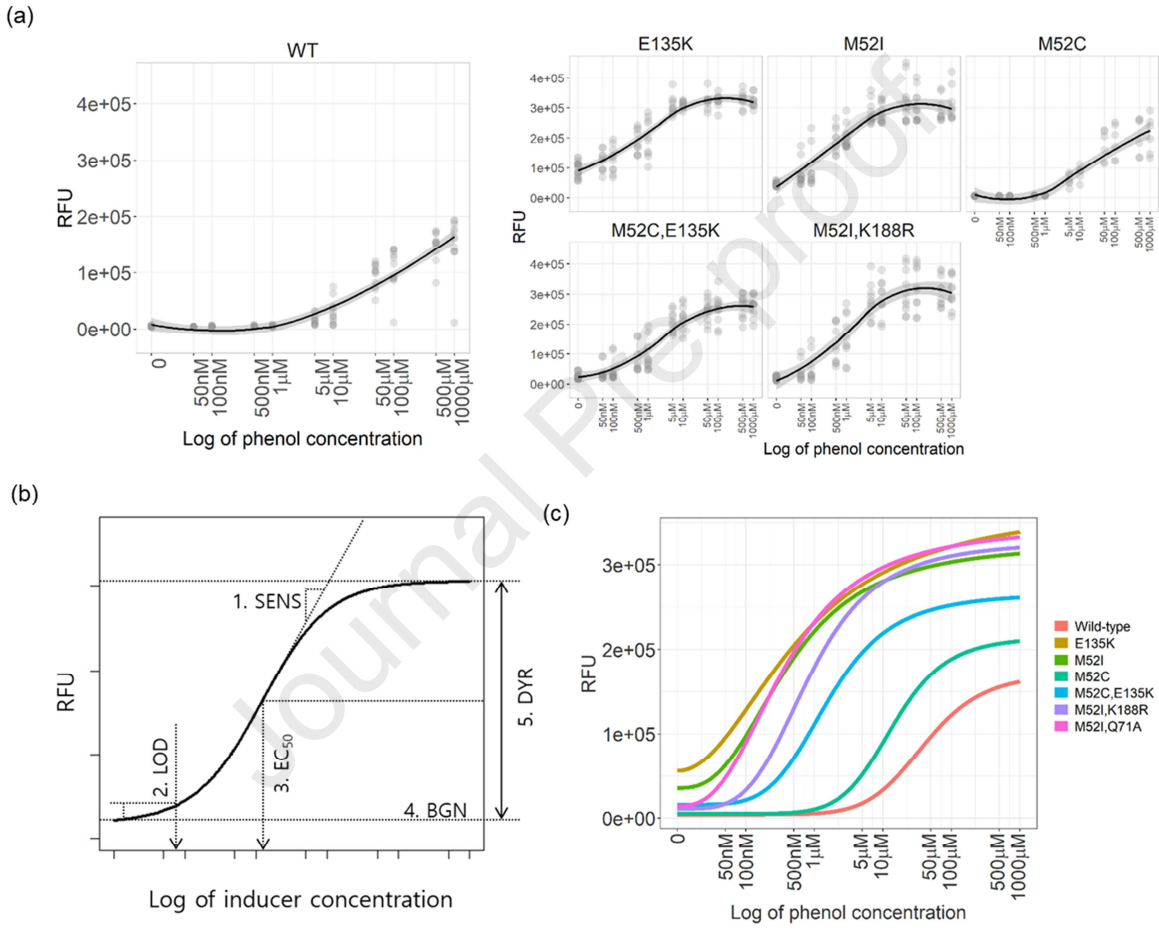


Figure 3 Characteristics of DmpR mutant-based biosensors. (a) Phenol response curves of the seven DmpR mutants, including WT dmpR. A total of 11 different final phenol concentrations from 0 to 1 mM were used to test the DmpR mutant activities. In each plot, X-axis represents the log of phenol concentrations, and Y-axis shows the relative fluorescence unit, which was computed by measuring the fluorescence values (excitation 488 nm and emission 515 nm) divided by the optical density 600 nm. Between six and twelve replications were evaluated at each phenol concentration, and the Loess curve was fitted to the smooth line (black) with standard error (grey area). (b) Biosensor characteristics with five measures including 1. Sensitivity (SENS), 2. Limit of detection (LOD), 3. Half-maximal effective concentration 50 (EC_{50}), 4. Basal noise (BGN), and 5. Dynamic range (DYR). (c) Sigmoidal response curves of the seven DmpR mutant-based biosensors, WT; E135K; M52I; M52C; M52I,E135K; M52I,Q71A; and M52I,K188R. X- and Y-axes are the log of inducer (phenol) concentration and relative fluorescence units, respectively. The Hill equation was fitted for the sigmoidal curve (black line in Fig. 3b) using the "cobyla" function of the "nloptr" R package.

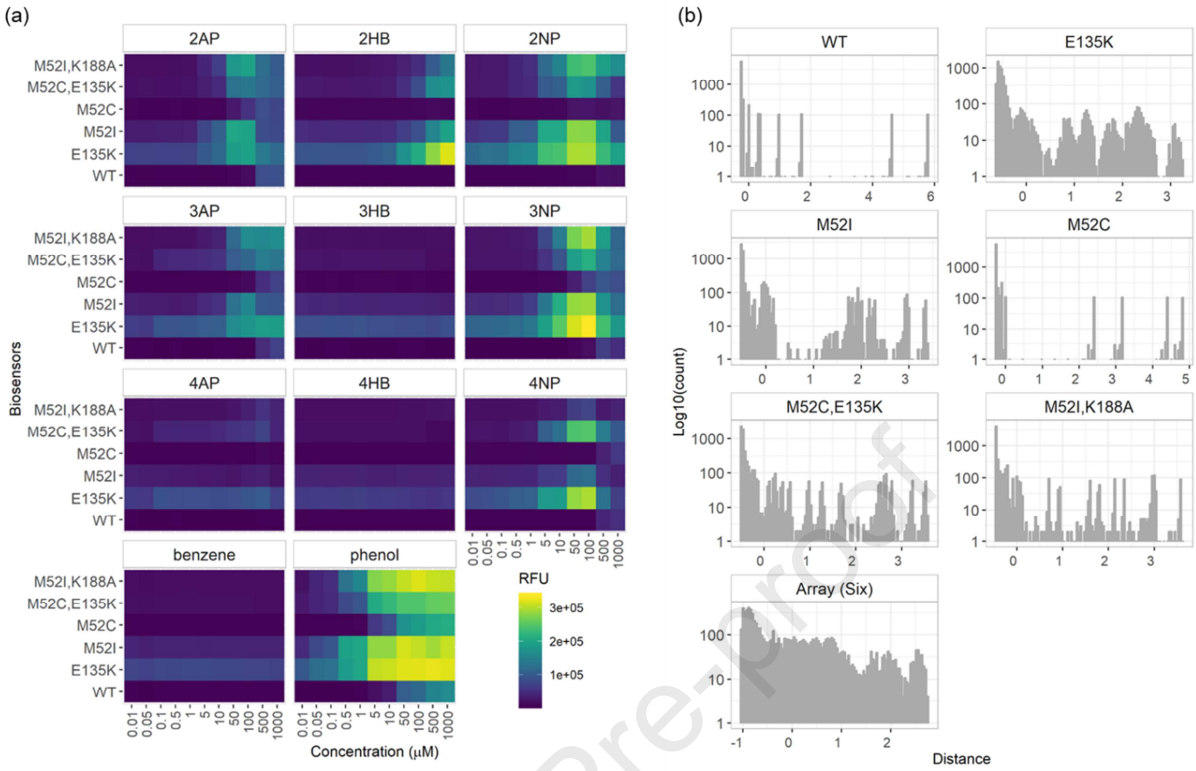


Figure 4 Characterization of biosensor array responses (a) Biosensor array finger printings (11 substrates and 11 concentration points of each substrate for six biosensors). Each pixel is colored with the mean RFU value, which can be calculated by measuring fluorescence values with excitation 488 nm and emission 515 nm divided by the value of optical density 600 nm. Y-axis represents the six biosensors of the biosensor array, and X-axis represents the 11 concentration points of each substrate. (b) Distributions of the distances among the 121 fingerprinting patterns for each biosensor and biosensor array (the bottom left). In each plot, X-axis represents the possible range of the distance, and the Y-axis is the log10 of count belonging to each unit range of the distances in the X-axis.

3.3. Collection of large-scale training dataset for chemical fingerprinting

To achieve the underlying goal of this study, it was necessary to build an ML model and to be trained with a dataset, which ideally consists of a large-scale biosensor responding data labelled with a chemical name and its concentration. However, this type of labelled dataset is not normally available due to the difficulty of chemical identification using low-throughput equipment, such as GC-MS/LC-MS. Alternatively, we simulated test conditions by treating with a fixed concentration of chemical substrates to the biosensor array and measuring their fluorescence responses. For the six biosensors (WT, E135K, M52I, M52C, CK, and IR), the fluorescence response data for a total of 11 substrates (2-aminophenol, 2AP; 3-aminophenol, 3AP; 4-aminophenol, 4AP; 2-hydroxybenzoic acid, 2HB; 3-hydroxybenzoic acid, 3HB; 4-hydroxybenzoic acid, 4HB; 2-nitrophenol, 2NP; 3-nitrophenol, 3NP; 4-nitrophenol, 4NP; benzene; phenol) were obtained at 11 different concentrations from 0-1000 μM. For the rapid production of the data, we used a semi-automatic liquid handler that can dilute the substrate

of the solution in a 96-well plate. Experiments were repeated 6-12 times for each substrate, resulting in a total of 7,324 different sample conditions. Each replicated experiment is depicted with different colors in Fig. S5 to show the variations among the repeats. All mutant biosensors except WT and M52C were induced by 2AP, 2HB, 2NP, 3NP, and 4NP, which was consistent with previous studies reporting the lack of WT biosensor response to these substrates (Choi et al., 2014). These broad substrate specificities may cause ambiguity in identifying chemicals when a single mutant biosensor is used. In Figure 4a, the biosensor array response patterns are clearly distinguished from each other, depending on the type of substrate. Note that each pixel of Figure 4a is coloured by averaging all the values of replications of the condition. There is a total of 121 patterns (11 substrates $\times$ 11 concentration points) generated by the six biosensors. Although the overlap patterns at low substrate concentrations were hardly distinguished between each other (Fig. S6a and b), the mutant array could generate more diverse patterns than any single biosensor. This diversity can be directly compared by plotting the distribution of distances among the 121 patterns in Figure 4b. The distance between two points of the x^{th} and y^{th} patterns was calculated by the Euclidean distance, $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ where i is the index of the biosensors (n = 6). In Figure 4b, the biosensor array (Six) clearly shows that the 121 patterns distributed widely across the distances, whereas single biosensors were sparsely distributed. This means that a biosensor array that generates diverse response patterns is more desirable for monitoring environmental samples.

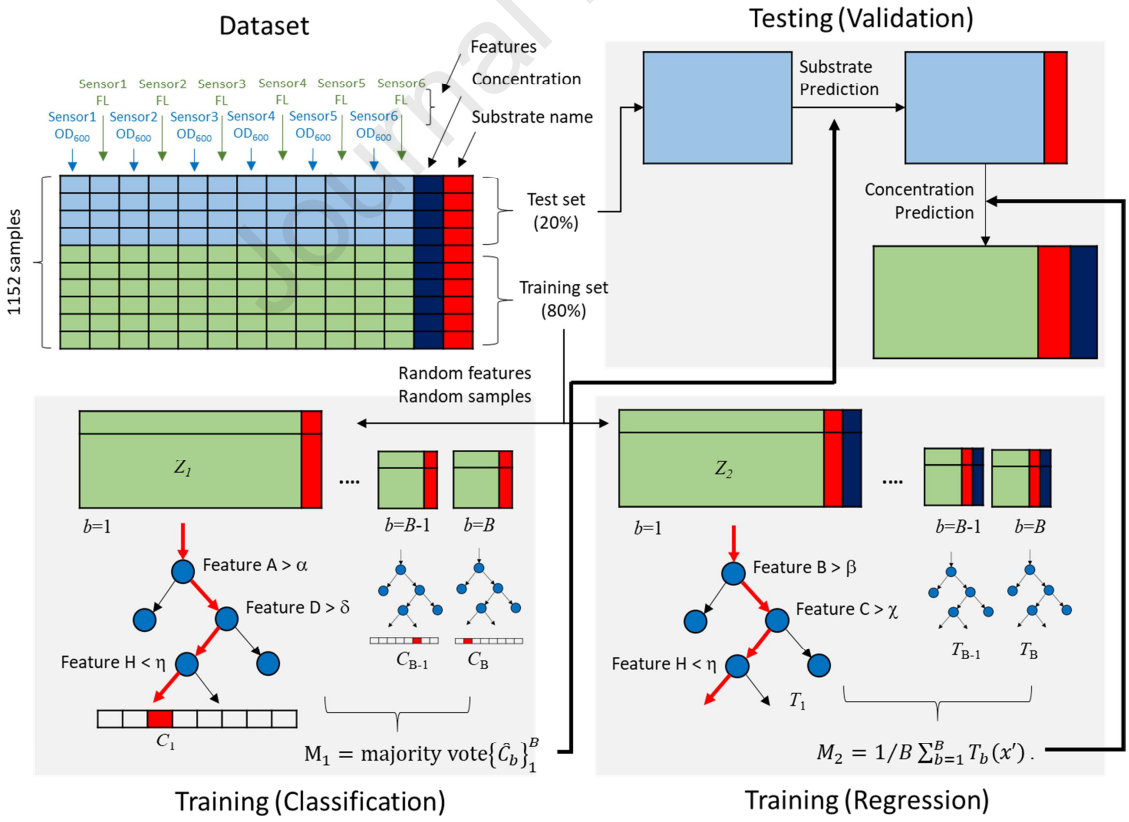


Figure 5 Schematic diagram of random forest models used in this study. The dataset in the top left consists of 1152 samples, 12 features (fluorescence (FL) and optical density 600 nm (OD_{600}) values of the six biosensors),

and 2 labels (concentration, substrate type). Random forests model for classification (bottom left) were fitted with the training dataset (80% of the dataset) and predicted the type of the substrates. With the predicted class of the substrate, the regression models (bottom right) were used to estimate the concentration of the predicted substrate (top right).

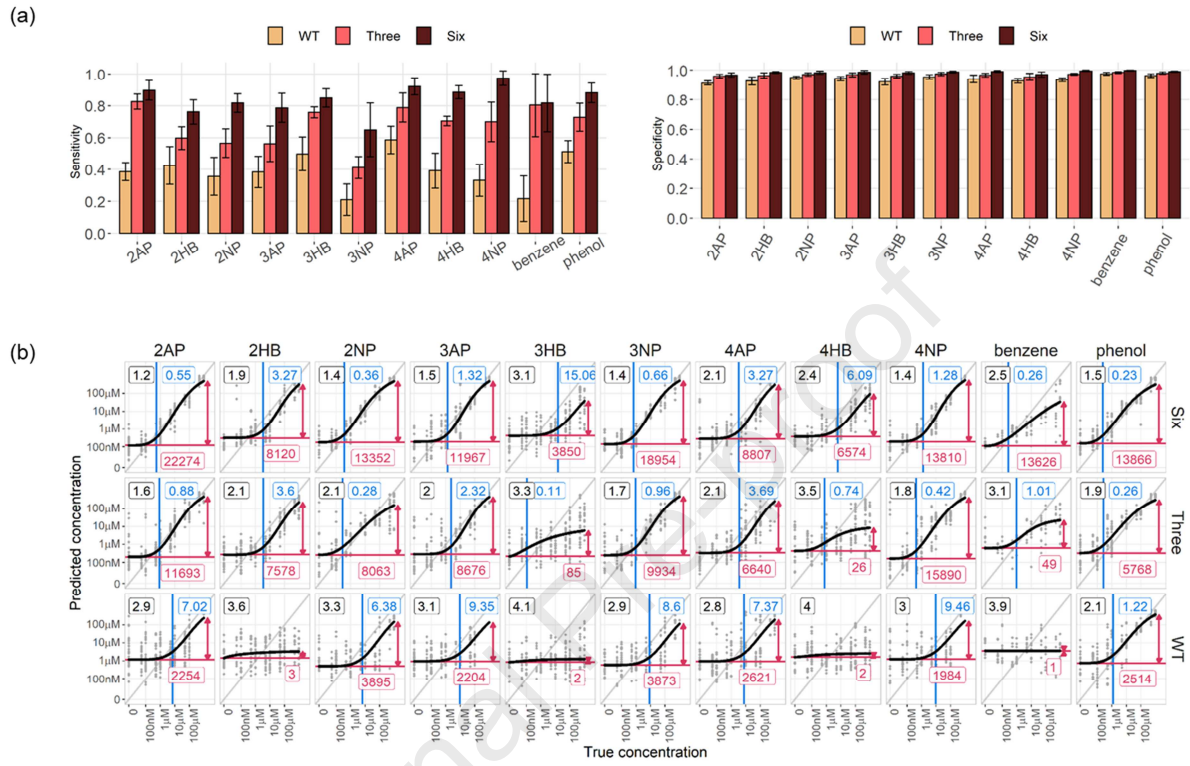


Figure 6 Performance comparison of the biosensor array. (a) Sensitivity (true positive rate, upper left) and specificity (true negative rate, upper right) of the biosensor array (Six, dark brown) in predicting substrate name compared to that of WT biosensor (WT, yellow) and the three biosensor array with WT, E135K, and M52I DmpR mutants (Three, orange). In each plot, X- and Y-axes represent the 11 substrates and sensitivity/specificity, respectively. The error bars indicate the standard deviation of each group of measures. (b) Concentration prediction of the biosensor array (Six, top) compared to that of WT biosensor (WT, bottom) and three biosensor arrays with WT, E135K, and M52I DmpR mutants (Three, middle) in all the 11 cases of substrates. In each plot, the X-axis represents the true concentration value in the test dataset, and the Y-axis indicates the predicted concentration by the random forest model. The points on the diagonal line indicate that the prediction accurately represents the true value of the test sample. Loess curve was fitted for the smooth line (black). Black text label in the upper-left corner in each plot is the root mean squared error (RMSE). Blue text label with vertical line represents the limit of detection (LOD), and red text label with horizontal line indicate the dynamic range (DYR).

3.4. Machine learning-based molecular fingerprinting

To effectively capture the response patterns from the biosensor array, we used random forest, one of the widely used decisions making ML algorithms (Breiman, 2001). It is similar to an ensemble technique involving bagging and randomized node optimization (Berk, 2016), which is applicable for classification and regression problems. Figure 5 depicts the schematic diagram of our learning

Journal Pre-proof

procedure with the descriptions of the datasets. We defined a biosensor array consisting of WT and five DmpR mutant biosensors (E135K, M52I, M52C, CK, and IR). The number of total samples was 7,324, which included two types of response values (fluorescence and OD_{600}). After filtering and scaling the dataset, we obtained a dataset of 6,912 samples which was transformed into 1,152 observations with 12 features (Two observations of fluorescence and OD_{600} for each biosensor) and two labels (concentration and chemical substrate name) for ML application. Twenty percent of the observations were used as test data, and the others were used for training. Among the two labels, only substrate name was used as a response variable for model training in classification analysis. The total B decision trees were construed by drawing Z_b (b = 1, 2, …, B) bootstrap samples randomly from the training dataset, and the final model was built by voting from the B decision trees. Predictions of substrate names for substrate identification were conducted, and the true positive rate (sensitivity) and true negative rate (specificity) were measured for prediction evaluation.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad \text{Eq. (4)}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \quad \text{Eq. (5)}$$

When the single WT biosensor was used, its true positive rate for substrate prediction was below 50% for most cases, whereas that of the three (WT, E135K, and M52I), and six (WT, E135K, M52I, M52C, CK, and IR) biosensors increased up to 0.953 in 4NP case, as shown in the left panel of Figure 6a. The specificities of all cases were sufficiently high, as there was a much larger number of false cases than true cases (Fig. 6a right panel). Based on the predicted substrate name, we further estimated the concentrations of the substrate. In this regression problem, the random forest model was trained with the same training dataset as that used in the previous classification case except that the dataset included previously predicted substrate class and used the concentration values as response variables (Fig. 5 bottom right panel). Figure 6b depicts the predicted concentrations of the three cases (WT only, three, and six biosensors) for 11 substrates. The diagonal line of each plot represents the perfect prediction; a greater number of dots distributed close to the diagonal line indicated better prediction performance. To quantify the prediction performance in the regression problem, root mean squared error (RMSE) was introduced. The computed RMSE values are presented with black colored text in the left-top corner of each plot in Figure 6b. Sigmoid curve of Eq. (3) was fitted with the predicted and true concentration data as response and predicted variables, respectively. As in Fig. 3c, LOD and DYR of biosensor arrays were obtained by fitting the sigmoid curve. In Figure 6b, blue text labels represent LOD scores and red text labels indicate DYR. Both LOD and DYR of the Six-biosensor array show better scores than those of WT and Three biosensor arrays for all substrates. This result indicates that the combination of mutant biosensors with ML can significantly improve biosensor performance that cannot be achieved with the single biosensor engineering approach. For instance, in Table S3, WT-biosensor has the best DYR score among the mutant based six biosensors while the LOD scores of all the mutant biosensors outperformed WT-biosensor. Note that the LOD

Journal Pre-proof

and DYR in Fig 6b are not directly comparable to those of measures in Table S3 since the response values in Fig 6b are the predicted concentration while the response values in Table S6 are specific fluorescence. Also, the LOD and DYR values of WT-biosensor with high RMSE could be less reliable than those values of the Six biosensor array.

4. Discussion

TF-based genetically-encoded biosensors have been widely studied for detecting and quantifying small molecules by engineering sensitivity, specificity, or LOD of the TF. However, it is necessary to consider the limited number of available TFs and the broad substrate specificity of the engineered TFs for practical biosensor applications. This study showed that the high-throughput GESS and NGS techniques, providing a high-resolution TF engineering tool, were applied to overcome the limited availability of TF, and thus, high specificity for molecular recognition can be achieved by using multiple biosensor arrays with ML. Even a single point mutation of DmpR could significantly contribute to improving the biosensor array performance, as DmpR requires ATP hydrolysis that accompanies oligomerization and triggers RNA polymerase activation. This energy-dependent mechano-transcriptional regulator enables its biosensor to generate a large phenotypic variation. For example, the M52I biosensor showed significant improvements in LOD and sensitivity compared to WT. The methionine residue at position 52 of DmpR may be responsible for DmpR dimerization, which could affect the oligomer formation of the protein (Fig. S2). The multimerization of proteins is tightly linked to protein activity, particularly among NtrC family proteins (Lee et al., 2003). If the residue is too long for lysine or too short for alanine, as shown in Fig. 1c, the activity disappears. Dimerization of DmpR protein may affect its activity, as implied by the increase of basal noise in the order of WT (methionine), cysteine, and isoleucine. This suggests the existence of a disulfide bond or a hydrogen bond between two A domains of the dimerized DmpR; and dynamic interactions between these domains may affect the activity of DmpR mutants via the mechanistic behavior of the mutant. Additionally, cells harboring the M52I mutant showed impaired growth, as depicted in the bottom right graph in Fig. 2c. Although proofing these concepts is beyond the scope of our research, the unknown mechanical complexity helps to generate informational variations, even via a single residue change in the protein. Our test condition is limited by the use of one chemical substrate per sample. A mixture containing two substrates results in completely different biosensor signals in comparison with the single substrate cases. For example, 4-nitrophenol acts as an antagonist of DmpR; it cannot trigger the TF activation (Choi et al., 2014). Antagonists can bind to the active site of the protein, and decrease its activity by interfering with regulator-ligand complex formation. Hence, the WT biosensor might not show any responses to a mixture with 4-nitrophenol and phenol. This would mean that the ML model trained by the single substrate data will not be suitable for substrate prediction in a mixture containing more than two chemicals. Nonetheless, it is feasible to collect data responding to a mixture with multiple substrates. Therefore one can apply the same framework with our biosensor array and ML algorithms for more practical testing conditions.

Journal Pre-proof

5. Conclusions

The use of multiple biosensors guarantees to increase reliability and accuracy in molecular identification. This study shows employing multiple variants of a single transcription factor can effectively circumvent the limited availability of the transcription factors. In addition, we obtained the labeled data by testing simulated conditions of the samples, which provided an effective way to collect the large-scale labelled data for training the machine learning models. Highly complex biological system is naturally compatible with ML or deep learning algorithms that can efficiently capture hidden patterns in the complex system. Better prediction power may be achieved with increasing studies on the responses to more diverse substrates and concentrations. Recent developments of Biofoundry could provide a high-throughput platform for rapid and high-quality data generation (Hillson et al., 2019). Our ML study with mutant biosensors is expected to contribute to the development of point-of-care testing tools that are easy to use, inexpensive, and have high specificity. More importantly, it will serve as a new method of practical biosensor applications that function in a complex environment beyond laboratory conditions.

ACKNOWLEDGEMENT

The authors would like to thank the members of the Synthetic Biology Laboratory in the Synthetic Biology and Bioengineering Center at KRIBB for their valuable comments and helpful discussions.

Funding: This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2017R1E1A1A03070884), the Bio & Medical Technology Development Program (2018M3A9H3024746), the Next-Generation Biogreen 21 Program (PJ01330902), and the KRIBB Research Initiative Program.

CONFLICT OF INTEREST

The authors declare no competing interests

REFERENCES

Berk, R.A., 2016. Statistical Learning from a Regression Perspective. https://doi.org/10.1007/978-3-319-44048-4

Beyer, M.K., Clausen-Schaumann, H., 2005. Mechanochemistry: the mechanical activation of covalent bonds. Chem. Rev. 105, 2921–2948.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Journal Pre-proof

Bucur, B., Fournier, D., Danet, A., Marty, J.-L., 2006. Biosensors based on highly sensitive acetylcholinesterases for enhanced carbamate insecticides detection. Anal. Chim. Acta 562, 115–121.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., 2014. Nbclust: An R package for determining the relevant number of clusters in a data set. J. Stat. Softw. 61, 1–36. https://doi.org/10.18637/jss.v061.i06

Cherepanov, P.P., Wackernagel, W., 1995. Gene disruption in Escherichia coli: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. Gene 158, 9–14.

Choi, S.L., Rha, E., Lee, S.J., Kim, H., Kwon, K., Jeong, Y.S., Rhee, Y.H., Song, J.J., Kim, H.S., Lee, S.G., 2014. Toward a generalized and high-throughput enzyme screening system based on artificial genetic circuits. ACS Synth. Biol. 3, 163–171. https://doi.org/10.1021/sb400112u

Chong, H.H., Ching, C.B., 2016. Development of colorimetric-based whole-cell biosensor for organophosphorus compounds by engineering transcription regulator DmpR. ACS Synth. Biol. 5, acssynbio.6b00061. https://doi.org/10.1021/acssynbio.6b00061

Datsenko, K.A., Wanner, B.L., 2000. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc. Natl. Acad. Sci. 97, 6640–6645.

Dickert, F.L., Hayden, O., Zenkel, M.E., 1999. Detection of volatile compounds with mass-sensitive sensor arrays in the presence of variable ambient humidity. Anal. Chem. 71, 1338–1341. https://doi.org/10.1021/ac981014e

Elad, T., Belkin, S., 2013. Broad spectrum detection and “barcoding” of water pollutants by a genome-wide bacterial sensor array. Water Res. 47, 3782–3790. https://doi.org/10.1016/j.watres.2013.04.011

Fernandez-López, R., Ruiz, R., de la Cruz, F., Moncalián, G., 2015. Transcription factor-based biosensors enlightened by the analyte. Front. Microbiol. 6, 1–21. https://doi.org/10.3389/fmicb.2015.00648

Gupta, S., Saxena, M., Saini, N., Mahmooduzzafar, Kumar, R., Kumar, A., 2012. An effective strategy for a whole-cell biosensor based on putative effector interaction site of the regulatory DmpR protein. PLoS One 7, e43527. https://doi.org/10.1371/journal.pone.0043527

Hillson, N., Caddick, M., Cai, Y., Carrasco, J.A., Chang, M.W., Curach, N.C., Bell, D.J., Le Feuvre, R., Friedman, D.C., Fu, X., Gold, N.D., Herrgård, M.J., Holowko, M.B., Johnson, J.R., Johnson, R.A., Keasling, J.D., Kitney, R.I., Kondo, A., Liu, C., Martin, V.J.J., Menolascina, F., Ogino, C., Patron, N.J., Pavan, M., Poh, C.L., Pretorius, I.S., Rosser, S.J., Scrutton, N.S., Storch, M., Tekotte, H., Travnik, E., Vickers, C.E., Yew, W.S., Yuan, Y., Zhao, H., Freemont, P.S., 2019. Building a global alliance of biofoundries. Nat. Commun. 10, 1–4. https://doi.org/10.1038/s41467-019-10079-2

Journal Pre-proof

Jha, R.K., Kern, T.L., Kim, Y., Tesar, C., Jedrzejczak, R., Joachimiak, A., Strauss, C.E.M., 2016. A microbial sensor for organophosphate hydrolysis exploiting an engineered specificity switch in a transcription factor. Nucleic Acids Res. 44, gkw687. https://doi.org/10.1093/nar/gkw687

Jin, H.L., Mitchell, R.J., Byoung, C.K., Cullen, D.C., Man, B.G., 2005. A cell array biosensor for environmental toxicity analysis. Biosens. Bioelectron. 21, 500–507. https://doi.org/10.1016/j.bios.2004.12.015

Juribašić, M., Užarević, K., Gracin, D., Ćurić, M., 2014. Mechanochemical C–H bond activation: rapid and regioselective double cyclopalladation monitored by in situ Raman spectroscopy. Chem. Commun. 50, 10287–10290.

Kim, H., Rha, E., Seong, W., Yeom, S.-J.J., Lee, D.-H.H., Lee, S.-G.G., 2016. A Cell–Cell Communication-Based Screening System for Novel Microbes with Target Enzyme Activities. ACS Synth. Biol. 5, 1231–1238. https://doi.org/10.1021/acssynbio.5b00287

Kwon, K.K., Lee, D.-H., Kim, S.J., Choi, S.-L., Rha, E., Yeom, S.-J., Subhadra, B., Lee, J., Jeong, K.J., Lee, S.-G., 2018. Evolution of enzymes with new specificity by high-throughput screening using DmpR-based genetic circuits and multiple flow cytometry rounds. Sci. Rep. 8, 2659.

Lee, S., Torre, A.D. La, Yan, D., Kustu, S., Nixon, B.T., Wemmer, D.E., 2003. Regulation of the transcriptional activator NtrC1 : structural studies of the regulatory and AAA + ATPase domains 1, 2552–2563. https://doi.org/10.1101/gad.1125603.main

Mandal, S., Koirala, D., Selvam, S., Ghimire, C., Mao, H., 2015. A Molecular Tuning Fork in Single□Molecule Mechanochemical Sensing. Angew. Chemie 127, 7717–7721.

Meyer, A.J., Segall-Shapiro, T.H., Glassey, E., Zhang, J., Voigt, C.A., 2018. Escherichia coli "Marionette" strains with 12 highly optimized small-molecule sensors. Nat. Chem. Biol. 15, 196. https://doi.org/10.1038/s41589-018-0168-3

O'Neill, E., Sze, C.C., Shingler, V., 1999. Novel effector control through modulation of a preexisting binding site of the aromatic-responsive ??54-dependent regulator DmpR. J. Biol. Chem. 274, 32425–32432. https://doi.org/10.1074/jbc.274.45.32425

Patil, V.V., Park, K.H., Lee, S.G., Woo, E., 2016. Structural Analysis of the Phenol-Responsive Sensory Domain of the Transcription Activator PoxR. Structure 24, 624–630. https://doi.org/10.1016/j.str.2016.03.006

Pavel, H., Forsman, M., Shingler, V., 1994. An aromatic effector specificity mutant of the transcriptional regulator DmpR overcomes the growth constraints of Pseudomonas sp. strain CF600 on para-substituted methylphenols. J. Bacteriol. 176, 7550–7.

Psuj, G., 2018. Multi-sensor data integration using deep learning for characterization of defects in steel elements. Sensors (Switzerland) 18, 1–15. https://doi.org/10.3390/s18010292

Journal Pre-proof

Ray, S., Gunzburg, M.J., Wilce, M., Panjikar, S., Anand, R., 2016. Structural Basis of Selective Aromatic Pollutant Sensing by the Effector Binding Domain of MopR, an NtrC Family Transcriptional Regulator. ACS Chem. Biol. https://doi.org/10.1021/acschembio.6b00020

Rogers, J.K., Taylor, N.D., Church, G.M., 2016. Biosensor-based engineering of biosynthetic pathways. Curr. Opin. Biotechnol. 42, 84–91. https://doi.org/10.1016/j.copbio.2016.03.005

Shin, H.J., Park, H.H., Lim, W.K., 2005. Freeze-dried recombinant bacteria for on-site detection of phenolic compounds by color change. J. Biotechnol. 119, 36–43. https://doi.org/10.1016/j.jbiotec.2005.06.002

Shingler, V., Moore, T., 1994. Sensing of aromatic compounds by the DmpR transcriptional activator of phenol-catabolizing Pseudomonas sp. strain CF600. J. Bacteriol. 176, 1555–60.

Wan, X., Volpetti, F., Petrova, E., French, C., Maerkl, S.J., Wang, B., 2018. Cascaded amplifying circuits enable ultrasensitive cellular sensors for toxic metals. Nat. Chem. Biol. https://doi.org/10.1038/s41589-019-0244-3

Webb, B., Sali, A., 2014. Protein structure modeling with MODELLER, in: Protein Structure Prediction. Springer, pp. 1–15.

Wise, a a, Kuske, C.R., 2000. Generation of novel bacterial regulatory proteins that detect priority pollutant phenols. Appl. Environ. Microbiol. 66, 163–9.

Xue, H., Shi, H., Yu, Z., He, S., Liu, S., Hou, Y., Pan, X., Wang, H., Zheng, P., Cui, C., Viets, H., Liang, J., Zhang, Y., Chen, S., Zhang, H.M., Ouyang, Q., 2014. Design, construction, and characterization of a set of biosensors for aromatic compounds. ACS Synth. Biol. 3, 1011–1014. https://doi.org/10.1021/sb500023f

Yeom, S.J., Kim, M., Kwon, K.K., Fu, Y., Rha, E., Park, S.H., Lee, H., Kim, H., Lee, D.H., Kim, D.M., Lee, S.G., 2018. A synthetic microbial biosensor for high-throughput screening of lactam biocatalysts. Nat. Commun. 9, 1–12. https://doi.org/10.1038/s41467-018-07488-0

Zhang, F., Keasling, J., 2011. Biosensors and their applications in microbial metabolic engineering. Trends Microbiol. 19, 323–9. https://doi.org/10.1016/j.tim.2011.05.003

Journal Pre-proof

Highlights

- Employing multiple variants of a single transcription factor for biosensor array
- Machine-learning technique with biosensor array significantly improves small molecule identification
- Collecting large-scale labeled data via simulated conditions of target chemicals
- High-throughput screening with NGS technique for high-resolution protein engineering

Journal Pre-proof

Journal Pre-proof

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof