Computational and Structural Biotechnology Journal 21 (2023) 2613–2620

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Predicting the effects of cultivation condition on gene regulation in *Escherichia coli* by using deep learning

Mun Su Kwon [a], Joshua Julio Adidjaja [a], Hyun Uk Kim [a,b,*]

[a] Systems Biology and Medicine Laboratory, Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea
[b] BioProcess Engineering Research Center and BioInformatics Research Center, KAIST, Daejeon 34141, Republic of Korea

ARTICLE INFO

Article history:
Received 9 December 2022
Received in revised form 2 April 2023
Accepted 12 April 2023
Available online 13 April 2023

Keywords:
Deep learning
Transcriptional response
Gene regulation
Cultivation condition
Escherichia coli

ABSTRACT

Cell's physiology is affected by cultivation conditions at varying degrees, including carbon sources and inorganic nutrients in growth medium, and the presence or absence of aeration. When examining the effects of cultivation conditions on the cell, the cell's transcriptional response is often examined first among other phenotypes (e.g., proteome and metabolome). In this regard, we developed DeepMGR, a deep learning model that predicts the effects of culture *m*edia on *g*ene *r*egulation in *Escherichia coli*. DeepMGR specifically classifies the direction of gene regulation (i.e., upregulation, no regulation, or downregulation) for an input gene in comparison with M9 minimal medium with glucose as a control condition. For this classification task, DeepMGR uses a feedforward neural network to process: i) DNA sequence of a target gene, ii) presence or absence of aeration and trace elements, and iii) concentration and structural information (SMILES) of up to ten nutrients. The complete DeepMGR showed accuracy of 0.867 and F1 score of 0.703 for a test set from the gold standard dataset. DeepMGR was further subjected to simulation studies for validation where regulation directions for groups of homologous genes were predicted, and the DeepMGR results were compared with the literature with focus on carbon sources that upregulate specific genes. DeepMGR will be useful for designing experiments to understand gene regulations, especially in the context of metabolic engineering.

1. Introduction

Cells show a wide range of different phenotypes in response to cultivation conditions, including carbon sources and various inorganic nutrients in culture medium as well as the presence or absence of aeration. Studying the cell's response to a cultivation condition is important for understanding the cell's gene regulation [1], especially for designing a cultivation condition [2,3]. An optimal medium is associated with the optimal growth of a chemical-producing microbial strain and its production performance for a target chemical. The cell's response to cultivation conditions can be examined through various techniques, but transcriptional response would be one of the preferred targets to study, for example by RNA sequencing [4,5]. Here, the process of studying the cell's transcriptional response can be facilitated if it becomes possible to predict directions of gene regulation (i.e., upregulation, no regulation, or downregulation) under a specific cultivation condition. Being able to predict the direction of gene regulation can be particularly useful in metabolic engineering or bioprocess engineering where a range of genes, including those less studied, need to be examined under specific conditions of interest before conducting experiments. Such a resource would help design and/or minimize time-consuming experiments (e.g., RNA-seq) to examine gene regulations.

The importance of predicting the cell's transcriptional response under a specific condition has been well recognized, and relevant prediction models have been developed [6,7]. Representative examples include: host response model (HRM) that predicts transcriptional responses of *Escherichia coli* and *Bacillus subtilis* upon addition of inducers (e.g., IPTG and arabinose) [6]; DeepCOP that predicts the effects of small molecules on gene regulation in cancer cell lines [8]; and DeepCE that predicts gene expression profiles for chemicals (i.e., drug candidates) [9]. These studies were focused on very specific additions in a medium, depending on the study

[*] Corresponding author at: Systems Biology and Medicine Laboratory, Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea.
E-mail address: ehukim@kaist.ac.kr (H.U. Kim).

https://doi.org/10.1016/j.csbj.2023.04.010

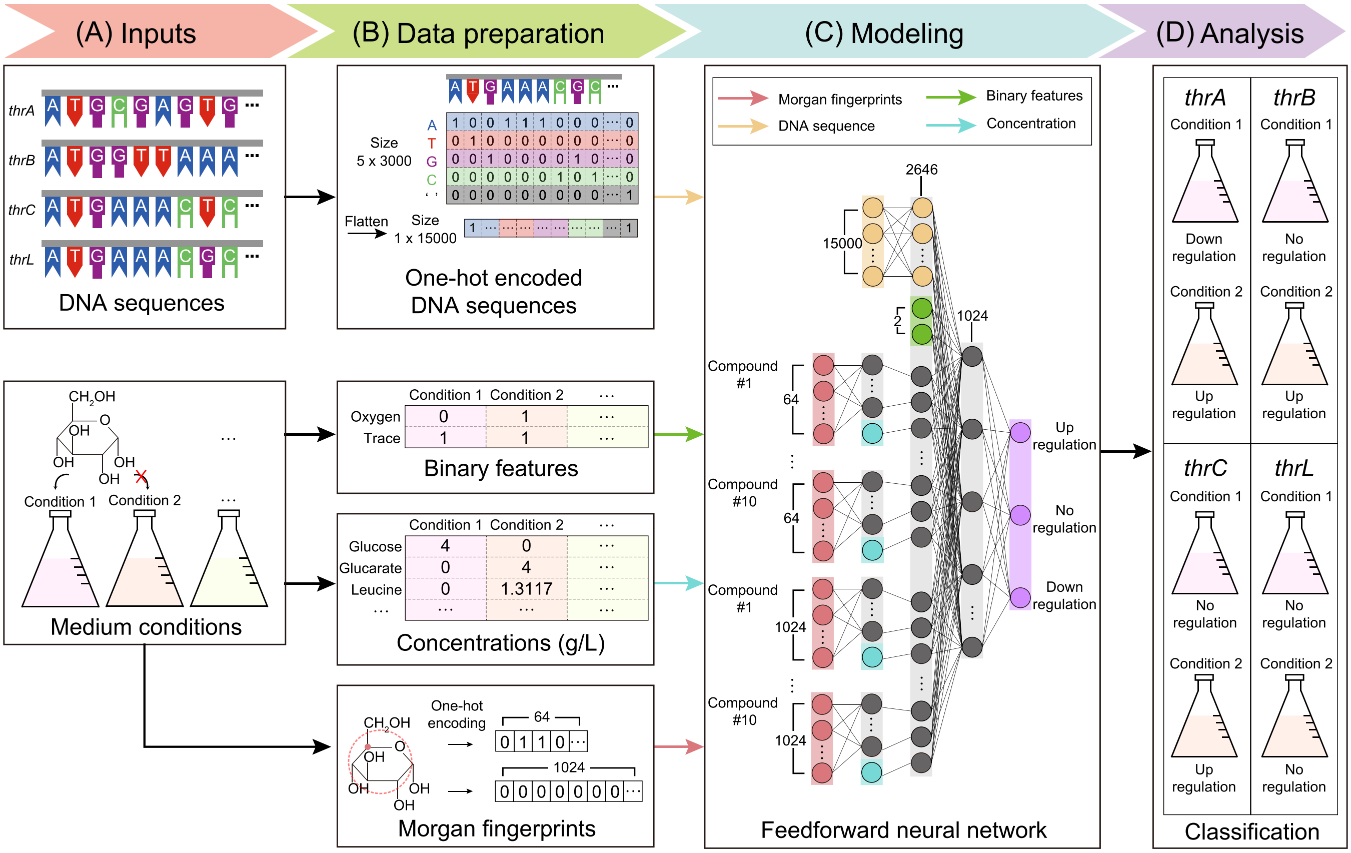


Fig. 1. Overall scheme of DeepMGR that predicts the direction of gene regulation in *E. coli* under a specific cultivation condition. (A) Inputs of DeepMGR include DNA sequence of a target gene, presence (‘1′) or absence (‘0′) of aeration and trace elements, and concentration and structural information (SMILES) of up to ten nutrients. DeepMGR accepts information on a single gene and one cultivation condition at a time. (B) All the input data first need to be represented (or featurized) as numerical vectors. A one-hot encoded DNA sequence vector has a length of 15,000 (= 5 ×3000) because each base requires five elements to represent ‘A′, ‘T′, ‘G′, ‘C′, or an empty space, and genes with up to 3000 base pairs were considered in this study. Presence or absence of aeration and trace elements are indicated as binary features. Real numbers are used for concentration of up to ten nutrients. For structural information, SMILES of up to ten nutrients are encoded as Morgan fingerprints with radius of 3 and bits of 64 and 1024. (C) Feedforward neural network receives the featurized input data through various nodes, and classifies whether the target gene is upregulated, downregulated, or not regulated under a cultivation condition. (D) Classification results from DeepMGR can be further analyzed in comparison with experimental data.

objectives (e.g., drug screening). Equally important question would be, especially in the context of metabolic engineering, to consider the overall effects of the entire cultivation condition, including the type and concentration of nutrients in growth medium as well as the presence or absence of aeration and trace elements (i.e., a set of inorganic compounds). To the best of our knowledge, such prediction models have not been developed that consider the overall effects of cultivation condition on the cell’s transcriptional response.

In this study, we report the development of DeepMGR, a deep learning model that predicts the effects of culture *media* on gene regulation (hence, ‘MGR’) in *E. coli* (Fig. 1). DeepMGR takes DNA sequence of a target gene and information on cultivation condition (e.g., medium composition) as inputs, and classifies the direction of regulation (i.e., upregulation, no regulation, or downregulation) for the input gene as output (Fig. 1). DeepMGR is an addition to a suite of machine learning models developed for metabolic engineering, and will be useful for designing experiments, in particular growth media, for microbial biotechnology [10–12].

2. Materials and methods

2.1. Data collection

GenBank files of two *E. coli* strains, K-12 MG1655 and BW25113, were obtained from NCBI Nucleotide (accession numbers: NC_000913.3 and NZ_CP009273.1, respectively). DNA sequence of each gene was retrieved from these GenBank files to generate the gold standard dataset for DeepMGR development. A total of 80 RNA-seq data (70 for *E. coli* K-12 MG1655, and 10 for *E. coli* K-12 BW25113) were collected in this study. Among the 80 RNA-seq data, 77 data covering both the MG1655 and BW25113 strains were obtained from a Github repository (https://github.com/SBRG/precise-db) [13], and 3 additional data on the MG1655 strain were obtained from NCBI Gene Expression Omnibus (GEO; accession numbers: GSM3463601, GSM3463602 and GSM1581602). Information on the 77 RNA-seq data is also available at NCBI GEO. Growth conditions corresponding to the 80 RNA-seq data were available in ‘Growth protocol’ of NCBI GEO.

Structural information of nutrients (e.g., main carbon sources) available in growth media was obtained from PubChem [14] through PUG-REST (https://pubchem.ncbi.nlm.nih.gov/docs/pug-rest; [15]). The structural information was subsequently converted to simplified molecular-input line-entry system (SMILES) with an isomeric form for each chemical to distinguish isomers. The collected SMILES were converted to Morgan fingerprints with radius of 3 and bits of 64 and 1024 by using `GetMorganFingerprintAsBitVec` in RDKit 2021.03.3. Here, Morgan fingerprint is a molecular fingerprinting method used in cheminformatics to represent the structural features of molecules [16,17]. In this study, Morgan fingerprints were considered as input for DeepMGR to process nutrients beyond those available in the gold standard dataset.

Information on Biolog PM1 and PM2 was obtained from EcoCyc (https://biocyc.org/ECOLI/NEW-IMAGE?object=Growth-Media; [18]). Growth conditions in Biolog PM1 and PM2 involve 192 different

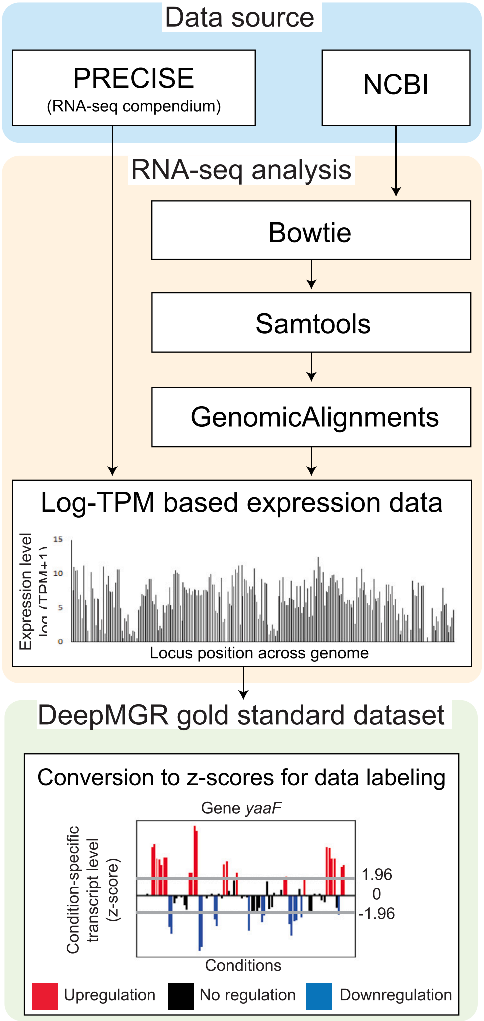


Fig. 2. Preprocessing of 80 RNA-seq data collected from a Github repository (https://github.com/SBRG/precise-db) [13] and NCBI GEO for preparation of the gold standard dataset. The Github repository provides the already processed expression data that are presented in $\log_2$(TPM+1). RNA-seq data from NCBI GEO were quantified in $\log_2$(TPM+1) by using Bowtie 1.3.0 [19], Samtools 1.11 [20] and GenomicAlignments 3.14 [21]. The quantified expression levels were converted to z-scores, which were subsequently used to classify upregulation, no regulation, or downregulation for a target gene under a specific cultivation condition (Materials and methods and 'Prediction for the 4020 genes' of Supplementary Data 3). The resulting z-scores higher than 1.96 and those lower than − 1.96 were considered as upregulation and downregulation, respectively.

carbon sources, which showed varied growths of *E. coli*: growth observed under 80 conditions; low growth under 2 conditions; no growth under 101 conditions; and 9 inconsistent growths. Based on this information, 82 conditions were selected for simulation studies of DeepMGR in this study.

2.2. Quantification and normalization of RNA-seq data

Quantification and normalization were conducted for the three RNA-seq data (GSM3463601, GSM3463602 and GSM1581602) from NCBI GEO by using the following software programs, which were also used to process the 77 RNA-seq data available at the Github repository (Fig. 2): Bowtie 1.3.0 [19], Samtools 1.11 [20] and GenomicAlignments 3.14 [21]. Briefly, Bowtie is a short-read aligner that maps RNA-seq reads to the two reference genomes (i.e., NC_000913.3 and NZ_CP009273.1, in this study), and generates Sequence Alignment Map (SAM) files. In Bowtie, reference genome data for NC_000913.3 and NZ_CP009273.1 were first created using `bowtie-build -f sequence_NC_000913.fasta ref_sequence_NC_000913` and `bowtie-build -f sequence_NZ_CP009273.fasta ref_sequence_NZ_CP009273`, respectively. Next, SAM files for NC_000913.3 and NZ_CP009273.1 were obtained by using `bowtie -X 1000 -n 2 -3 3 -x ref_sequence_NC_000913 -1 alignment1.fastq -2 alignment2.fastq -S result_NC_000913.sam`, and `bowtie -X 1000 -n 2 -3 3 -x ref_sequence_NZ_CP009273 -1 alignment1.fastq -2 alignment2.fastq -S result_NZ_CP009273.sam`, respectively. These SAM files were converted to BAM files by implementing Samtools with `samtools view -Sb result_NC_000913.sam -o result_NC_000913.bam` and `samtools view -Sb result_NZ_CP009273.sam -o result_NZ_CP009273.bam`. GenomicAlignments, an R-based package to count the pre-aligned short reads, was used to calculate gene expression levels in $\log_2$(TPM+1) with the following arguments for `summarizeOverlaps`: `mode = "IntersectionStrict"`; `singleEnd = FALSE`; and `ignore.strand = FALSE`, `preprocess.reads = invertStrand`.

2.3. Labeling of RNA-seq data

To label each gene as upregulation, no regulation, or downregulation, the RNA-seq processing protocol implemented by Sebestyén et al. [22] was adopted in this study. For this, 13 cultivation data, all involving M9 minimal medium with 2 g/L glucose as a single carbon source [23,24], were used as a 'control condition'. Next, median absolute deviation (MedianAD) was calculated for each gene from these 13 RNA-seq data. Finally, z-score was calculated for each gene expression level from all the other 67 cultivation conditions according to the following formula from Sebestyén et al. [22]:

$$z = \begin{cases} \frac{Gene\ expression\ level - Median}{1.486 \times MedianAD} \&(if\ MedianAD \neq 0) \\ \frac{Gene\ expression\ level - Median}{1.253314 \times MeanAD} \&(if\ MedianAD = 0) \end{cases}$$

z-scores higher than 1.96 and those lower than − 1.96 were considered as upregulation and downregulation, respectively; z-scores between − 1.96 and 1.96 correspond to no regulation. If MedianAD was zero, mean absolute deviation (MeanAD) with a scale factor of 1.253314 was used instead [22].

2.4. Training and optimization of DeepMGR

Feedforward neural network (FNN) within DeepMGR was developed by using Keras 2.4.0 with TensorFlow backend 2.3.1 [25]. Scikit-learn 0.23.2 [26] was used to split the gold standard dataset, and evaluate the model's classification performance with respect to accuracy, precision, recall, and F1 score. Here, F1 score was considered important in this study because the gold standard dataset used in this study is highly imbalanced (i.e., 32,072 upregulated genes and 27,639 downregulated genes in contrast to 261,889 genes with average expression levels among a total of 321,600 genes from 4020 genes across the 80 conditions). The dataset was split into training (60%), validation (20%) and test (20%) sets for the model development. For the optimization of hyperparameters, `BayesianOptimization` class in KerasTuner 1.0.1 was used to maximize F1 score for the validation set. Hyperparameters were selected, which gave the greatest F1 scores for the validation set (Fig. S1 and Table S1).

A UMAP plot for the DeepMGR results using the test set of the gold standard dataset was generated by using umap-learn 0.5.0 [27]. Two hyperparameters `n_neighbors` and `min_dist` were set to 15 and 0.2, respectively, after examining several different values: [5, 10, 15, 20, 25] for `n_neighbors` and [0.1, 0.2, 0.3, 0.4, 0.5] for `n_neighbors`.

Table 1
Classification performance of DeepMGR with different machine learning methods. The best performance is presented in bold.

Machine learning methods	Accuracy	F1 score	Precision	Recall
DeepMGR	**0.867**	**0.703**	0.744	**0.671**
DeepMGR with convolution	0.830	0.632	0.640	0.625
DeepMGR with *k*-nearest neighbors	0.813	0.544	0.600	0.512
DeepMGR with random forest	0.864	0.675	**0.762**	0.623

2.5. Use of different machine learning methods for DeepMGR

The classification performance of different versions of DeepMGR was examined by using different machine learning methods, including: DeepMGR with FNN, but with additional use of convolution and 1-max pooling layers (hereafter, 'DeepMGR with convolution'); DeepMGR with random forest (RF) in place of FNN; and DeepMGR with *k*-nearest neighbors (*k*NN) in place of FNN. RF and *k*NN were implemented using scikit-learn. For DeepMGR with convolution, the same optimization process as DeepMGR with FNN only was followed, but additional hyperparameters, such as kernel size and pool size, were also optimized (Table S2). For RF, `n_estimators` of 150 and `max_depth` of 40 gave the best macro F1 score after examining [50, 100, 150, 200] for `n_estimators` and [15, 20, 25, 30, 35, 40, 45] for `max_depth` based on grid search (Table S3). For *k*NN, `n_neighbors` of 3 led to the best Macro F1 score after examining [2, 3, 5, 10] via grid search (Table S4). It should be noted that the results presented in Table 1 come from the parameters that gave the best macro F1 score.

3. Results

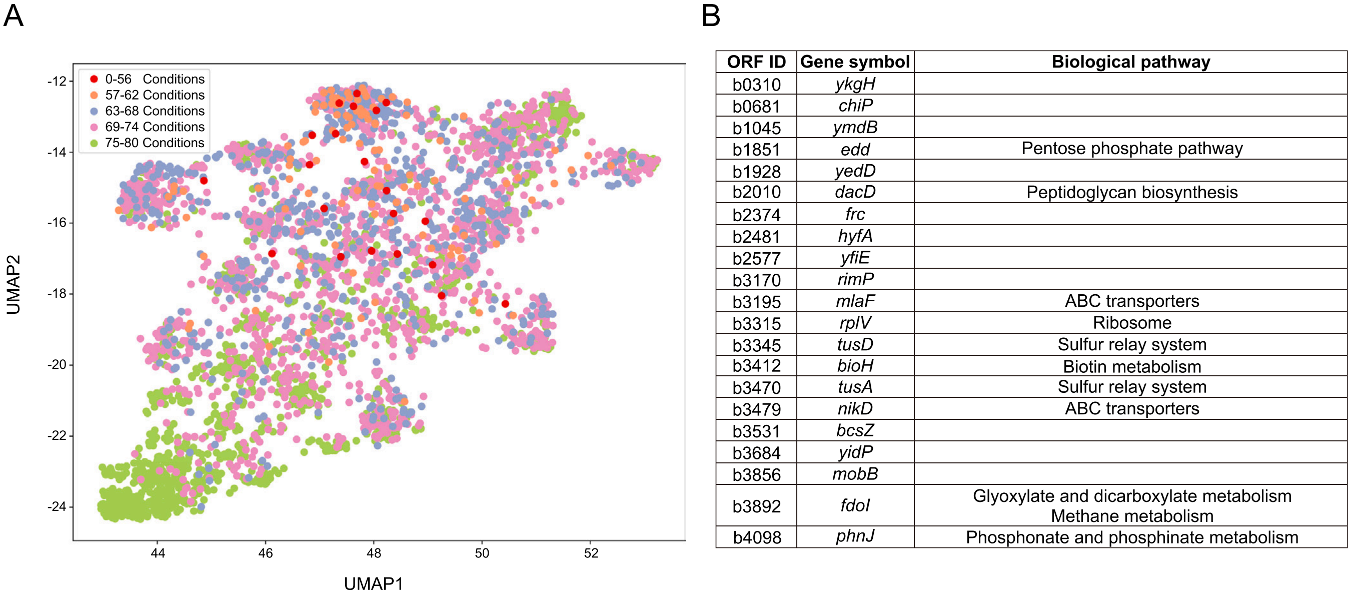
3.1. Preparation of a gold standard dataset

Information on cultivation conditions of *E. coli* was considered as input data for DeepMGR, including DNA sequence of a target gene, presence or absence of aeration and trace elements, and concentration and structural information (SMILES) of up to ten nutrients (Supplementary Data 1). In order to systematically examine the effects of cultivation conditions on gene regulations, RNA-seq data of the two *E. coli* strains, K-12 MG1655 and BW25113, were collected, which all involved the use of defined media. Experimental conditions that involved the use of antibiotics and/or complex media were not considered because the number of such samples was too small for model training. As a result, the collected 80 RNA-seq data were considered to prepare the DeepMGR gold standard dataset.

The gold standard dataset also requires DNA sequences of genes. For this, DNA sequences of 4389 genes were extracted from the GenBank files of both *E. coli* K-12 MG1655 and BW25113; 4389 genes are those commonly available in the two *E. coli* strains. DNA sequences of the 4389 genes were subjected to one-hot encoding method for featurization. Here, 369 of the 4389 genes were removed based on the following four criteria (Supplementary Data 2): 1) 55 genes with more than 3000 base pairs, which correspond to 1.25% of the entire 4389 genes [28]; 2) 34 genes with either identical DNA sequence for multiple genes or multiple DNA sequences per gene in the GenBank file; 3) 267 genes that were either not expressed in all the 80 RNA-seq data, or not present in both the *E. coli* strains K-12 MG1655 and BW25113; and 4) 13 genes with highly homologous DNA sequences, having pairwise distance values of less than 0.1 in comparison with homologs of these 13 genes according to Clustal Omega 1.2.2 [29]. Finally, the resulting 4020 genes were labeled as upregulation, no regulation, or downregulation for each of the 67 RNA-seq data in comparison with a control condition (Materials and methods, and 'Labeling of the 4020 genes' of Supplementary Data 3); 13 out of the 80 RNA-seq data, all involving M9 minimal medium with 2 g/L glucose as a single carbon source [22–24], were used as a 'control condition'. The resulting gold standard dataset includes 32,072 upregulated genes, 27,639 downregulated genes, and 261,889 genes with no regulation.

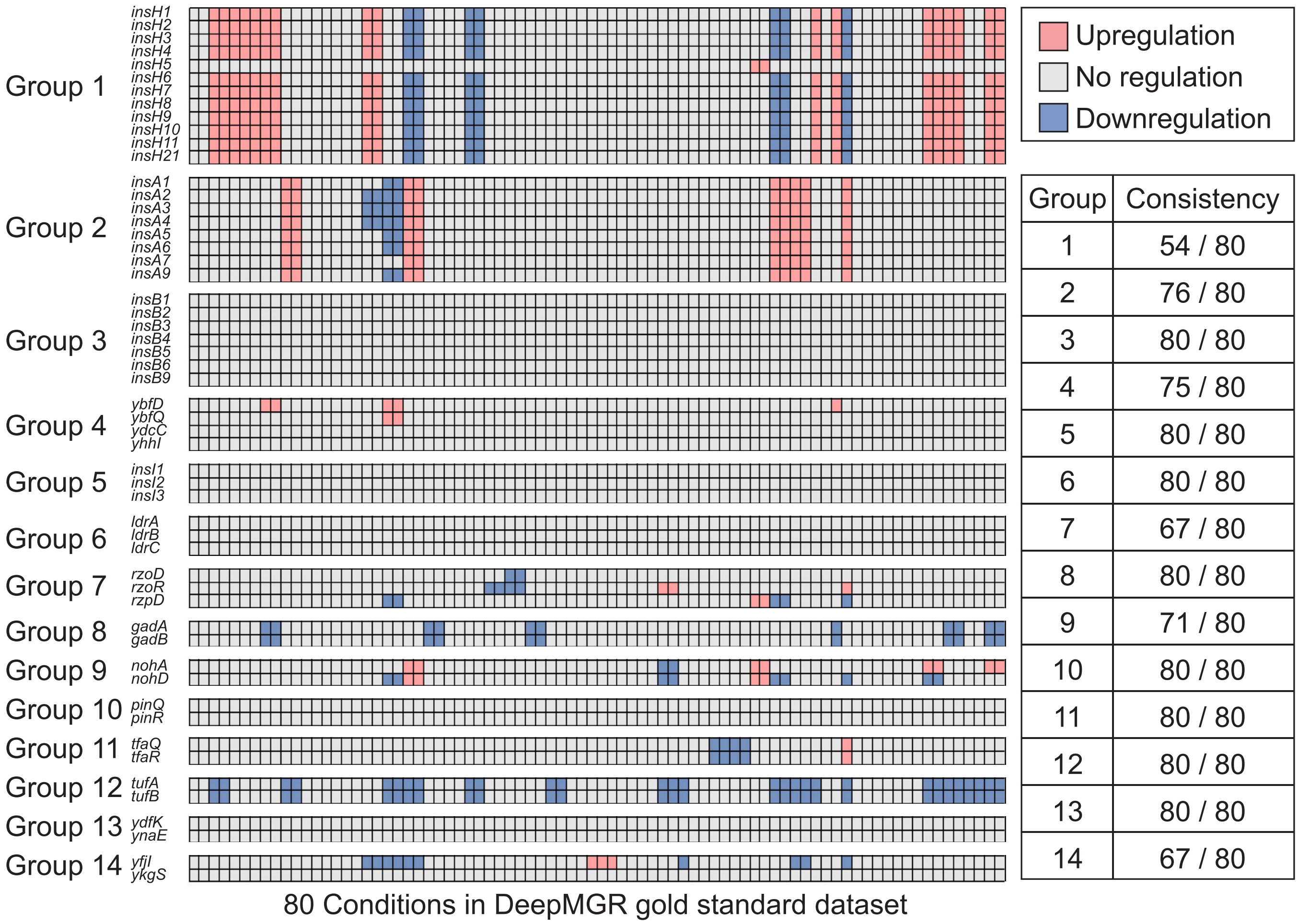
3.2. Development of DeepMGR to predict the direction of a gene regulation

DeepMGR uses an FNN to predict the direction of gene regulation (i.e., upregulation, no regulation, or downregulation) under a given



ORF ID	Gene symbol	Biological pathway
b0310	*ykgH*	
b0681	*chiP*	
b1045	*ymdB*	
b1851	*edd*	Pentose phosphate pathway
b1928	*yedD*	
b2010	*dacD*	Peptidoglycan biosynthesis
b2374	*frc*	
b2481	*hyfA*	
b2577	*yfiE*	
b3170	*rimP*	
b3195	*mlaF*	ABC transporters
b3315	*rplV*	Ribosome
b3345	*tusD*	Sulfur relay system
b3412	*bioH*	Biotin metabolism
b3470	*tusA*	Sulfur relay system
b3479	*nikD*	ABC transporters
b3531	*bcsZ*	
b3684	*yidP*	
b3856	*mobB*	
b3892	*fdol*	Glyoxylate and dicarboxylate metabolism Methane metabolism
b4098	*phnJ*	Phosphonate and phosphinate metabolism

Fig. 3. Overview of the DeepMGR results using the gold standard dataset. (A) UMAP plot of the DeepMGR results for the 4020 genes. Each dot is a single gene, and its dot color indicates the number of correct predictions (i.e., direction of gene regulation) made among the 80 conditions from the gold standard dataset. For example, 21 red dots indicate 21 genes, for which the number of correct predictions made was between 0 and 56 cultivation conditions. These 21 genes are: *bcsZ*, *bioH*, *chiP*, *dacD*, *edd*, *fdo*, *lfrc*, *hyfA*, *mlaF*, *mobB*, *nikD*, *phnJ*, *rimP*, *rplV*, *tusA*, *tusD*, *yedD*, *yfiE*, *yidP*, *ykgH*, and *ymdB*. (B) Biological pathways associated with the 21 genes (red dots in the UMAP plot). Empty lines indicate that there are no biological pathways known to be associated with the gene.



Group	Consistency
1	54 / 80
2	76 / 80
3	80 / 80
4	75 / 80
5	80 / 80
6	80 / 80
7	67 / 80
8	80 / 80
9	71 / 80
10	80 / 80
11	80 / 80
12	80 / 80
13	80 / 80
14	67 / 80

Fig. 4. DeepMGR results for predicting the regulation directions for groups of homologous genes. A total of 54 homologous genes were clustered as 14 groups on the basis of analysis using Clustal Omega. 'Consistency' indicates the number of consistent predictions made for each group across the 80 different cultivation conditions from the gold standard dataset.

cultivation condition (Figs. 1 and S2). To run the FNN, DNA sequence of a target gene and information on a cultivation condition were used as input (Fig. 1A). For DNA sequence, it was first subjected to one-hot encoding and padding, which resulted in a matrix having a dimension of 5 × 3000 (each row for 'A′, 'T′, 'G′, 'C′, and an empty space; Fig. 1B). This matrix was subsequently flattened as a single vector (1 ×15,000) to be used as input for the FNN. Here, it is important to note that, in DeepMGR, DNA sequence of a gene was considered without regulatory elements (e.g., promoter and enhancer) because DNA sequence has been reported to contain the majority of information on gene expression levels [30]. Also, our knowledge on the regulatory components for genes in response to a variety of nutrients is limited, even in *E. coli* [31]. Therefore, we attempted to predict the directions of gene regulation by using the well-defined DNA sequence of genes.

For the information on a cultivation condition, it includes the presence or absence of oxygen and trace elements, and concentration and SMILES of up to ten nutrients (Fig. 1B). Although information on 38 nutrients (Supplementary Data 1) was used to develop DeepMGR, information on nutrients other than these 38 nutrients can be used for a gene regulation prediction. Also, the maximum number of nutrients considered for the FNN input was determined to be ten because the maximum number of main nutrients used across the collected 80 RNA-seq data was nine. Additionally, the presence or absence of oxygen and trace elements was indicated as binary features because: 1) dissolved oxygen concentration was not available in the RNA-seq data descriptions at NCBI GEO; and 2) the amounts of trace elements were very small, compared with the 38 nutrients (Supplementary Data 1). For SMILES of the nutrients, the corresponding Morgan fingerprints were used to represent each main nutrient in a medium with 64 and 1024 bits [16,17]. Use of Morgan fingerprints with both 64 and 1024 bits appeared to show the best classification performance after examining the performance of various combinations of Morgan fingerprints with 64, 128, 256, 512 and 1024 bits (Table S5).

FNN of DeepMGR has a total of five layers. Its first and second layers receive Morgan fingerprints (both 64 and 1024 bits, and the radius of 3; red nodes in Fig. 1C) and concentration (light blue nodes in Fig. 1C) of each nutrient, respectively. Its third layer with 2646 nodes additionally accepts one-hot encoded DNA sequence (light orange nodes in Fig. 1C) and binary information on aeration and trace elements (green nodes in Fig. 1C). The fourth layer was added as a result of the model optimization (Table S1). The output layer classifies the direction of gene regulation as upregulation, no regulation, or downregulation (Fig. 1 C,D). The best-performing model was determined by implementing Bayesian optimization [32] with 20 different hyperparameter sets (Table S1 and 'Prediction for the 4020 genes' of Supplementary Data 3). Using the test set, DeepMGR showed accuracy and F1 score of 0.867 and 0.703, respectively. Macro F1 scores for predicting the upregulated and downregulated genes were 0.591 and 0.595, respectively.

To further evaluate the prediction performance of DeepMGR, a UMAP plot was prepared that displays the number of correct predictions (i.e., direction of gene regulation) made for the 4020 genes

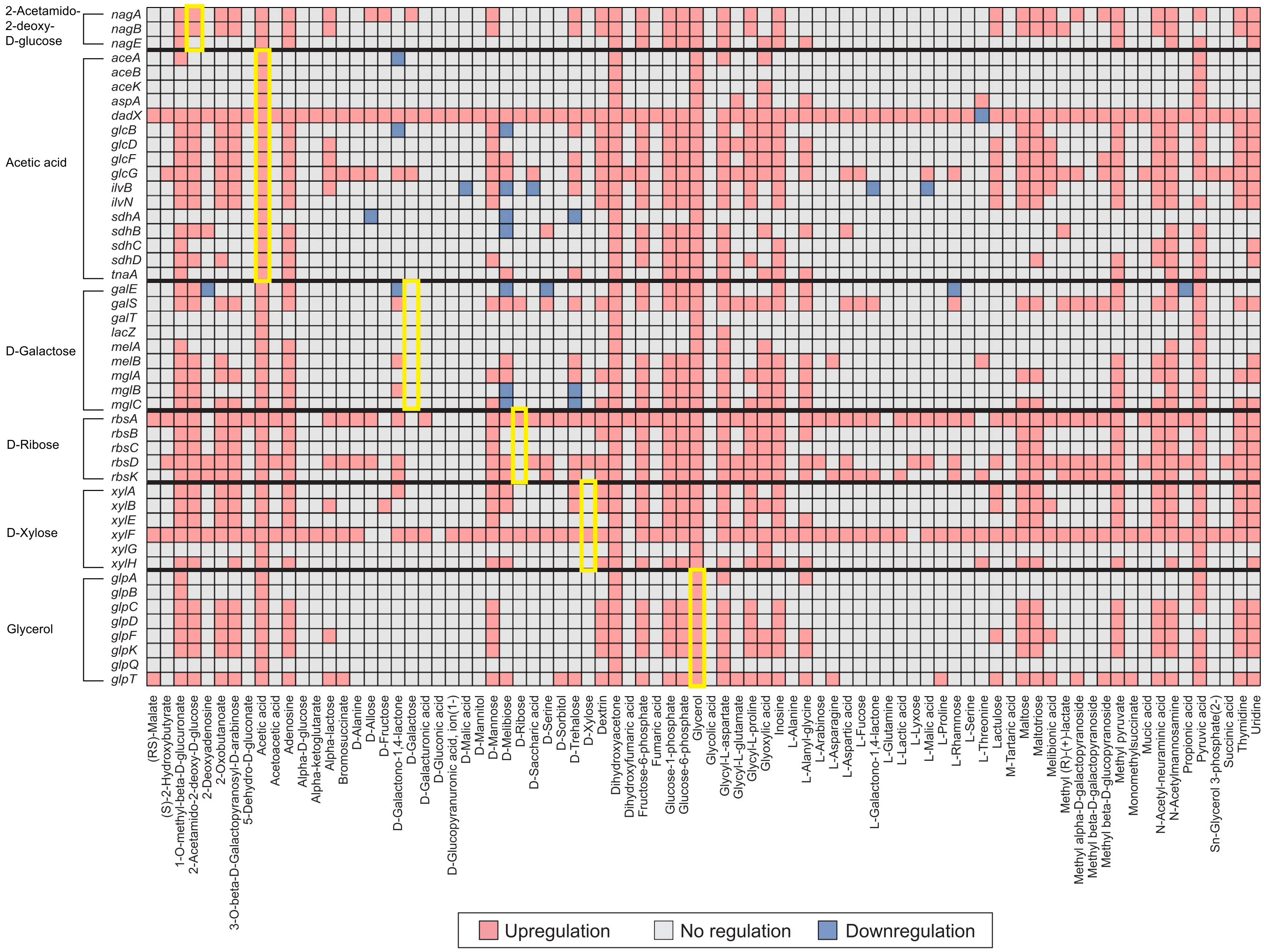


Fig. 5. DeepMGR results for 47 genes across the 82 Biolog conditions. For example, *nagA*, *nagB* and *nagE* become overexpressed by 2-acetamido-2-deoxy-D-glucose present in a medium [34], and additional five groups are also presented likewise (i.e., acetic acid [35]; D-galactose [34]; D-ribose [36]; D-xylose [37]; and glycerol [38]). (a) These 47 genes were reported to be upregulated by carbon sources presented on the y-axis. (b) Yellow boxes indicate the DeepMGR results for the 47 genes when they were exposed to their regulating carbon sources.

across the 80 conditions (Fig. 3A). The results showed that the correct predictions were made for 3999 genes in at least 57 or more cultivation conditions. The remaining 21 genes were the ones that showed relatively poor predictions; for these 21 genes, correct predictions were made for only 0–56 cultivation conditions (red dots in Fig. 3A). Interestingly, 11 out of the 21 genes were not assigned with biological pathways, and the remaining 10 genes were all associated with different pathways (Fig. 3B). Taken together, DeepMGR appeared to make correct predictions for most of the genes (i.e., 3999 genes for 57 or more cultivation conditions), and no particular biological features were found among the 21 genes that led to the relatively poor predictions.

3.3. Classification performance of DeepMGR with different machine learning methods

Next, the classification performance of different versions of DeepMGR was examined by using different machine learning methods in DeepMGR in order to justify the use of current architecture of the FNN. In addition to the original DeepMGR considered above, following versions were additionally considered in this analysis (Fig. S2): DeepMGR with convolution; DeepMGR with RF in place of FNN; and DeepMGR with *k*NN in place of FNN (Table 1). For DeepMGR with convolution, convolution and 1-max pooling layers were additionally considered because they were reported to effectively capture the sequence information [33]. These variant models were trained with the same training and test sets used for the original version of DeepMGR with FNN, but the hyperparameter sets were independently optimized for each version (Tables S1–S4); Bayesian optimization was used for DeepMGR with convolution, and a set of hyperparameters were examined for DeepMGR with RF (Table S3) and DeepMGR with *k*NN (Table S4). As a result, the original DeepMGR showed the highest accuracy and F1 score, followed by DeepMGR with RF, DeepMGR with convolution, and DeepMGR with *k*NN (Table 1). This analysis partly justifies the use of DeepMGR with FNN for further subsequent analyses, and other DeepMGR versions were no longer considered.

3.4. Predicting directions of regulation for groups of homologous genes

First, we challenged DeepMGR to determine if it can classify a group of homologous genes into the same direction of regulation, based on the assumption that the homologous genes would show the same response under a given condition. For this, we obtained 14 groups of genes, involving a total of 54 genes, where each group contains multiple homologous genes according to Clustal Omega (Fig. 4). These 54 genes were subjected to DeepMGR under 80 different conditions that were available in the gold standard dataset

(Fig. 4 and Supplementary Data 4). As a result, genes in the same group overall showed similar directions of regulation, while different groups of genes showed notably different directions of regulation under 80 different conditions; the consistent prediction results ranged from 54 (from group 1 in Fig. 4) to 80 out of the 80 different conditions (from groups 3, 5, 6, 8, 10, 11, 12 and 13 in Fig. 4). These prediction results partly showed that DeepMGR classifies directions of gene regulation by taking into account DNA sequences.

3.5. Predicting directions of regulation for genes from Biolog data

Next, DeepMGR was implemented under 82 different Biolog conditions. Among the growth media in Biolog PM1 and PM2, which were all conducted using M63 minimal medium with 192 different carbon sources (https://biocyc.org/ECOLI/NEW-IMAGE?object=Growth-Media; [18]), *E. coli* was reported to survive using 82 carbon sources, and *E. coli* did not grow under other remaining conditions. In this analysis, the regulation directions were predicted for a total of 47 genes under the 82 Biolog conditions, and they were compared with the literature (Fig. 5 and Supplementary Data 5). These 47 genes were selected because these genes were reported to be upregulated by specific carbon sources, which also correspond to the Biolog conditions [34–38]; the carbon sources upregulating these 47 genes include 2-acetamido-2-deoxy-D-glucose, acetic acid, D-galactose, D-ribose, D-xylose, and glycerol (Fig. 5). Also, 69 out of the 82 Biolog conditions include nutrients not covered by the gold standard dataset (e.g., acetoacetic acid, dextrin, and maltose), providing an opportunity to rigorously validate DeepMGR. As a result of the predictions using DeepMGR, many of these genes (28 out of 47 genes) were predicted to be upregulated when their upregulating carbon source was available in the cultivation condition. This analysis partially validates DeepMGR's performance; however, obtaining additional experimental data on carbon sources that regulate specific genes will enable a more rigorous validation and updating of DeepMGR.

4. Discussion and conclusion

In this study, we developed a deep learning model DeepMGR that predicts the direction of gene regulation under a specific cultivation condition in comparison with M9 minimal medium with glucose as a control condition. DeepMGR takes DNA sequence of a target gene, presence or absence of aeration and trace elements, and concentration and SMILES of up to ten nutrients as inputs. Once developed, DeepMGR underwent two distinct simulation studies: predicting the regulation directions for groups of homologous genes (Fig. 4), and comparing the simulation results to the literature, focusing on six carbon sources that regulate specific genes (Fig. 5). While the simulation studies showed overall positive results for DeepMGR, further research opportunities were also clearly observed.

First, the most critical limiting factor in developing DeepMGR was the insufficient volume of RNA-seq data obtained from a variety of clearly defined cultivation conditions. Such data were somewhat available for *E. coli*, but they were not available for other representative organisms, such as *Corynebacterium glutamicum* and *Bacillus* species. This problem consequently resulted in a specific architecture of DeepMGR, for example consideration of up to ten nutrients in the input layer of the FNN. Upon more availability of RNA-seq data from diverse cultivation conditions and also from other organisms, the architecture of DeepMGR could be updated for wider applications. Next, to improve the biological explainability of the model prediction results, gene regulatory elements such as promoters and 5′-UTRs can be additionally considered. These challenges became evident during the development of DeepMGR. By successfully addressing them, it will be possible to develop a more robust model suitable for various biotechnology applications.

CRediT authorship contribution statement

Mun Su Kwon: Conceptualization, Methodology, Investigation, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Figure drawing. **Joshua Julio Adidjaja:** Methodology, Investigation, Validation, Formal analysis, Data curation. **Hyun Uk Kim:** Conceptualization, Resources, Data curation, Funding acquisition, Supervision, Project administration, Writing – review & editing.

Declaration of Competing Interest

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.

Acknowledgments

This work was supported by the Bio & Medical Technology Development Program (2018M3A9F3079664), by the Development of next-generation biorefinery platform technologies for leading bio-based chemicals industry project (2022M3J5A1056072), and by the Development of platform technologies of microbial cell factories for the next-generation biorefineries project (2022M3J5A1056117) from National Research Foundation supported by the Korean Ministry of Science and ICT.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.04.010.

References

[1] Imdahl F, Vafadarnejad E, Homberger C, Saliba A-E, Vogel J. Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria. Nat Microbiol 2020;5:1202–6. https://doi.org/10.1038/s41564-020-0774-1

[2] Cardoso VM, Campani G, Santos MP, Silva GG, Pires MC, Gonçalves VM, et al. Cost analysis based on bioreactor cultivation conditions: production of a soluble recombinant protein using *Escherichia coli* BL21(DE3). Biotechnol Rep 2020;26:e00441. https://doi.org/10.1016/j.btre.2020.e00441

[3] Song H, Kim TY, Choi B-K, Choi SJ, Nielsen LK, Chang HN, et al. Development of chemically defined medium for *Mannheimia succiniciproducens* based on its genome sequence. Appl Microbiol Biotechnol 2008;79:263–72. https://doi.org/10.1007/s00253-008-1425-2

[4] Machas M, Kurgan G, Abed OA, Shapiro A, Wang X, Nielsen D. Characterizing *Escherichia coli*'s transcriptional response to different styrene exposure modes reveals novel toxicity and tolerance insights. kuab019 J Ind Microbiol Biotechnol 2021;48. https://doi.org/10.1093/jimb/kuab019

[5] LaVoie SP, Summers AO. Transcriptional responses of *Escherichia coli* during recovery from inorganic or organic mercury exposure. BMC Genom 2018;19:52. https://doi.org/10.1186/s12864-017-4413-z

[6] Eslami M, Borujeni AE, Eramian H, Weston M, Zheng G, Urrutia J, et al. Prediction of whole-cell transcriptional response with machine learning. Bioinformatics 2022;38:404–9. https://doi.org/10.1093/bioinformatics/btab676

[7] Kwon MS, Lee BT, Lee SY, Kim HU. Modeling regulatory networks using machine learning for systems metabolic engineering. Curr Opin Biotechnol 2020;65:163–70. https://doi.org/10.1016/j.copbio.2020.02.014

[8] Woo G, Fernandez M, Hsing M, Lack NA, Cavga AD, Cherkasov A. DeepCOP: deep learning-based approach to predict gene regulating effects of small molecules. Bioinformatics 2020;36:813–8. https://doi.org/10.1093/bioinformatics/btz645

[9] Pham T-H, Qiu Y, Zeng J, Xie L, Zhang P. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. Nat Mach Intell 2021;3:247–57. https://doi.org/10.1038/s42256-020-00285-9

[10] Kim GB, Kim WJ, Kim HU, Lee SY. Machine learning applications in systems metabolic engineering. Curr Opin Biotechnol 2020;64:1–9. https://doi.org/10.1016/j.copbio.2019.08.010

[11] Lawson CE, Martí JM, Radivojevic T, Jonnalagadda SVR, Gentz R, Hillson NJ, et al. Machine learning for metabolic engineering: A review. Metab Eng 2021;63:34–60. https://doi.org/10.1016/j.ymben.2020.10.005

[12] Li F, Yuan L, Lu H, Li G, Chen Y, Engqvist MKM, et al. Deep learning-based k_{cat} prediction enables improved enzyme-constrained model reconstruction. Nat Catal 2022;5:662–72. https://doi.org/10.1038/s41929-022-00798-z

[13] Sastry AV, Gao Y, Szubin R, Hefner Y, Xu S, Kim D, et al. The *Escherichia coli* transcriptome mostly consists of independently regulated modules. Nat Commun 2019;10:5536. https://doi.org/10.1038/s41467-019-13483-w

[14] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2023 update. Nucleic Acids Res 2023;51:D1373–80. https://doi.org/10.1093/nar/gkac956

[15] Kim S, Thiessen PA, Bolton EE, Bryant SH. PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem. Nucleic Acids Res 2015;43:W605–11. https://doi.org/10.1093/nar/gkv396

[16] Morgan HL. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. J Chem Doc 1965;5:107–13. https://doi.org/10.1021/c160017a018

[17] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50:742–54. https://doi.org/10.1021/ci100050t

[18] Keseler IM, Gama-Castro S, Mackie A, Billington R, Bonavides-Martínez C, Caspi R, et al. The EcoCyc database in 2021. Front Microbiol 2021:12.

[19] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009;10:R25. https://doi.org/10.1186/gb-2009-10-3-r25

[20] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9. https://doi.org/10.1093/bioinformatics/btp352

[21] Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol 2013;9:e1003118. https://doi.org/10.1371/journal.pcbi.1003118

[22] Sebestyén E, Singh B, Miñana B, Pagès A, Mateo F, Pujana MA, et al. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Res 2016;26:732–44. https://doi.org/10.1101/gr.199935.115

[23] Bedson P, Farrant TJD. Practical Statistics for the Analytical Scientist: A Bench Guide. Royal Society of Chemistry,; 2021.

[24] Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. J Am Stat Assoc 1993;88:1273–83. https://doi.org/10.1080/01621459.1993.10476408

[25] Abadi M., Barham P., Chen J., Chen Z., Davis A., Dean J., et al. TensorFlow: A system for large-scale machine learning. 12th USENIX Symp. Oper. Syst. Des. Implement. OSDI 16, 2016, p. 265–83.

[26] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825–30.

[27] McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. J Open Source Softw 2018;3:861. https://doi.org/10.21105/joss.00861

[28] Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci 2019;116:13996–4001. https://doi.org/10.1073/pnas.1821905116

[29] Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci Publ Protein Soc 2018;27:135–45. https://doi.org/10.1002/pro.3290

[30] Zrimec J, Börlin CS, Buric F, Muhammad AS, Chen R, Siewers V, et al. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nat Commun 2020;11:6141. https://doi.org/10.1038/s41467-020-19921-4

[31] Ireland WT, Beeler SM, Flores-Bautista E, McCarty NS, Röschinger T, Belliveau NM, et al. Deciphering the regulatory genome of *Escherichia coli*, one hundred promoters at a time. ELife 2020;9:e55308. https://doi.org/10.7554/eLife.55308

[32] Snoek J., Larochelle H., Adams R.P. Practical Bayesian optimization of machine learning Algorithms 2012. https://doi.org/10.48550/arXiv.1206.2944.

[33] Zeng H, Edwards MD, Liu G, Gifford DK. Convolutional neural network architectures for predicting DNA–protein binding. Bioinformatics 2016;32:i121–7. https://doi.org/10.1093/bioinformatics/btw255

[34] Soupene E, van Heeswijk WC, Plumbridge J, Stewart V, Bertenthal D, Lee H, et al. Physiological studies of *Escherichia coli* strain MG1655: growth defects and apparent cross-regulation of gene expression. J Bacteriol 2003;185:5611–26. https://doi.org/10.1128/JB.185.18.5611-5626.2003

[35] Oh M-K, Rohlin L, Kao KC, Liao JC. Global expression profiling of acetate-grown *Escherichia coli*∗. J Biol Chem 2002;277:13175–83. https://doi.org/10.1074/jbc.M110809200

[36] Chang D-E, Smalley DJ, Tucker DL, Leatham MP, Norris WE, Stevenson SJ, et al. Carbon nutrition of *Escherichia coli* in the mouse intestine. Proc Natl Acad Sci 2004;101:7427–32. https://doi.org/10.1073/pnas.0307888101

[37] Khankal R, Luziatelli F, Chin JW, Frei CS, Cirino PC. Comparison between *Escherichia coli* K-12 strains W3110 and MG1655 and wild-type *E. coli* B as platforms for xylitol production. Biotechnol Lett 2008;30:1645–53. https://doi.org/10.1007/s10529-008-9720-7

[38] Matamouros S, Hayden HS, Hager KR, Brittnacher MJ, Lachance K, Weiss EJ, et al. Adaptation of commensal proliferating *Escherichia coli* to the intestinal tract of young children with cystic fibrosis. Proc Natl Acad Sci 2018;115:1605–10. https://doi.org/10.1073/pnas.1714373115