History-Driven Genetic Modification Design Technique Using a Domain-Specific Lexical Model for the Acceleration of DBTL Cycles for Microbial Cell Factories

Shiori Nakazawa, Osamu Imaichi, Takahisa Kogure, Takeshi Kubota, Koichi Toyoda, Masako Suda, Masayuki Inui, Kiyoto Ito,* Tomokazu Shirai, and Michihiro Araki

ABSTRACT: The development of microbes for conducting bioprocessing via synthetic biology involves design—build—test—learn (DBTL) cycles. To aid the designing step, we developed a computational technique that suggests next genetic modifications on the basis of relatedness to the user’s design history of genetic modifications accumulated through former DBTL cycles conducted by the user. This technique, which comprehensively retrieves well-known designs related to the history, involves searching text for previous literature and then mining genes that frequently co-occur in the literature with those modified genes. We further developed a domain-specific lexical model that weights literature that is more related to the domain of metabolic engineering to emphasize genes modified for bioprocessing. Our technique made a suggestion by using a history of creating a Corynebacterium glutamicum strain producing shikimic acid that had 18 genetic modifications. Inspired by the suggestion, eight genes were considered by biologists for further modification, and modifying four of these genes proved experimentally efficient in increasing the production of shikimic acid. These results indicated that our proposed technique successfully utilized the former cycles to suggest relevant designs that biologists considered worth testing. Comprehensive retrieval of well-tested designs will help less-experienced researchers overcome the entry barrier as well as inspire experienced researchers to formulate design concepts that have been overlooked or suspended. This technique will aid DBTL cycles by feeding histories back to the next genetic design, thereby complementing the designing step.

KEYWORDS: genetic modification design, DBTL cycle, text mining, similarity search, bioinformatics, microbial cell factories

With increasing demand for sustainable manufacturing, material production done using microbial cell factories through synthetic biology has attracted considerable attention in various industrial fields because of the low environmental burden of these factories. The development of microbes that synthesize target compounds at the required productivity, which largely depends on metabolic engineering via genetic modification, involves substantial time and effort. Therefore, the concept of the design—build—test—learn (DBTL) cycle has been advocated in recent years to accelerate the development of such microbes with a systematic approach.1−4

The DBTL cycle is a repetitive cycle consisting of four steps: designing the metabolic pathway, genetic modification, and assembly that may enhance the productivity of the current strain, building these designs, testing the resultant strains for productivity and other relevant profiles, and learning these results to feed them back to the design for the next modification. Developing a desired microbial strain involves considerable repetitive optimization through DBTL cycles until the desired productivity is achieved, which requires substantial time and effort. Therefore, the smooth circulation of the steps in the DBTL cycle is key to rapidly developing microbes to satisfy the required productivity. Techniques for replacing manual procedures in every step of the DBTL cycle are required to accelerate the creation of microbial cell factories.

Designing at the first or earlier rounds of DBTL cycles is well-addressed by computational techniques, including those for metabolic pathways,5−9 genetic modifications,10−12 and regulatory elements on gene expression.13−15 Flux balance analysis,10−12 for example, can suggest a stoichiometrically optimized set of genetic modifications under arbitrary objective functions based on well-curated models of metabolic networks. Methodologies for the learning step, i.e., feeding information acquired in former cycles back to the next design for improved productivity, continue to be developed as well. The learning step consists of storing information acquired in former steps and proposing suggestions by using the stored information for improved design in the next round.3 Storage has been well-addressed by repositories, such as JBEI-ICE,16 the iGEM...
Significant endeavors have been made these days to optimize designs by learning in-house data on design factors and experimental results, such as EDD\textsuperscript{17} and those described in refs 20 and 21. Moreover, attempts to inventory design factors and experimental results from literature across the globe and then predict the experimental results of untested combinations of designs have been covered by tools such as LASER.\textsuperscript{22,23} LASER, for example, assists in creating better designs by formalizing manually curated data, including the designs and results of previous experiments. These learning tools can leverage well-curated data from in-house experiments and literature to suggest improvements within the models made of those data, thus helping to improve designs with solid support from experimental observations.

However, existing tools are less likely to expand toward designs which have not been considered by the user. Expanding designs, that is, helping to focus on designs that are related to the user’s purpose but not limited only to models that are based on solid experimental data, remains one of the challenges with DBTL cycles. For example, considering a situation in which all early designs of genetic modifications proved ineffective, a review of the design history, consisting of genes modified through former cycles, would help with the next designs. In this case, expanded concepts of next designs are sought after, such as modifications to the synthesis, consumption, transportation, or regulation of metabolites not yet considered in the history. However, reviewing a history to conceive of next designs is still dependent on humans and thus incomprehensive. A computational technique that helps expand design concepts that are related to the histories of users but have been overlooked or suspended by the users will drive DBTL cycles further.

To address the above challenge, we aimed at computer-aided retrieval of previously reported genetic modifications that are related to a history. This will help avoid overlooking well-tested approaches for designs and complement the next designing step, which is skill-dependent. We focused on the history accumulated through former rounds of DBTL cycles to design the next genetic modifications. One possible concept, that is, making suggestions on the basis of a history in order to guide and perform a next action, involves evaluating the “relatedness” of a possible next action to a history of former actions. This concept is utilized in “also-bought” approaches in e-retailing as well. Also-bought systems help predict items a customer will purchase next on the basis of the customer’s purchase history by extracting items bought frequently by other customers with similar purchase histories. Likewise, we hypothesized that genes to be modified next could be predicted from a user’s history of modifications by extracting genes modified frequently in studies that were “related” to the genes in the user’s history.

We set biomedical literature as the source for next designs since information on genes mentioned in published studies, including those neglected in models and public databases, would be best found there. Published literature contains the most valuable information on the functions or effects of genes that have been tested; thus, such literature can be a source of useful information for discovering new gene modification strategies. Text mining is a powerful approach to acquiring valuable information from literature. However, a considerable amount of biomedical literature, e.g., more than 30 million citations in PubMed, also contains information outside the domain of metabolic engineering. Studies from different domains may lead to noise in the suggestions, including genes that are important for understanding physiology or pathology but are ineffective for the development of productive microbes. Thus, literature in the domain of metabolic engineering should be emphasized more than that published in other domains, including basic biology and medical science. Hence, we also developed a domain-specific lexical model that evaluates the “relatedness” of
literature to that in the domain of metabolic engineering to emphasize genes mentioned in the literature that are more “related” to the domain.

Here, we propose a computational technique that aids in the design of next genetic modifications (Figure 1) by comprehensively retrieving data on previously published designs. The technique suggests a list of genes for next modifications that are ranked by relatedness on two axes:

1) The relatedness of a gene to the user’s history of modifications, which is determined by the co-occurrence of the gene and the genes composing the history with the literature collected by using history-based queries.

2) The relatedness of source literature for a gene to the domain of metabolic engineering, which is determined by our domain-specific lexical model.

This technique serves as a way of designing by feeding a history of former DBTL cycles back to the design for the next cycles. It works in a different way from previous learning tools that use experimental data to suggest designs from among former designs by learning relations between the designs and experimental results obtained by testing. Relatedness-based suggestions will help broaden the range of options for design concepts by retrieving designs that are related to but not limited to the available data, whereas suggestions made on the basis of the co-occurrence of each extracted gene against the design history. In this step, the technique will complement designs made on the basis of human intuition and existing computational design or learning techniques to drive DBTL cycles further. Comprehensive retrieval of well-tested designs will also help less-experienced researchers overcome the entry barrier. Moreover, it will help even experienced researchers to rediscover design concepts that have been overlooked, suspended, or that are out of their current focus.

**RESULTS AND DISCUSSION**

1. **Overview of Design Suggestion.** Figure 2 illustrates the overall framework of our proposed technique. In this technique, a design history is considered as a set of gene modification records. Each record includes the name of a gene that was modified in a parent strain in the previous DBTL cycle. Our technique provides a gene suggestion for the next DBTL cycle on the basis of the history by using three functional modules, that is, a literature searcher, a gene name extractor, and a gene scorer, as follows.

   First, the literature searcher collects literature related to a gene or genes of the same EC number, if applicable, for each record of modification by querying a literature database, such as PubMed. We opted for EC numbers, which imply the user’s purpose for metabolic engineering and the relevant metabolic pathways, over using exact gene names as queries with the intention of broadening the scope of information that can be retrieved. Then, the literature searcher determines a literature score $L$ for each piece of collected literature by using a domain-specific lexical model. The model is a collection of terms that are characteristic to the literature on metabolic engineering with weights. The $L$ is computed on the basis of the occurrence frequency of the terms in the model as an index that represents the relatedness of a piece of literature to metabolic engineering (details are given in the Methods section). The literature searcher repeats these procedures for collecting and scoring literature for all genes in the design history. Next, the gene name extractor collects gene names that appear in the titles and abstracts of the collected literature. Then, the gene scorer computes the relatedness of each extracted gene against the design history. In this step, the co-occurrence between an extracted gene and genes in the design history is determined as the relatedness to the design history. The co-occurrence in each piece of literature is weighted by multiplying the number of co-occurrences of each piece of literature by the literature score $L$. Finally, the gene scorer determines a gene score $G$ for each extracted gene by summation of the co-occurrence count, weighted by literature score, for each piece of literature in which the gene was reported (details are also given in the Methods section). As a result, gene suggestions are provided as a ranking of the extracted genes done in accordance with the computed gene score $G$. In this manner, our technique can extract gene names that have been mentioned in the same literature as those mentioned in the design history for
Table 1. Top 10 Literature for (“pyruvate” AND “acetolactate synthase”) Sorted by Literature Score

| rank | literature score L | PubMed ID | title | journal |
|------|-------------------|-----------|-------|---------|
| 1    | 1.88              | 7654314   | Metabolic engineering of Escherichia coli to enhance recombinant protein production through acetate reduction. | Biotechnology Progress |
| 2    | 1.76              | 28204883  | Metabolic engineering of Saccharomyces cerevisiae for 2,3-butanediol production. | Applied Microbiology and Biotechnology |
| 3    | 1.70              | 24614328  | Metabolic engineering of Candida glabrata for diacetyl production. | PloS One |
| 4    | 1.59              | 27594916  | Metabolic engineering of Zymomonas mobilis for 2,3-butanediol production from lignocellulosic biomass sugars. | Biotechnology for Biofuels |
| 5    | 1.52              | 20213636  | Metabolic engineering of acetoin and meso-2,3-butanediol biosynthesis in E. coli. | Biotechnology Journal |
| 6    | 1.49              | 25007652  | System metabolic engineering strategies for 2,3-butanediol production by Torulopsis glabrata. | Wei sheng wu xue bao = Acta Microbiologica Sinica |
| 7    | 1.47              | 24725668  | Enhancement of acetoin production in Candida glabrata by in silico metabolic engineering. | Microbial Cell Factories |
| 8    | 1.44              | 26511723  | Engineering Corynebacterium glutamicum for the production of 2,3-butanediol. | Microbial Cell Factories |
| 9    | 1.40              | 17586670  | Metabolic engineering of Bacillus subtilis for ethanol production. | Applied and Environmental Microbiology |
| 10   | 1.39              | 29330665  | A synthetic pathway for the production of 2-hydroxyisovaleric acid in Escherichia coli. | Journal of Industrial Microbiology & Biotechnology |

Figure 3. History for development of shikimic acid-producing C. glutamicum strain as input for design suggestion. (A) The 18 genetic modifications composing the input history of the development of the shikimic acid-producing strain of C. glutamicum described in ref 24. Each modification is represented by the gene name, the EC number of the gene product if applicable, and the direction of modification (up for introduction and enhancement, down for disruption and repression). (B) The genetic modifications in (A) illustrated on a metabolic pathway map. [Adapted from ref 24, Copyright 2016, Elsevier]. Glucose 6-phosphate (G6P), fructose 6-phosphate (F6P), fructose 1,6-bisphosphate (FBP), glyceraldehyde 3-phosphate (GAP), dihydroxyacetone phosphate (DHAP), 3-deoxy-D-arabino-heptulosonic acid 7-phosphate (DAHP), 3-dehydroquinate acid (DHQ), 3-dehydroshikimic acid (DHQ), quinic acid (QA), shikimic acid 3-phosphate (S3P).
successfully aggregated across journals independent of a journal’s scope. The R-precision, a measure for evaluating information retrieval, was 92.8%, indicating that literature in the domain of metabolic engineering ranked well above that in other domains that would generate noise.

These results demonstrate the efficiency of using literature scores determined by the domain-specific lexical model in order to evaluate relatedness to the domain of metabolic engineering. Hence, we assumed that the model gave higher scores to literature that was more abundant with terms specific to the domain of metabolic engineering, that is, more related to the domain.

3. Evaluating Use of Relatedness to History and Domain. We assumed that genes to be modified after a design history would also be mentioned in literature that both mentions the genes in a history and that belongs to the domain of metabolic engineering. Therefore, we utilized both the co-occurrence of genes in literature to measure the relatedness to a user’s history and the literature score to measure the relatedness to the domain of metabolic engineering for conducting gene scoring.

To validate the assumption of suggesting genes on the basis of relatedness to both a history and the above domain, we applied the proposed technique, with and without consideration of these elements, to a history consisting of 18 genes for shikimic acid production by Corynebacterium glutamicum (Figure 3A,B).

Here, we refer to the 18 “genes in the history” and 7 “genes selected by experts” together as 25 “relevant genes.” The term “genes selected by experts” refers to genes manually selected from among those suggested (Table S1) by experienced biologists to be candidates for the next modifications that are worth testing in order to enhance shikimic acid production.

The effects of considering the relatedness to the history and the domain were evaluated in terms of the capability of suggesting relevant genes. Table 2 shows the ranking of relevant genes suggested with or without relatedness to the domain (literature score) and history. See Table S2 for all rankings.

Literature was collected by querying the PubMed database on the target compound name (“shikimate” or “shikimic-acid”), followed by gene ranking by the occurrence counts of the genes under (-/-) conditions. These counts were weighted with the source literature scores under the [-/D (domain)] condition. Literature was also collected by querying the PubMed database on the genes present in the user’s history with our query expansion (see Methods), followed by gene ranking by the co-occurrence of the extracted genes and those in the history under [H (history)/-] conditions. The co-occurrences of the genes were weighted with the source literature scores under (H/D) conditions.

The relevant genes that could be easily found in the shikimate pathway (i.e., aroG, aroB, aroD, aroE, and aroK) ranked over 30 when the suggestion was proposed on the basis of the number of the collected pieces of literature in which the genes were observed (-/- in Table 2).

When the literature scores were applied to gene scoring (-/D in Table 2), 12 out of the 25 relevant genes ranked higher than the ranking observed under the (-/-) condition. These increases were observed more for genes involved in glycolysis or sugar uptake (ldhA, ptsG, glk, ppsA, galP, and pykF). This indicates that the literature score determined by our lexical model emphasizes genes that are related to the purpose of adjusting the metabolic network to enable productivity to be indirectly improved.

| Table 2. Rank of Relevant Genes Suggested with or without Relatedness to History and Domain |
| Relatedness (H: history; D: domain) | -/- | -/D | H/- | H/D |
|-----|-----|-----|-----|-----|
| aroG | 11  | 4  | 15 | 9  |
| aroB | 5  | 5  | 7  | 5  |
| aroD | 28 | 17 | 54 | 48 |
| aroE | 2  | 3  | 10 | 6  |
| ptsG | 48 | 24 | 82 | 82 |
| ptsI | 82 | 2  | 11 | 8  |
| ptsH | 198 | 198 | 276 | 276 |
| ptsD | 108 | 39 | 2  | 4  |
| ptsI | 198 | 198 | 96  | 49  |
| ptsH | 198 | 154 | 202 | 165 |
| ptsG | 106 | 52 | 24 | 18 |
| ptsD | 84  | 72  | 129 | 82 |
| ptsH | 198 | 198 | 198 | 198 |
| ptsI | 198 | 198 | 198 | 198 |
| ptsG | 108 | 52 | 24 | 18 |
| ptsD | 84  | 72  | 129 | 82 |
| ptsH | 198 | 198 | 198 | 198 |
| ptsI | 198 | 198 | 198 | 198 |
| ptsG | 108 | 52 | 24 | 18 |
| ptsD | 84  | 72  | 129 | 82 |
| ptsH | 198 | 198 | 198 | 198 |
| ptsI | 198 | 198 | 198 | 198 |
| ptsG | 108 | 52 | 24 | 18 |
| ptsD | 84  | 72  | 129 | 82 |
| ptsH | 198 | 198 | 198 | 198 |
| ptsI | 198 | 198 | 198 | 198 |
| ptsG | 108 | 52 | 24 | 18 |
| ptsD | 84  | 72  | 129 | 82 |
| ptsH | 198 | 198 | 198 | 198 |
| ptsI | 198 | 198 | 198 | 198 |

Likewise, 13 out of the 25 relevant genes ranked higher when the history was used for the suggestion process (H/- in Table 2) than the ranking observed under the (-/-) condition. Moreover, 11 genes (ldhA, ptsG, ptsH, ptsI, glk, ppsA, galP, pykF, pfkA, pgi, and pps) among these 13 were observed in the rank of the top 30 genes. This indicates that the history enabled relevant genes to be further detected that were presumably more related to the user’s history.

When both the history and the literature scores were considered (H/D in Table 2), the ranks of 15 genes (ldhA, tkt, tal, ptsG, ptsH, ptsI, glk, gapA, hdpA, ppsA, galP, pykF, pfkA, pgi, and pps) increased more than 10 compared with those observed under the (-/-) condition, 11 of which (ldhA, ptsG, ptsH, ptsI, glk, ppsA, galP, pykF, pfkA, pgi, and pps) were present in the top 30. These results indicate that suggestion based on both the relatedness to domain and history resulted in genes more relevant to the user’s purpose being pushed higher in the ranking.

Genes aroD, aroE, aroK, and shiA, which were already detected under the (-/-) condition, ranked lower under the (H/D) condition when this relatedness to the domain and the history was considered. Our technique can emphasize genes that may be overlooked when they are simply collected by querying on a target compound name.
4. Effects of Modifications on Suggested Genes in Shikimic Acid Production. To validate the suggestion of genes by our proposed technique in designing next modifications, we determined how the modifications of suggested genes would affect the production of shikimic acid. A suggestion was put forth for the history (Figure 3) of the development of a state-of-the-art (SOTA) strain of shikimic acid-producing C. glutamicum (unpublished, see Methods). Shown in Figure 4A are the top 30 genes in the list (the ranking information is given in Table S1). A few nongene terms were also detected as annotated in Figure 4A because the same strings were designated as gene names in UniProt.

Biologists experienced in the field of synthetic biology manually selected the next genes to be modified from the suggested genes after omitting those involved in downstream reactions of shikimic acid in the endogenous metabolic pathway; the genes selected were ppsA, galP, pfkA, pykF, pgi, ppc, and shiA (Figure 4B). Inspired by the proposal of shiA in the suggestion list, qsuA—a C. glutamicum homologue of shiA—was also selected. Then, the titer and yield of shikimic acid in culture medium were evaluated after conducting fed-batch fermentation over a period of 32 h as per previously described methods reported in ref 24, after the strains harboring the respective genetic modifications were constructed. Among the constructed strains, applying the strain with qsuA overexpression, that with pgi overexpression, that with pykF deletion, and that with overexpression of exogenous ppsA of Escherichia coli improved the shikimic acid production titer (Figure 4C) without affecting the yield from glucose (Figure 4D) (see Figure S1 for the results of applying the tested strains). This finding adds experimental
validation indicating that our proposed technique may be used to suggest genes worth modification subsequently after substantial efforts have been made by experts to maximize productivity through multiple rounds of DBTL cycles.

5. Effect of Updating History on Aiding Design Suggestion. Next, we discuss the effect that accumulating design histories through iterations of DBTL cycles has on the gene suggestions made by our technique. As a model case of iterative improvement of a microbe through DBTL cycles, a series of partial histories was made from a design history for developing a shikimic acid-producing strain, starting with a partial history consisting of just one modified gene, *aroG*, for the first gene modification and then adding *aroB* to the partial history for the next gene modification. In this manner, one gene was added to the next partial history at every iteration in the order shown in Figure 3A to mimic a series of iterative DBTL cycles.

As in Table 2 (H/D), 22 relevant genes including 15 genes in the history and 7 genes selected by experts for further improving the current strain were suggested when the full history, which consisted of 18 genes, was input. Figure 5 shows the change in mean rank of the 22 relevant genes with the iterative addition of genes to the history (see Table S4 for the full rank table).

When only the first 3 genes, all of which are in the shikimic acid pathway, were given, 20 of 22 relevant genes appeared in the suggestion, in which 13 ( *tkt*, *tal*, *ptsG*, *ptsH*, *ptsL*, *glk*, *gapA*, *hdpA*, *ppsA*, *galP*, *pykF*, *pgi*, *ppc*) were nonshikimic acid pathway genes (Table S4). The nonshikimic acid pathway genes that were retrieved contained genes encoding enzymes that mediate the synthesis or consumption of metabolites that join or branch from the shikimic acid pathways and transporters. This means that this technique could retrieve nonshikimic acid pathway genes on the basis of only shikimic acid pathway genes, indicating its potential to suggest genes of pathways that users have not yet worked on. Thus, this technique would be beneficial during the earlier rounds of DBTL cycles by broadening the scope of design concepts.

Moreover, those 22 relevant genes that were suggested tended to be ranked higher on average when genes were added to the history at least until the 12th iteration in this case. This suggests that updating the history resulted in updating the suggestion, at least when a gene added to the history co-occurred with another relevant gene. Adding genes to the history, i.e., *ldhA*, *tkt*, and *ptsG* in this case, impacted the ranking of relevant genes more significantly than others. Nonshikimic acid pathway genes would have expanded the range of literature and genes related to the history, and thus, further affected the relatedness-based scoring of genes. Adding designs based on different concepts could expand the variety of retrieved literature and genes, which would help in retrieving overlooked or suspended designs.

These results demonstrate that our technique may help in designing in order to improve the next iterations of DBTL cycles by updating its suggestions when a history is updated at the latest iteration.

6. Conclusion. Feeding former rounds of DBTL cycles back to design of subsequent genetic modifications after a series of DBTL cycles still depends largely on nonsystematic methods. Here, we developed a technique that takes design histories to suggest genes to be modified next. This technique mines those genes from a considerable amount of text in published literature that describes previously tested modifications. It suggests genes on the basis of relatedness to both the domain of metabolic engineering and the history of a user’s former DBTL cycles. In an experiment, the suggestion proposed by our technique in consideration of a design history for shikimic acid-producing *C. glutamicum* provided eight candidate genes that experts regarded as being worth subsequent analysis. Six of those eight genes ranked higher in the suggestion list put forth by our technique than those observed in the occurrence ranking for literature collected from a search on shikimic acid. Among the tested genes, four proved effective in increasing the productivity of the SOTA shikimic acid-producing strain. Our technique can aid in the next designing step by feeding a user’s history of DBTL cycles back to the design step and by suggesting subsequent relevant genes to be modified. It can also provide researchers with design concepts that have been overlooked or those that have not been used through the comprehensive retrieval of well-tested designs from previous studies. Therefore, the suggestions provided by this technique will drive DBTL cycles further.

7. Possible Future Studies. One of the downsides of this technique is echo-chambering, that is, suggesting genes that peer researchers in the same domain have considered may increase the risk of overlooking those that they have ignored. Solutions to this could be using alternative suggestions made by other domain-specific lexical models or by inverting the literature scores to emphasize nonmetabolic engineering genes. Using a combination of learning and designing techniques that are based on other concepts or simply searching literature without biasing metabolic engineering terms will also help overcome this downside. Learning from failure, that is, feeding negative
histories that proved to be counterproductive back to the next
designs, will also help, especially in avoiding repeating failed
attempts. Learning from failure may involve mining negative
remarks in literature or linking with other learning tools that
consider experimental results as well. Expanding the concepts explored for suggesting from genes
coexpressed genes, that have been appended to the suggested
genes can be further weighted on the basis of their
further enhance DBTL cycles. For example, the scores of
suggests genes of the relatedness to the production of compounds
will provide broader ideas for developing
well. Expanding the concepts explored for suggesting from genes
coexpressed genes, that have been appended to the suggested
genes will be a step toward automated design.

Minor improvements such as the following would benefit as
well. Expanding the concepts explored for suggesting from genes
to enzymes, protein complexes, metabolic pathways, or
compounds will provide broader ideas for developing
subsequent designs. Policies for evaluating relatedness can be
modified or added in order to further meet user-specific
purposes. For example, the information domain of the domain-
specific lexical model can be shifted to another domain of
interest other than metabolic engineering, or the directions or
purpose of a genetic modification can be integrated into the
calculation of the relatedness to a history. Expanding the source of
suggestions from titles and abstracts to full text will help widen
the variety of suggestions, although further sophistication on
scoring is necessary to nullify the increased generation of noise
that may occur. Suggestions could be further improved and
refined, for example, by refining the definition of gene names, by
focusing on relevant taxonomic ranges, or by unifying synonyms
and isoforms. Suggestions may also be deepened by including
details on modifications extracted from the source literature.

■ METHODS

Design History. A history is a series of records on genetic
modifications conducted on an original strain of microbes
through DBTL cycles in order to produce a target compound.
Each record consists of genes that have already been modified,
the EC number of the products of these genes, if applicable, and
the direction of modification of these genes in terms of
upregulation (e.g., introduction or enhancement) or down-
regulation (e.g., disruption or repression).

Literature Searcher and Domain-Specific Lexical
Model. The literature searcher queries the literature available
in the PubMed database by using the information from a design
database after query expansion, and it then scores the collected
literature by using the domain-specific lexical model. At query
expansion, the EC numbers are expanded to the corresponding
domain in the UniProt database within a specified kingdom;
bacteria in the present study. The direction of each modification
is replaced with a keyword list (Table S3) depending on the
intended direction of the modification. Then, the domain-
specific lexical model scores the pieces of literature on the basis
of the relatedness to the production of compounds via metabolic
engineering.

The domain-specific lexical model used in this study consisted of
domain-specific terms extracted from literature related to the
domain of metabolic engineering by using the SMART score. The
SMART score evaluates the domain-specificity of each term
by comparing the frequency of a term within a literature set
related to the specified domain (S) versus that within the
comprehensive literature set (W). All 34,939,161 pieces of
literature published in or before June 2019 and available on
PubMed were considered W, and those published in the journal
Metabolic Engineering were considered S, and they were both
integrated into the SMART score. The lexical model
(Supporting Information S7) consisted of 20 domain-specific
terms extracted from titles and 80 from abstracts of the latest
1000 papers in Metabolic Engineering in or before December
2018. Moreover, we omitted terms that appeared 1,000,000
times or more in W to avoid being influenced by highly common
terms (including “of”, “in”, and “and”). The resulting lexical
model was composed of weighted domain-specific terms.

The literature score L was calculated as a lexical similarity
between the vocabularies of the literature and the domain-
specific terms. The similarity was defined as the cosine similarity
between the frequency list of the domain-specific terms
observed in the titles and abstracts of the literature and the set of
the terms, and the similarity was normalized with the total
term count of the literature.

Gene Name Extractor. The gene name extractor extracts
gene names appearing in the titles and abstracts of literature
collected by the literature searcher by following the rules below:

1. A gene name is comprised of a string designated as the
   “Name” or one of the “Synonyms” in the “Gene name”
   section of a reviewed protein in the Archaea, Bacteria,
   Fungi, or Plant section of the UniProt database (accessed on February 27, 2020) except the following.
   1.i. Strings with a length of less than three letters were omitted (e.g., “SA”, “KO”).
   1.ii. Strings in the common word list we prepared were omitted. The list consisted of 121 words that occurred at a ratio of more than 1/100 000 in the unlemmatized written words list of the British National Corpus (e.g., “are”, “can”, and “for”).
2. A gene name is case-sensitive except for the first letter of a sentence.
3. Modified forms of gene names matching one of the patterns described below are also extracted as gene names after converting to the supposedly original form:
   3.i. Abbreviations of overlapping gene names when the names of two to five genes except for the last letter, are identical (e.g., “aroGBD”, where “aroG”, “aroB”, and “aroD” are gene names).
   3.ii. A gene name hyphenated with another string (e.g., “pUC19-ilhdA”, where “ilhdA” is a gene name).
   3.iii. Different gene names delimited with a slash (“/”) or a dual colon (“::”) (e.g., “ptsH/ptsI”, where “ptsH” and “ptsI” are gene names).
   3.iv. A gene name prefixed or suffixed with “Delta”, “delta”, or “Delta” (e.g., “DeltaK” or “aroK”, where “aroK” is a gene name).
   3.v. A gene name prefixed with a hyphen (e.g., “-IlhdA”, where “IlhdA” is a gene name).
   3.vi. A gene name directly concatenated with a parenthesized phrase (e.g., “aroG(FBR)”, where “aroG” is a gene name).
   3.vii. A gene name marked with an italic tag of hypertext markup language (e.g., “ikt”, where “ikt” is a gene name).

Gene Scorer. The gene scorer calculates gene scores as
shown in Figure 6. Gene scoring is based on both the co-
ocurrence of the extracted genes and the already modified
genes in a history as well as the literature scores of the literature
that mentions them. It adds the score of a piece of literature that
ments an extracted gene and a gene in the history appearing in the title and abstract of the piece of literature to the gene score of the extracted gene. Therefore, the gene score herein represents the sum of the scores of literature in which an extracted gene is mentioned together with genes present in the user’s history.

**Preparing Design Histories.** The design history used in making the suggestions for the production of shikimic acid in this study was based on attempts to establish a *C. glutamicum* strain previously reported after biologists had attempted to maximize the production yield of shikimic acid. The input history for the shikimic acid consisted of 18 genes modified to enhance its production, as shown in Figure 3AB.

**Producing Shikimic Acid in *C. glutamicum* Transformed with Suggested Genes.** The SOTA shikimic acid-producing strain (unpublished) is derived from *C. glutamicum* R (wild-type) and possesses the markerless integration of PgpaA-E. coli *aroE*G<sup>1807</sup>, PgpaA-*aroB*, PgpaA-*aroD*, PgpaA-*aroE*, PgpaA-*tkt-tal*, PgpaA-*iolTI*, and PgpaA-*ppgk*, an attenuated mutant of *aroK*, as well as markerless deletion in *qsuB*, *qsuD*, and *ldhA*. This strain was further modified by introducing or deleting the suggested genes and was evaluated for shikimic acid production. Biologists determined the direction of the modification for each gene, i.e., gene introduction or deletion. Genes selected for overexpression were chromosomally introduced under the control of a strong *gapA* promoter as per previously described methods. Chromosomal deletion of genes was achieved with a markerless insertion system by using pCRA725, which carries the *sacB* gene as per protocols described previously. Fed-batch fermentation experiments were performed by using the resulting strains. The titer and the yield (from glucose) of shikimic acid in the culture medium were evaluated after 32 h by liquid chromatography as per previously described methods.

### ASSOCIATED CONTENT

| Supporting Information |
|------------------------|
| The Supporting Information is available free of charge at [https://pubs.acs.org/doi/10.1021/acssynbio.1c00234](https://pubs.acs.org/doi/10.1021/acssynbio.1c00234). |

Document S1: Literature score ranking example (PDF)

Table S1: Suggestions for shikimic acid production; Table S2: Suggestions for shikimic acid production with or without inclusion of literature scores and history; Table S3: Keywords for the directions of modifications; Table S4: Ranks of suggested relevant genes suggested for the partial histories; Table S5: Complete list of the domain-specific lexical model terms; Figure S1: Shikimic acid production using the tested strains of *Corynebacterium glutamicum* modified at the genes selected by the experts (XLSX)

### AUTHOR INFORMATION

**Corresponding Author**

Kiyoto Ito — Center for Exploratory Research, Research and Development Group, Hitachi, Ltd., Kokubunji-shi, Tokyo 185-8601, Japan; orcid.org/0000-0001-5953-7740

Koichi Toyoda — Research Institute of Innovative Technology for Earth, Kizugawa-shi, Kyoto 619-0292, Japan; orcid.org/0000-0002-2007-0503

Masayo Inui — Research Institute of Innovative Technology for Earth, Kizugawa-shi, Kyoto 619-0292, Japan

Complete contact information is available at: [https://pubs.acs.org/10.1021/acssynbio.1c00234](https://pubs.acs.org/10.1021/acssynbio.1c00234)

**Author Contributions**

SN, OI, and KI designed the scheme and interpreted the data. SN and OI wrote the code and conducted in silico experiments. TKo, TKu, KT, and MS designed the wet-lab experiments. TKo and MS conducted the wet-lab experiments. MI conceptualized and supervised the wet-lab experimentation aspects of this study. KI, TS, and MA conceptualized the design-suggestion technique. SN drafted the manuscript, on which KI performed substantial revisions.

**Notes**

The authors declare the following competing financial interest(s): Part of this work was financially supported by the New Energy and Industrial Technology Development Organ-
ization (NEDO). Hitachi, Ltd. has patents pending for the literature-search technique using the domain-specific lexical model (P2020-045980, application on March 17, 2020) and the design-suggestion technique (P2020-218498, application on December 28, 2020).

**ACKNOWLEDGMENTS**

This work is based on results obtained from a project named “Development of Production Techniques for Highly Functional Biomaterials Using Plant and Other Organism Smart Cells” commissioned by the New Energy and Industrial Technology Development Organization (NEDO). We thank Dr. Fumio Matsuda (Osaka University) and Dr. Yoshiki Niwa (Hitachi, Ltd.) for their helpful comments.

**ABBREVIATIONS**

DBTL, design—build—test—learn; SOTA, state-of-the-art.

**REFERENCES**

(1) Poust, S., Hagen, A., Katz, L., and Keasling, J. D. (2014) Narrowing the gap between the promise and reality of polyketide synthases as a synthetic biology platform. *Curr. Opin. Biotechnol.* 30, 32–39.

(2) Liu, R., Bassalo, M. C., Zeitoun, R. I., and Gill, R. T. (2015) Genome scale engineering techniques for metabolic engineering. *Metab. Eng.* 32, 143–154.

(3) Nielsen, J., and Keasling, J. D. (2016) Engineering Cellular Metabolism. *Cell* 164, 1185–1197.

(4) Carbonell, P., Le Feuvre, R., Takano, E., and Scrrutton, N. S. (2020) In silico design and automated learning to boost next-generation smart biomanufacturing. *Synth. Biol.*, 5 S.

(5) Carbonell, P., Parutto, P., Herisson, J., Pandit, S. B., and Faulon, J.-L. (2014) XTMS: pathway design in an eXTended metabolic space. *Nucleic Acids Res.* 42, W389–W394.

(6) Araki, M., Cox, R. S., III, Makiguchi, H., Ogawa, T., Taniguchi, T., Miyaoku, K., Nakatsui, M., Hara, K. Y., and Kondo, A. (2015) M-path: a compass for navigating potential metabolic pathways. *Bioinformatics* 31, 905–911.

(7) Delépine, B., Duigou, T., Carbonell, P., and Faulon, J.-L. (2018) RetroPath2.0: A retrosynthesis workflow for metabolic engineers. *Metab. Eng.* 45, 158–170.

(8) Motwalli, O., Uludag, M., Mijakovic, I., Alazmi, M., Bajic, V. B., Gojobori, T., Gao, X., and Essack, M. (2020) PATHcre8: A Tool That Facilitates the Searching for Heterologous Biosynthetic Routes. *ACS Synth. Biol.* 9, 3217.

(9) Wang, L., Dash, S., Ng, C. Y., and Maranas, C. D. (2017) A review of computational tools for design and reconstruction of metabolic pathways. *Synth. Syst. Biotechnol.* 2, 243–252.

(10) Orth, J. D., Thiele, I., and Palsson, B. O. (2010) What is flux balance analysis? *Nat. Biotechnol.* 28, 245–249.

(11) Lakshmanan, M., Koh, G., Chung, B. K. S., and Lee, D.-Y. (2014) Software applications for flux balance analysis. *Briefings Bioinf.* 15, 108–122.

(12) Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., and Lee, S. Y. (2019) Current status and applications of genome-scale metabolic models. *Genome Biol.* 20, 121.

(13) Bedoya, O., and Bustamante, S. (2011) CNN-Promoter, New Consensus Promoter Prediction Program Based on Neural Networks. *Rev. EIA* 153–164.

(14) Gilman, J., Singleton, C., Tennant, R. K., James, P., Howard, T. P., Lux, T., Parker, D. A., and Love, J. (2019) Rapid, Heuristic Discovery and Design of Promoter Collections in Non-Model Microbes for Industrial Applications. *ACS Synth. Biol.* 8, 1175–1186.

(15) Umarov, R. K., and Solovyev, V. V. (2017) Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS One* 12, No. e0171410.

(16) Ham, T. S., Dmytryev, Z., Plahar, H., Chen, J., Hillson, N. J., and Keasling, J. D. (2012) Design, implementation and practice of JBEI-ICE: an open source biological part registry platform and tools. *Nucleic Acids Res.* 40, No. e141.

(17) iGEM Registry of Standard Biological PartsiGEM Registry of Standard Biological Parts, http://parts.igem.org/Main_Page (accessed Feb 2, 2021).

(18) McLaughlin, J. A., Myers, C. J., Zundel, Z., Misri, G., Zhang, M., Oiferu, I. D., Goñi-Moreno, A., and Wipat, A. (2018) SynBioHub: A Standards-Enabled Design Repository for Synthetic Biology. *ACS Synth. Biol.* 7, 682–688.

(19) Morrell, W. C., Bierk, G. W., Forrer, M., Lopes, T., Backman, T. W. H., Dussault, M., Petzold, C. J., Baidoo, E. E. K., Costello, Z., Ando, D., Alonso-Gutierrez, J., George, K. W., Mukhopadhyay, A., Vaino, L, Keasling, J. D., Adams, P. D., Hillson, N. J., and Garcia Martin, H. (2017) The Experiment Data Depot: A Web-Based Software Tool for Biological Experimental Data Storage, Sharing, and Visualization. *ACS Synth. Biol.* 6, 2248–2259.

(20) Carbonell, P., Jervis, A. J., Robinson, C. J., Yan, C., Dunstan, M., Swainston, N., Vinaixa, M., Hollywood, K. A., Curin, A., Rattray, N. J. W., Taylor, S., Spiess, R., Sung, R., Williams, A. R., Fellows, D., Stanford, N. J., Mulherin, P., Le Feuvre, R., Barran, P., Goodacre, R., Turner, N. J., Goble, C., Chen, G. G., Kell, D. B., Micklefield, J., Breiling, R., Takano, E., Faulon, J.-L., and Scrrutton, N. S. (2018) An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals. *Commun. Biol.* 1, 1–10.

(21) Oyetunde, T., Liu, D., Martin, H. G., and Tang, Y. J. (2019) Machine learning framework for assessment of microbial factory performance. *PLoS One* 14, No. e0210558.

(22) Winkler, J. D., Halweg-Edwards, A. L., and Gill, R. T. (2015) The LASER database: Formalizing design rules for metabolic engineering. *Metab. Eng. Commun.* 2, 30–38.

(23) Winkler, J. D., Halweg-Edwards, A. L., and Gill, R. T. (2016) Quantifying complexity in metabolic engineering using the LASER database. *Metab. Eng. Commun.* 3, 227–233.

(24) Kogure, T., Kubota, T., Suda, M., Hiraga, K., and Inui, M. (2016) Metabolic engineering of Corynebacterium glutamicum for shikimate overproduction by growth-arrested cell reaction. *Metab. Eng.* 38, 204–216.

(25) Singhal, A., Buckley, C., and Mitra, M. (1996) Ivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 21–29, Association for Computing Machinery, New York, NY, DOI: 10.1145/243199.243206.

(26) Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Mazumder, R., O’Donovan, C., Redaschi, N., and Suzek, B. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* 34, D187–D191.

(27) Kilgarriff, A. (1995) BNC word frequency lists, http://www.kilgarriff.co.uk/bnc-readme.html#raw (accessed Dec 18, 2020).

(28) Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. (2007) *The British National Corpus*, version 3 (BNC XML ed.), http://www.natcorp.ox.ac.uk/ (accessed Jan 15, 2021).

(29) Inui, M., Kawaguchi, H., Murakami, S., Vertès, A. A., and Yukawa, H. (2005) Metabolic Engineering of Corynebacterium glutamicum for Fuel Ethanol Production under Oxygen-Deprivation Conditions. *Microb. Physiol.* 8, 243–254.