

MPS: A WEB SERVER FOR METABOLISM PATHWAY SYNTHESIS ON E.COLI

Anjun Zhu^{1,†}, Haoran Zheng^{1,2,3,*†}, Shijun Wu¹, Jiong Hong², Jia Zhang², Xuwei Liu¹

¹School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, PR China

²School of Life Science, University of Science and Technology of China, Hefei 230026, PR China

³Anhui Key Laboratory of Software Engineering in Computing and Communication, University of Science and Technology of China, Hefei 230026, PR China

hrzheng@ustc.edu.cn, zhuanjun@mail.ustc.edu.cn

Keywords: *de novo* gene synthesis, systems biology, integrated system, web server.

Abstract

An integrated system MPS has been developed which includes three modules: MRSD (Metabolic Route Search and Design), GD (Gene Design) and OD (Oligonucleotide Design). It can complete the process of from designing and searching a route to designing DNA oligonucleotides (oligos) for *de novo* gene synthesis. And each module can also be accessed and used individually. MRSD submodule searches and designs routes based on data from KEGG. GD submodule reversely translates multiple catalytic enzymes contained in chosen pathway into DNA sequence combining with operons for expression. OD submodule cuts resultant multiple DNA sequences into a series of oligonucleotide sequences characterized by highly homogeneous melting temperatures and a minimized tendency for secondary structure formation and mis-hybridization. The ideal software developed for biologists to *de novo* sequences design synthesizes new metabolic route for target compound and provides a good interactive interface. **Availability:** <http://bioinfo.ustc.edu.cn/software/MPS/>.

1 Introduction

With the advent of post-genomic era and increasing use of computational techniques, a lot of computer-aided tools which have numerous applications in systems biology and synthetic biology were developed. In this field, it's a new challenge to design new metabolic route for the product that native organisms cannot produce. Currently, if researchers want to make all preparation for experiments, they firstly need aids like FMM [1], Retro-Path [2], MRSD [3] some others to design or search a metabolic route, then collect sequences of enzymes, construct the operons or expression cassette, and combine them into one DNA sequence manually, and finally cut them into oligonucleotides for *de novo* gene DNA synthesis. Related works include OPTIMIZER [4], GeneDesigner [5], TmPrime [6], DNAworks [7], Eugene [8] and so on.

Overall, this process is not integrated, there are currently no tools can help researches getting the whole work done without any letup. Also from every step, most software do not support multi-gene synthesis and assembly with operons. Our web service tool MPS integrated with MRSD could make up these shortcomings. MPS sets *E.coli* as the host organism, as it's popular in microbial metabolism.

Herein we present an integrated system MPS including three interrelated modules, which can automate the process as mentioned above seamlessly. GD (gene design module) enables users to reversely translate multiple catalytic enzymes into gene sequences based on input of users. The genes sequences are codon optimized and the secondary structure formation and mis-hybridization in sequence are avoided. In the pathway, the protein should not be expressed in same levels. Here GD provides various promoter and RBS to design operons for users. OD (oligonucleotide design module) outputs a series of oligonucleotide sequences for *de novo* gene synthesis. Multiplex gene synthesis will form ultra-long DNA sequence in current software, and OD implements a method of splitting long sequences into short fragments for multi-gene synthesis. Each module can be used individually. And any two of them can also be used in combination.

2 METHODS

MPS contains three major functions: MRSD (Metabolic Route Search and Design), GD (Gene Design) and OD (Oligonucleotide Design). These three functions are inter-related to fulfil a process of getting metabolic pathways, reverse translation of catalytic enzymes, and the design of oligonucleotides for *de novo* gene synthesis. Also each function can be used individually. As MRSD has been introduced previously [3], the method of the latter two modules will be introduced in detail.

In the gene design module, to perform subsequent experiments perfectly, the algorithm of reverse translation considers the potential mis-hybridization and secondary structures among oligonucleotides to ensure the oligonucleotides highly specific to their targets to avoid incorrect assembly. An approach is employed to screen sequences

*Corresponding author. E-mail: hrzheng@ustc.edu.cn.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

and replace codons by sequence alignment. The algorithm of reverse translation considers the following factors:

$$\left\{ \begin{array}{l} (GC)\% \leq Threshold \\ secondary\ structure\ as\ few\ as\ possible \\ mis-hybridization\ as\ few\ as\ possible \\ codon\ usage \geq Threshold \\ avoiding\ restriction\ sites\ as\ required \end{array} \right.$$

where (GC)% means G+C content. For gene's high expression, the program selects codon with usage greater than the threshold value in the process of reverse translation. Meanwhile, for the success of gene synthesis and subsequent cloning, some restriction sites should be removed. The min, max and threshold in the equation above or below are all defined by users to meet their experiment conditions.

After the gene DNA sequences have been obtained, to express the pathway in *E.coli*, the operons are designed. The promoter and RBS with various strength can be selected in this module to construct the pathway expression operons. Then DNA sequences containing operons for selected route are cut into oligos for *de novo* gene DNA synthesis.

Oligonucleotide design module can handle ultra-long DNA sequence. It cuts long sequences into short fragments with overlap and then cuts fragments into oligonucleotides, which aims to reduce errors in the synthetic gene. Thermodynamic properties (i.e. melting temperature) between all oligonucleotides should be uniformized to aid the hybridization during assembly and maximize the assembly efficiency. Meanwhile, in order to improve the oligo synthesis efficiency, the length also needs to be limited, so the oligonucleotide and fragments should fulfil the following constraints:

$$\left\{ \begin{array}{l} min \leq fragment.length \leq max \\ oligonucleotide.length\ as\ close\ as\ target\ length \\ |overlap.Tm - Tm.threshold| \leq 2 \end{array} \right.$$

where 'fragment.length' and 'oligonucleotide.length' are the length of fragment and oligonucleotide, respectively. Tm is the abbreviation of melting temperature. 'overlap.Tm' represents the Tm value of overlap between two fragments or oligonucleotides. Tm value is relevant with the length of sequence and G+C content. In our program, the distribution of G+C content is maintained uniform, so the deviation in length among designed oligonucleotide fragments is small while the Tm is kept consistent. The algorithm ensures Tm value priority and no gap between adjacent oligonucleotides in order to reduce the difficulty and error rate during DNA assembly.

3 RESULTS

3.1 The data source

MRSD uses data from KEGG database to perform route search and design. The main reasons behind selecting KEGG as the data source for MRSD are the versatility of data presented in KEGG and its popularity as a data source for different metabolic network. GD uses bacteria data (including EC number, product, source, amino acid

sequence, and so on) from Genetic Sequence Data Bank (<ftp://ftp.ncbi.nih.gov/genbank>); Promoters, RBS and terminators for *E.coli* are all from the iGEM Registry of Standard Biological Parts.

3.2 Applications and web interface

MPS is built upon C++ and PHP, and supported by MySQL database to store data. The graph visualization tool Graphviz (<http://www.graphviz.org/>) is used to depict the searching results of metabolic pathways.

MPS consists of three basic components: (i) MRSD: a web-based user interface for constructing route queries and displaying results; (ii) GD: a schema for gene design; (iii) OD: a web interface for setting parameters to perform oligonucleotide design. The detail of web interface is given in the MPS 'Tutorial' page. Below briefly introduction to the web interface is along with case study.

Here the core of MRSD [3] System is used based on KEGG database. It can give a metabolic route as users' need. We perform a search on the metabolic network of the 'superbacteria' to find Glycolysis, in which the metabolic pathway converts glucose into pyruvate. A screen shot of the query interface and query refinement page is shown in Fig.1 (A) and Fig.1 (B). Source metabolite and product metabolite of interest (keyword, formula or KEGG compound ID) can be input. And the refinement settings including whether the query using the intermediate metabolites or not are optional. Continue with the steps below.

First, select a required route from search result of Glycolysis, from glucose into pyruvate. Here the NO.1 route highlighted with red line is selected as shown in Fig.1 (C). Users can also achieve the desired metabolic pathways from design module step by step. In 'Route Details' part of page given in Fig.1 (D), users can perform some operations on corresponding route, such as 'highlight route', 'route design' and 'Gene Design'.

Second, gene design module reversely translates the enzymes contained in chosen pathway in the first step. From web interface users can set parameters (G+C content and codon frequency) and choose regulatory elements (promoter, RBS and stop codon) to construct the pathway expression operon. For each enzyme or coding sequence, users can input manually when you have your own needs. The procedure outputs gene sequence based on user's set, completing codon optimization and avoiding restriction sites, secondary structure and mis-hybridization. As the absolute expression level of the genes and the relative expression level between each gene are very important for producing target product, the output is a complete DNA sequence containing promoter-[RBS-CDS-stop codon]_n-terminator in order as the users' needs which can express in selected host [9] as shown in Fig.1 (E). Finally, five gene sequences corresponding to the route of choice are achieved.

Third, in order to obtain DNA sequence through chemical synthesis, users cut it into oligonucleotide suitable for DNA *de novo* synthesis. From the parameters setting interface as shown in Fig.1 (F), researchers need set all the parameters fulfilling their experiment conditions. Then program processes and returns fragments and oligonucleo-

tides based on parameters set by users. The final result will be sent to you by email.

4 CONCLUSION

The MPS is a user-friendly, integrated and versatile tool that can aid researchers seamlessly to obtain the genes from pathway design to *de novo* synthesis. It sets *E.coli* as the host organism. It includes three modules and each module can be used individually. In the future, as the host organism is a key choice during new metabolic route synthesis, MPS will provide more host organisms for users to select.

Acknowledgments

The *Funding*: This work was supported by the National Key Technologies R&D Program [2011CB910200] and Anhui Provincial Natural Science Foundation [1508085MF128].

Conflict of interest: none to declare.

References

- [1] Chou, C.H., *et al.* (2009) FMM: a web server for metabolic pathway reconstruction and comparative analysis, *Nucleic acids research*, **37**, W129-134.
- [2] Carbonell, P., *et al.* (2011) A retrosynthetic biology approach to metabolic pathway design for therapeutic production, *Bmc Syst Biol*, **5**, 122.
- [3] Xia, D.G., *et al.* (2011) MRSD: a web server for Metabolic Route Search and Design, *Bioinformatics*, **27**, 1581-1582.
- [4] Puigbo, P., *et al.* (2007) OPTIMIZER: a web server for optimizing the codon usage of DNA sequences, *Nucleic acids research*, **35**, W126-131.
- [5] Villalobos, A., *et al.* (2006) Gene Designer: a synthetic biology tool for constructing artificial DNA segments, *BMC bioinformatics*, **7**.
- [6] Bode, M., *et al.* (2009) TmPrime: fast, flexible oligonucleotide design software for gene synthesis, *Nucleic acids research*, **37**, W214-221.
- [7] Hoover, D.M. and Lubkowski, J.(2002) DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis, *Nucleic acids research*, **30**.
- [8] Gaspar, P., *et al.* (2012) EuGene: maximizing synthetic gene design for heterologous expression, *Bioinformatics*, **28**, 2683-2684.
- [9] Medema, M.H., *et al.* (2012) Computational tools for the synthetic design of biochemical pathways, *Nat Rev Microbiol*, **10**, 191-202.

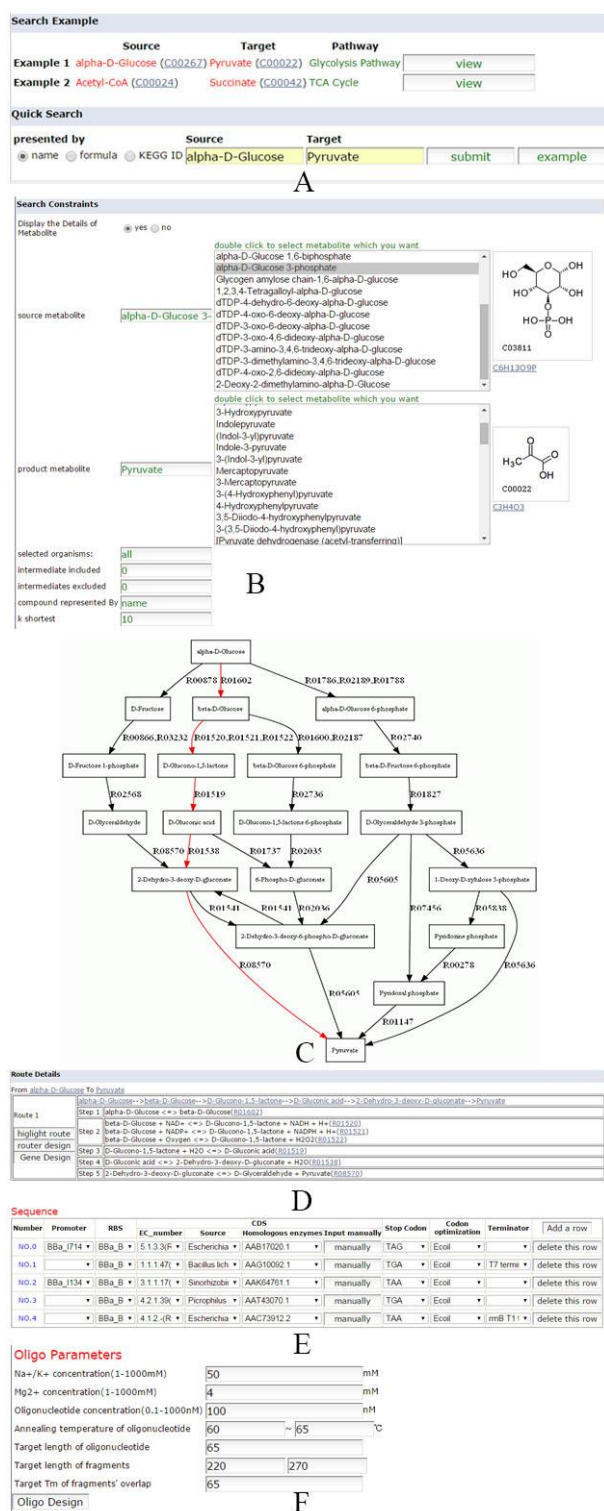


Figure 1: (A) the route query interface; (B) the query refinement page; (C) the shortest routes highlighted by red line between glucose and pyruvate in the metabolic network; (D) the details of the routes; (E) a schema for gene design: a complete DNA sequence containing promoter-[RBS-CDS-stop codon]_n-terminator in order as the users' needs; (F) a web interface for setting parameters to perform oligonucleotide design.