Deliverable D5.3

Funding Scheme: THEME [ICT-2007.8.0] [FET Open]

Paving the Way for Future Emerging DNA-based Technologies:
Computer-Aided Design and Manufacturing of DNA libraries

Grant Agreement number: 265505
Project acronym: CADMAD

Deliverable number: D5.3

Deliverable name: Report on defining user specification and requirements from DNApl and vDNApl (GUI)

Contractual Date\(^1\) of Delivery to the CEC: M12

Actual Date of Delivery to the CEC:

Author(s)\(^2\): Sandra Meyer, Frank Edenhofer

Participant(s)\(^3\): UKB

Work Package: WP5

Security\(^4\): Pub

Nature\(^5\): R

Version\(^6\): 0.0

Total number of pages: 10

---

1 As specified in Annex I
2 i.e. name of the person(s) responsible for the preparation of the document
3 Short name of partner(s) responsible for the deliverable
4 The Technical Annex of the project provides a list of deliverables to be submitted, with the following classification level:
   Pub - Public document; No restrictions on access; may be given freely to any interested party or published openly on the web, provided the author and source are mentioned and the content is not altered.
   Rest - Restricted circulation list (including Commission Project Officer). This circulation list will be designated in agreement with the source project. May not be given to persons or bodies not listed.
   Int - Internal circulation within project (and Commission Project Officer). The deliverable cannot be disclosed to any third party outside the project.
5 R (Report): the deliverables consists in a document reporting the results of interest.
   P (Prototype): the deliverable is actually consisting in a physical prototype, whose location and functionalities are described in the submitted document (however, the actual deliverable must be available for inspection and/or audit in the indicated place)
   D (Demonstrator): the deliverable is a software program, a device or a physical set-up aimed to demonstrate a concept and described in the submitted document (however, the actual deliverable must be available for inspection and/or audit in the indicated place)
   O (Other): the deliverable described in the submitted document can not be classified as one of the above (e.g. specification, tools, tests, etc.)
6 Two digits separated by a dot:
   The first digit is 0 for draft, 1 for project approved document, 2 or more for further revisions (e.g. in case of non acceptance by the Commission) requiring explicit approval by the project itself.
   The second digit is a number indicating minor changes to the document not requiring an explicit approval by the project.
Abstract

The goal of this work package is to provide a basis for generation of an interface between computer scientists and biologists, thereby bringing together developers of the DNA programming tools and potential end users of the system. This deliverable deals with the specification of requirements in the DNA programming language from an end users point of view by designing several DNA libraries using the newly developed software.

Keywords:

DNA Library design, DNalid, requirements, specifications

1. Introduction

a. Aim / Objectives

CADMAD aims at replacing conventional *de novo* DNA synthesis not only by high throughput computer-aided and automated DNA processing but also by exploiting DNA reuse on a large scale. The success of CADMAD technology will critically depend on a powerful and user-friendly interface and a high level of end user compliance. Work package 5 is intended to provide a profound basis for the generation of a reliable and robust interface between computer scientists, who develop the DNA processing tools, and biologists, as potential end users of the developed system. This deliverable deals with the definition and compilation of user specifications and requirements. In particular by this we will ensure on the one hand that the development efforts are focused and not deviated from what a diverse group of potential end user scientists sees as the future of DNA programming. On the other hand we aim to test and challenge the CADMAD system by constructing several DNA libraries. These libraries will, at a later stage of the project, be used to verify the breakthrough that computer-aided design and manufacturing can be effectively employed in DNA-based research and development.

b. State of the Art

---

7 Keywords that would serve as search label for information retrieval
At present the direct synthesis of genes is the most efficient way to generate functional genetic constructs. For this de novo synthesis activated monomers (protonated deoxyribonucleoside 3'-phosphoramidites) are sequentially added to a growing chain that is linked to an insoluble substrate (Letsinger and Mahadevan, 1965; Caruthers et al., 1987). In brief, the 3' phosphorus atom of the monomer is joined to the 5' oxygen atom of the growing chain, whereas a protecting group blocks the 5'-OH group of the monomer. In the next step the generated phosphite triester is oxidized to form a phosphotriester. Finally, the protecting group on the 5'-OH is removed, so that another monomer can be added to the growing chain in the next cycle. Single-stranded oligonucleotides with a length up to 100 monomers can be generated with this method. These oligonucleotides can then be enzymatically assembled to form complete synthetic genes.

In practice, several companies offer conventional gene synthesis (e.g. GenScript, GENEWIZ, Life Technologies and various others). Generally, they serve the end user with an internet-based order interface, in which the desired DNA-sequences can be uploaded by straightforward “copy and paste” action. Some companies (e.g. GenScript, Life Technologies and others) provide tools to further customize the DNA-synthesis to fulfill the needs of scientists such as codon optimization. Additionally, the DNA-sequence can be modified by removal of cryptic splice sites and RNA destabilizing sequence elements. These gene optimization steps appear to result in a maximized expression of the synthetic gene in the desired expression systems. The desired DNA-sequence is then synthesized and is cloned in an output vector of choice before delivery to the end user. However, besides these obvious advantages in this method also limitations become evident, that should be eliminated by using automated DNA processing. In this regard, the most important point is the time- and money-consumption of conventional gene synthesis. CADMAD aims at improving this point by reassembly and rearrangement of existing DNA fragments instead of conventional de novo DNA synthesis. For this, it is crucial to develop a robust graphical user interface for convenient use of CADMAD DNA programming. Additionally, CADMAD DNA processing language should include a possibility to improve the synthetic genes by codon optimization, removal of cryptic splice sites and unwanted restriction sites.

c. Innovation

To achieve the goal of automated DNA-processing a new textual DNA programming language and a new user interface as well as biochemistry and algorithms for computer-based approaches have to be developed. Further processing via robotics and microfluidics has to be automated. Afterwards, the output DNA has to be further analysed and verified by applying the newly developed textual DNA programming language as well as the newly developed user interface. For this, potential end users involved in this work package drafted eight different DNA libraries. These libraries cover various independent research fields and complexities and will serve as contributors to the development of the
system and to its validation. The diversity of the DNA libraries and their descriptions are presented in the results section of this report. The design of these libraries should help to identify further relevant specifications and requirements that need to be included in the CADMAD DNA programming approaches. A list summarizing these requirements can be found in the conclusions section of this report.

2. Implementation

To analyze the user friendliness of the newly developed DNA programming language (DNA library designer (DNald)), eight libraries have been drafted from end users from various independent research fields and scientific backgrounds:

<table>
<thead>
<tr>
<th>Number of library</th>
<th>Short name of partner</th>
<th>Title of library</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UEVE</td>
<td>Exploring the parameter space for the development of combinatorial genetic circuits with computationally predicted behavior and their utilization in bacterial computation</td>
</tr>
<tr>
<td>2</td>
<td>UEVE</td>
<td>Library of reshuffled bacterial chromosomes for the development of rewired genomes with computationally predicted behavior</td>
</tr>
<tr>
<td>3</td>
<td>UEVE</td>
<td>Developing a highly efficient antimicrobial peptide cocktail</td>
</tr>
<tr>
<td>4</td>
<td>UNOTT</td>
<td>Investigating post-transcriptional regulation of <em>Pseudomonas aeruginosa</em> azurin by RsmA</td>
</tr>
<tr>
<td>5</td>
<td>ETHZ</td>
<td>Quasi-combinatorial synthetic operon library</td>
</tr>
<tr>
<td>6</td>
<td>FMI</td>
<td>Sequence replacement library to identify determinants of CpG islands methylation states</td>
</tr>
<tr>
<td>7</td>
<td>UKB</td>
<td>Derivation of Oct4 expression library to identify enhanced variants of Oct4 reprogramming factor</td>
</tr>
<tr>
<td>8</td>
<td>UH</td>
<td>Library of RET-oncogene receptor based on combination of distinct domains from a set of organisms</td>
</tr>
</tbody>
</table>
The specificities of these libraries and the resulting requirements for later versions of the DNA programming language are described below. A comprehensive description of each library can be found in appendix A.

3. Results

Description of designed libraries

1. Exploring the parameter space for the development of combinatorial genetic circuits with computationally predicted behavior and their utilization in bacterial computation (UEVE – Jaramillo group)

This project aims at using bacterial populations for the combinatorial generation and selection for required behavior among random genetic networks. The CADMAD library will be used to generate a subset of those networks with predefined behavior and to explore optimal parameter combinations for their implementation. The library will be processed by generation of permutations in sensitive points of some of our designed network components and inclusion of them in appropriate plasmid backbones. The output library will be characterized \textit{in vivo} in E.coli, and the influence of certain parameters over the behavior will be studied. In addition, it will be used as a template for the \textit{in vivo} selection of working randomly combined genetic circuits in E. coli.

The library is generally based on a “copy and paste”-strategy. However, it is necessary to exclude unwanted restriction sites for subsequent cloning. Additionally, generation of a graphical user interface in the DNA programming language would facilitate generation of the library by enabling a simple “drag-and-drop”-mechanism. Additionally, this library requires the possibility to simplify the definition of a reverse complement of a sequence, preferably without the need to include the sequence in the input section of the software.

2. Library of reshuffled bacterial chromosomes for the development of rewired genomes with computationally predicted behavior (UEVE – Kepes group)

In the context of synthetic biology, the design of an organism that can respond in a directed way to variations in its environment has been a particularly interesting and challenging problem. This design requires the reengineering of suitable signal transduction and regulation systems. Because transcriptional regulation is the most well studied regulatory system in bacteria, it represents an optimal starting point for the design of rewired genomes. The aim of the study is the validation of a computational design methodology integrating currently known transcriptomic and signaling data to refactorize the E. coli genome. Therefore, new genomes will be generated by inserting seven endogenous transcription factors into seven operons according to various permutations. The
reshufflings with the highest transcriptomic fitness were selected for this. The output library will be used to insert the constructions in the genome of E. coli MG1655 and characterize the growth rate.

The specific requirements of this application are the inclusion of a graphical user interface to visually present the library, as well as the labeling and exclusion of unwanted sequences, like restriction sites.

3. Developing a highly efficient antimicrobial peptide cocktail (UEVE – Faulon group)

Due to the emergence of pathogens that are resistant to antibiotics, new therapeutic alternatives must be developed. To this aim, computational tools are developed and an experimental setup is designed in which the anti-bacterial activity of peptides is tested. The CADMAD library will be used to generate this combinatorial library of peptides. Peptides with known anti-bacterial activity and new peptides with predicted antimicrobial activity will be tested in E.coli. The toxicity parameters obtained will be used to build a model for the design of peptide cocktails.

For the design of a combinatorial library of a vast amount of peptides it is helpful to be able to break the scale in a potential graphical user interface, thereby enabling to zoom in and out. This would further facilitate library construction via simple drag-and-drop.

4. Investigating post-transcriptional regulation of *Pseudomonas aeruginosa* azurin by RsmA (UNOTT)

Azurin production in *Pseudomonas aeruginosa* is positively controlled by the RNA-binding protein RsmA. This control is not exerted upstream of the ATG start codon, and cannot be exerted downstream of the rho-independent transcriptional terminator. It could be possible that RsmA enhances the stability of the azurin mRNA by binding to it, thus enhancing the quantity of protein that is translated from it. RsmA usually binds to mRNA at stem-loop structures having the following sequence: (U/A)CANGGANG(A/U). To bind, AGGA or AGGGA have almost always been found on the single-stranded loops. The azurin ORF has three sites that could correspond to RsmA binding sites. Thus seven parts could be defined, of which three could be altered to remove the AGGA sequences and the stem-loop structures, and then recombined to form new variants, which may lose the positive regulation, by RsmA. Therefore, the aim of the study is to investigate the contributions of each potential AGGA stem-loop to azurin transcript stability. The output library will be cloned in *Pseudomonas aeruginosa* PAO1 and the Azurin levels will be measured and related to the combinations of AGGA stem-loops present.

In this library variants were produced using standard translation tools to back-translate sequences, filter these for AGGA and detect and filter stem-loop structures. Therefore, the generation of this library would be facilitated by integrating not only these functions in DNald, but also integrate a function that backtranslation accepts nucleotide sequences which are forward translated prior to backtranslation.
5. Quasi-combinatorial synthetic operon library (ETHZ)

The current objective is to optimize the carbon flow through glycolysis in order to increase dihydroxyacetone phosphate (DHAP) production in *E. coli* by fine-tuning of the relative and absolute amount of enzymes in the system. The approach will be focused on testing the effect of gene dosage (plasmid copy number), ribosome binding site (RBS) strength and promoter strength in a synthetic polycistronic operon encoding the three main bottlenecks of DHAP biosynthesis. Basically, five modules should be quasi-combinatorially recombined: one for plasmid ori (five copy-number options), one for promoter (five strength options), one for RBS (five strength options) and three for enzyme-codifying genes (one gene-option per enzyme. In order to select the best variants from each series of variations a powerful analysis technique will be developed to rapidly perform detailed measurements for each pathway intermediate: particularly focusing on how it is influenced by the variation of the expression of a specific gene. In this regard, enzyme activity analysis and DHAP quantification both based on cell-free extract assays on microplates coupled to TECAN robot handling will be performed. Besides the basic requirements already mentioned above, the subsequent use of this library would be facilitated by being able to select the delivery format of the library output. E.g. in this case a linear DNA fragment in a 96-well microplate is the desired output format.

6. Sequence replacement library to identify determinants of CpG islands methylation states (FMI)

DNA methylation at CpG dinucleotides is a heritable epigenetic modification associated with transcriptional silencing and essential for development. Methylation levels are mainly controlled at the level of the DNA sequence, with little or no influence from neither chromatin, nor transcriptional environment. However, little is known about the identity of the regulatory sequences able to tune DNA methylation states. The aim of the library is to obtain a quantitative assessment of the contribution of different DNA determinants in the establishment of DNA methylation patterns. For this approach, the iterative replacement of the non CpG DNA composing the CpG island by CpG free prokaryotic DNA (E. coli) will be performed. This library then will be cloned in a recombination vector, electroporated in mammalian cells for genomic integration and tested for methylation. Since the DNA library will be cloned in different vectors prior to further use, it is here again important to be able to highlight and exclude restriction enzyme sites. Another helpful alternative is the possibility to choose the delivery format of the library output; in this case distribution in different recombination vectors would be desired.

7. Derivation of Oct4 expression library to identify enhanced variants of Oct4 reprogramming factor (UKB)

The Oct4 transcription factor plays a major role in changing cellular fate of mammalian cells. Reprogramming of somatic cells to a pluripotent state involves viral transduction of Oct4. The
designed library aims at identification of modified variants of Oct4 that allow enhanced reprogramming of target cells. This combinatorial library should be designed in a way that expression constructs encoding Oct4 are linked with various cell-penetrating peptides, linkers and stabilizing peptides. This expression library will then be used a) for the expression of fusion proteins in E.coli, and b) viral transduction of mammalian cells.

To enable expression of this library in various host cells, a possibility to perform a codon optimization for a variety of organisms should be integrated in the DNA programming language. For this, the embedding of DNA optimizer algorithms is highly desired.

8. Library of RET-oncogene receptor based on combination of distinct domains from a set of organisms (UH)

Ret-oncogene receptor is involved in signaling pathway that is essential to maintain the neurons affected in Parkinson's disease. In order to understand the mechanism of signaling the atomic structure is required. The main obstacle here is the successful production of RET protein. So, the aim of the study is the successful expression and crystallization of the complex of RET and its co-receptor and a ligand. For this, a combinatorial library of RET expression constructs was designed in which four extracellular domains (CLD1-4) are shuffled from the set of organisms. The short initial library contains four RET genes from four organisms (256 variants). The expression of successful clones will be examined in a high throughput assay in mammalian expression systems. Proteins will also be functionally screened for ability to bind co-receptor and ligand.

Since this library also aims at the expression of genes from different organisms in a mammalian system, the specific requirement of this application is, as mentioned above, the possibility of codon optimization during design of the library.

4. Conclusions

Specific requirements

Based on the specificities of the libraries described above the need for various basic functions in the DNA programming language becomes evident. These requirements arise in tandem with those presented in D1.1.

1. Integration of a graphical interface enabling “drag-and-drop” of parts of combinatorial DNA.
2. Possibility to break the scale in the visual presentation of the library in order to facilitate “drag-and-drop” (zoom-in and zoom-out).
3. The possibility to view the output in a visual presentation is already included in DNald. However, this function is not working properly at the moment. This should be improved in future versions of the software.

4. Restriction enzyme sites should be highlighted in the software’s output section in order to facilitate subsequent cloning of the libraries.

5. “Toxic sequences” such as unwanted restriction sites or cryptic splice sites have to be excluded in the library output. Such sequences could be integrated from external text files, e.g. derived from web-based databases or publications.

6. A possibility to filter and select against DNA secondary structures, like stem-loops, should be integrated.

7. A function to define the reverse complement of a sequence without the need to integrate the reverse sequence in DNald’s input section should be included.

8. Integration of a function for backtranslation in DNald that accepts not only amino acid but also nucleotide sequences.

9. Integration of codon optimization tables for expression in various host cells.

10. Integration of already existing DNA optimizer algorithms such as GeneOptimizer® (Raab et al., 2010).

11. Compatibility with open source DNA construction software, in particular ApE.

12. Platform-independent DNA programming language, e.g. JAVA based application.

13. The delivery format (e.g. 96 well plate) and the form of the library output (e.g. linear DNA fragment) should be selectable.

14. Quality control of the output DNAs has to be assured, e.g. by verification of sequence.

5 References


### 6. Abbreviations

List all abbreviations used in the document arranged alphabetically.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>DHAP</td>
<td>dihydroxyacetone phosphate</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DNald</td>
<td>DNA library designer</td>
</tr>
<tr>
<td>E.coli</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>e.g.</td>
<td>for example (exempli gratia)</td>
</tr>
<tr>
<td>et al</td>
<td>and others (et alii)</td>
</tr>
<tr>
<td>ETHZ</td>
<td>Eidgenoessische Technische Hochschule Zuerich</td>
</tr>
<tr>
<td>FMI</td>
<td>Friedrich Miescher Institut</td>
</tr>
<tr>
<td>RBS</td>
<td>ribosome binding site</td>
</tr>
<tr>
<td>UEVE</td>
<td>Universite d'Evry Val d'Essonne</td>
</tr>
<tr>
<td>UH</td>
<td>University of Helsinki</td>
</tr>
<tr>
<td>UKB</td>
<td>Universitaetsklinikum Bonn</td>
</tr>
</tbody>
</table>