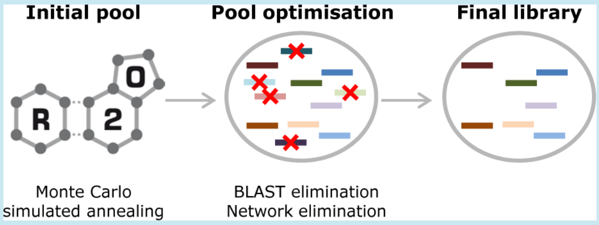
R2oDNA Designer: Computational Design of Biologically Neutral Synthetic DNA Sequences

Arturo Casini,[†,‡] Georgia Christodoulou,[‡] Paul S. Freemont,[†,‡] Geoff S. Baldwin,[†,‡] Tom Ellis,[†,§] and James T. MacDonald*,[†,‡]

[†]Centre for Synthetic Biology and Innovation, Imperial College London, London SW7 2AZ, United Kingdom
[‡]Department of Life Sciences, Imperial College London, London SW7 2AZ, United Kingdom
[§]Department of Bioengineering, Imperial College London, London SW7 2AZ, United Kingdom



ABSTRACT: R2oDNA Designer is a web application that stochastically generates orthogonal sets of synthetic DNA sequences designed to be biologically neutral. Biologically neutral sequences may be used for directing efficient DNA assembly by overlap-directed methods, as a negative control for functional DNA, as barcodes, or potentially as spacer regions to insulate biological parts from local context. The software creates optimized sequences using a Monte Carlo simulated annealing approach followed by the elimination of sequences homologous to host genomes and commonly used biological parts. An orthogonal set is finally determined by using a network elimination algorithm. Design constraints can be defined using either a web-based graphical user interface (GUI) or uploading a file containing a set of text commands.

KEYWORDS: *software, synthetic biology, biologically neutral sequences, linkers*

A key feature of engineering disciplines is the use of modular, well-characterized, and composable parts. Coupled with computer aided design tools and assembly standards, this enables the automated design and assembly of genetic devices, ideally using robotics.[1] Truly modular features are currently lacking in the field of synthetic biology, despite considerable recent effort to address this.[2,3] In recent years, overlap-directed DNA assembly methods such as Gibson Isothermal Assembly have become widely used, accompanied by a move away from standards such as BioBricks, toward bespoke 'one-off' DNA assembly approaches.[4] To address issues arising from this, various strategies have been developed to address specific context dependency issues of small but critical DNA parts, including (i) the Ribosome Binding Site (RBS) calculator,[5] which helps define the RBS-Open Reading Frame (ORF) junction to regulate protein translation initiation, (ii) surrounding minimal promoters with insulating sequences,[6] and (iii) the use of RNA processing of mRNA sequences to leave easily accessible RBS regions.[7,8] In many instances, however, there is a requirement for rationally designed generic nonfunctional DNA that does not lead to unwanted biological effects, such as genetic instability due to recombination, or unwanted changes in gene expression due to unintended incorporation of sequence motifs. Such biologically neutral DNA could be valuable as inter- or intragenic spacer regions that insulate parts from one another and can also be used as a mechanism to direct DNA assembly. Synthetic orthogonal DNA sequences lacking in function could also find use as barcode sequences for high-throughput experiments or simply as negative control DNA, when the function of a natural sequence needs to be compared to that of a sequence with no function. Similar approaches have been proposed independently by other groups[9,10] where the synthetic sequences were referred to as unique nucleotide sequences (UNSs) but their design software had not been made available for external users. Earlier work on the design of 12mer structure-free DNA word sets required a large pre-existing pool of sequences from which to select sequences (e.g., a pool of all possible 12mers), but this method would be computationally intractable for longer sequences.[11,12]

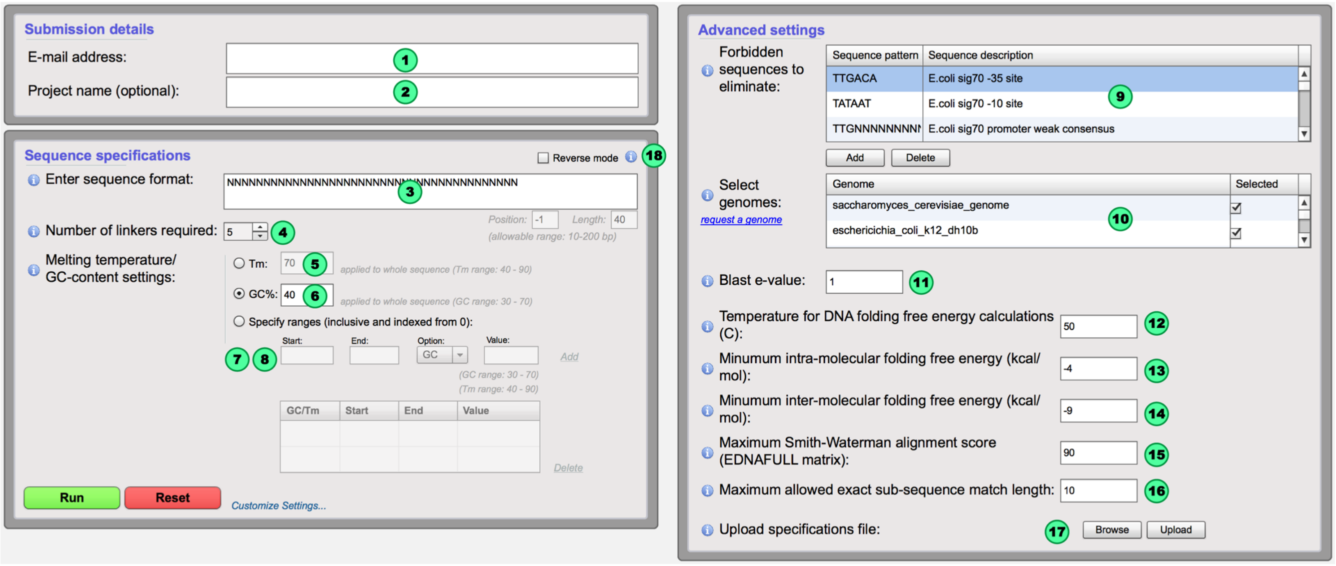
Here, we present a web-based tool to computationally design orthogonal synthetic DNA sequences. In addition, the algorithm is optimized to remove sequences that may be biologically active in an effort to reduce context-dependence. While R2oDNA Designer is intended to design neutral DNA sequences for a variety of needs, its use in designing libraries of linker sequences to act as reliable and efficient adapters for modular overlap-directed DNA assembly reactions has been recently described.[13] The Web tool can be accessed at www.r2odna.com. The software was implemented using Java/Apache Tomcat on the server side and Adobe Flash/Apache Flex on the client side.

HOW DOES R2ODNA DESIGNER WORK?

R2oDNA designer is an online software tool that allows the user to define the parameters and constraints desired for a set of synthetic DNA sequences (e.g., specific GC-content,

Special Issue: IWBDA 2013

Received: September 6, 2013
Published: January 31, 2014



Submission details

E-mail address:

Project name (optional):

Sequence specifications

□ Reverse mode

Enter sequence format: NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

Position: -1 Length: 40 (allowable range: 10-200 bp)

Number of linkers required: 5

Melting temperature/ GC-content settings:

○ Tm: 70 applied to whole sequence (Tm range: 40 - 90)

● GC%: 40 applied to whole sequence (GC range: 30 - 70)

○ Specify ranges (inclusive and indexed from 0):

Start: End: Option: GC Value:

(GC range: 30 - 70)
(Tm range: 40 - 90)

GC/Tm	Start	End	Value

Advanced settings

Forbidden sequences to eliminate:

Sequence pattern	Sequence description
TTGACA	E.coli sig70 -35 site
TATAAT	E.coli sig70 -10 site
TTGNNNNNNNNNN	E.coli sig70 promoter weak consensus

Select genomes:

Genome	Selected
saccharomyces_cerevisiae_genome	☑
eschericichia_coli_k12_dh10b	☑

Blast e-value: 1

Temperature for DNA folding free energy calculations (C): 50

Minumum intra-molecular folding free energy (kcal/mol): -4

Minumum inter-molecular folding free energy (kcal/mol): -9

Maximum Smith-Waterman alignment score (EDNAFULL matrix): 90

Maximum allowed exact sub-sequence match length: 10

Upload specifications file:

Figure 1. R2oDNA Designer graphical user interface. See Table 1 for the key to the number labels.

forbidden restriction enzyme sites, etc.) in a browser-based GUI. The specifications are then sent to a server, which submits a job to a queuing system on a cluster. The final results are a set of orthogonal linker sequences that satisfy the input constraints. These are emailed to the user in the form of a ZIP archive containing a FASTA format file with the orthogonal sequences (sequencesFinalFile.fa), a text file that contains the parameters (jobSpecifications.txt) used to generate these sequences, and a text file showing the number of sequences remaining at each step in the design process (adminStatFile.txt). The parameters file can be edited by the user, uploaded, and parsed by the web tool to relieve the user of having to enter long lists of parameters in the GUI. The adminStatFile.txt file is useful for users to refine design parameters if too many sequences were eliminated at any stage in the design process.

Parameters. Design specifications are entered either in a GUI interface or by uploading a text file containing the formatted parameters (Figure 1 and Table 1). These allow the user to specify sequence constraints including overall length, the bases permitted at each position, GC content, or T_m targets for all or subsections of the sequence, disallowed sequence motifs, and the parameters used to produce an orthogonal set.

Initial Sequence Generation Using Monte Carlo Simulated Annealing. A simple scoring function consisting of a linear weighted sum of five terms was defined: $S = w_{hp}S_{hp} + w_rS_r + w_fS_f + w_{GC}S_{GC} + w_{T_m}S_{T_m}$, where S_{hp} is the sum of the length of inverted repeats longer than a minimum length 4, S_r is the sum of the length of repeats over a minimum length 6, S_f is the sum of the number of matches to a list of forbidden sequence motifs in the form of Java regular expressions, S_{GC} penalizes GC-content straying from a specified target value and is equal to the length of the specified region multiplied by the square of the difference between the target GC-content and the actual GC-content. S_{T_m} is the same as S_{GC} but for T_m rather than GC. T_m is calculated using the base stacking algorithm of SantaLucia[14] with 50 mM monovalent salt concentration, 0 mM divalent salt concentration, and 200 nM primer concentration. The weights w_{hp}, w_r, and w_f are set to 10.0, and the weights w_{GC} and w_{T_m} are set to 1.0. This scoring function is minimized using Monte Carlo simulated annealing (MCSA) starting from initial random sequences multiple times to generate a pool of initially optimized sequences. Only sequences that minimize the terms S_{hp}, S_r, and S_f to zero and keep GC-content and T_m within a certain tolerance are accepted into this initial pool. The initial pool is then further culled by eliminating sequences with DNA secondary structures or self-annealing (including the self-annealing of the reverse complement) minimum free energies (MFEs) below user defined cut-offs (Table 1). The MFE values are calculated using PairFold.[15]

BLAST and Network Elimination. The initial sequence pool is further optimized by BLASTing against possible chassis organism genomes and commonly used biological parts (Table 1) using the software program BLASTN[16] with default settings but with the word size parameter reduced to 8. Any linkers with hits below a defined E-value cutoff (default 1.0) are eliminated. This has the dual aim of reducing the chance of any undesirable homologous recombination events between designed sequences and host genomes and to help remove sequences with possible biological function. Finally, in order to ensure sequences are orthogonal to one another, those with the potential to mis-anneal to other sequences in the generated set are removed using a network elimination algorithm.[17] This has the added benefit of reducing downstream evolutionary instability by minimizing the likelihood of homologous recombination within the sequences generated within a set, as well as ensuring orthogonality for other purposes, such as eliminating cross-annealing of DNA sequences during DNA assembly or for PCR primer binding sites.

The network elimination algorithm works by creating an undirected graph based on the results of all-against-all pairwise alignments and pairwise MFEs of the sequences in the pool. Edges are drawn between two sequences if (i) there is any exact subsequence match between the sequences above a defined length, if (ii) the Smith-Waterman local sequence alignment score is above a defined value, or if (iii) the pairwise DNA folding free energy calculated using PairFold is below a defined value. Alignments and free energy calculations are conducted in all possible orientations, and all parameters can be user defined (Table 1). A random sequence is picked, and all connected sequences were eliminated. This process is iterated until a

Table 1. R2oDNA Designer Parameters[a]

location on GUI (Figure 1)	text command: keyword <argument1> <argument2> ...	description
1	`Email <email address>`	Email address to send results.
2	`ProjectName <user project name>`	Arbitrary user defined name for the set of linkers.
3	`Format <IUPAC degenerate sequence format of sequences required>`	This allows the user to specify the total length of the linker and positions that are allowed to vary or must be fixed in IUPAC degenerate nucleotide format. For example the command "Format NNNNNNNNNN" would result in a set of linkers 10 bp long with all positions allowed to vary freely. The command "Format NNNNNNNNTS" would result in a set of linkers where the 9th base position is fixed as T and the last position may be a G or a C. Note that this feature can be used to add 5′ and 3′ sequence context around variable regions.
4	`Seqnum <number of linkers required>`	The number of orthogonal sequences required by the user.
5	`tm <value>`	Target T_m for whole sequence. This command and the commands tm_range and gc_range commands are mutually exclusive.
6	`gc <value>`	Target GC content for whole sequence. This command and the commands tm_range and gc_range commands are mutually exclusive.
7	`tm_range <value> <start position> <end position>`	Target T_m for a specific range over the sequence. More than one of these ranges may be specified. Positions are inclusive and indexed from 0.
8	`gc_range <value> <start position> <end position>`	Target GC content for a specific range over the sequence. More than one of these ranges may be specified. Positions are inclusive and indexed from 0.
9	`Genomes <genome1> <genome2> ...`	Names of sequence databases to be searched using BLAST. Currently these are limited to *saccharomyces_cerevisiae_genome, eschericichia_coli_k12_dh10b, eschericichia_coli_k12_mg1655, eschericichia_coli_k12_w3110, bacillus_subtilis_168 and igem_all_parts_082013*. The iGEM database was built the FASTA file obtained from http://parts.igem.org/fasta/parts/All.Parts as downloaded on Aug. 30, 2013.
10	`ForbSeq <Java regex pattern> <description>`	Sequence motifs that are forbidden in the designed sequences. These are specified in the form of Java regular expressions. IUPAC degenerate codes are automatically translated into equivalent regular expressions internally; e.g., *W* is translated into *[AT]*.
11	`Evalue <value>`	BLASTN E-value. Any linker sequences with BLAST hits below this value are eliminated.
12	`FoldTemp <value>`	Temperature (°C) at which DNA secondary structure minimum free energies (MFE) are calculated using the software PairFold. This affects the MinIntraEnergy and MinInterEnergy MFE calculations. All MFE calculations are carried out at 1 M NaCl.
13	`MinIntraEnergy <value>`	Minimum allowed intramolecular MFE allowed calculated using PairFold. This is used to filter sequences with residual secondary structure generated by MCSA step and before the BLAST elimination step. Both forward and reverse strands MFEs are calculated.
14	`MinInterEnergy <value>`	Minimum allowed intermolecular MFE allowed calculated using PairFold. This is used to filter self-annealing sequences generated by MCSA step and before the BLAST elimination step. Additionally, this is used during the network elimination step. MFEs for all forward and reverse strand combinations are calculated at all stages.
15	`MaxSWScore <value>`	Maximum Smith−Waterman (SW) local sequence alignment score allowed. This is used during the network elimination step to eliminate homologous sequences from the final set. SW scores are calculated for all forward and reverse strand combinations using the EDNAFULL matrix.
16	`MaxSubMatch <value>`	Maximum length of any exact subsequence match allowed. This is used during the network elimination step. This is calculated for all forward and reverse strand combinations.
17	`N/A`	Upload specifications file with text commands described in this table.
18	`ReverseMode <true/false>`	Enables reverse mode. If this option is selected then comma separated sequences may be entered into the "sequence format" box (see location 3). On submitting the job, these sequences are scored and emailed back to the user. IUPAC degenerate bases are not permitted in this mode.
N/A	`JobID <unique ID>`	Nonuser definable unique job ID. The user needs this to track job progress.
N/A	`ClusterID <unique ID>`	Nonuser definable unique ID of job on the cluster queuing system. For internal use only.

[a] All parameters can be set using the GUI or by uploading a text file containing the corresponding commands. All results are returned with a corresponding text command file (jobSpecifications.text).

completely disconnected graph is obtained. The remaining sequences then formed the final set of orthogonal sequences.

Job Progress Tracking. On pressing the 'Run' button, a unique job ID tracking code is displayed on a pop-up window and also emailed to the user. Clicking on the "Track jobs" tab in the top right of the GUI allows the user to enter the job ID and track progress of the job on the queuing system. Results are emailed to the user on completion.

Reverse Mode. This mode scores existing sequences using the scoring functions described above rather than design new sequences. This is a useful functionality for assessing the biological neutrality of existing sequences, and may be used to assess their suitability for use in designing biological systems. In this mode, lists of comma separated sequences are entered by the user in the sequence format box. Results are emailed to the user as a ZIP archive containing a comma separated file containing the scores and BLAST hit counts for each sequence (AllScores.csv) and a file containing the raw pairwise scores used in the network elimination step (NetworkScores.csv).

■ FUTURE DEVELOPMENT

Designing nonfunctional synthetic DNA sequences, whether to act as spacer regions or reliable linker sequences for DNA assembly, requires removal of DNA sequences known to have function. As more is understood about sequence-to-function relationships in different organisms and genetic contexts, the list of sequences to remove will increase. A future development for R2oDNA Designer is to crowd-source undesired sequence motifs by maintaining a registry of those sequences specified by users of the software as ones to not include. Future users will therefore be able to learn from one another what sequences to specify as unwanted.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: j.macdonald@imperial.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the EPSRC [EP/G036004/1, EP/H019154/1] and EU Framework Program 7 [ST-FLOW; KBBE.2011.5-289326]. Valentina Gnassi is thanked for creating the R2o logo.

■ REFERENCES

(1) MacDonald, J. T., Barnes, C., Kitney, R. I., Freemont, P. S., and Stan, G. B. (2011) Computational design approaches and tools for synthetic biology. *Integr. Biol. (Camb.) 3*, 97−108.

(2) Mutalik, V. K., Guimaraes, J. C., Cambray, G., Lam, C., Christoffersen, M. J., Mai, Q. A., Tran, A. B., Paull, M., Keasling, J. D., Arkin, A. P., and Endy, D. (2013) Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods 10*, 354−360.

(3) Mutalik, V. K., Guimaraes, J. C., Cambray, G., Mai, Q. A., Christoffersen, M. J., Martin, L., Yu, A., Lam, C., Rodriguez, C., Bennett, G., Keasling, J. D., Endy, D., and Arkin, A. P. (2013) Quantitative estimation of activity and quality for collections of functional genetic elements. *Nat. Methods 10*, 347−353.

(4) Ellis, T., Adie, T., and Baldwin, G. S. (2011) DNA assembly for synthetic biology: From parts to pathways and beyond. *Integr. Biol. (Camb.) 3*, 109−118.

(5) Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol. 27*, 946−950.

(6) Davis, J. H., Rubin, A. J., and Sauer, R. T. (2011) Design, construction and characterization of a set of insulated bacterial promoters. *Nucleic Acids Res. 39*, 1131−1141.

(7) Lou, C., Stanton, B., Chen, Y. J., Munsky, B., and Voigt, C. A. (2012) Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat. Biotechnol. 30*, 1137−1142.

(8) Qi, L., Haurwitz, R. E., Shao, W., Doudna, J. A., and Arkin, A. P. (2012) RNA processing enables predictable programming of gene expression. *Nat. Biotechnol. 30*, 1002−1006.

(9) Guye, P., Li, Y., Wroblewska, L., Duportet, X., and Weiss, R. (2013) Rapid, modular, and reliable construction of complex mammalian gene circuits. *Nucleic Acids Res. 41*, e156.

(10) Torella, J. P., Boehm, C. R., Lienert, F., Chen, J. H., Way, J. C., and Silver, P. A. (2014) Rapid construction of insulated genetic circuits via synthetic sequence-guided isothermal assembly. *Nucleic Acids Res. 42*, 681−689.

(11) Tulpan, D., Andronescu, M., Chang, S. B., Shortreed, M. R., Condon, A., Hoos, H. H., and Smith, L. M. (2005) Thermodynamically based DNA strand design. *Nucleic Acids Res. 33*, 4951−4964.

(12) Shortreed, M. R., Chang, S. B., Hong, D., Phillips, M., Campion, B., Tulpan, D. C., Andronescu, M., Condon, A., Hoos, H. H., and Smith, L. M. (2005) A thermodynamic approach to designing structure-free combinatorial DNA word sets. *Nucleic Acids Res. 33*, 4965−4977.

(13) Casini, A., MacDonald, J. T., Jonghe, J. D., Christodoulou, G., Freemont, P. S., Baldwin, G. S., and Ellis, T. (2013) One-pot DNA construction for synthetic biology: The Modular Overlap-Directed Assembly with Linkers (MODAL) strategy. *Nucleic Acids Res. 42*, e7 DOI: 10.1093/nar/gkt915.

(14) SantaLucia, J., Jr. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. U.S.A. 95*, 1460−1465.

(15) Andronescu, M., Aguirre-Hernandez, R., Condon, A., and Hoos, H. H. (2003) RNAsoft: A suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Res. 31*, 3416−3422.

(16) Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol. 215*, 403−410.

(17) Xu, Q., Schlabach, M. R., Hannon, G. J., and Elledge, S. J. (2009) Design of 240 000 orthogonal 25mer DNA barcode probes. *Proc. Natl. Acad. Sci. U.S.A. 106*, 2289−2294.