3. About R2oDNA Designer

Please read these publications for more details:

Casini A, Christodoulou G, Freemont PS, Baldwin GS, Ellis T, MacDonald JT. R2oDNA Designer: Computational design of biologically-neutral synthetic DNA sequences. (2014). Under review


R2oDNA Designer generates neutral synthetic DNA sequences to user preferences and also assesses existing sequences for suitability as spacers and linkers in DNA assembly. R2oDNA Designer was developed in parallel with the Modular Overlap-Directed Assembly with Linkers (MODAL) Strategy [Citation] for DNA Assembly.

**How it works:**

1. Create the first pool of sequences

![Diagram of creating the first pool of sequences]

The user enters a seed sequence between 10 and 200 bases in IUPAC form, where e.g. N = random base. From this a pool of random sequences that match the user requirements (GC content, melting temperature, forbidden sequences, etc) is generated by a Monte Carlo Simulated Annealing algorithm. This results in an initial pool of sequences.

2. Eliminate sequences that match against genomes to create the second pool of sequences

![Diagram of eliminating sequences]

The first pool of sequences can then be compared to key genome sequences (as defined in the advanced settings) and the sequences of all parts in the 2013 iGEM Parts Registry [LINK] using
BLASTN [citation]. Any sequences with significant hits are removed, leaving the second pool of sequences.

3. Eliminate sequences that cross-anneal to leave an orthogonal set of sequences

A network is created using the second pool of sequences, and a network elimination algorithm is applied to this to remove all sequences that could cross-anneal to one another.

**Results** are emailed to the user in the form of a ZIP archive containing the following 3 files:

- sequencesFinalFile.fa contains the final set of generated sequences in FASTA format.
- jobSpecifications.txt contains the list of parameters used to generate the sequences and this can be used to re-run the job without the user having to re-enter the parameters manually in the user interface.
- adminStatFile.txt is a text file containing statistics on the number of sequences remaining at each step in the design process and is provided to help the user debug jobs that return fewer sequences than required.

**Reverse mode:**

In addition to designing *de novo* sequences, R2oDNA Designer can be used to assess sets of existing sequences using the same scoring functions and criteria. This can be used to troubleshoot problems with existing sets of orthogonal sequences. This mode can be entered by clicking the ReverseMode tickbox. A comma separated list of non-degenerate sequences to be scored can be entered in the Sequence Format box.

**Results** are returned to the user as a ZIP archive containing the following four files:

- scoredSequences.fa contains the input comma-separated sequences reformatted into FASTA format and sequentially labelled seq1, seq2, seq3, etc.
- jobSpecifications.txt contains the list of parameters used to generate the sequences and this can be used to re-run the job without the user having to re-enter the parameters manually in the user interface.
- AllScores.csv is a comma-separated file (CSV) containing all single-sequence scores. The first line contains the heading for each column. The GC content and Tm headings are variable and depend on the input parameters. Typically headings are:
  - name – the name of the sequence
  - sequence – the actual sequence
total_score – the total weighted sum of all scores as used in the Monte Carlo optimisation step (see R2o and MODAL papers)

Delta_GC:all_50.0% - the difference from the specified GC content (in this case 50%)

HairpinScore – the weighted inverse repeat score (see R2o and MODAL papers)

RepeatScore – the weighted repeat score (see R2o and MODAL papers)

TabooScore – the forbidden sequence score (see R2o and MODAL papers)

seq_seq_mfe – the self-annealing dimer minimum free energy (MFE)

seq_inv_seq_inv_mfe - the self-annealing dimer MFE for the inverse complement sequence.

seq_intra_mfe,name – the monomer folding MFE

seq_inv_intra_mfe,name – the monomer folding MFE of the inverse complement sequence

BLASThitCounts – the number of BLAST hits to the selected genome or iGEM database sequences.

- NetworkScores.csv is a CSV file all pairwise scores used in network elimination. The first line contains the headings for each column.

seq1_label – this is the name of the first sequence in the pairwise comparison

seq2_label – this is the name of the second sequence in the pairwise comparison

is_connected – this is a Boolean value and is set to true if the two sequences would be connected in the network elimination graph given the current parameters (i.e. these two sequences should NOT be used in the same orthogonal set).

seq1_seq2_SWScore – Smith-Waterman alignment score between seq1 and seq2

seq1_seq2_inv_SWScore – Smith-Waterman alignment score between seq1 and the inverse complement of seq2

seq_seq_inv_mfe – MFE of the dimer of seq1 and the inverse complement of seq2

seq_seq_mfe - MFE of the dimer of seq1 and seq2

seq_inv_seq_mfe – MFE of the dimer of the inverse complement of seq1 and seq2

seq_inv_seq_inv_mfe – MFE of the dimer of the inverse complement of seq1 and the inverse complement of seq2

has_submatch – a Boolean value that is true if the sequences share an exact sequence submatch above the cutoff length defined by the parameters MaxSubMatch
FAQs

Q: Why did my job return more/fewer sequences than requested?

A: The software generates ten times the requested number of sequences at the initial Monte Carlo simulated annealing (MCSA) step. Sequences are then eliminated at the BLAST and network elimination steps. Due to the nature of the algorithm it is impossible to predict the exact number that will be eliminated in advance. Some parameter sets may result in no sequences being returned. For example, if input sequence format contains a fixed sequence motif that always results in a BLAST hit to one of the selected genomes then no sequences will be returned. It may be necessary to adjust the input parameters in order get sequences designed.

Q: Why did I get a message saying my job had been killed by the queuing system?

A: The most likely explanation is that your job took longer to run than the maximum run time allowed. It is possible that the MCSA algorithm was unable to generate any sequences that satisfy the design constraints given the input parameters and so remained stuck. If this is the case you should check the sequence format and the GC content and Tm parameters.

Q: The sequences that were designed for me didn’t work in my DNA Assembly!

A: We don’t yet know everything there is to know about biology yet and so there is a chance that the sequences generated contained as-yet-unknown motifs that lead to unanticipated effects either in DNA assembly or in DNA replication and/or gene expression. Thousands of smart researchers around the world are working right now on making biology more predictable and hopefully over time our understanding of how every base of DNA encodes biological function will improve. Meanwhile, you may want to consider looking at the local sequence context in which you’ve used the designed sequences. If you take the designed sequence along with 50 to 100 bases of upstream and downstream flanking sequence and upload this to sequence analysis software, it will tell you if your designed sequence forms unintended DNA/RNA folds with its neighbouring sequence or creates a region that act as a ribosome binding site for bacterial gene expression. Recommended sequence analysis sites are: denovoDNA (RBS Calculator), UNAFOLD (DNA/RNA folding)... etc...

Q: This is great! How can I help improve it?

A: