**OptMAGEv0.9 Documentation**

Written by Harris H. Wang

[hhwang@genetics.med.harvard.edu](mailto:hhwang@genetics.med.harvard.edu)

Laboratory of Prof. George Church

Department of Genetics

Harvard Medical School

New Research Building, Room 238

77 Avenue Louis Pasteur

Boston, MA 02115, USA

Last updated: April 23, 2010

**1. OptMAGE General Description**

Using the Multiplex Automated Genome Engineering (MAGE) technology, we can rapidly manipulate the genome of an organism through single-stranded oligonucleotide-mediated allelic replacement (Wang HH, Nature 2009). Various genomic modifications can be made through MAGE including insertion of new sequences (insertions), deletion of genomic sequences (deletions), and replacement of genomic sequences with desired sequences (mismatches). These genomic modifications can be directed to protein-coding sequences (*e.g.* genes) or non-protein-coding sequences such as regulatory units (e.g. ribosomal binding sites, transcriptional binding sites, promoters) and other sequences (e.g. insertion sequences, transposable elements). Episomal sequences may also be targeted. Sequence diversity of genes or operons can be serially and combinatorially generated using the MAGE method.

Optimization of the design of single-stranded oligonucleotides (ss-oligos) used in MAGE will increase the replacement efficiency and expand the capability of MAGE in size, scale, and number of modifications. Various parameters that increase the efficiency of allelic replacement govern the design of MAGE oligos including:

1. base-pair length of oligos
2. base-pair size of the modifications or mutations desired
3. oligo pool complexity (for multiple targets and combinatoric variant generation)
4. oligo degeneracy (for single genomic target but sequence diversity generation)
5. oligo secondary structure due to internal hybridizations or foldings
6. chemical modifications to oligo to increase half-life *in vivo* (*i.e.* terminal 5’ phosphorothioated bonds)
7. targeting leading or lagging strand at the replication fork (lagging strand preferred)
8. local secondary genomic secondary structure

OptMAGE attempts to optimize oligos for maximal efficiency based on these design criteria.

**2. Detailed descriptions of optimization parameters**

This section describes each optimization parameter in detail, how they impact allelic replacement efficiency, and how they are used in OptMAGE.

1. **Oligo length**

For oligo with four terminal 5’ phosphorothioated bonds, the replacement efficiency fit is:

*y* = -0.00000310*x*4 + 0.00059523*x*3 - 0.03321546*x*2 + 0.66738907*x*

(R2=0.985)

For oligo with two terminal 5’ phosphorothioated bonds, the replacement efficiency fit is:

*y* = -0.00000255*x*4 + 0.00049119*x*3 - 0.02740964*x*2 + 0.55073438*x*

where *x* is the oligo length in basepairs and *y* is the % replacement efficiency

Note: the ratio of the fit of 4mods to 2mods is 1.2 (see section E).



1. **Size of modification or mutation desired**
2. Mismatches

The mismatch size is logarithmically related to the replacement efficiency. The coefficient A0 is not important as its impact on efficiency scales with the oligo length, concentration, and chemical modifications. The important parameter is the exponential decay which is -0.1352.

The fit function is

*y* = *A*0 *e*-0.1352*x* (R2=0.972)

where *A*0 for a 2mod is 22.1, *x* is the size of mismatch in basepairs, and *y* is the replacement efficiency in % of population converted.



1. Insertions

The insertion size is logarithmically related to the replacement efficiency. The coefficient A0 is not important as its impact on efficiency scales with the oligo length, concentration, and chemical modifications. The important parameter is the exponential decay which is -0.0751.

The fit function is

*y* = *A*0 *e*-0.0751*x* (R2=0.856)

where *A*0 for a 2mod is 12.2, *x* is the size of mismatch in basepairs, and *y* is the replacement efficiency in % of population converted.



1. Deletions

For deletion size of >30bps, the logarithm of the deletion size is logarithmically related to the replacement efficiency. The coefficient A0 is not important as its impact on efficiency scales with the oligo length, concentration, and chemical modifications. The important parameter is the exponential decay which is -1.3667.

The fit function is

*y* = *A*0 *e*-1.3667*x* (R2=0.944)

where *A*0 for a 2mod is 33.342, *x* is the log10 size of deletion in basepairs, and *y* is the replacement efficiency in % of population converted.



For deletion size of <30bps, the replacement efficiency can be modeled as an exponential decay function of the deletion size.



The fit function is

*y* = *A*0 *e*-0.0579x (R2=0.846)

where *A*0 for a 2mod is 18.93, *x* is the size of deletion in basepairs, and *y* is the replacement efficiency in % of population converted.

1. Hybridization free energy of oligo and genomic sequence

The hybridization free energy of the oligo and the genomic sequence is linearly correlated. This data is based on oligos with mismatches distributed across the entire length of the oligo.

The optimization fit is

*y* = -0.288*x* -13.66 (R2=0.799)

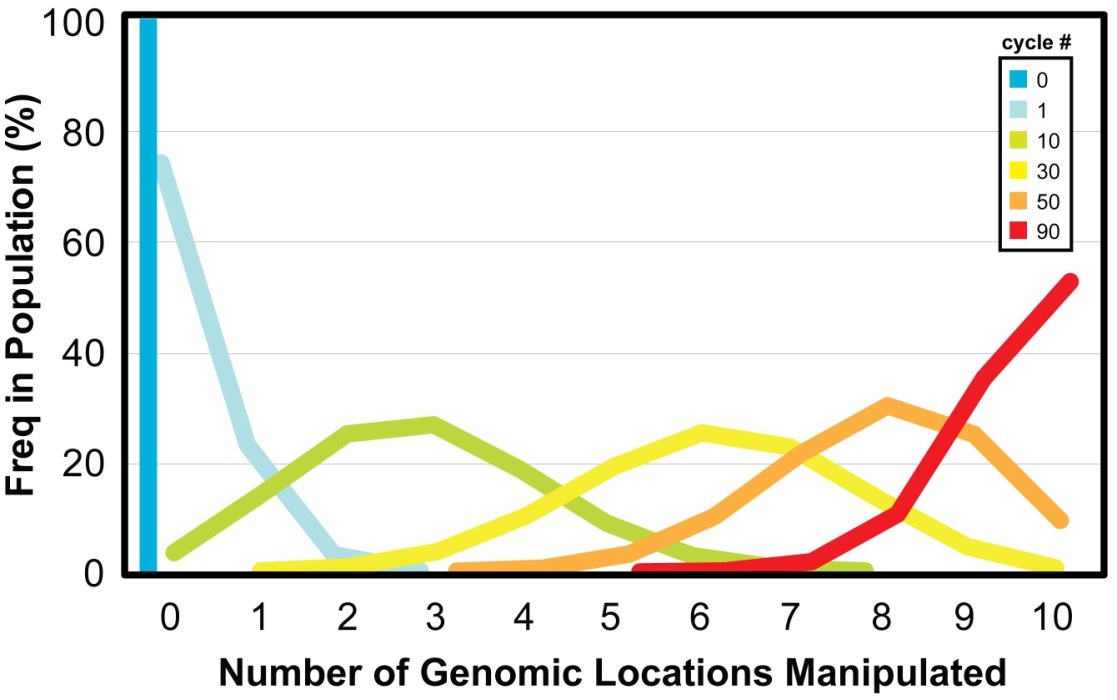
where *x* is the size of mismatches in basepairs, and *y* is the % replacement efficiency



1. **Oligo pool complexity**

When multiple targets are directed using oligos, we can quantitatively assess the distribution of the combinatorial variants generated through the MAGE process after each cycle. The MAGE cell population is binomially distributed according to the equation: ,

where *K* is the number of loci simultaneously targeted, *N* is the number of MAGE cycles, and *M* is the mutation rate at any individual locus.



The above figure shows the predicted distribution of genetic variants in a population that has undergone simultaneous allelic manipulation at 10 different genomic locations at 30% overall replacement efficiency. In this case, 10 different genes are simultaneously targeted for inactivation. Each colored solid line represents the histographic distribution of variants containing different numbers of knockouts (KO) across the population as a function of MAGE cycle number. As MAGE cycles increase, population evolves towards acquiring all 10 gene KO’s.

1. **Oligo degeneracy**

When a degenerate oligo sequence is used (e.g. for RBS or promoter tuning or sequence space exploration), we can calculate potential bias of the system toward particular sequences. For example, a degenerate sequence of 10 base pairs will have some sequences that are much more similar (i.e. more homology) to the target sequence and thus will have a higher replacement efficiency than a sequence whose 10 base pairs are all non-homologous to the target sequence. Thus we can calculate biases in the sampling distribution initially and how quickly those biases disappear as MAGE cycles increase.

For a degenerate oligo with *N* bps of degenerate sequences (2 to 4 base degeneracy) in a 90mer, what is the distribution of homologous sequences to the wild-type sequence?

Analytical solution:



and



where the power series coefficient *am* represents the frequency of having *m* homologous sequences to the wild-type in the degenerate sequence, and  (that is *m* is an integer between 0 and (*n*4 + *n*3 + *n*2)). The converse of *bl* is that it represents the frequency of having *l* non-homologous sequences (1-*cm*). Here *n*4, *n*3 or *n*2 are the number of bases of degeneracy of 4, 3, or 2 respectively and *x* and *y* are place holders for the coefficients.

For *m* = *n*4, the frequency is simply (1/4)^*n*4

For *m* = *n*3, the frequency is simply (1/3)^*n*3

For *m* = *n*2, the frequency is simply (1/2)^*n*2

To solve this equation fully for all other combinations of nonzero solutions to *n*4, *n*3 and *n*2, we take the Taylor series of *g*(*x*) to obtain the coefficient *cm* analytically.

, … for all *m*.

This procedure can be accomplished using MatLab symbolics toolbox or any equivalent symbolic manipulation toolkit.

*For example:*

wild-type sequence tgggataagacataacaaa

degenerate sequence tgggDDRRRRRDDDDcaaa

the oligo degeneracy is 23328 and the homology breakdown is:

|  |  |
| --- | --- |
| deg | freq |
| 4 | 0 |
| 3 | 2 |
| 2 | 7 |
| 0 | 1 |

This yields the following homology distribution with a mean of 4.166 bps of homology in the degenerate oligo



Combining this analysis with the efficiency of recombination for oligos with different number of mismatched bases (see Section B), we can quantitatively predict the bias of the degenerate oligo and track its potential sequence trajectory through the cycle process.

1. **Secondary structure**

The secondary structure due to internal folding of the oligo can inhibit binding to β (the single-stranding DNA-binding protein that help mediate the allelic replacement process) or prevent the exposure of single-stranded bases from targeting the genomic targets. The secondary structure appears to affect replacement efficiency in a linear fashion with more dramatic inhibitory effects beyond the -12.5 kcal/mol range. The linear fit is

*y* = 0.99*x* + 22.5 (R2=0.3695)

where *x* is the optimal folding energy in kcal/mol as predicted by UNAFold and *y* is replacement frequency in arbitrary units.



1. **Chemical modification of oligo (5’ phosphorothioated bonds)**

We showed previously that terminal 5’ modification on the oligo such as phosphorothioated bonds (but not 3’) can increase the replacement efficiency likely due to decreasing degradation by exonucleases and increasing oligo half-life inside the cell.

For oligos at 2uM concentration, the replacement efficiency fit is:

*y* = 0.0405*x*3 - 1.0950*x*2 + 8.4697*x* + 13.0651 (R2 = 0.9979)

where *x* is the number of terminal 5’ phosphorothioated and *y* is the replacement efficiency in %.



1. **Leading vs. lagging strand targeting**

Lagging strand targeting is highly preferred than leading strand when designing oligos. The leading or lagging strand is determined in *E.coli* by assessing the replichore (1 or 2) and genome strand (+ or -) of the target location. If the target lies on replichore 1 and (+) strand or on replichore 2 and on the (-) strand, then the sense strand is always targeted (oligo is complementary to the sense of the target). . If the target lies on replichore 1 and (-) or on replichore 2 and on the (+) strand, then the anti-sense strand is always targeted (oligo is complementary to the anti-sense of the target). See illustration below of an *E.coli* genome (MG1655) with its replichore representation and all relevant target scenarios.



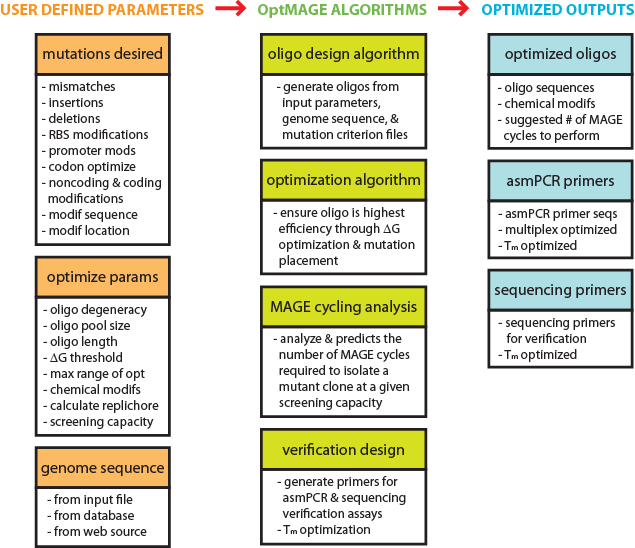
**3. OptMAGE Implementation**

OptMAGE is written from scratch in Perl programming language and utilizes two additional module packages:

- bioperl (<http://www.bioperl.org/>) and

- UNAFold (<http://dinamelt.bioinfo.rpi.edu/download.php>)

The figure below summarizes the functional components of OptMAGE



OptMAGE will be available as a downloadable package or can be used through a web service. In either format, the user is asked to input 3 essential sets of parameters:

1. Genomic sequence of the organism being manipulated
   1. From file or database (local or web-based)
2. Optimization parameters of MAGE
   1. Degeneracy of oligos
   2. Pool size of oligos or the number of genomic targets
   3. Length of oligos
   4. G threshold cut off to initiate secondary structure optimization
   5. Maximum basepair range of optimization
   6. Presence of chemical modifications (i.e. phosphorothioated bonds)
   7. Automatic calculation of replichore information to determine leading/lagging strand targeting
   8. Screening capacity
3. Genomic mutations or modifications desired:
   1. Identifier for each entry
   2. Desired target modification to genomic sequence
   3. Location of target modification
   4. Strand of target site (+ or -)
   5. Replichore information of target site (1 or 2)
   6. Amplicon size of asPCR product of the target site during verification
   7. Amplicon size of product for sequencing of the target site during verification
   8. Desired Tm of asPCR and sequencing primers

**OptMAGE Operation**

The user is first asked to enter the essential parameters. Once these parameters are entered into program, the *oligo design* *algorithm* is called which determines the initial oligo sequences for each of the given targets based on user-defined modification sequence, oligo length, target location and replichore information. The *optimization algorithm* is then called to assess the replacement efficiency of the oligo sequence designed based on its predicted *G* metric (UNAFold). If the oligo sequence has a *G* that is lower than the defined threshold, the algorithm iteratively reconstructs the oligo by shifting the location of the target mutation along the oligo towards the 3’ terminus until it reaches the defined maximal distance. This list of oligos is then assessed for the lowest *G* and is chosen as the optimized oligo sequence. If the initial oligo sequence has a *G* that is higher than the defined threshold, no optimization is done. In case no optimized oligo sequences can be found, the initial sequence is taken as the final optimized sequence.

*Example:*

Initial oligo sequence:

5’ AGACCGCCATACCAGAAACTAGCTCCTCCGACTACGGGCCAGAAG**C**TCAGCAGGTCAGTCAGTCACAGCGGCGCA**A**ACCCTCTGCGACCA 3’

*G*: -20 kcal/mol

G threshold: -12.5 kcal/mol

Mutation location (from 3’): **45**

Max mutation shift (from 3’): **15**

Optimized oligo sequence:

5’ CTCCTTAAATAGGCTCTACCGAAATCTCCGAGACCGCCATACCAGAAACTAGCTCCTCCGACTACGGGCCAGAAG**C**TCAGCAGGTCAGTC 3’

*G*: -7 kcal/mol

Mutation location (from 3’): **15**

Max mutation shift (from 3’): **15**

The optimized sequence is then sent to the *MAGE cycling analysis algorithm*. A predefined “ideal optimized oligo” with replacement efficiency (*RE*) of 32% is used for the efficiency analysis. This default oligo contains parameters:

* + - *Gss*: -5 kcal/mol
    - Chemical modification: four 5’ phosphorothioated bonds
    - Oligo length: 90 bps
    - *Ghyb*: -117 kcal/mol
    - Mismatch of 1 bp
    - Insertion of 1bp
    - Deletion of 1 bp

The algorithm takes the current optimized oligo parameters and compares it with the ideal parameters along the following criterion:

secondary structure linear penalty: 0.991

hybridization linear penalty: 0.288

mismatch exponential decay: -0.1352

insertion exponential decay: -0.0751

deletion exponential decay: -1.3667 (exponential *x* variable)

chemical modification factor: *y* = 0.0405*x*3 - 1.0950*x*2 + 8.4697*x* + 13.0651

oligo length factor:  *y* = -0.00000255*x*4 + 0.00049119*x*3 - 0.02740964*x*2 + 0.55073438*x*

The predicted *RE* is thus:



which reduces to



where *REgen* is the predicted general replacement efficiency, *Gss* is the difference in oligo secondary structure free energy *G*’s between the current and ideal oligos, *Gss* is the difference in hybridization free energy *G*’s between the current and ideal oligos, *a* is the integer number of 5’ terminal phosphorothioate bonds, and *b* is the length of the oligo in integer base pairs.

Alternatively for specific modifications (i.e. mismatches, insertions, or deletions) a more defined predicted RE can be calculated based on the modification type.

For mismatch modifications:



where *REmm* is the predicted general replacement efficiency, *Gss* is the difference in oligo secondary structure free energy *G*’s between the current and ideal oligos, *m* is the number of basepairs of mismatch modification introduced by the oligo, *a* is the integer number of 5’ terminal phosphorothioate bonds, and *b* is the length of the oligo in integer base pairs.

For insertion modifications:



where *REins* is the predicted general replacement efficiency, *Gss* is the difference in oligo secondary structure free energy *G*’s between the current and ideal oligos, *n* is the number of basepairs of insertion modification introduced by the oligo, *a* is the integer number of 5’ terminal phosphorothioate bonds, and *b* is the length of the oligo in integer base pairs.

For insertion modifications:



where *REins* is the predicted general replacement efficiency, *Gss* is the difference in oligo secondary structure free energy *G*’s between the current and ideal oligos, *d* is the number of basepairs of target deletions made by the oligo, *a* is the integer number of 5’ terminal phosphorothioate bonds, and *b* is the length of the oligo in integer base pairs.

The predicted *RE* for each target is then either a) averaged to produce a mean *RE* for the oligo pool used in the analytical solution outlined below or b) individually analyzed in a pooled Monte-Carlos simulation. To predict the number of MAGE cycles required, we take the binomial distribution:

 where  and 

where *N* is the number of MAGE cycles, *K* is the oligo pool complexity or the number of genomic targets (if only one target exist, then *K* = 1), and *M* is the average conversion rate of individual targets and is the ratio of *RE* to *K*,

For a screening capacity of 100 clones, the population enrichment needs to be at 2% (1% x 2 for conservative estimation). Thus to determine the number of cycles, we set need to solve for *N* in the function *P*(*K*) = 0.02.

Once the predicted RE and predicted MAGE cycle # is determine, the *verification design* algorithm is used to generate allele specific multiplex PCR primers for verification of the mutations and to generate sequencing primers for sequence verification of the mutations. Primers are designed to optimize for the melting temperature (Tm) and to maximize resolution of each amplicon which span 100, 150, 200, 250, 300, 400, 500, 600, 700, 850 bps.

Finally, OptMAGE outputs the results of the algorithm for display or into a text file. The output results include:

1. optimized oligo sequence including modifications
2. predicted Gss
3. predicted Ghyb
4. predicted *REgen*, *REmm*, *REins*, and/or *REdel*
5. suggested number of MAGE cycles required
6. asmPCR primer sequences
7. sequencing primer sequences