CASTER 2.0. A tool to design and evaluate saturation mutagenesis libraries for directed evolution

CASTER is a user-friendly computer aid of great utility when applying one of the most efficient methods in directed evolution of enzymes, namely the Combinatorial Active-Site-Saturation Test (CAST), useful for evolving enhanced activity, widening substrate scope, increasing enantio- or diastereoselectivity and controlling regioselectivity of enzymes as catalysts in organic chemistry and biotechnology. CASTER 2.0 is an improved version of the original form. It can be applied in the formation of initial CAST saturation mutagenesis libraries, but also in iterative CASTing steps as an embodiment of Iterative Saturation Mutagenesis (ISM). A recent review of ISM can be found in: M. T. Reetz, *Angew. Chem. Int. Ed.* **2011**, *50*, 138-174.

CASTER is a computer tool programmed using the Visual Basic Editor included in Microsoft Excel.
CASTER is divided in six excel worksheets (A-F):
A. Select Degenerate Codon. Evaluation of the effects of a single codon randomization using a selected DNA degeneracy.
B. Diverse degenerate codons. Evaluation of the effects of the simultaneous randomization of different codons using different DNA degeneracies.
C. AA encoded – Structures. When degeneracy is set in Worksheet A, the selected amino acids are depicted with their structures in worksheet C.
D. Links. Useful links related to mutant library creation.
E. References. Selection of bibliographic references about library creation strategies and directed evolution methods.
F. About. Authorship and description of the modifications added in each release.
The user can only input data in sheets A and B in the light green colored cells.

A. Select Degenerate Codon

Worksheet A is as well divided in six different sections. The user has to enter data in section I: to select the codon DNA degeneracy to use according the IUBMB Degeneracy Code (International Union of Biochemistry and Molecular Biology). The user has to select the degeneracy desired in the three nucleotides of the codon to randomize using the drop down menus as appearing in the image below.

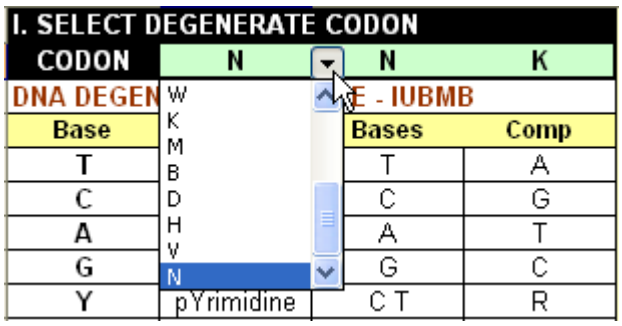

Figure 1. Section I: Select Degenerate Codon. The user has to use the drop down menus to select the desired degeneracy in each of the three positions of the codon to be randomized.

Automatically the consequences of this selection are described in sections II, III, IV and V as well as in worksheet C. AA encoded – structures where the chemical structures of the amino acids encoded are shown, classified according to the chemical properties of their side-chains.

In section II the automatic calculation shows the number of transformed colonies to be screened to cover a 95% of all possible sequence variants without considering any wild-type background. The results shown below refer to 1, 2, 3, 4, 5 and 10 amino acid sites (codons) randomized simultaneously with the same degeneracy set in Section I. It needs to be pointed out that such numbers are only approximations due to various sources of amino acid bias, but they do in fact serve as a general trend and as a useful guide for searching in protein sequence space. For example, these calculations are combinatorial calculations that do not consider the exponential deviation that occurs in every experimental PCR using different primer species, leading to amino acid over-representation of some amplification products.
In CASTER 2.0 the user can set a desired % of coverage and an expected wild-type background according to his experimental experience being automatically considered in the calculations. For that purpose the user has to write down the appropriate numbers in the corresponding light-green cells.

II. SET % COVERAGE	95
and % WT background	0

AA Sites	Codons*	Colonies**
1	32	94
2	1024	3066
3	32768	98163
4	1048576	3141251
5	33554432	100520093
10	1.13E+15	3.4E+15

Figure 2. Section II: Number of codons encoded and colonies to be screened for 95 % coverage considering 0 % of wild-type background. Example shows data corresponding to NNK degeneracy set in section I (default in CASTER).

Section III reflects the amino acid encoded by a given codon degeneracy. The number of codons encoding each amino acid are highlighted in yellow.

III. Amino acids [AA]	
	Codons
Ala [A]	2
Arg [R]	3
Asn [N]	1
Asp [D]	1
Cys [C]	1
Gln [Q]	1
Glu [E]	1
Gly [G]	2
His [H]	1
Ile [I]	1
Leu [L]	3
Lys [K]	1
Met [M]	1
Phe [F]	1
Pro [P]	2
Ser [S]	3
Thr [T]	2
Trp [W]	1
Tyr [Y]	1
Val [V]	2
Stop	1
Codons	32
AA	20

Figure 3. Section III: Number of codons encoding each amino acid using a given degeneracy. In the example, NNK degeneracy, that encodes 32 codons and covers all the amino acid species but with certain bias, Arg, Leu and Ser are encoded by 3 codons included in NNK and Ala, Gly, Pro, Thr and Val are encoded by 2, i.e., those codons would be already over-represented in the oligonucleotides used as primers before starting the in vitro reaction (PCR).

More restricted degenerate codons like NDT or NHT encode 12 amino acids, one codon per amino acid, and they cover amino acids from all chemical types. The number of colonies to be screened is also highly reduced as well as the possibilities of having PCR products representing all encoded variants.

In section 4 the codons covered by a certain codon degeneracy can be seen highlighted in yellow in the global genetic code.

IV. Genetic Code - Codon Distribution											
		SECOND POSITION OF THE CODON									
		T		C		A		G			
FIRST POSITION	T	TTT	Phe [F]	TCT	Ser [S]	TAT	Tyr [Y]	TGT	Cys [C]	T	THIRD POSITION
		TTC	Phe [F]	TCC	Ser [S]	TAC	Tyr [Y]	TGC	Cys [C]	C	
		TTA	Leu [L]	TCA	Ser [S]	TAA	*Ter* [end]	TGA	*Ter* [end]	A	
		TTG	Leu [L]	TCG	Ser [S]	TAG	*Ter* [end]	TGG	Trp [W]	G	
	C	CTT	Leu [L]	CCT	Pro [P]	CAT	His [H]	CGT	Arg [R]	T	
		CTC	Leu [L]	CCC	Pro [P]	CAC	His [H]	CGC	Arg [R]	C	
		CTA	Leu [L]	CCA	Pro [P]	CAA	Gln [Q]	CGA	Arg [R]	A	
		CTG	Leu [L]	CCG	Pro [P]	CAG	Gln [Q]	CGG	Arg [R]	G	
	A	ATT	Ile [I]	ACT	Thr [T]	AAT	Asn [N]	AGT	Ser [S]	T	
		ATC	Ile [I]	ACC	Thr [T]	AAC	Asn [N]	AGC	Ser [S]	C	
		ATA	Ile [I]	ACA	Thr [T]	AAA	Lys [K]	AGA	Arg [R]	A	
		ATG	Met [M]	ACG	Thr [T]	AAG	Lys [K]	AGG	Arg [R]	G	
	G	GTT	Val [V]	GCT	Ala [A]	GAT	Asp [D]	GGT	Gly [G]	T	
		GTC	Val [V]	GCC	Ala [A]	GAC	Asp [D]	GGC	Gly [G]	C	
		GTA	Val [V]	GCA	Ala [A]	GAA	Glu [E]	GGA	Gly [G]	A	
		GTG	Val [V]	GCG	Ala [A]	GAG	Glu [E]	GGG	Gly [G]	G	

Figure 4. Section IV: The cells highlighted in yellow shows the codons and corresponding amino acids encoded by the degeneracy set by the user in section I.

In Section V, the chemical properties of the amino acids encoded by the degeneracy set in section I are displayed in a reference table.

V. Amino acid Distribution - Chemical Classification		
Side Chain	**Amino acids**	%
Acidic [-]	Asp, Glu	6.3
Basic [+]	Arg, His, Lys	15.6
Non-polar aliphatic	Ala, Ile, Leu, Met, Val	28.1
Aromatic	Phe, Trp, Tyr	9.4
Polar	Asn, Cys, Gln, Ser, Thr	25.0
Special Features	Gly, Pro	12.5

Figure 5. Section V: Amino acid distribution of degeneracy set in section I according to the chemical properties of the amino acid side-chains.

Below in section VI the chemical structures of the nitrogenated bases and the canonical Watson & Crick DNA base pairs that can be formed are shown.

Worksheet B. Diverse Degenerate Codons

This worksheet is specially designed for the cases where the researcher wants to randomize more than an amino acid site and wants to use different codon degeneracies at each site. It is important to consider/understand the consequences of the encoding (or not) of the wild-type amino acid in the chosen degeneracy for each site as explained in the worksheet.

The user can also see the % of coverage desired and the wild-type background expected.

I. SIMULTANEOUS RANDOMIZATION OF DIFFERENT POSITIONS USING DIFFERENT DEGENERATE CODONS

Position 1	N	D	T
Position 2	N	H	T
Position 3			
Position 4			
Position 5			

II. SET % COVERAGE 95
and % WT background 0

Positions Randomized	Codons	Colonies
1		
1 + 2	144	430
1 + 2 + 3		
1 + 2 + 3 + 4		
1 + 2 + 3 + 4 + 5		

IMPORTANT: When different positions are randomized simultaneously it is important to select in each position a degeneracy that contains the correspondent wild-type aminoacid, otherwise the evaluation of the possible effects of the individual libraries would not be properly analyzed, as one of the positions is forced to be mutated.

The creation of libraries randomizing more than one codon using "**mutation forced**" **degeneracies** is as well interesting as the diversity generated differs strongly to the wild-type amino acid sequence.

Amino acids [AA] encoded in each position

	1	2	3	4	5
Ala [A]	0	1	0	0	0
Arg [R]	1	0	0	0	0
Asn [N]	1	1	0	0	0
Asp [D]	1	1	0	0	0
Cys [C]	1	0	0	0	0
Gln [Q]	0	0	0	0	0
Glu [E]	0	0	0	0	0
Gly [G]	1	0	0	0	0
His [H]	1	1	0	0	0
Ile [I]	1	1	0	0	0
Leu [L]	1	1	0	0	0
Lys [K]	0	0	0	0	0
Met [M]	0	0	0	0	0
Phe [F]	1	1	0	0	0
Pro [P]	0	1	0	0	0
Ser [S]	1	1	0	0	0
Thr [T]	0	1	0	0	0
Trp [W]	0	0	0	0	0
Tyr [Y]	1	1	0	0	0
Val [V]	1	1	0	0	0
Stop	0	0	0	0	0
Codons	12	12	0	0	0
AA	12	12	0	0	0

Figure 6. Worksheet B. Diverse Codon Degeneracies.

Worksheet C – AA selected – Structures

Displays the chemical structures of the amino acids selected by the degeneracy set in worksheet A section I. The classification is made using a color code that classifies the amino acids in acidic, basic, non-polar aliphatic, aromatic, polar and amino acids with special features (as in section V).

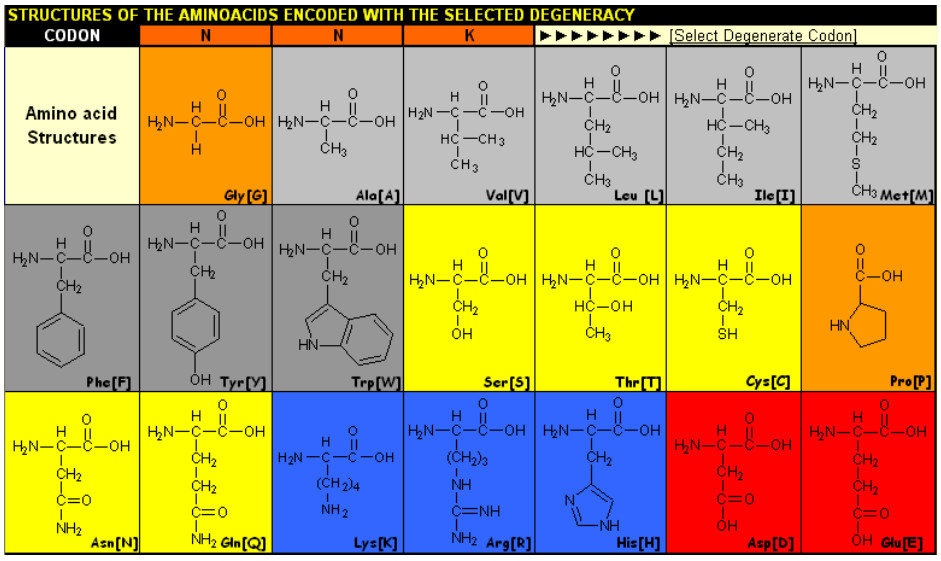

Figure 7.- Worksheet C: AA selected – Structures