International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:4, No:3, 2010

In silico Simulations for DNA Shuffling Experiments

Luciana Montera

***Abstract*—**DNA shuffling is a powerful method used for *in vitro* evolute molecules with specific functions and has application in areas such as, for example, pharmaceutical, medical and agricultural research. The success of such experiments is dependent on a variety of parameters and conditions that, sometimes, can not be properly pre-established. Here, two computational models predicting DNA shuffling results is presented and their use and results are evaluated against an empirical experiment. The *in silico* and *in vitro* results show agreement indicating the importance of these two models and motivating the study and development of new models.

***Keywords*—**Computer simulation, DNA shuffling, *in silico* and *in vitro* comparison.

I. Introduction

DIRECTED molecular evolution is an *in vitro* technique that tries to mimic the natural process of selection and evolution according to Darwin, aiming to produce proteins with improved properties [1]. In such experiments, diversity is created through mutagenesis or recombination and the resulting library is screened for improvements in properties of interest [2]. Several methods for *in vitro* evolutions has been proposed such as error prone PCR [3], staggered extension process (StEP) [4], random priming recombination [5] and DNA Shuffling [6], [7]. The Stemmer method is one of the most used protocols; many works using it can be found in the literature [8]-[15]. The basic protocol involves the following steps:

1. Selection of the parental sequences;
2. Fragmentation of the parental sequences by enzymatic digestion;
3. Reassembly of the fragments by Polymerase Chain Reaction (PCR) cycles;
4. Amplification by PCR of the full-length[1] sequences reassembled.

The parental selection step is particularly important and can determine the success of the method. The parental sequences must share sequence similarities in order to have their fragments reassembled during the PCR cycles. The parental fragmentation is usually done by using the Dnase I enzyme, which produce random cuts through a DNA molecule. Before being reassembled, the resulting fragments are purified (or isolated) by agarose gel electrophoresis so that those with sizes, measured in pairs of bases, within an interval of interest, are selected to give continuity to the process. The selected fragments are then submitted to PCR cycles that include three temperature-controlled reactions: denaturing, annealing and extension that can be described as follows:

1. *Denaturing*: double-stranded DNA molecules are heated to a specific temperature (named denaturation temperature (around 94ºC)), so that the double-stranded DNA molecules are separated into two single-stranded sequences;
2. *Annealing*: the temperature is lowered to a specific temperature (named annealing temperature) such that the single-stranded fragments sharing complementary bases anneal each other;
3. *Extension*: the temperature is raised to the optimum temperature for the polymerase enzyme used in this reaction to extend the annealed fragments to reproduce double-stranded DNA fragments.

After a number n of PCR cycles, recombinant sequences are formed, which can be seen as a 'mixture' from parental sequences. The recombinants that have the same parental length are amplified. Fig. 1 shows a scheme of how DNA Shuffling works to produce recombinant sequences.

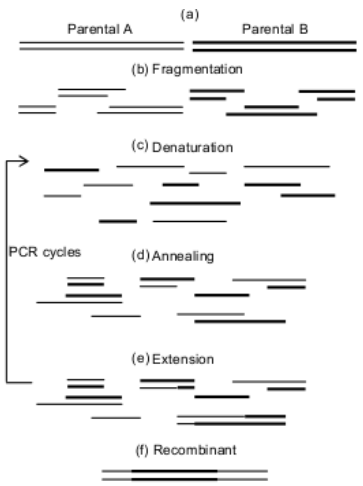


Fig. 1 Shuffling scheme. Two parental sequences are fragmented by enzymatic digestion. The resulting fragments are submitted to PCR cycles to be reassembled. As the annealed fragments are extended, a new sequence (recombinant or not) is formed. The recombinant sequence showed in (f) can be considered as a result of the occurrence of two crossovers between the parental A and B.

Looking to a recombinant, the sequence is composed of fragments originally from distinct parental**;** it is common to

Luciana Montera is with Faculty of Computing of Federal University of Mato Grosso do Sul, Brazil (e-mail: montera@facom.ufms.br).

[1] A full-length is a reassembled sequence which has, approximately, the same length as the parental sequences.

say that crossover(s) occurred between parental to result on these recombinant. The efficiency of a directed evolution method can be measured by the average number of recombination events that occur in the reassembled sequences [16]; Although directed evolution experiments have largely been guided by empirical information and experience without a quantitative understanding of the recombination step and subsequent optimization of the experimental setup [17], some computation models have been proposed and used as tools to support and, in many cases, direct *in vitro* experiments. In this paper the models proposed in [18] and [19] are used to evaluate a published DNA shuffling experiment [20].

Following this introductory section, Section 2 describes the basics concepts involved in the model proposed by Moore and co-workers [18] and Patrick and co-workers [19] to predict results for DNA shuffling experiments. Section 3 compares the *in vitro* with the *in silico* results. Finally, Section 4 presents conclusions and highlights the scope for future work.

II. Computer Models to Predict DNA Shuffling Results

This section presents two models for *in silico* simulations of DNA shuffling. The use of these models can help researchers conducting such experiments. The first model**,** proposed by Moore in 2001 [18]**,** was implemented as a Fortran program named eShuffle. The second model**,** proposed by Patrick in 2003 [19], was also implemented as a Fortran program named DRIVeR.

A. eShuffle

The software eShuffle is able to run in three different models, each predicting a specific metric:

1. Forward Crossover: predicts the percentage of library which has from one, two to ten crossover per sequence, as well as the average number of crossovers per sequence in the library. To predict this results, the parental sequences are used in the same direction as inputted, and it is assumed that the sequences are stored in 5'→3' direction;

2. Reverse Crossover: makes the same predictions as the Forward Crossover mode, however the parental sequences are used in the opposite direction as inputted, i.e., the complement of the parental sequences (the sequences in the direction 3'→5') are calculated by the software before being used.

3. Crossover Profile: checks, for each position on the sequence, its potential to serve as a crossover point among the parental. For each point, a value between zero and one is returned to represent this probability.

The model uses thermodynamic concepts and complete parental nucleotide sequences as a basis to model the annealing and fragments reassembly events that occur during the DNA shuffling process. Following are described how the annealing and the reassembly events are modeled.

Annealing

Given a set of DNA fragments which are competing with each other to anneal, the more stable DNA pair, i.e., that with the higher free energy**,** is more likely to occur. As proposed in [21], "the stability of a DNA duplex appears to depend primarily on the identity of the nearest neighbor bases", indicates that not only the base pair itself contributes to the stability, but so do the nearest neighbor bases in an annealed region. The four bases that compose a DNA molecule allow sixteen different pairwise nearest neighbor possibilities that can be used to predict the stability of a duplex. It was demonstrated by [21], that "the DNA duplex structures thermodynamically can be considered the sum of their nearest neighbor pairwise interaction". Table I below lists the nearest neighbor interaction values for the enthalpy (ΔH) and entropy (ΔS) variations used by eShuffle and described in [22], calculated for a duplex at 1M NaCl, 37ºC (or 310K) and pH = 7. Values for mismatched nearest neighbor pairs are also defined, but they are not presented here (see [17]).

TABLE I
Values to Some Nearest Neighbor Pair (NN Pair)

Nearest Neighbor Pair	ΔH	ΔS
AA/TT, TT/AA	-7.9	-22.2
AT/TA	-7.2	-20.4
TA/AT	-7.2	-21.3
CA/GT, TG/AC	-8.5	-22.7
GT/CA, AC/TG	-8.4	-22.4
CT/GA, AG/TC	-7.8	-21.0
GA/CT, TC/AG	-8.2	-22.2
CG/GC	-10.6	-27.2
GC/CG	-9.8	-24.4
GG/CC, CC/GG	-8.0	-19.9

Knowing the ΔH and ΔS values to each nearest neighbor pair, it is possible to estimate the free energy (ΔG) resulting from two annealed fragments, which can be approximated by the sum of the free energy associated with each nearest neighbor pair in the annealing region. For a DNA sequence X = x_1, x_2, ... x_n, ΔH, ΔS and ΔG are calculated as showed in (1), (2) and (3), respectively:

$$\Delta H_{total} = \sum_{i=1}^{n-1} \Delta H [x_i, x_{i+1}] \quad (1)$$

$$\Delta S_{total} = \sum_{i=1}^{n-1} \Delta S [x_i, x_{i+1}] \quad (2)$$

$$\Delta G = \Delta H_{total} - \Delta S_{total} \quad (3)$$

Given a set of DNA fragments competing to anneal to a specific DNA fragment F, originally from the parental m and named template, there are many possibilities for the annealing between the template and any fragment from the set. The

annealing between fragments A and the template F is due to an overlap region (see Fig. 1) of size v.

As we have a set of fragments competing for the annealing with a template F, the selectivity of a fragment A at a specific temperature (T) depends on the concentration of all other fragments in the mixture available to anneal with the same template F, as expressed by the (4).

$$s_{mv}(T) = \frac{X_{AF_{mv}}}{\sum_{m'v'} x_{F_{m'v'}}} \quad (4)$$

where $X_{AF_{MV}}$ is the concentration of fragment A that anneals to F (from parental m) with overlap size v and $X_{F_{m'v'}}$ is the concentration of all other fragments that can anneal to F by any overlap size.

As the annealing selectivity is dependent on the temperature, and considering that after the denaturation during the cycles of PCR without primer (see Introduction) the temperature is lowed to the annealing temperature chosen, the annealing events must be considered to occur in the entire range of values from the denaturation temperature to the annealing temperature, instead of a fixed temperature. So, the proposed model calculates the duplex contribution in the entire interval varying from 94 ºC to 55 ºC.

Moore and co-workers evaluated the length effect of the overlap on the selectivity of a fragment. In spite of some annealing involving short overlap regions, the results show that there is a strong preference toward annealing involving a larger overlap region.

Regarding selectivity, given a set of DNA fragments competing to anneal with a specific DNA fragment F named template, in the proposed model, the DNA fragment from the set annealing with template F that results in higher free energy will be chosen.

Assembly

The reassembly procedure addresses the following question: what is the probability of a reassembled sequence with B nucleotides having x crossover? To answer this question, the model assumes that a full-length reassembled sequence, i.e., a sequence reassembled that has the same parental length, is a result of successive annealing events. The assumptions considered by the model are described as follows.

Let one annealing event join a template fragment F_1 to a fragment F_2. Even though the resulting fragment is formed by $F_1 + F_2$, only fragment F_2 is considered as a template to the following annealing event. Lets still consider the size of this reassembled fragment ($F_1 + F_2$) to be i – 1, so the next annealed fragment will be added at position i. The probability that reassembly from position i to the end (B) of the full-length DNA sequence will be formed with exactly x crossovers, given that the last added fragment (ending at position i – 1) is originated from parental k**,** is expressed by P_{ik}^{x} .

Notice that when a fragment of length i – 1 anneals to other of length L by a overlap region of size v**,** the resulting fragment will have size (i – 1) + (L – v), and this point will be considered the new i – 1 point of the next annealing event. Fig. 2, adapted from supplementary material supplied by [18], illustrates the reassembly procedure.

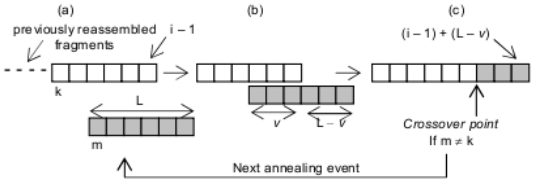


Fig. 2 Reassemble procedure. Adapted from [18]

Regarding Fig. 2(a), if the parental m is different from the parental k, a crossover occurred during the reassembly. So, if one crossover has occurred at position i, then one must calculate the probability that another x – 1 crossovers occur during the reassembly processes to estimate P_{ik}^{x} . For given parental sequences, eShuffle estimates the mean number of crossovers in the resulting sequences as well as the ratio of sequences that have 0, 1, 2,..., 10 crossovers. It is important to mention that some crossovers do not result in diversity. These crossovers are named silent crossovers where the extension of the annealed fragments results in a fragment identical to a specific region from both parental sequences.

B.DRIVeR

The software DRIVeR[2] (Diversity Resulting from *In vitro* Recombination) [19] is a program that implements a statistical model that estimates the expected number of distinct sequences in a library created by random crossovers between two parental highly homologous sequences (i.e.**,** differing from each other in only a few (e.g.**,** 20) base pair positions). DRIVeR also returns the probability of each distinct shuffled sequence (or variant) occurring in the library. We previous described and used the DRIVeR for assessing the adequacy of three different pairs of sequences as parental sequences under different software settings [1].

In contrast to eShuffle, DRIVeR does not consider all sequence information to estimate the average number of crossovers in the sequences resulting from a DNA shuffling experiment. Only the distance between consecutive differences (or mutations) existent between the parental sequences are considered. It is assumed that the number x of crossovers that can occur between two consecutive mutations follows the Poisson distribution:

$$P(x) = \frac{e^{\lambda^{true}}(\lambda^{true})^{x}}{x!}, x = 0,1,2,3,...$$

where λ^{true} is the real number of crossovers that is observed in a sample of the resulting sequences. The real number of crossovers excludes the number of silent crossovers.

[2] available for download at www.bio.cam.ac.uk/~blackburn/stats.html

DRIVeR also estimates the probability of occurrence of each possible distinct variant resulting from the shuffling of the parental. Having the parental sequences distinct from each other by m base pairs (mutations), 2^m distinct variants can appear in the resulting library. Each variant is represented by a binary sequence, where the digit 0 indicates that an even number of crossovers occur between two consecutive mutations and the digit 1 indicates that an odd number of crossover occur between two consecutive mutations[3]. Fig. 3 shows two parental sequences A and B where the mutations between them are represented by black and white circles, respectively, for a possible variant resulting from a shuffling experiment. The binary representation of the variant sequence is also shown.

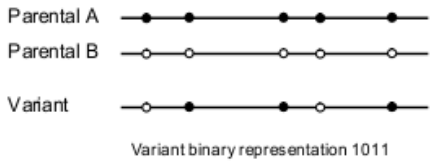


Fig. 3 A variant resulting from a DNA shuffling experiment between parental A and B and its binary representation

III. Comparing *In Silico* and *In Vitro* Results

To properly evaluate the predicted results from eShuffle, information about the parental sequences (composition, mutations between them), experimental conditions (fragment size, annealing temperature) as well as information about the resulting sequences (crossover number and crossover distribution per sequence) are needed. Unfortunately, few data are available on the composition of the resulting sequences from a DNA shuffling experiments (shuffled library) from which the efficiency of the method can be assessed [23].

Raillard [20] reports a DNA shuffling experiment of two highly homologous triazine hydrolases genes. The amino acid composition of some shuffled sequences are presented, allowing this experiment to be used to evaluate the eShuffle and DRIVeR predictions.

Raillard and co-workers used two highly homologous triazine hydrolases genes as parental in a shuffling experiment and the resulting sequences were explored to verify the substrate specificities in order to evaluate their improvement in relation to their parental. As reported by the authors "the shuffled library contained enzymes with up to 150-fold greater transformation rates than either parent", so that, it can be said the shuffling was successful.

The genes atzA and triA were used**;** their DNA sequences can be found at GenBank by the accession number P72156 and AAG41202, respectively. Both, atzA and triA sequences codify to a protein with 475 amino acids which differs from each other by 9 amino acids located at positions 84, 92, 125, 217, 219, 253, 255, 328 and 331. Looking to the nucleotide sequence, with a length of 1425, the differences are located at positions 250, 274, 375, 650, 655, 757, 763, 982 and 991. The 25 active variants reported, as well as the parental sequences used, are shown in Table II. Each sequence is represented by only the nine variants amino acids instead of the complete amino acids sequence and its origin is highlighted by the cell color – gray for amino acids from atzA and white for amino acids from triA. In addition, the DRIVeR binary representation and the number of crossovers for each sequence are also represented in the table.

Fig. 4 shows schematically the alignment between the nucleotide sequence of the parental atzA and triA. To simplify, only the mutations are represented (white square for sequence atzA and black circle for sequence triA). This figure also shows the distance (in number of base pairs) between each consecutive and distinct base[4].

An important consideration should be made in relation to the fragment size used during the PCR cycles in a shuffling experiment. Consider the distances that separate the consecutive mutations shown in Fig. 4. The minimum distance between two consecutive mutations is 4 and the maximum is 208. If only fragments larger than the maximum were used in a shuffling experiment, the likelihood that consecutive mutations remain together at the same fragment increase. Theoretically, the optimum fragment size should not be larger than the minimum distance between two consecutive mutations observed between the parental sequences. However, while "preparing shorter fragments increases the recombination frequency, small fragments will be inefficiently reassembled" [16].

TABLE II
VARIANTS RESULTING FROM DNA SHUFFLING BETWEEN ATZA AND TRIA

	Amino Acid Position									Variant Representation	Number of Cros.
	84	92	125	217	219	253	255	328	331		
atzA	F	V	E	T	T	I	G	N	S		
triA	L	L	D	I	P	L	W	D	C		
1	F	V	E	T	T	L	W	N	S	00001010	2
2	F	V	E	T	T	L	G	N	S	00001100	2
3	F	V	D	T	T	L	W	N	S	01101010	4
4	L	V	E	T	T	L	W	N	S	10001010	3
5	L	V	E	T	T	L	W	N	S	10001010	3
6	F	L	E	T	T	I	G	D	C	11000010	3
7	F	V	E	T	T	I	G	D	C	00000010	1
8	L	L	E	T	T	L	G	D	C	01001110	4
9	L	L	E	T	T	L	W	D	S	01001001	3
10	L	L	E	T	T	L	W	D	C	01001000	2
11	L	V	E	T	T	L	W	D	C	10001000	2
12	L	A	E	T	T	L	W	D	C	01001000	2
13	L	V	E	T	T	I	G	D	S	10000011	3
14	L	V	E	T	T	I	G	D	C	10000010	2
15	L	V	E	T	T	I	G	D	C	10000010	2
16	L	L	D	T	T	L	W	D	S	00101001	3
17	L	L	E	T	T	I	G	D	C	01000010	2
18	L	L	D	T	T	I	W	D	C	00100100	2
19	F	V	D	T	T	I	G	N	S	01100000	2
20	L	V	E	T	T	I	G	N	S	00100000	1
21	F	V	E	T	T	L	W	D	C	00001000	1
22	F	V	E	T	T	I	G	D	C	00000010	1
23	L	V	E	T	T	I	G	D	S	10000011	3
24	F	L	D	T	T	I	G	D	C	10100010	3
25	L	F	D	I	T	L	W	D	C	00011000	2
											Average 2.32

[3] Observe that, the occurrence of any odd number of crossovers (1, 3, 5, ...) between two consecutive mutations has the same effect on the resulting sequence. The same occurs for any even number of crossovers.

[4] The reader is encouraged to download these sequences at GenBank and construct the alignment between them to properly see how far the differences between the parental sequences are located.

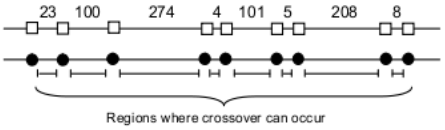


Fig. 4 Distinct bases between the parental sequences. A crossover can occurs at any region between two consecutive distinct bases

To execute eShuffle, the following parameters are required:

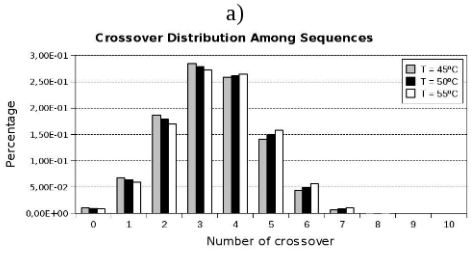
F : fragment size (number of nucleotides)
T : annealing temperature (ºC)
L : parental length
N : number of parental sequences
File_name: file name where the parental sequences are stored

Considering any parental sequences A and B**,** the file storing these sequences must adhere to the following pattern: each line must contain exactly and alternately 60 nucleotides each from one parental, i.e., the first line must contain the first 60 nucleotides from parental A, the second line must contain the first 60 nucleotide from parental A, the third line must contain the 61 to 120 nucleotides from parental A, and so on.

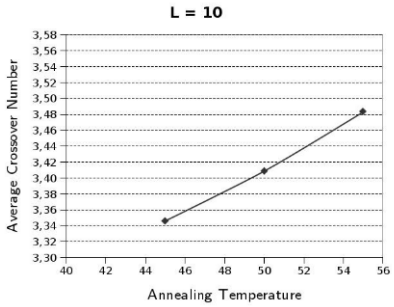
After preparing the input file for the parental sequences atzA and triA, the three program modes were executed. As expected, the Forward Crossover Distribution and the Reverse Crossover Distribution produced similar results**;** therefore only the results from the Forward Crossover Distribution are reported here. Different executions were made considering distinct parameters values.

Raillard reports that the shuffling experiment was conducted as described by Stemmer [7], where fragments from just 10 to 50 base pairs in length were used and the annealing temperature ranged from 50 ºC to 55 ºC. The simulations using the eShuffle program were done accordingly to match the conditions reported. Additionally, simulations considering annealing temperature equal to 45 ºC were done**.** Graphs 1 through 6 show the eShuffle results of the mode Forward Crossover where (a) represent the percentage of recombinants containing 0 to 10 crossovers and (b) the average crossover number considering annealing temperatures of 45, 50 and 55 ºC.

a)

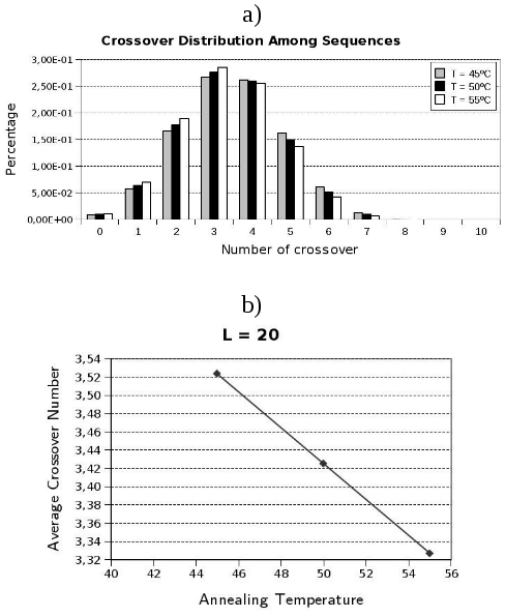


b)

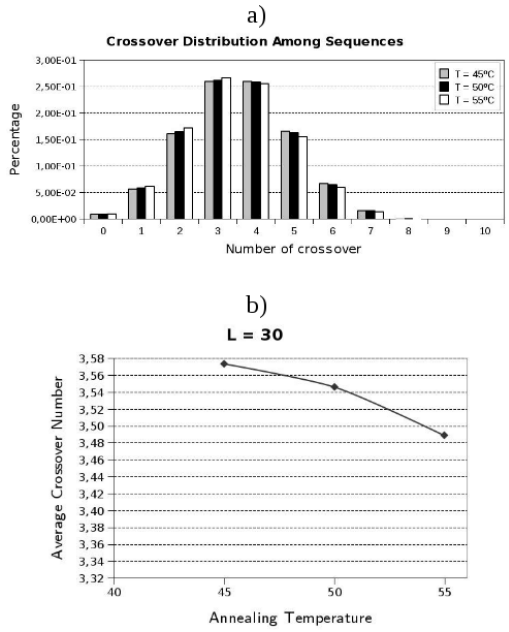


Graphic 1. Simulations results for atzA and triaZ fragment size equal to 10

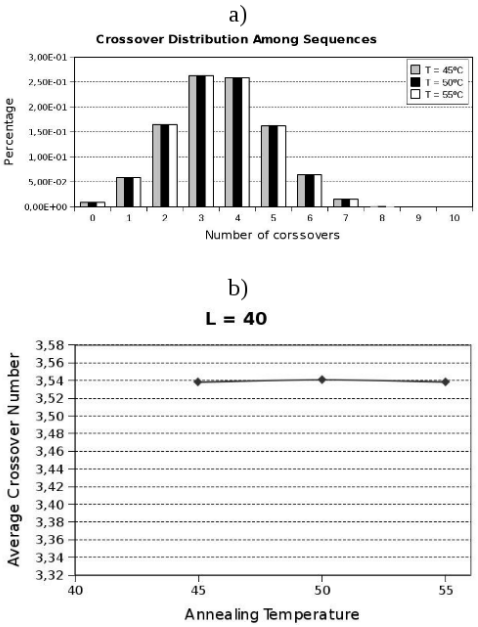
Regarding Graph 1(a), where the fragment length under consideration is 10, it can be concluded that the eShuffle predicts that, independent of the annealing temperature (45 ºC, 50 ºC or 55 ºC), the huge portion of the reassembled sequence is the result of three or four crossovers between the parental. The average crossover number for this simulation is equal to 3.4. Similar conclusions can be inferred from the results showed by graphs 2(a), 3(a), 4(a) and 5(a). Regarding how the annealing temperature and the fragment length influence the average crossover number, a pattern could not be detected. It is known that lower temperatures favor the annealing between fragments sharing few complementary bases, i.e., small overlap, and, in addition, favor mismatched annealing, could explain the behavior shown by graphs 2 and 3. However, in addition to temperature, fragments composition determines the annealing occurrence, which can explain the behavior observed at graphics 1(b) to 5(b), since different sizes produce fragments with different extremity composition.



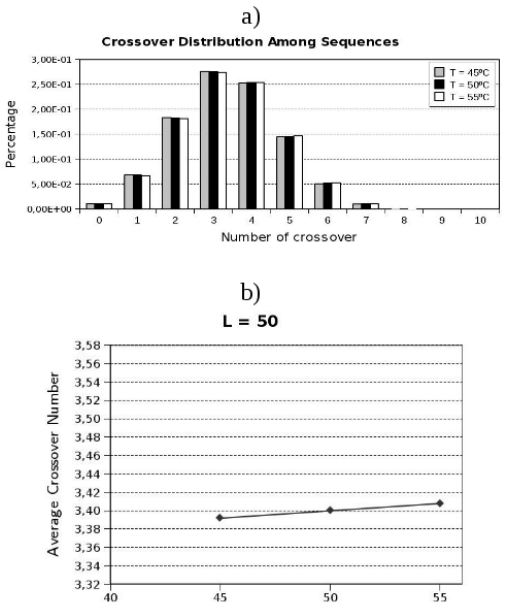
Graphic 2. Simulations results for atzA and triaZ fragment size equal to 20



Graphic 3. Simulations results for atzA and triaZ fragment size equal to 30



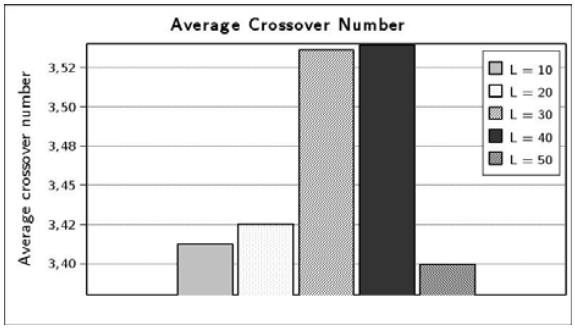
Graphic 4. Simulations results for atzA and triaZ fragment size equal to 40



Graphic 5. Simulations results for atzA and triaZ fragment size equal to 50

Graph 6 shows the relationship between the fragment length and the average crossover number. The higher crossover number is achieved when fragments with 30 or 40 bases in length are used. In fact, this is an expected result, once smaller fragments are inefficiently reassembled [16] and longer fragments remain consecutive mutation together

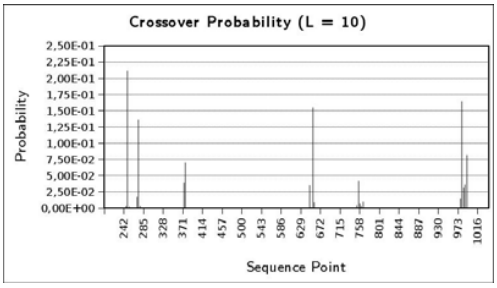
and decreasing thereby the possibility of crossovers occurrence.



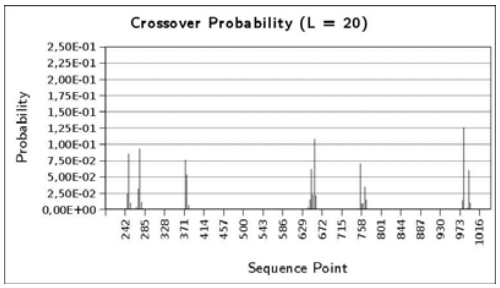
Graphic 6. Average crossover number predicted by eShuffle for the DNA shuffling between atzA and triaZ considering different fragment sizes

Analyzing the recombinant work reported at Raillard (see Table II), it can be concluded that the mean crossover number achieved by the experiment was 2.3. With eShuffle simulations considering fragments size 10, 20, ..., 50, the average predicted is 3.5. It should be noticed, however, that while the *in vitro* experiment uses fragments ranging from around 10 to 50 all together within the same reaction, the *in silico* experiment was done separately to each specific fragment size and the mean calculated.

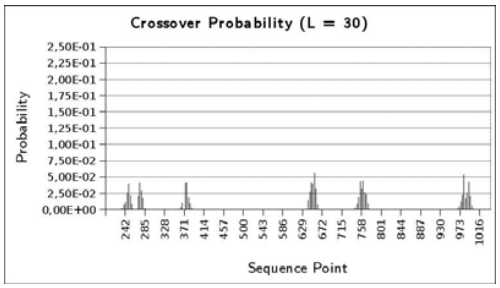
The probability at each position in the parental sequences of being a crossover point is also estimated (eShuffle Crossover Profile mode). As can be inferred by Fig. 4, a crossover point is any point between to consecutive mutations where a crossover can occur. Graphs 7 to 11 show the estimated probabilities considering the use of fragments varying from 10 to 50 in length.



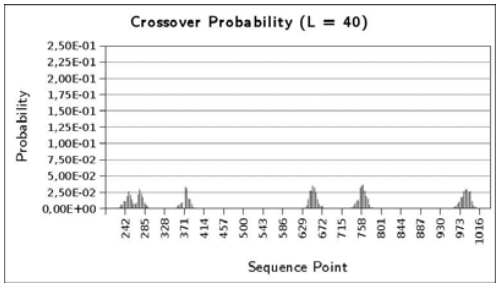
Graphic 7. Probability of each point along the parental sequences being a crossover point considering fragments of size 10 pb



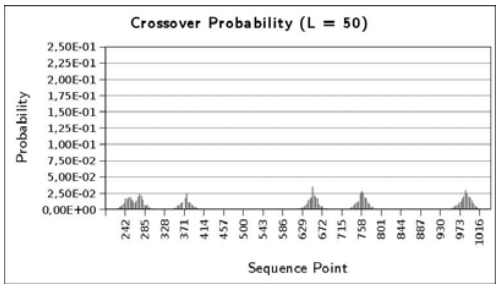
Graphic 8. Probability of each point along the parental sequences being a crossover point considering fragments of size 20 pb



Graphic 9. Probability of each point along the parental sequences being a crossover point considering fragments of size 30 pb



Graphic 10. Probability of each point along the parental sequences being a crossover point considering fragments of size 40 pb



Graphic 11. Probability of each point along the parental sequences being a crossover point considering fragments of size 50 pb

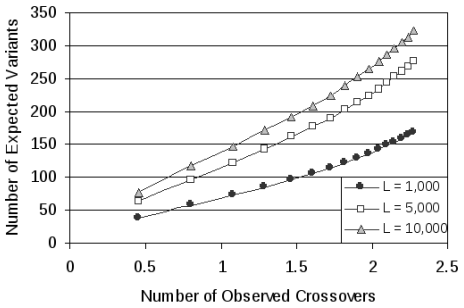
To execute DRIVeR, the following parameters are required:

I. N: parental sequence length (in number of nucleotides);
II. λ^{true}: mean number of real crossover per sequence;
III. L: library size;
IV. M: number of mutation pairs between parental;

V. m_i, $1 < i \leq M$, representing the positions of mutation points.

DRIVeR was executed for three distinct library sizes: 1,000, 5,000 and 10,000. In each simulation, values for λ^{true} varying from 1 to 16 were used. For a given λ^{true}, DRIVeR estimated the corresponding λ^{obs} (number of observed crossovers, i.e., total number of crossovers excluding the silent crossovers). Graph 12 summarizes the DRIVeR results.

Consider the execution where L = 1,000, and λ^{obs} = 2.236738 (which means λ^{true} = 15). Looking at the file where the probability of each distinct variant is stored, it was found that all of the 25 variants reported by Raillard (Table II) were within the 83 more probable variants predicted by the software. The probability of each variant predicted by the software and its order within the 83 more probable variants are shown in Table III.



Graphic 12. DRIVeR results

IV. Conclusion

Directed molecular evolution by DNA shuffling is a powerful technique, first proposed by Stemmer in 1994 and continuously developed, to generate (*in vitro*) new and improved molecules with the more varied purpose (alter enzyme substrate specificity, improve enzymes stability, drug resistance, etc.). The experiment is labor and time consuming so that the use of a computational model to first simulate (*in silico*) the experiment and then evaluate the results is an important pre-processing step that can guide conducting successful experiments.

The theoretical results presented here using eShuffle and DRIVeR software show some agreement with empirical results which suggests their use by a specialist can help achieve better results when conducting such experiments.

Acknowledgment

The author would like to thanks the founding agency CAPES by supporting this research and also Dr. Moore by gently making eShuffle available for this research.

Table III
Ranking The Most Probable Variants

Variant	Binary Representation (DRIVeR)	Probability	Order
1	00001010	2.08E-02	9
2	00001100	1.12E-03	55
3	01101010	1.62E-02	16
4	10001010	4.97E-03	25
5	10001010	4.97E-03	25
6	11000010	4.95E-03	26
7	00000010	2.63E-02	3
8	01001110	8.65E-04	75
9	01001001	1.41E-03	47
10	01001000	1.67E-02	13
11	10001000	5.07E-03	21
13	10000011	5.33E-04	83
14	10000010	6.30E-03	19
15	10000010	6.30E-03	19
16	00101001	1.78E-03	39
17	01000010	2.07E-02	10
18	00100100	1.41E-03	46
19	01100000	2.10E-02	8
20	00100000	2.67E-02	2
21	00001000	2.12E-02	5
22	00000010	2.63E-02	3
23	10000011	5.33E-04	83
24	10100010	6.26E-03	20
25	00011000	8.99E-04	65

References

[1] Montera, L.; Nicoletti, M.C.; Silva, F.H. "Computer Assisted Parental Sequences Analysis as a Previous Step to DNA Shuffling Process". IEEE Congress on Evolutionary Computation, 8079–8086, 2006.

[2] Voigt, C.A., Mayo, S.L., Arnold, F.H. and Wang, ZG. "Computationally Focusing the Directed Evolution of Proteins". Journal of Cellular Biochemistry Supplement 37, 58–63, 2001.

[3] Cadwell, RC.; Joyce, G;F. "Randomization of genes by PCR mutagenesis". PCR Method Appl., 2, 28 33, 1992.

[4] Zhao, H., Giver, L., Shao, Z., Affholter, A., Arnold, F. H. "Molecular evolution by staggered extension process (StEP) *in vitro* recombination". Nature Biotechnol. 16, pp. 258–261, 1998.

[5] Shao, Z., Zhao, H., Giver, L., Arnold, F.H. "Random-priming *in vitro* recombination: an effective tool for directed evolution". Nucleic Acids Research 26, 681–683, 1998.

[6] Stemmer,W.P.C. "Rapid evolution of a protein *in vitro* by DNA shuffling". Nature 370, 389–391, 1994.

[7] Stemmer,W.P.C. "DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution". Proc. Natl. Acad. Sci. USA 91, 10747–10751, 1994.

[8] Patnaik, R.; Louie, S.; Gavrilovic, V.; Perry, K.; Stemmer, W.P.C.; Ryan, C.M.; Cardayré, S. "Genome shuffling of Lactobacillus for improve acid tolerance". Nature Biotechnology 20, 707–712, 2002.

[9] Christians, F.C.; Scapozza, L.; Crameri, A.; Folkers, G.; Stemmer, W.P.C. "Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling". Nature Biotechnology 17, 259–264, 1999.

[10] Chang, C.C.; Chen, T.T.; Cox, B.W.; Dawes, G.N.; Stemmer, W.P.C.; Punnonen, J.; Patten, P.A. "Evolution of a cytokine using DNA family shuffling". Nature Biotchnology 17(8), 793–797, 1999.

[11] Ness, J.E.; Welch, M.; Giver, L.; Bueno, M.; Cherry, J.R.; Borchert, T.V.; Stemmer, W.P.C.; Minshull, J. "DNA shuffling of DNA subgenomic sequences of subtilisin". Nature Biotechnology 17, 893–896, 1999.

[12] Yanga, L.; Jianga, J.; Drouinb, L.M.; Agbandje-Mckennab, M.; Chena, C.; Qiaoa, C.; Pua, D.; Huc, X.; Wangc, D.; Lia, J.; Xiaoa, X. "A myocardium tropic adeno-associated virus (AAV) evolved by DNA shuffling and in vivo selection". PNAS 106(10), 3946–3951, 2009.

[13] Ryu, K.; Hwang, S.Y.; Kim, K.H.; Kang, J.H.; Lee, E.K. "Functionality improvement of fungal lignin peroxidase by DNA shuffling for 2,4-dichlorophenol degradability and H_2O_2 stability". Journal of Biotechnology 133(1), 110–115, 2008.

[14] Koerber, J.T.; , Jang, J.H; Schaffer, D.V. "DNA Shuffling of Adeno-associated Virus Yields Functionally Diverse Viral Progeny". Molecular Therapy 16(10), 1703–1709, 2008.
[15] Maheshri, N.; Koerber, J.T.; Kaspar, B.K.; Schaffer, D.V. "Directed evolution of adeno- associated virus yields enhanced gene delivery vectors". Nat. Biotech. 24, 198 –204, 2006.
[16] Volkov, A.A.; Arnold, F.H. "Methods for *in vitro* DNA Recombination and Random Chimeragenesis". Methods in Enzymology 328, 447–456, 2000.
[17] Moore, G.L.; Maranas, C.D.; Gutshall, K.R.; Brenchley, J.E. "Modeling and optimization of DNA recombination". Computers & Chemical Engineering 24, 693–699, 2000.
[18] Moore, G.L.; Maranas, C.D.; Lutz, S.; Benkovic, S.L. "Predicting crossover generation in DNA shuffling". PNAS 98, 3226–3231, 2001.
[19] Patrick, W.M.; Firth, A.E.; Blackburn, J.M. "User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries". Protein Engineering 16(6), 451–457, 2003.
[20] Raillard, S.; Krebber, A.; Chen, Y.; Ness, J.E.; Bermudez, E.; Trinidad, R.; Fullem, R.; Davis, C.; Welch, M.; Seffernick, J.; Wackett, L.P.; Stemmer, W.P.C.; Minshull, J. "Novel enzyme activities and functional plasticity revealed by recombining highly homologous enzymes". Chemistry & Biology 8, 891–898, 2001.
[21] Breslauer,K.J.; Frank,R.; Blocker, H.; Marky, L.A. "Predicting DNA duplex stability from the base sequence". Proc Natl Acad Sci USA 83(11), 3746–3750, 1986.
[22] Allawi, H.T. & SantaLucia, J.Jr. "Thermodynamics and NMR of Internal G-T Mismatches in DNA", Biochemistry 36, 10581-10594, 1997.
[23] Joern, J.M.; Meinhold, P.; Arnold, F.H. "Analysis of shuffled gene libraries". Journal of Molecular Biology, 316 (3), 643–56, 2002.