Deliverable D2.3

Funding Scheme: THEME [ICT-2007.8.0] [FET Open]

Paving the Way for Future Emerging DNA-based Technologies: Computer-Aided Design and Manufacturing of DNA libraries

Grant Agreement number: 265505

Project acronym: CADMAD

Deliverable number: D2.3

Deliverable name: Report on the features of algorithm for library risk assessment

Contractual Date[1] of Delivery to the CEC: M24
Actual Date of Delivery to the CEC: M25
Author(s)[2]: Tuval ben-Yehezkel
Participant(s)[3]: WEIZMANN
Work Package: WP3
Security[4]: Pub
Nature[5]: R
Version[6]: 0.0
Total number of pages:

1 As specified in Annex I

2 i.e. name of the person(s) responsible for the preparation of the document

3 Short name of partner(s) responsible for the deliverable

4 The Technical Annex of the project provides a list of deliverables to be submitted, with the following classification level:
Pub - Public document; No restrictions on access; may be given freely to any interested party or published openly on the web, provided the author and source are mentioned and the content is not altered.
Rest - Restricted circulation list (including Commission Project Officer). This circulation list will be designated in agreement with the source project. May not be given to persons or bodies not listed.
Int - Internal circulation within project (and Commission Project Officer). The deliverable cannot be disclosed to any third party outside the project.

5 **R (Report):** the deliverables consists in a document reporting the results of interest.
P (Prototype): the deliverable is actually consisting in a physical prototype**,** whose location and functionalities are described in the submitted document (however, the actual deliverable must be available for inspection and/or audit in the indicated place)
D (Demonstrator): the deliverable is a software program, a device or a physical set-up aimed to demonstrate a concept and described in the submitted document (however, the actual deliverable must be available for inspection and/or audit in the indicated place)
O (**Other):** the deliverable described in the submitted document can not be classified as one of the above (e.g. specification, tools, tests, etc.)

6 Two digits separated by a dot:
The first digit is 0 for draft, 1 for project approved document, 2 or more for further revisions (e.g. in case of non acceptance by the Commission) requiring explicit approval by the project itself;
The second digit is a number indicating minor changes to the document not requiring an explicit approval by the project.

Abstract

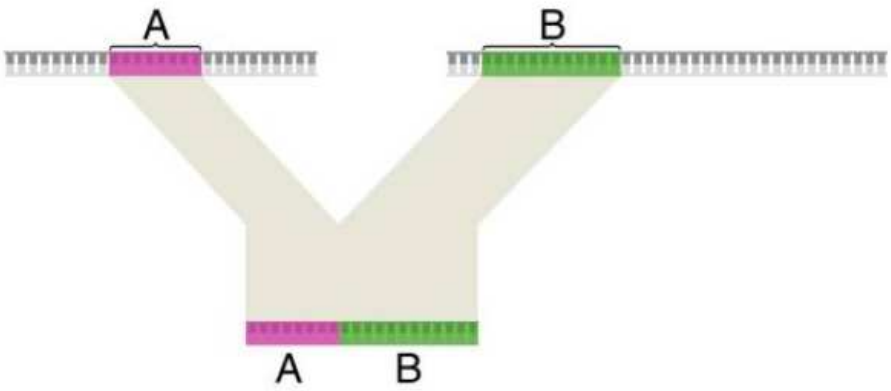
Most methods of DNA sequences concatenations use PCR to amplify the concatenated products. We consider this PCR step as the most likely point of failure and therefore concentrate our efforts on minimizing its use, predicting its outcome and preventing possible errors.
PCR failures can be divided into two major causes, primer dimers and mispriming. We've addressed these two separately on both the experimental and the computational fronts. First, we've paired together 9216 pairs of primers that were computationally enriched for higher chances of diming and examined the resulting dimers. Second, we designed a combinatorial library based on a simple PCR test system of a short single stranded template and two primers. Using that combinatorial library, we produced almost 80,000 different PCR triplets of template and two primers in vitro and trained software classifiers according to the resulting Ct (cycle threshold) values.

Keywords[7]**:**

Machine learning, PCR, Fluidigm, Mispriming, Primer dimers, BioHEL

Our basic scar-less DNA editing step, regardless of the assembly method, requires heavy use of PCR. The process usually starts with a PCR step for amplifying the parent sequences and appending overlapping sequences for the assembly:

Figure A: Concatenating two DNA sub-sequences



[7] Keywords that would serve as search label for information retrieval

The overlapping fragments are then assembled together using either the Y operation or Gibson assembly:

Figure B: assembly using Y operation

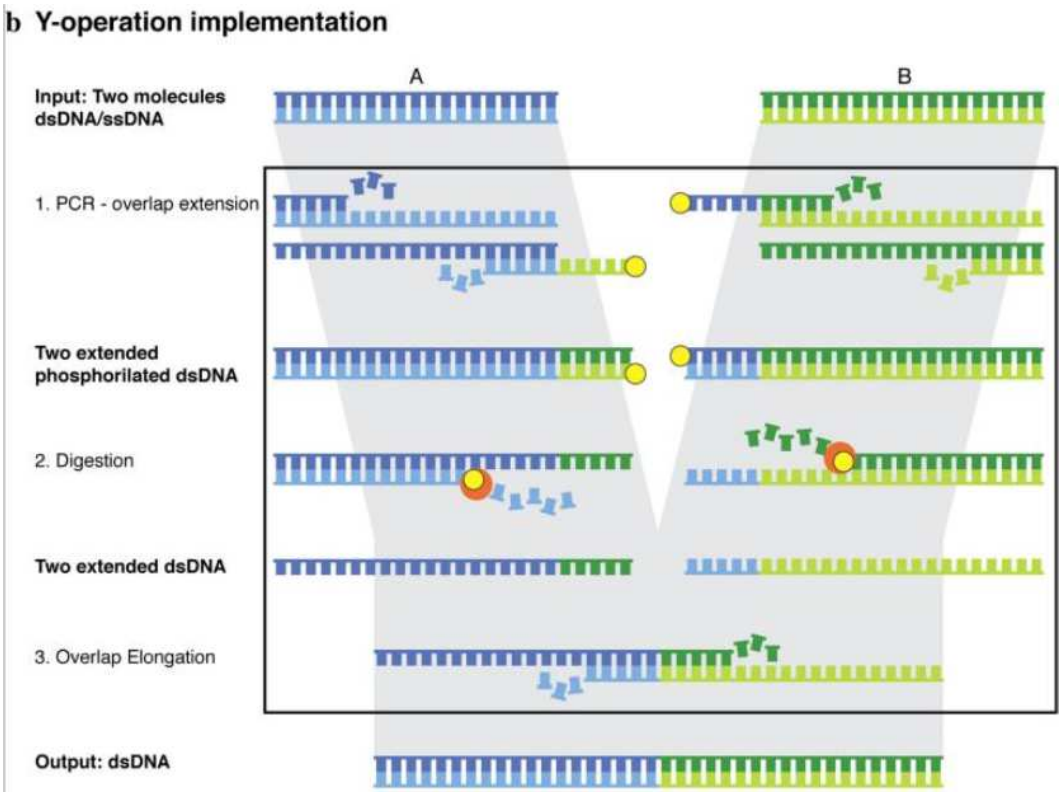
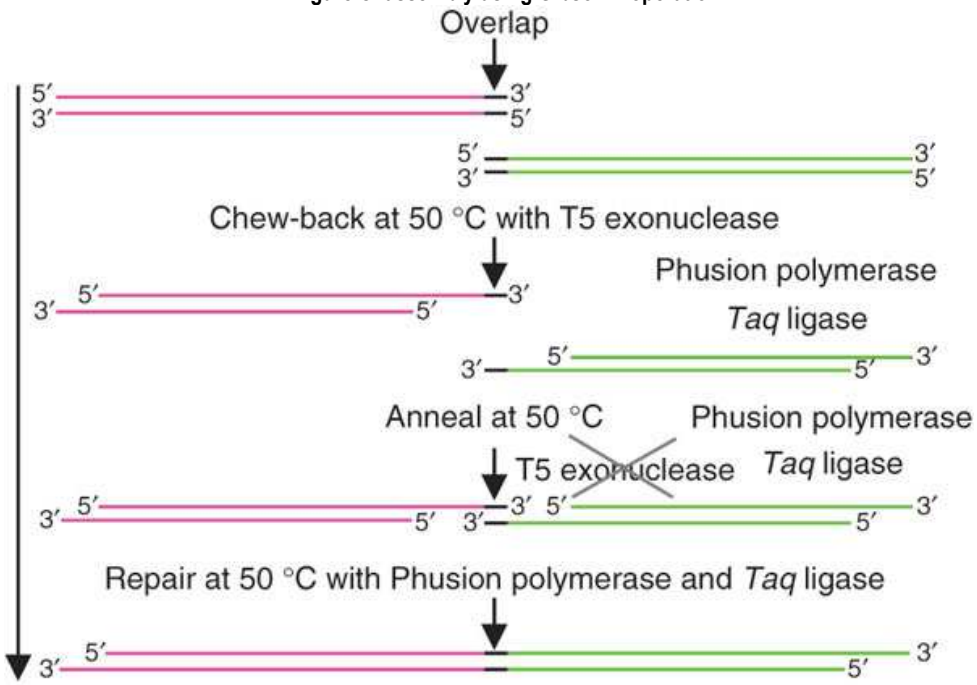


Figure C: assembly using Gibson™ operation



The resulting assembled product is usually amplified using PCR.

Yet despite its global popularity, PCR's failure rate is much higher than desirable, leading to numerous problems in library construction. Predicting such problems and computationally avoiding them with careful design has been the motive of extensive research.
Designed primers should be unique to their intended binding site on the template to be amplified and anneal at a high enough temprature. If this is not met, a longer primer that is unique or has a higher Tm will be proposed.
Next alternative binding sites that are not perfect matches, but might be competitive are examined. The template's sequence is scanned using a sliding window along its length, when for each windows instance, a competitiveness measure is calculated.
According to traditional primer design considerations, primers should also meet a plethora of standards like length, GC content, relative GC content to the indices from the 3 prime end, equal melting temperature, homodimer / heterodimer interactions and so on.

Our goal is to thoroughly examine the points of failure of PCR and develop and calibrate means of predicting problematic primers and even whole concatenations.

1. Implementation

Throughout the PCR related research, we use Fluidigm's microfluidics platform "96x96 Dynamic Array IFC for Real-Time Quantitative PCR", allowing us to monitor 9216 separate PCRs.

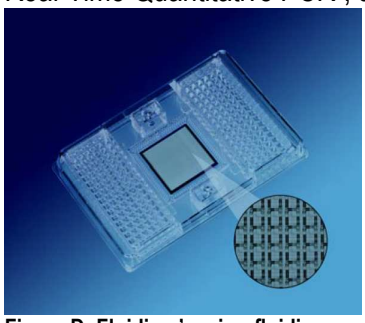


Figure D: Fluidigm's microfluidics array

Primer dimers

To study the primers dimers phenomena, we've identified a total of 2196 obtainable primers and calculated their score based on our existing methods for primers rating (figure E). To investigate the worst case/scenario, we've selected a group of 192 primers, enriched for primer pairs that are lowest scoring (figure F), expanding the portion of primer pairs expected to produce dimers from 2% to 20%.

Figure E: Full primers set scores (higher is better)

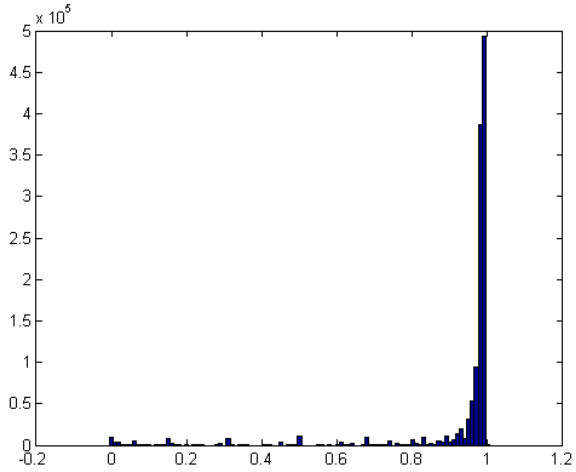
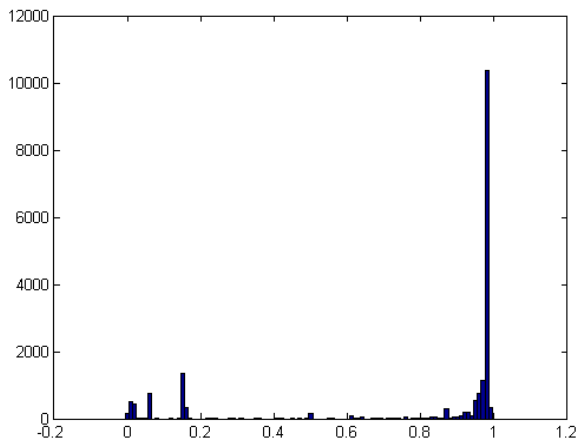


Figure F: Enriched primers set scores (higher is better)



Despite our great efforts to generate multiple cases of primer dimers, only 10 pairs out of 9216 produced exponential amplification that we will consider a risk. That is 0.1% rather than the expected 20%.

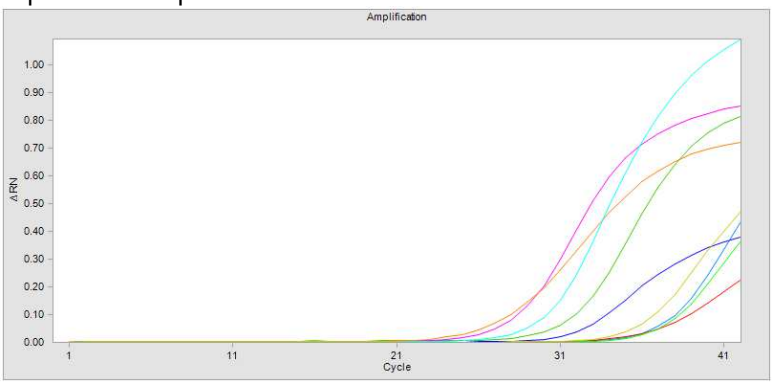


Figure G: Amplification curves of the 10 hetero - diming primer pairs.

In a duplicate run of the primers group, only 3 out of the ten amplification recurred. Additional 5 pairs were amplified on the second run that did not amplify the first time around. Three of the first run's heterodiming pairs were run together 13 times in the second chip to check for determinism. And only two of those reproduced their diming.
The rarity of primer dimers precludes the use of machine learning for a lack of instances to train on and test against. On top of that, if the effect is as week as it seems, the risk posed by primer dimers seems to be overestimated and its research, not as urgent as expected.

Mispriming

To study the mispriming phenomena we've created an all synthetic PCR test system comprised of a random 50bp long single stranded template and two perfectly matching primers 20bp long.
Our goal is to simulate a primer's alternative binding site while thoroughly examining the tolerance of PCR for mismatches. By calibrating such measure, we will be able to scan the template for alternative priming sites for our two primers and assess the risk they pose.

Figure H: Classic PCR reaction amplifying dsDNA

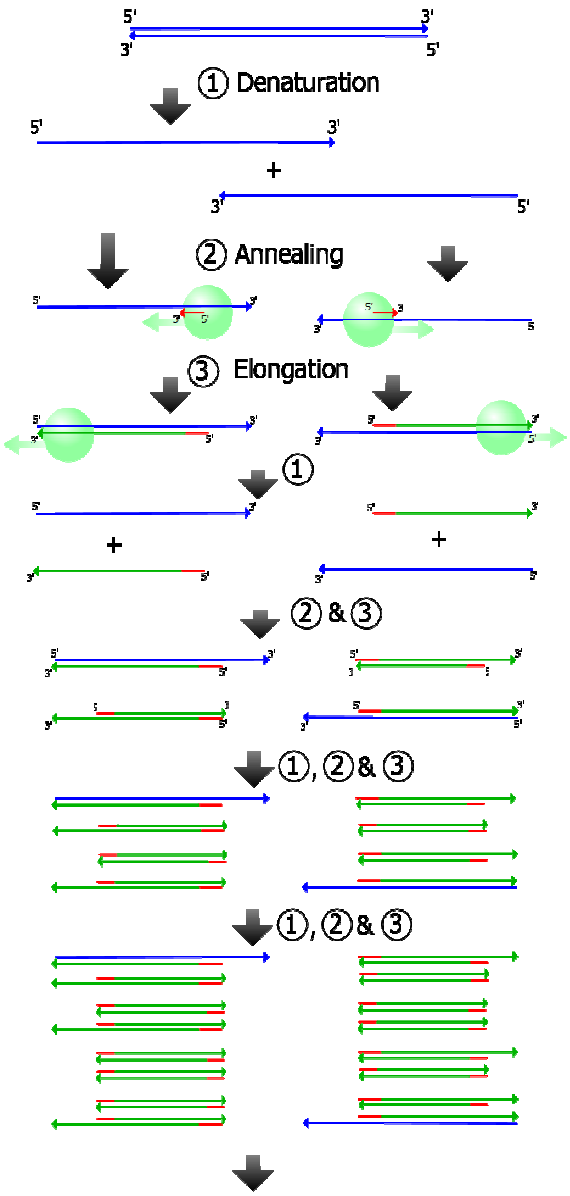
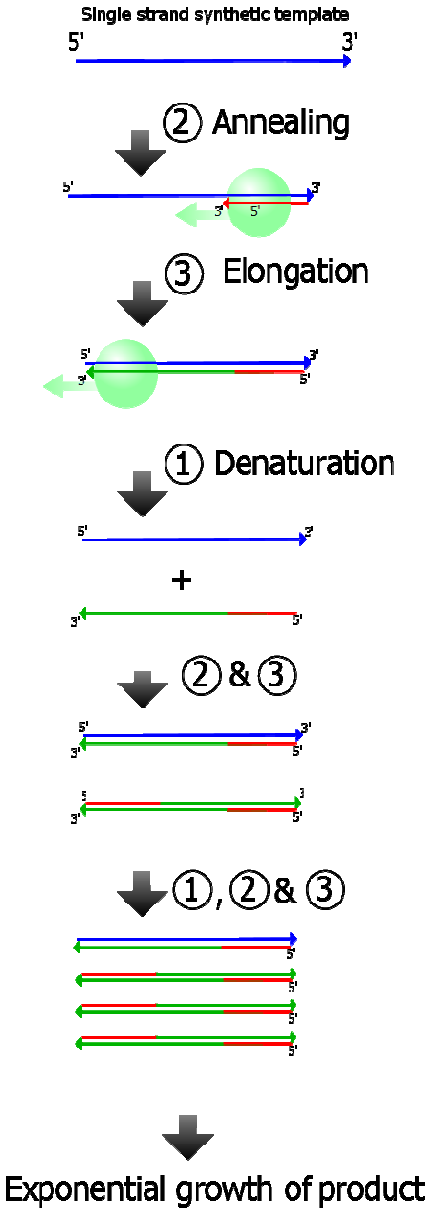


Figure I: Our simplistic priming model



To do that, we left the reverse primer of our test system intact throughout the experiment, while modifying the forward primer's sequence and its binding site's complementary sequence. These modifications allowed us to thoroughly scan the space of possible mismatches using all combinations of 96 primers and 96 templates in a Fluidigm chip.

Figure J: The experimental binding site

2. Results

Primer Dimers

We've managed to take the edge off the primers dimers hazard to some extent. The restrictiveness of existing primer dimer filters seem exaggerated and we moved most of our efforts to the mispriming research.

Mispriming

By manually analysing the results, we've managed to solidify some of the theory regarding PCR and get a good measure for just how much each of the primer design consideration really matter.

The reproducibility of the results decrease as the Ct increases, rendering most of the results that amplify after the 30th cycle, too noisy.

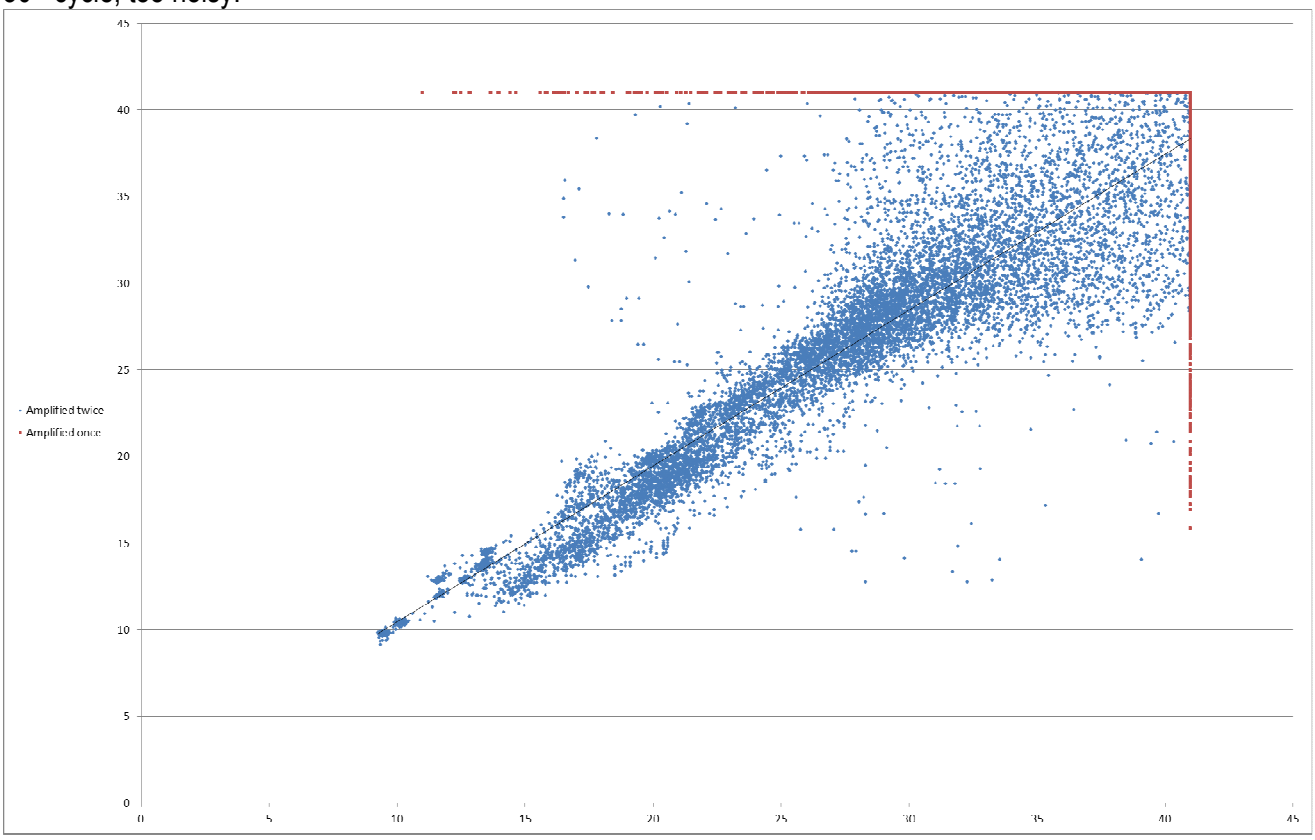


Figure K: PCR Ct values repetitiveness

This chart compares the resulting Ct values of PCR reactions based on identical primers and template.

We've solidified the theory that the Ct value of a PCR reaction scatters and gets less deterministic with prolonged cycling.

We've also manually analysed the results for correlation between the Ct value and the proximity of mismatches to the 3 prime end of the primer.

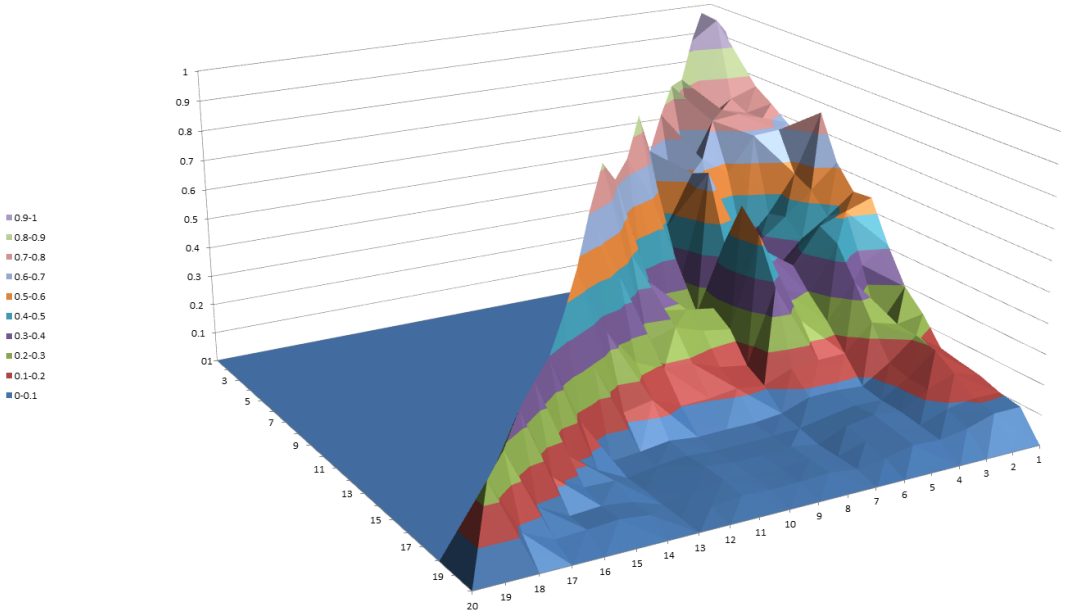


Figure L: chances of amplification per location of mismatches windows

This chart shows the chances for the successful amplification of the PCR reaction (vertical axis) and the location of possible mismatches between the primer and the template (1-2 mismatches within a 2 basepairs window, one window for each horizontal axis).
*The higher ground on the diagonal spine is a result of the two windows overlapping and does not represent an evidential trail.

Machine learning of the microfluidic miss-priming experiment results

Determining competitive binding sites using Machine Learning

The full library of synthetic PCR tests was analysed using machine learning techniques. Since each template-primer pair in the dataset has a related Ct-value, it is possible to model the mispriming problem as a binary supervised learning problem. This is done by selecting a Ct-threshold to divide the data in two classes:

- If Ct-threshold ≤ x (i.e. 20,30,40) then CLASS=PASS (amplification under x cycles occurs which means that the corresponding template-primer pair expresses a competitive binding site)
- If Ct-threshold > x then CLASS=FAIL (amplification does not occur until after x cycles. Therefore, this pair is not a competitive binding site and the possible amplification can be disregarded)

Many different Ct-thresholds have been tested being the most reliable Ct-threshold 20. This is due to the increasing noise presented by the Ct-values as its value increases, which was shown in the previous section.

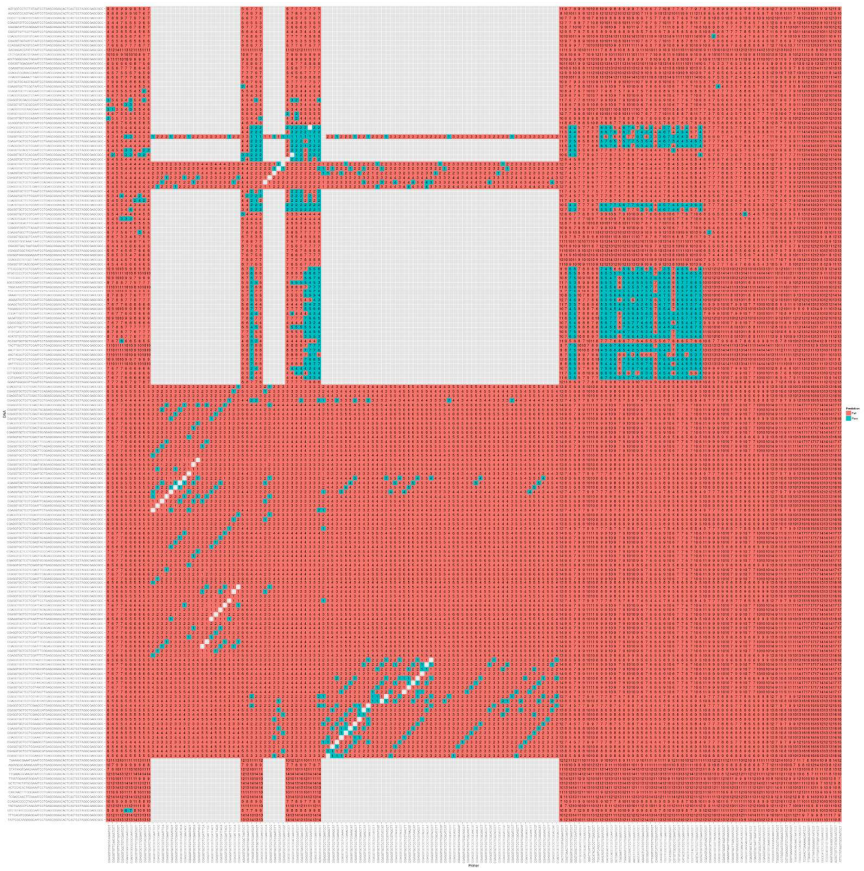


Figure M: Problem landscape using Ct-threshold 20. The vertical axis shows the different templates tested and the horizontal axis the different primers tested. Cells marked in blue show the PASS cases (where a competitive amplification can occur) and the cells marked with red show the FAIL cases.

Problem representation

Regarding the representation of the problem in attributes that are interpretable for the machine learning techniques, many different representations have been tested. In our preliminary research we have scanned many different possibilities from the usage of mispriming positions (without any information about the template or the primer), to representing changes by pairs and using the whole template and primer encoding.

The best results so far have been obtained with a representation in which each position of the template and the primer correspond to a single discrete attribute that can have the values {A,C,G,T}. This produces a total of 70 discrete attributes (50bp + 20bp).

Also the usage of a redundant attribute (attribute 71st) which expresses the number of changes between the template and the primer has been analysed in terms of accuracy of the models and interpretability of the solutions.

The attribute then are numbered as follows:

- Attributes 1-50: Template encoding
- Attributes 51-70: Primer encoding
- Attribute 71: Changes
- Class

Machine Learning algorithms

The different machine learning algorithms tested to perform data mining on this domain are:

1 Neural Networks[8]
2 Nearest Neighbours[9]
3 C4.5 Desicion trees[10]
4 BioHEL (Evolutionary Learning System)[11] [12]
5 Support Vector Machines[13]

Neural Networks, Nearest Neighbours and C4.5 Desicion Trees have been discarded at early stages because of the lack of competitiveness in the results obtained. In this report we present results with the most competitive algorithms UNOTT's BioHEL and Support Vector Machines. The following paragraph explain these two machine learning techniques briefly.

- Support Vector Machines is an exact machine learning approach which tries to separate two classes by determining the hyperplane that best separates the two classes. The models generated with this technique are not interpretable by the finals users.

- UNOTT's BioHEL is an Iterative Evolutionary Learning System that generates a set of rules of the type CONDITION -> CLASS. These rules are evolved by means of a standard GA and they are interpretable by final users providing an extra advantage in the usage of this technique. BioHEL can also be trained to generate a set of rules that explicitly explain either the class PASS or the class FAIL.[14]

[8] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. In Anderson, J. A. and Rosenfeld, E., editors, Neurocomputing: foundations of research, chapter A logical calculus of the ideas immanent in nervous activity, pages 15–27. MIT Press, Cambridge, MA, USA.

[9] Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance based learning algorithms. Mach. Learn., 6:37–66.

[10] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[11] Bacardit, J., Burke, E. K., and Krasnogor, N. (2009a). Improving the scalability of rule-based evolutionary learning. Memetic Computing, 1(1):55–67.

[12] Franco, M. A., Krasnogor, N., and Bacardit, J. (2010b). Speeding up the evaluation of evolutionary learning systems using GPGPUs. In GECCO '10: Proceedings of the 12th annual conference on Genetic and evolutionary computation, pages 1039–1046, New York, NY, USA. ACM.

[13] Vapnik, V. N. (1995). The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA.

[14] Bacardit, J., Goldberg, D. E., and Butz, M. V. (2007b). Improving the performance of a pittsburgh learning classifier system using a default rule. In Learning Classifier Systems, Revised Selected Papers of the International Workshop on Learning Classifier Systems 2003-2005, pages 291–307. Springer-Verlag, LNCS 4399.

Results

Figure N shows the results obtained with the best configurations of BioHEL and Support Vector Machines over the two variations of the dataset (using and not using the attribute 71 - # of changes). For BioHEL we show the best results with the model that explicitly explains the class PASS (first column) and explicitly explains the class FAIL (second column). For Support Vector Machines we show the best results using polynomial kernels (third column) and gaussian kernels (fourth column).

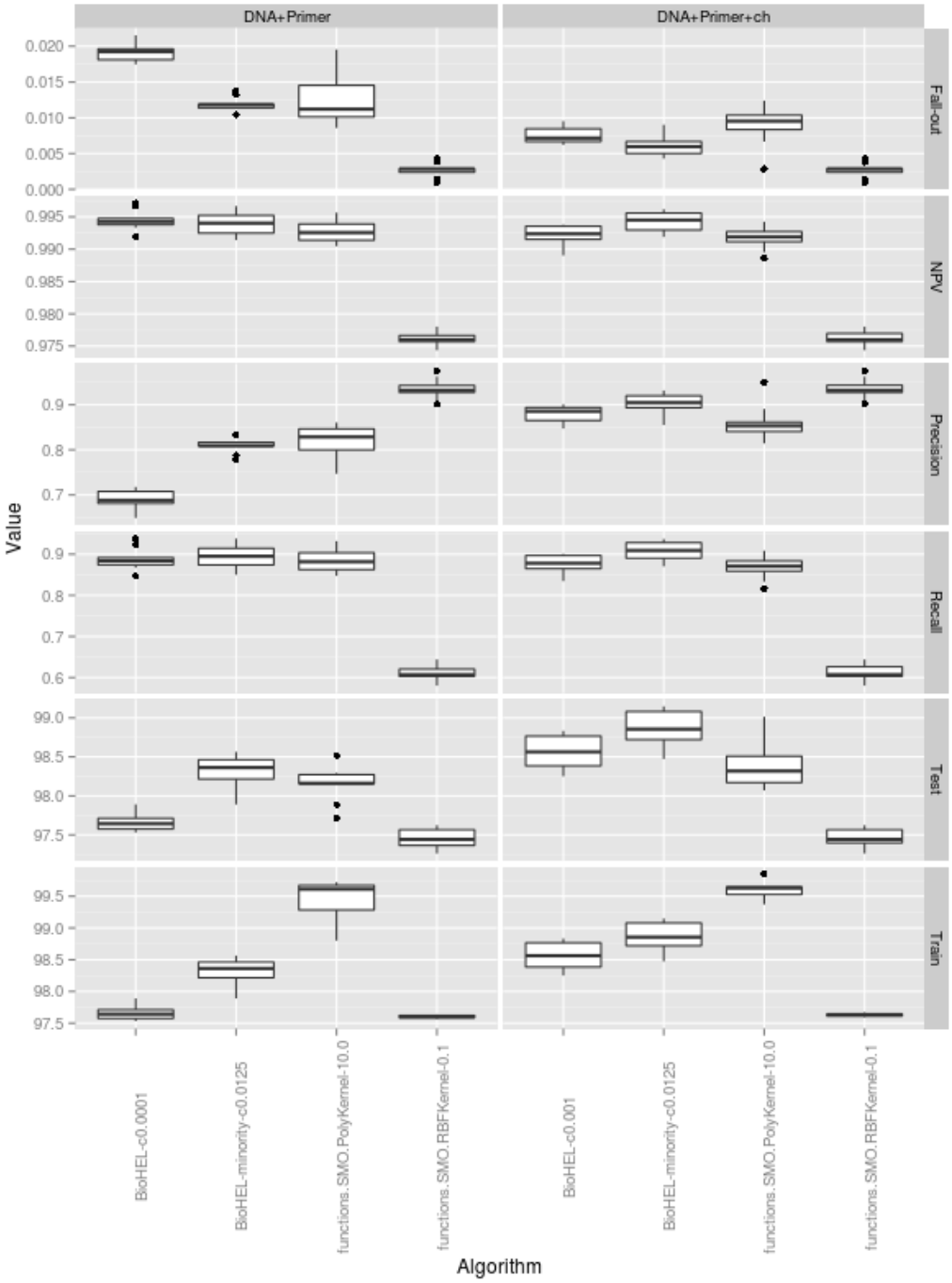


Figure N: results obtained with the best configurations of BioHEL and Support Vector Machines

In this figure we can observe that BioHEL-FAIL produces the best results for both datasets. Using the attribute 71 the accuracies are in general higher and BioHEL-FAIL obtains 98.87% accuracy on unseen instances. Without using the attribute changes the BioHEL-FAIL obtains an accuracy of 98.31%.

Considering the fall-out and the precision, the models generated by BioHEL-FAIL are more strict and produce less misclassification of the class PASS (which is the type of error that we wish to avoid). However, analysing the models produced with BioHEL-PASS might be more beneficial, as they explain the reasons why the competitive mispriming occurs.

Learned rules

The following is an example of the rule-sets generated by BioHEL-FAIL:

Att a16 is C,T,G|Att changes is [>3.000000]|Fail
Att a16 is C,T,G|Att a63 is A,T,G|Fail
Att a14 is A,C,T|Att a55 is A,C,T|Fail
Att a07 is A,C,T|Att a08 is A,C,G|Att a59 is A,C,G|Fail
Att a06 is A,C,G|Att a63 is A,T,G|Att changes is [4.000000,14.000000]|Fail
Att a66 is C,T,G|Att a68 is A,C,G|Att changes is [2.000000,6.000000]|Fail
Att a13 is A,T,G|Att a60 is A,C,T|Att changes is [4.000000,14.000000]|Fail
Att a16 is C,T|Att a62 is A,C,G|Fail
Att a15 is C,T,G|Att changes is [>3.000000]|Fail
Att a10 is A,T,G|Att changes is [4.000000,14.000000]|Fail
Att a19 is A,T,G|Att a51 is C,T,G|Att a65 is A,C,T|Att changes is [>2.000000]|Fail
Att a16 is T,G|Att changes is [>2.000000]|Fail
Att a20 is A,C|Att a69 is T,G|Fail
Att a15 is C,T,G|Att a65 is C,T|Att changes is [2.000000,3.000000]|Fail
Att changes is [>8.000000]|Fail
…
Defaul rule -> Pass

It is worth noticing that a recurring rule appears in most of the datasets expressing that if more than 8 changes occur between the template and the primer competitive amplification does not occur.

The following is an example of a ruleset generated by BioHEL-PASS:

Att a53 is C,T,G|Att changes is [<3.000000]|Pass
Att a06 is A,C,G|Att changes is [<2.000000]|Pass
Att a01 is A,C,G|Att a02 is A,T|Att changes is [2.000000,4.000000]|Pass
Att a07 is A,C,T|Att a57 is A,T,G|Att changes is [<3.000000]|Pass
Att a01 is A,C,G|Att a04 is C,T|Att a55 is A,C,T|Att changes is [2.000000,5.000000]|Pass
Att a13 is C,G|Att a14 is A,C,T|Att changes is [2.000000,2.000000]|Pass
Att a01 is A,T,G|Att a53 is C,T,G|Att changes is [<4.000000]|Pass
Att a02 is A,T|Att a53 is T,G|Att a64 is A,T,G|Att changes is [3.000000,5.000000]|Pass
Att a02 is A,C,G|Att a04 is A,C,T|Att a52 is A,C,T|Att changes is [<5.000000]|Pass
Att a51 is A,T,G|Att a54 is A,C,T|Att changes is [<4.000000]|Pass
Att a01 is A,T,G|Att a62 is A,C,T|Att changes is [2.000000,3.000000]|Pass
…
Default rule -> Fail

Understanding the BioHEL's Rule Sets

To extract new insights from the models generated with BioHEL different techniques have been applied. First post-processing mechanisms[15] have been used to push forward the generality of the rule sets and compress them. The post-processing stage involves 3 operators:

- Rule pruning: Deletes unnecessary attributes for the rules. Attributes that can be deleted without degrading the accuracy.
- Rule cleaning: Deactivates values in the attribute's discrete predicate that belong to areas of the search space where a) there are negative examples or b) there are no examples.
- Rule swapping: Is an heuristic that swaps the order or the rules according to a similarity metric to try to find and order that produces a more general set of rules and allows to delete some of the rules, hence making the rule set smaller.

Figure O shows the attribute dependency before (left) and after applying (right) the post-processing mechanisms to the rule sets generated with the attribute 71 (# of changes).

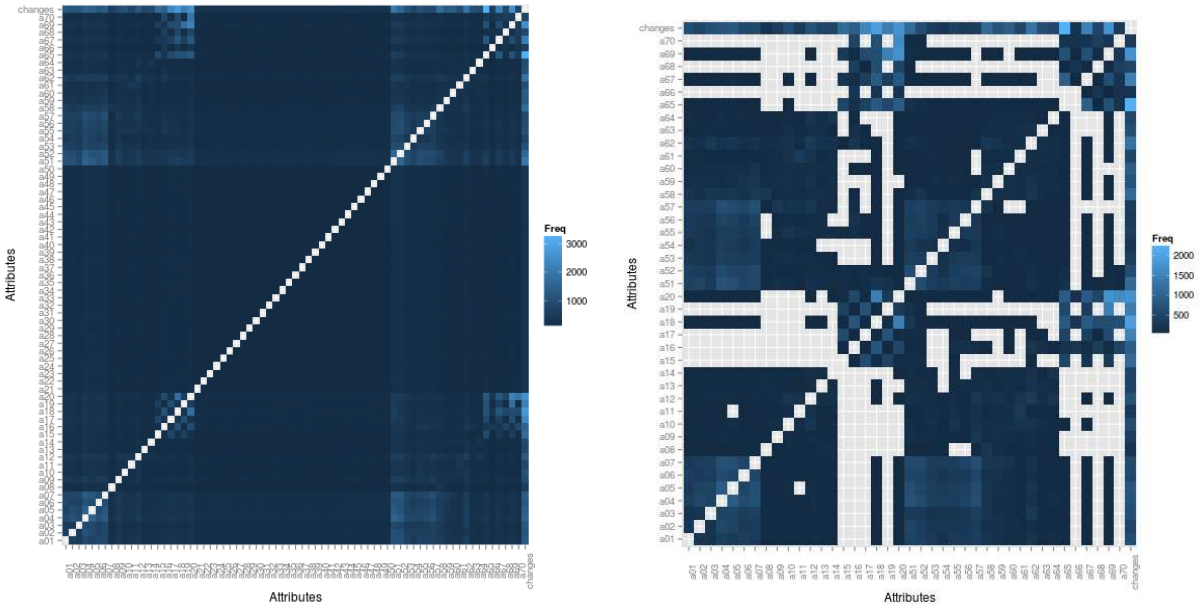


Figure O: attribute dependency before (left) and after applying (right) BioHEL's post-processing

In these figures we can observe that the post-processing mechanisms discard irrelevant dependencies and produce a more structured set of relationships between attributes (positions). It is worth noticing that these post-processing operators discard any dependency within the tail of the template (attributes 21-50) as these values remain constant in the dataset. The image on the left also shows a frequent dependency between the 6 right-most positions of the primer and the 6 right-most positions of the template. Also there is a frequent interaction between

[15] Franco, M. A., Krasnogor, N., and Bacardit, J. (2012d). Post-processing operators for decision lists. In Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference, GECCO '12, pages 847–854, New York, NY, USA. ACM Press.

the 7 left-most positions of the template and the 7 left-most positions of the primer. Moreover, the attribute # of changes as expected is related to all the positions in both the template and the primer.

The usage of the attribute # of changes can produce more compact rule sets as it has been observed empirically (not shown) but can also reduce the expressivity of the rules if we want to interpret them, as the rules are not forced to show which type of changes should occur for competitive amplification to happen. Considering this, we also analysed the rule sets generated with the datasets without using the attribute # of changes.

Figure P shows the attribute dependency without using the attribute # of changes. In this figure we can observe that:

- The 6 right-most positions of the primer are only dependent with themselves or with the 6 right-most positions of the template.
- The 14 left-most positions of the template are only related to the 14 left-most positions of the primer.
- The 14 left-most positions of the primer have also a relationship with the rest of the primer positions.
- The most frequent dependencies observed are between the last 2 positions of the primer and the template which is a phenomenon already considered in the PCR theory.

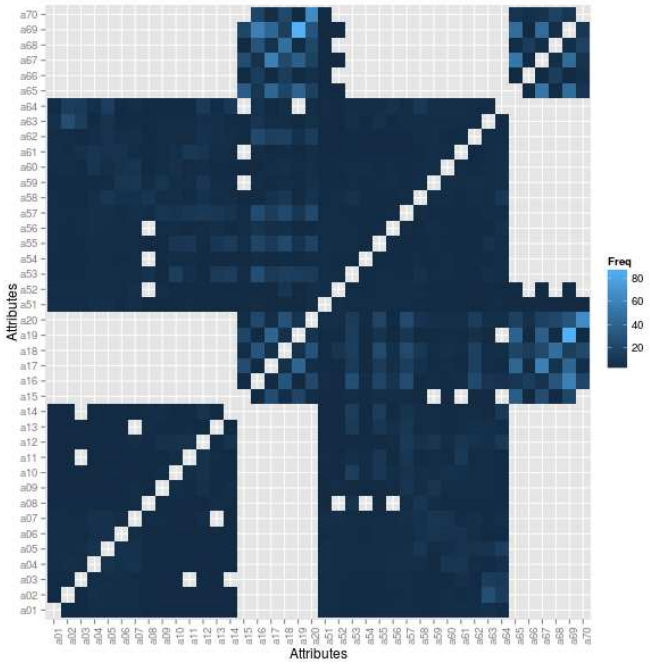


Figure P: attribute dependency without the # changes attribute

Figure M: performance comparison of amplification classifiers for ct=20

3. Conclusions

Primer dimers are not a common cause for PCR failures.
Mispriming can be predicted with very good accuracy by scanning the along the template for a primer-size window alternative binding site.
Current methods for primer design suffer from a plethora of strict rules, usually without experimental justification. When addressing the problems of library design, these strict sets of rules become a serious burden as every junction that is “impossible to PCR” results in a plan modification for the worse. This can lead to higher library costs and less reliable construction plan as full construction generation are added.

4. References

BioHEL:
http://icos.cs.nott.ac.uk/data/papers/Bacardit2009.pdf
http://icos.cs.nott.ac.uk/data/papers/Franco2010.pdf