PHD THESIS:

"INNOVATIVE SIGNAL PROCESSING TECHNIQUES IN BIOENGINEERING: COMPRESSED SENSING AND HIGH RESOLUTION DNA MELTING ANALYSIS"

STUDENT:

CÉSAR HUGO PIMENTEL ROMERO

ADVISOR:

PROF. GIANLUCA SETTI

Università Degli Studi di Ferrara

July 2020

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God for giving me the health and strength to keep going in these difficult times. To my beautiful family, my parents Hugo and Consuelo for their love, encouragement, sacrifice and support throughout my life, thank you for always believing me. To my beloved sister Naty, for those words of encouragement in difficult times, that moments of crisis when I stopped believing in myself, thank you for your unconditional love and support. Dear sister, this last months has been very difficult for the family. I pray to God that every new day brings you more strength for a speedy recovery. I love you so much and we trust God for your fast recovery. I would also like to thank my sisters Karen and Gaby as well as my beautiful nieces.

Many thanks to the University of Ferrara (UNIFE) for giving me the opportunity to study the PhD course. Thanks for having me in this beautiful and fascinating country, its places and friendly people will live forever in my heart. My great admiration and sincere gratitude to my advisors Prof. Gianluca Setti and Prof. Fabio Pareschi for your patience and knowledge.

Besides my advisors, I would like to thank the ARCES research group in Bologna, Prof. Mauro Mangia and Prof. Riccardo Rovatti, thank you for this wonderful learning experience. Also, I would like to express my gratitude to my thesis referees Prof. Pamela Abshire and Prof. Maurizio Martina for all the valuable and insightful comments.

To all my friends, thank you for your kindness, encouragement and support in the most difficult moments. I thank the live for blessing me with your friendship. It is impossible to make a list of all of you, but you are always in my heart.

Life is beautiful.
It's about giving. It's about *family*.

— Walt Disney

To mi family, especially my sister Naty.
We trust God for your recovery.

PREFACE

In the last years, the revolution of digital platforms and energy consumption issues have forced the development of new strategies to store, process and transmit information. Currently, the amount of data stored and shared has been experiencing a tremendous growth day by day. The international Data Corporation (IDC) asserts that the summation of the data created, captured or replicated will grow from 33 Zettabytes (ZB) in 2018 to 175 ZB by 2025. This impressive growth is due in part to the evolution of electronic devices. For example, the Internet of things (IoT) has emerged as a system to collect, process and send a lot of data gathered by many devices over a network. IoT allows to intercommunicate devices by using data collected without human help. Some applications include health monitoring, infrastructure applications, transportation, smart home among others. Acquire all these digital information creates a really challenge for analog to digital interfaces mainly for two reasons. First, if a signal has a very broad bandwidth the acquisition of the data requires a denser sampling. Second, even if the data could be acquired, it will be processed and stored efficiently.

Understanding the challenge to deal with this vast amount of data, new techniques to process and store the information efficiently have been developed, becoming important allies of hardware improvements. In the field of data compression, one of the most significant new approach is the Compressed Sensing (CS) [10, 20], a powerful mathematical tool that, under certain conditions is capable to acquire and compress an input signal using a sub-Nyquist sampling.

Although it is true the effectiveness of standard CS technique, there are methods in the literature that can improve performance of the standard CS in therms of compression ratio in some kinds of signals. [49, 50]. These methods based in the CS are reviewed in this work and a new approach is proposed.

Punctually, a detailed analysis of two sensing matrix adaptation techniques is presented, the *Rakeness-based CS* Rak-CS and *Nearly Orthogonal-based CS* NeO-CS approaches, where the strengths and limitations of each of them are discussed. Based on these analysis, an algorithmic solution is proposed with the aim to overcome these limitations.

On the other hand, the strategies to encoding, processing and storing the data have been explored from other areas such as biotechnology. In the last decade, the idea to store digital information in synthetic DNA molecules has gained great interest by many researchers for two reasons, density and durability. In theory, a DNA molecule offers a density of 1 exabyte per cubic millimeter and an average durability of 500 years, if the information could be packaged with the same density as the genes of the bacterium Escherichia coli, all existing data could be contained in about one kilogram of DNA. The technological advances in synthesis (write), sequencing (read), and editing techniques of DNA molecules have allowed for example to store $\tilde{7}39$ kb of five compressed files including 154 Shakespeare's sonnets in ASCII text, a scientific paper (PDF), a piece of the famous speech "I have a dream" of Martin Luther King (MP3) and a photograph of the European Bioinformatics Institute (EBI) (JPEG200). Typically, the digital information is compressed and encoded over the the DNA code-words (A,C,G,T) constrained to error controls. However, all the approaches for data recordings have several drawbacks in all the steps, including the limited length of the DNA fragments

in the synthesis process, the errors and highly costs in synthesis and sequencing techniques. In addition, the difficulty to the partial access of the data, that means, it is necessary reconstruct the entire sequence to read even a single base. Despite DNA storage is an incredibly slow process compared with the timescales of the traditional data recording, it still a promising approach as is shown in [18, 33, 75].

With the idea of venturing into the interesting molecule of life world, a brief overview of DNA sequencing techniques is provided in this work. The importance of DNA to the life has encouraged the improvement of methods to obtaining the DNA information. This developments in bioinformatics use computational techniques to assembling an entire genome (*de novo*), algorithms to align the assembled data with a reference sequence to find variations: mutation, gene expression, or single nucleotide polymorphisms (SPNs), that are the most common genetic variations. However, if an experiment implies the identification of variants in regions of interest of the DNA it is not always necessary the DNA sequencing, in that case High Resolution Melting (HRM) analysis is a good choice because is cheaper and faster than sequencing and it is very reliable.

As first step to start with this interesting research, the second part of the thesis is focused in the study of the High Resolution Melting (HRM) analysis. Currently, commercial software of HRM analysis is expensive [1] [2], also there are few free software and commonly, they present among other problems: Exponential background subtraction not supported, grouping based in the melting temperature identification T_m not supported, average of targets not supported in the difference graphic, some software requires previous knowledge of programming language by the user, Bio-Rad CFX platforms not sup-

1 https://www.bio-rad.com/it-it/product/precision-melt-analysis-software
2 https://www.thermofisher.com/it/en/home/life-science/pcr/real-time-pcr/real-time-pcr-applications/genetic-variation-analysis-using-real-time/high-resolution-melting-hrm.html

ported, etc. For this reason, the researchers have to develop his own software. That is how the idea of designing a specialized software for HRM analysis arose: The Contribution with the design of a free High resolution Melting Analysis software covering the problems of the very available free software and capable to satisfy the requirements of students that cannot afford the commercial software. This software was developed with AppDesigner of Matlab, guaranteeing an user friendly software.

This thesis is organized in the following manner. The first part (Part I) gives an overview about the Compressed Sensing theory followed by an overview of the state of the art of CS optimization techniques. Numerical results are presented in Chapter 3 testing the proposed method with the Rak-CS and NeO-CS using synthetic signals, electrocardiograph (ECG) and electroencephalograph (EEG) signals. Finally, Chapter 4 concludes the first part of the thesis. The second part (Part II) provides general concepts of biology, followed by an overview of sequencing techniques in the Chapter 5. In the Chapterch:HRMan, the DNA analysis based in High Resolution Curves (HRM) is explained to finally present the results of the software developed in this work. Additional information is provided by the Appendixes including the user manual.

PUBLICATIONS

- Geometric Constraints in Sensing Matrix Design for Compressed Sensing.

  This paper provides a critical review of the state-of-the art of some Compressed Sensing (CS) adaptations in the sensing stage to identify the strengths and limitations of each of them. Based on these analysis, an algorithmic solution that overcomes these limitations was proposed. The new method named Nearly Orthogonal Rakeness-based CS (NOR-CS) is proposed exploiting as much as possible the geometric constraints used in the Nearly Orthogonal-based CS (NeO-CS) approach by adapting the localization value (Rakeness-based CS (Rak-CS)).

  The methods were tested with synthetic low pass signals and Electroencephalographic (EEG) signals. Results shows a remarkably better performance in therms of PCR of the proposed approach (NOR-CS) using synthetic low pass signals. Also, NOR-CS approach presents the best quality regardless the compression ratio to identify Evoked Potentials.

  Citation:

  C.H. Pimentel-Romero, M. Mangia, F. Pareschi, R. Rovatti, G. Setti, "Geometric Constraints in Sensing Matrix Design for Compressed Sensing", Signal Processing, Volume 171, 2020, ISSN 0165-1684, https://doi.org/10.1016/j.sigpro.2020.107498.

- Resource Redistribution in Internet of Things applications by Compressed Sensing: a Survey.

  This paper provides an overview and comparison of advanced CS approaches. These techniques are capable of exploiting some

additional priors of the input signal to improve the standard CS performance. In this paper, a survey of the most promising ones is presented. A classification of these techniques is proposed, according to the signal prior that is used and which processing block is modified with respect to the standard CS. First class: Signal model improvement. Second class: Sensing Matrix adaptation and Third class Reconstruction algorithm modification.

Citation:

A. Marchioni, C. H. Pimentel-Romero, F. Pareschi, M. Mangia, R. Rovatti and G. Setti, "Resource Redistribution in Internet of Things applications by Compressed Sensing: A Survey," 2018 IEEE International Symposium on Circuits and Systems (ISCAS), Florence, 2018, pp. 1-5.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

CS Compressed Sensing

RPI Restricted Isometry Proprety

ECG Electrocardiogram signals

EEG Electroencephalograph signals

EP Evoked potentials

HGP Human Genome Project

HRM High Resolution Melting

DNA Deoxyribonucleic Acid

Part I

COMPRESSED SENSING

This first part starts with an introduction of the CS theory. After that, an overview of the state-of-the art of some CS adaptations by using additional priors in the different stages is presented in order to explain the new CS adaptation proposed in this work. Some of the CS adaptations reviewed are confronted using synthetic signals, synthetic electrocardiograph (ECG) signals and electroencephalographic (EEG) signals.

1

INTRODUCTION TO THE COMPRESSED SENSING

Compressed Sensing (CS) is a signal-processing technique able to reduce the amount of data needed to represent *sparse* signals [10, 20]. Basically, CS exploits a very common feature of the signals called *sparsity*. This property expresses the idea that a signal, when is expressed in a proper basis, has a much more compact representation than what obtained by means of a straightforward Nyquist-rate sampling. Quickly, this approach has become the center of interest in the development of Analog-to Information Converters (AICs) [47, 48, 58]. These devices based on CS allow to reduce energy consumption in comparison with the standard Analogic-to-Digital Converters (ADCs), where their effectiveness has been investigated in different areas. There are bio-medical prototypes [32, 58, 69] as well as AIC for the acquisition of radio-frequency signals [16, 76].

Mentioned prototypes are essentially CS based encoder blocks. Compressed information are transmitted at a decoder stage, that is able to recover the original signal with a mechanism that is based on the sparse assumption. The encoder/decoder processing scheme that follows the standard CS theory, is an asymmetric scheme. The encoder compresses chunks of input signal by linear projections on rows of

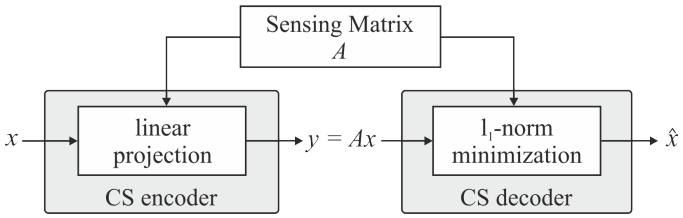


Figure 1: Std-CS based system block.

a properly designed sensing matrix A (usually drawn as instances of stochastic process) with a very low computational cost. The price expensive cost paid by the decoder, where a convex optimization problem has to be solved. Some proposed algorithms capable to recover the input signal are discussed in [22, 54]. However, for reasons of simplicity, ℓ_1 minimization-based algorithms are frequently adopted [8, 9]. Visually, the block scheme of a *Standard CS-based* (Std-CS) system is shown in Figure 1. Recent papers appeared in the literature aims to optimize the CS performance for both the encoder and decoder block.

1.1 COMPRESSED SENSING FUNDAMENTALS

Compressing Sensing theory insures that a signal can be recovered from fewer number of linear *measurements* than the number of Shannon-Nyquist rate samples. To allow it, CS is based on two fundamental premises: *sparsity* and *incoherence*.

In broad terms, the sparsity consists in the idea that the amount of information of a signal can be less than that suggested by his bandwidth, or, when a discrete-time signal depends on a number of degrees of freedom considerably smaller than its time length by an appropriate representation in a certain basis S. Formally, consider a signal conformed by n samples and represent it by the vector $x = (x_0, \ldots, x_{n-1})^\top \in \mathbb{R}^n$. We say that x is k-sparse if it can be expressed as a linear combination in a proper n-dimensional basis $S \in \mathbb{R}^{n\times n}$ as $x = S\xi$, where $\xi \in \mathbb{R}^n$ exhibits $k \ll n$ non-zero components for any possible instance x. CS asserts that any signal x can be captured in m-measurements collected in the vector $y = (y_0, y_1, \ldots, y_{m-1})^\top \in \mathbb{R}^m$, with $m < n$. This measure vector is the result of a linear operation between a *sensing matrix* $A \in \mathbb{R}^{m\times n}$ and the signal x, such that

$y = Ax = AS\xi$. On the other hand, incoherence means that the projection matrix A and the sparse representation matrix S should be as incoherent (uncorrelated) as possible.

The most practical choice to recover the signal represented by $\hat{x} = S\hat{\xi}$ is through algorithms based on the solution of the convex optimization problem

$$\begin{aligned} \hat{\xi} = \arg\min_{\xi \in \mathbb{R}^n} \ & \|\xi\|_1 \\ \text{s.t.} \ & \|AS\xi - y\|_2^2 \leqslant \varepsilon^2 \end{aligned} \tag{1}$$

where the 1-norm $\|\xi\|_1$ is used as sparsity-promoting function, while $\|AS\xi - y\|_2^2$ is the usual Euclidean norm indicating the accuracy with which the measurements y are matched by the solution, and $\varepsilon \geqslant 0$ should be chosen proportionally to the amount of noise expected on y. Such an approach is called basis pursuit with denoising (BPDN), and can be shown to be equivalent to the unconstrained problem

$$\hat{\xi} = \arg\min_{\xi \in \mathbb{R}^n} \left(\frac{1}{2} \|AS\xi - y\|_2 + \lambda \|\xi\|_1 \right) \tag{2}$$

for a proper value of the parameter λ.

There are two properties of A that guarantee a correct reconstruction: the *restricted isometry property* (RIP), i.e., A is able to approximately preserve the signal energy as $\|Ax\|_2 \approx \alpha \|x\|_2$ for some constant α as close as possible to 1 and for all k-sparse vectors x; and the *low-coherence property*, where coherence is defined as follows

$$\mu(A, S) = \sup_{j,k} |\langle A_{j\cdot}, S_{\cdot,k} \rangle| \tag{3}$$

where $A_{j\cdot}$, is the j-th row of A and $S_{\cdot,k}$ is the k-th column of S. Interestingly, both properties are satisfied when A is composed by instances of Gaussian or sub-Gaussian random variables with

$$m = \mathcal{O}(k \log n) \tag{4}$$

where m is the minimum number of measurements that guarantee a correct reconstruction [8, 9]. Conveniently for hardware applications [16, 58, 69, 76], A can be designed with an antipodal random process $A \in \{-1, +1\}^{m \times n}$ where $+1$ and -1 have the same probability to occur. This hardware-friendly matrix is chosen in this work for practical reasons (to reduce both hardware complexity and the energy necessary to encode the data), and from now on, it is implicitly assumed that A is antipodal. According to (1), encoder and decoder need to share the information of A, while the knowledge of S is required only at the decoding stage.

Theoretically, given a fixed sparsifying basis S, random matrices are largely incoherent. However, despite the advantages that the CS presents, an appropriate design of the matrix A can significantly increase the performance.

2

ADAPTED CS BY USING ADDITIONAL PRIORS

According to standard CS theory, m in (4) represents the optimum number of measurements required for a correct reconstruction. However, with an additional knowledge on the acquired class of input signals the CS can be adapted in order to reduce the value of m without compromise the reconstruction. A first step in this direction was presented in [26] and [23], where the matrix A is optimized by minimizing the mutual coherence between the columns of AS. The mutual coherence of the equivalent dictionary $D = AS$ can be measured considering the Gramm matrix $G = D^\top D$, where the mutual coherence is the off-diagonal entry with largest magnitude.

In particular in [26], given a fixed dictionary, the author tries to minimize the largest absolute values of the off-diagonal elements in the corresponding Gramm matrix by an iterative algorithm. As a result, the algorithm improves the performance of the Std-CS; however it needs many iterations to achieve a good performance. Authors in [23] aim to find A such that the corresponding Gramm matrix is as close as possible to the identity matrix. In addition, they propose a simultaneous optimization of the sensing matrix and the sparsifying dictionary adapted to image datasets. As a result, the images are represented by learned overcomplete dictionaries. Nevertheless, overcomplete learned dictionaries are not orthonormal bases and, in several applications the dictionary is restricted by physical considerations. These methods exploit only the knowledge of S and they have a limited impact on the overall system performance compared to other methods that work with additional priors as, for example, second or-

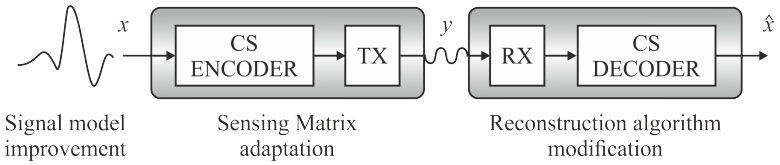


Figure 2: Block scheme of the CS processing chain with the clasification used in this work.

der statistic or structural sparsity. In this work, the CS adaptations are organized in three categories depending on which part of the system is modified with respect with the standard CS as illustrated in the Figure 2.

A different framework has been introduced in [50]. In this approach the correlation profile of the sensing stage is adapted with that of the input signal preserving the fundamentals requirements of the Std-CS for the reconstruction. The exploted prior is the second order statistic of the input signals. Here, the signal correlation matrix is estimated and analyzed to assess how much the process generating signal instances deviates from a purely random process with a flat spectral profile. This deviation is called *localization* [49, 50]. The idea of the localization is represented in the Figure 3, where the signals are points on the surface of a sphere. The Figure 3 (a) refers to a signal x that is uniformly distributed over the whole surface, while Figure 3 (b) refers to a localized signal, there the probability of a point on the surface to be a signal instance is not uniform.

Interestingly, the localization is not an unusual property. In fact, almost all real-life signals are localized [7].

Formally, consider the $n \times n$ input signal correlation matrix $\mathcal{X} = \mathbf{E}[xx^\top] = \sum_{j=0}^{n-1} \mu_j u_j u_j^\top$, with eigenvalues $\mu_0 \geqslant \mu_1 \geqslant \mu_2 \cdots \geqslant \mu_{n-1} \geqslant 0 \ \forall j \in \{0, 1, \ldots, n-1\}$ corresponding to the orthonormal eigenvectors $u_0, u_1, \ldots, u_{n-1}$. A signal is localized if the eigenvalues

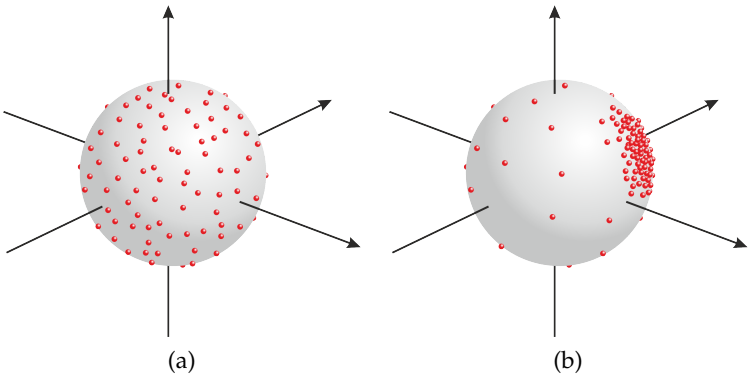


Figure 3: Instances of a) a purely random signal and b) a localized signal represented by points on the sphere surface.

μ_j are not equal to each other, and is quantified by computing the deviation of each eigenvalue from the isotropic case by

$$\mathcal{L}_x = \sum_{j=0}^{n-1} \left(\frac{\mu_j}{\mathrm{tr}\,(\mathcal{X})} - \frac{1}{n} \right)^2 = \frac{\mathrm{tr}\,(\mathcal{X})^2}{\mathrm{tr}^2\,(\mathcal{X})} - \frac{1}{n} \tag{5}$$

where $\mathrm{tr}(\cdot)$ stands for matrix trace. When the signal presents an uniform energy distribution (instances of a white random stochastic process) the localization is null, while the maximum localization ($\mathcal{L}_x \approx 1$) is reached when the signal energy is concentrated in one direction. Following this, the equation (5) provides a localization range of $\mathcal{L}_x \in \left[0, 1 - \frac{1}{n}\right]$. If x is localized, the directions where the signal is most probably concentrated are identified from $\mathcal{X}$. Nevertheless, it is important to highlight that also the less energetic directions have to be explored over the whole space signal. This is crucial to guarantee a correct reconstruction of any possible instance x, and this trade-off between focus-exploration is not a trivial issue.

2.0.1 *Rakeness CS-based*

In general terms, the Rak-CS approach discussed in [50] exploits this trade-off adapting the statistical distribution of the rows of the ma-

trix A to the features of the signal x. In particular, Rak-CS generates the rows of A randomly enough to redress the balance to the exploration side and not go against to the fundamental precepts of the standard CS theory. To formalize this concept, let us define a generic row of A as $a = (a_0, a_1, \ldots, a_{n-1})^\top$, with a correlation matrix $\mathcal{A} = \mathbf{E}[aa^\top]$. The aim of Rak-CS is to maximize the *rakeness*, defined as $\rho = \mathbf{E}_{a,x}[(a^\top x)^2]$, i.e., the average energy collected by a generic entry of the measurement vector $a^\top x$, under the assumption that the rows of A are still randomly enough.

The randomness of a is guaranteed by imposing a cap on the localization $\mathcal{L}_a$ of the generic row a by means of the parameter $l \geqslant 0$. Mathematically, the rakeness optimization problem [50] is analytically solved in terms of eigenvalues and eigenvectors of $\mathcal{A}$ (λ_j and v_j, respectively) as

$$\begin{aligned} v_j &= u_j \\ \lambda_j &= \frac{1}{J}\left[1 + \frac{J\mu_j - \Sigma_1(J)}{\sqrt{\frac{\Sigma_2(J) - \frac{1}{J}\Sigma_1^2(J)}{\gamma - \frac{1}{J}}}}\right] \end{aligned} \tag{6}$$

which holds for $j = 0, 1, 2, \ldots, J-1$, where J is an integer such that $J = \max\{j|\lambda_{j-1} > 0\}$ and $\lambda_j = 0 \quad \forall j \geqslant J$. Also, (6) depends to the partial sums $\Sigma_1(J) = \sum_{j=0}^{J-1} \mu_j$ and $\Sigma_2(J) = \sum_{j=0}^{J-1} \mu_j^2$, γ is defined as

$$\gamma = \frac{1}{n} + \frac{l^2 \mathcal{L}_x}{(1 - n\mu_j)^2} \tag{7}$$

A typical value for the localization scaling parameter l is $l = 0.5$ [7]. Finally, the correlation matrix $\mathcal{A}$ is given by

$$\mathcal{A} = \sum_{j=0}^{J-1} \lambda_j v_j v_j^\top \tag{8}$$

while the corresponding process generation rows of A possesses a localization equal to[1]

$$\mathcal{L}_a = \left(\frac{l}{1 - n\mu_{n-1}}\right)^2 \mathcal{L}_x.$$

The limitation of this approach is that the correlation profile of all the rows of A is adapted to the one of the input signal. So, dealing with highly localized signals could result in redundancies in terms of information of the measurement vector. In this case, the rows that conform A tends to be very similar to each other, resulting in measurements with almost the same information. This scenario has to be avoided by lowering the value of l with respect to the typical values suggested in [49, 50].

2.0.2 *Nearly Orthogonal CS-based*

As previously described, the key feature to deal with localized signals is the balance between focusing the projections to the most energetic directions and the exploration of the whole signal domain. Another solution to handle this trade-off is the NeO-CS approach discussed in [49] where the rows of the matrix A posses the same statistical distribution of x. This implies $\mathcal{L}_a = \mathcal{L}_x$.

To reduce redundancy in measurements, a lower bound is imposed on the angles between all couples of rows that compose the matrix A. Formally, indicate with a_j and a_k two rows of the matrix A. The cosine of the angle $\alpha = \widehat{a_j a_k}$ between them is $\cos(\alpha) = \frac{a_j^\top a_k}{||a_j||_2 ||a_k||_2}$. NeO-CS restricts the matrix A to be composed by couples of rows such that $|\cos(\alpha)| \leqslant c$ only. In [49] authors suggest to generate all m rows of A iteratively. Each new row is randomly obtained; if it

1 $\mathcal{L}_a$ is computed as in (5), where λ_j replaces μ_j and where $\mathrm{tr}(\mathcal{A}) = n$ to be compliant with the generation of antipodal sequences.

satisfies the geometric constraint with respect to all rows already generated, it is added to A. Otherwise, it discarded, and a new row is generated and tested.

An approximation of the probability to generate a new row whose angle with another row has a cosine smaller than c is explained in detail in [49] and given by

$$\Pr\{|\cos(\alpha)| \leqslant c\} = \operatorname{erf}\left(\frac{c}{\sqrt{2\left(\mathcal{L}_a + \frac{1}{n}\right)}}\right). \tag{9}$$

The probability to generate a new row that satisfies the geometric constraint when more than one row has already been generated rapidly decreases with the number of rows.

Let us indicate with Z the number of iterations necessary to generate all m rows of A. According to [49], when we deal with a high localized signal, a better performance in the reconstruction stage is achieved when a smaller value of c is imposed; however, the difficulty to generate the matrix A also increases considerably, and so, the expected value of Z. Although the matrix A can be generated offline and locally stored, it is important to consider if the increase in the performance justifies a possible huge computational effort. For instance, Figure 6 shows the average number of iterations $\mathbf{E}[Z]$ required to obtain a new row of A as a function of the total number of rows m with $c = 0.1875$ and for $n = \{128, 256\}$, obtained by montecarlo simulations. The figure also reports profiles obtained by data extrapolation (solid line) that evidences that $\mathbf{E}[Z] > 10^6$ is necessary to get a compression ratio $CR = n/m = 2$. If we limit Z to be less than a certain Z_{max}, NeO-CS could hardly be applicable with certain c values. From a geometric point of view, in the case of $\mathcal{L}_a = 0$, antipodal vectors can be represented by points uniformly distributed on the surface of a

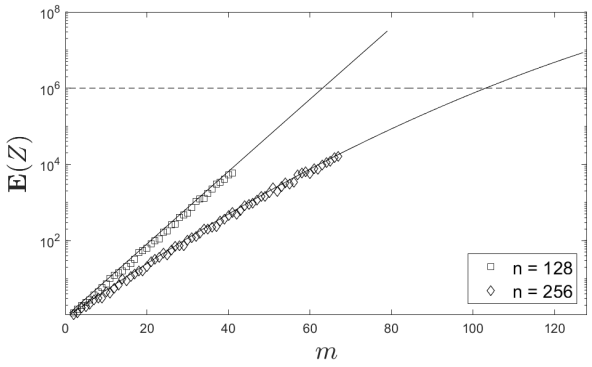


Figure 4: Average of the number of iterations $\mathbf{E}[Z]$ to generate the rows of the matrix A using NeO-CS with $c = 0.1875$ and two different values of n. The threshold value $\mathbf{E}[Z] = 10^6$ is also indicated.

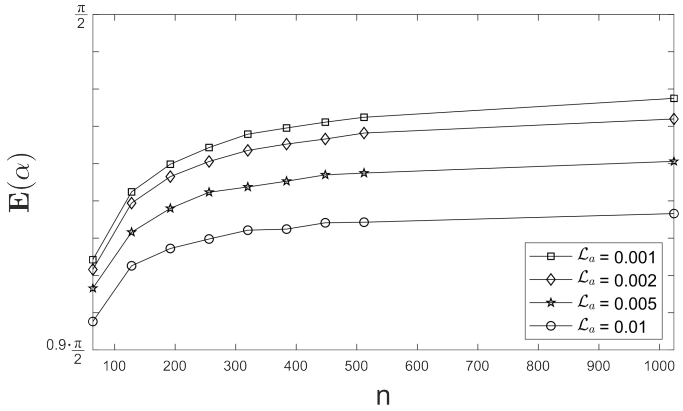


Figure 5: Average of the angle between two rows as a function of n.

multi-dimensional sphere. As a results, when n increases, the angle between a couple of rows increases up to $\frac{\pi}{2}$. However, for $\mathcal{L}_a > 0$, the points are not more uniformly distributed on the surface and their distribution concentrates according to the assigned correlation matrix. The observable effect is that the average value of α, $\mathbf{E}[\alpha]$, slowly approaches the angle $\frac{\pi}{2}$. This is the reason why when $\mathcal{L}_a$ increases the difficult to generate the rows of A increases as well. The impact of both localization and n is shown in Figure 5, that depicts the $\mathbf{E}[\alpha]$ profiles as a function of n for $\mathcal{L}_a = (0.001, 0.002, 0.005, 0.01)$. As one can observe, as n increases also $\mathbf{E}[\alpha]$ increases, but the speed of con-

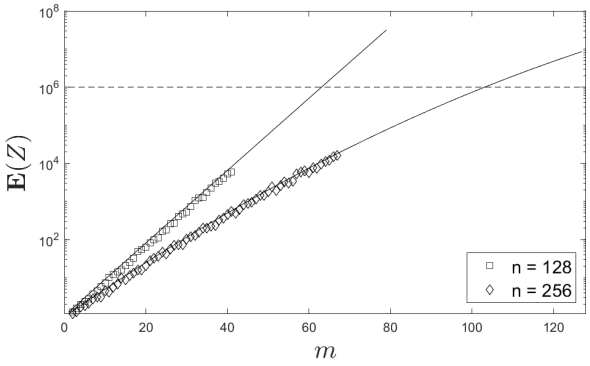


Figure 6: Average of the number of iterations $\mathbf{E}[Z]$ using NeO-CS approach with $c = 0.1875$ and two different values of n.

vergence of $\mathbf{E}[\alpha]$ to the $\frac{\pi}{2}$ asymptotic limit strongly depends on the localization value.

2.0.3 *Nearly Orthogonal Rakeness CS-based*

With the aim to propose a new focus/exploration trade-off, we introduce the Nearly Orthogonal Rakeness-based (NOR-CS) approach. This method exploits the geometric constraints that characterize NeO-CS, and mitigate the hardness in generating the m rows of A by exploiting the localization as in the Rak-CS approach mathematically described by (6), i.e., by re-introducing the l parameter that scales the localization imposed to the A rows.

With NeO-CS we limit the maximum number of iterations that are allowed to generate a new row of A by Z_{max}, NOR-CS tries to exploit as much as possible the same geometric constraint introduced before where the hardness of a new generation is imposed on average, i.e., $\mathbf{E}[Z] \leqslant Z_{max}$. If this cap cannot be ensured, the localization of the next generated row is reduced by scaling the value of l.

For a known class of signal, i.e., for a known $\mathcal{L}_x$, the hardness to generate the $\hat{m}$-th row of A with a given value of c has been modeled by the probability p that a process with a correlation profile evaluated

$n = 128$, $\hat{m} = 35$

	$c = 0.1094$	$c = 0.1563$	$c = 0.2031$	$c = 0.25$
$l = 0.163$	$5.48 \cdot 10^{-4}$	0.0897	0.5199	0.8731
$l = 0.281$	$2.82 \cdot 10^{-4}$	0.0609	0.4657	0.8345
$l = 0.415$	$1.00 \cdot 10^{-4}$	0.0360	0.3466	0.7711
$l = 0.490$	$4.55 \cdot 10^{-5}$	0.0229	0.2874	0.7147
$l = 0.616$	$9.61 \cdot 10^{-6}$	$8.50 \cdot 10^{-3}$	0.1779	0.5753
$l = 0.711$	$2.51 \cdot 10^{-6}$	$3.90 \cdot 10^{-3}$	0.1155	0.4763
$l = 0.817$	$4.72 \cdot 10^{-7}$	$1.30 \cdot 10^{-3}$	0.0621	0.3425
$l = 1.010$	$2.34 \cdot 10^{-8}$	$1.21 \cdot 10^{-4}$	0.0154	0.1620
$l = 1.370$	$1.36 \cdot 10^{-10}$	$5.07 \cdot 10^{-7}$	$4.68 \cdot 10^{-4}$	0.0185
$l = 1.493$	$8.23 \cdot 10^{-12}$	$1.25 \cdot 10^{-7}$	$1.07 \cdot 10^{-4}$	$7.10 \cdot 10^{-3}$

$n = 128$, $\hat{m} = 50$

	$c = 0.1094$	$c = 0.1563$	$c = 0.2031$	$c = 0.25$
$l = 0.163$	$1.94 \cdot 10^{-5}$	0.0308	0.3882	0.8223
$l = 0.281$	$7.43 \cdot 10^{-6}$	0.0173	0.3342	0.7697
$l = 0.415$	$1.69 \cdot 10^{-6}$	$8.30 \cdot 10^{-3}$	0.2159	0.6911
$l = 0.490$	$5.36 \cdot 10^{-7}$	$4.30 \cdot 10^{-3}$	0.1666	0.6190
$l = 0.616$	$5.63 \cdot 10^{-8}$	$1.00 \cdot 10^{-3}$	0.0831	0.4501
$l = 0.711$	$8.02 \cdot 10^{-9}$	$3.38 \cdot 10^{-4}$	0.0448	0.3423
$l = 0.817$	$7.06 \cdot 10^{-10}$	$7.05 \cdot 10^{-5}$	0.0183	0.2135
$l = 1.010$	$9.52 \cdot 10^{-12}$	$2.16 \cdot 10^{-6}$	$2.50 \cdot 10^{-3}$	0.0733
$l = 1.370$	$6.01 \cdot 10^{-15}$	$7.36 \cdot 10^{-10}$	$1.60 \cdot 10^{-5}$	$3.20 \cdot 10^{-3}$
$l = 1.493$	$1.01 \cdot 10^{-16}$	$1.06 \cdot 10^{-10}$	$1.90 \cdot 10^{-6}$	$8.24 \cdot 10^{-4}$

$n = 256$, $\hat{m} = 60$

	$c = 0.1094$	$c = 0.1563$	$c = 0.2031$	$c = 0.25$
$l = 0.154$	$8.00 \cdot 10^{-3}$	0.4887	0.9338	0.9972
$l = 0.273$	$2.50 \cdot 10^{-3}$	0.3373	0.8789	0.9869
$l = 0.405$	$2.75 \cdot 10^{-4}$	0.1500	0.7363	0.9606
$l = 0.500$	$3.45 \cdot 10^{-5}$	0.0670	0.5501	0.9147
$l = 0.600$	$2.75 \cdot 10^{-6}$	0.0191	0.3752	0.8306
$l = 0.708$	$1.22 \cdot 10^{-7}$	$4.00 \cdot 10^{-3}$	0.1867	0.6700
$l = 0.823$	$2.30 \cdot 10^{-9}$	$4.71 \cdot 10^{-4}$	0.0696	0.4362
$l = 0.972$	$1.17 \cdot 10^{-11}$	$1.93 \cdot 10^{-5}$	0.0131	0.2027
$l = 1.386$	$4.00 \cdot 10^{-19}$	$2.12 \cdot 10^{-10}$	$1.05 \cdot 10^{-5}$	$3.70 \cdot 10^{-3}$
$l = 1.603$	$2.05 \cdot 10^{-20}$	$8.35 \cdot 10^{-14}$	$7.30 \cdot 10^{-8}$	$1.68 \cdot 10^{-4}$

$n = 256$, $\hat{m} = 90$

	$c = 0.1094$	$c = 0.1563$	$c = 0.2031$	$c = 0.25$
$l = 0.154$	$6.82 \cdot 10^{-4}$	0.3413	0.9026	0.9958
$l = 0.273$	$1.19 \cdot 10^{-4}$	0.1946	0.8244	0.9798
$l = 0.405$	$4.23 \cdot 10^{-6}$	0.0569	0.6309	0.9409
$l = 0.500$	$1.83 \cdot 10^{-7}$	0.0169	0.4036	0.8739
$l = 0.600$	$4.01 \cdot 10^{-9}$	$2.50 \cdot 10^{-3}$	0.2272	0.7564
$l = 0.708$	$3.69 \cdot 10^{-11}$	$2.46 \cdot 10^{-4}$	0.0792	0.5472
$l = 0.823$	$8.91 \cdot 10^{-14}$	$9.74 \cdot 10^{-6}$	0.0181	0.2859
$l = 0.972$	$2.91 \cdot 10^{-17}$	$7.79 \cdot 10^{-8}$	$1.50 \cdot 10^{-3}$	0.0902
$l = 1.386$	$1.20 \cdot 10^{-28}$	$2.38 \cdot 10^{-15}$	$3.13 \cdot 10^{-8}$	$2.16 \cdot 10^{-4}$
$l = 1.603$	$1.94 \cdot 10^{-30}$	$1.47 \cdot 10^{-20}$	$1.66 \cdot 10^{-11}$	$2.10 \cdot 10^{-6}$

Figure 7: Values extracted from the $F_p(c, l, \hat{m})$ look-up table with $\mathcal{L}_x = 0.01$.

with (6)(7) will generate a row for which the geometric constraint holds with the $\hat{m} - 1$ already generated rows. Mathematically, fixed a-priori the values of $\mathcal{L}_x$ and n, p is a function of the cosine value c, of the localization scaling factor l, and of the number of step $\hat{m}$. Due to the complexity in computing a closed-expression for p, we evaluated it by means of a look-up table $p = F_p(c, l, \hat{m})$ that has been estimated by Montecarlo simulations for the two values $n = 128$ and $n = 256$. Some values of the lookup table have been reported in Figure 7.

The algorithm we propose to generate A according to the NOR-CS approach and limiting the (expected) number of iterations to Z_{max} is briefly described as follows, and summarized in Algorithm 1. The key feature is that the value of l is initially set to the desired value l_0. We use $l = l_0$ until the hardness of generating a single row exceeds a threshold p_{min} (computed according to the desired Z_{max}). After that,

by means of the proposed look-up table, l is reduced to the maximum value that still ensure $p > p_{min}$. In detail:

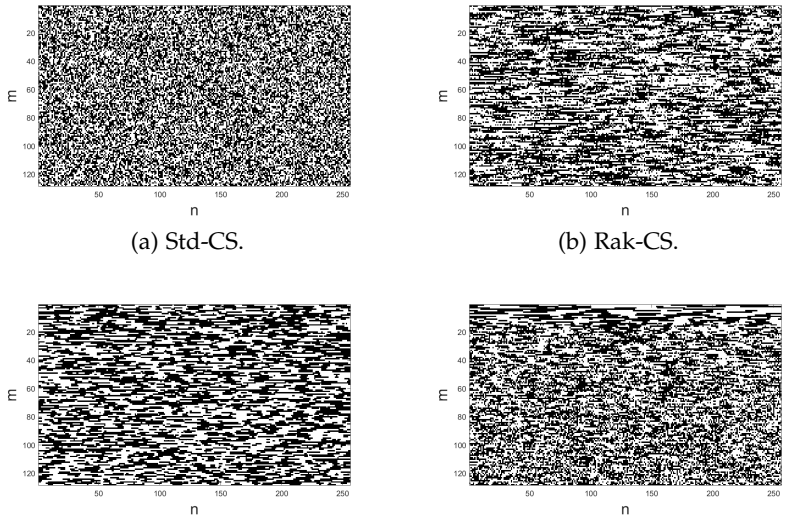
1. First, c and the hardness p_{min} to generate the rows are fixed.
2. The first antipodal row is generated with an initial correlation profile evaluated by (6)(7) where l is equal to l_0.
3. The probability p to generate the next row is obtained from the look-up table $p = F_p(c, l, \hat{m})$. If $p > p_{min}$, the row is generated with the same value of l as the previous row. Otherwise, a maximum value of l is calculated by exponential interpolations of the values in the look-up table such that $p > p_{min}$ holds. With this new value of l, (6)(7) is evaluated and the row is generated.

As an example, an instance of the matrix A generated by each method is showed in the Figure 8 trying to match the localization of a low-pass input signal x. The matrix generated with the Std-CS (Figure 8 a)) shows a purely random distribution. In the case of the matrix generated with Rak-CS and NeO-CS (Figure 8 b) and 8 c)) respectively), a low-pass profile is clearly identifiable. Yet, it is possible to see that sequences generates by NeO-CS is more localized than that obtained with Rak-CS due to higher localization. The Figure 8 d) shows the matrix generated by NOR-CS, where is possible to visualize how first rows are highly localized and then, gradually the localization decreases every certain numbers of rows.

Note that, among the many methods known to generate antipodal sequences with a certain correlation profile [13, 64, 65], the Rak-CS, the NeO-CS and the NOR-CS cases in the above example have been generated using the clipping of Gaussian instances [36, 72]. Basically, a $n \times n$ correlation matrix $\mathcal{G} = \sin\left(\frac{\pi}{2}\frac{n\mathcal{A}}{\mathrm{tr}(\mathcal{A})}\right)$ is used to generate a zero mean Gaussian vector g, such that antipodal a_j are computed by clipping the elements of g.

Input : l_0, c, p_{min}, m.
$l \leftarrow l_0$
  $\mathcal{A} \leftarrow$ equation (6)(7)
  $A(1,:) \leftarrow$ generate the first antipodal row with $\mathcal{A}$
  for $i = 2$ **to** m **do**
    $p = F_p(c, l, i)$
    if $p < p_{min}$ **then**
      $l \leftarrow \arg\max_{\hat{l}} F_p(c, \hat{l}, i) > p_{min}$
      $\mathcal{A} \leftarrow$ equation (6)(7)
    end
    $A(i,:) \leftarrow$ generate antipodal row with $\mathcal{A}$
  end
Output : A

Algorithmus 1 : Nearly Orthogonal Rakeness CS-based pseudocode



(a) Std-CS. (b) Rak-CS.

(c) Neo-CS. (d) NOR-CS.

Figure 8: Matrix A generated with the different approaches. Black dots correspond to +1, white dots to -1.

The reconstruction problem, either in the form of (1) or (2), can be easily solved by mapping it into a linear programming problem by using general purpose solvers as ℓ_1-MAGIC [12] and SPG-ℓ_1 [3].

3

NUMERIC RESULTS

3.1 NUMERIC RESULTS

Performance of all the approaches described above is provided through Montecarlo simulations. For each instance, we compute the Reconstructed Signal to Noise Ratio (RSNR) as the ratio between the energy of the input signal x and the energy of the difference between x and the reconstructed signal $\hat{x}$ expressed in dB. Starting from these data we evaluate the Average RSNR (ARSNR)

$$\text{ARSNR} = \mathbf{E}_{A,x} \left[20 \log_{10} \left(\frac{||x||_2}{||x - \hat{x}||_2} \right) \right] \tag{10}$$

and by the Probability of Correct Reconstruction (PCR)

$$\text{PCR} = \Pr\{\text{RSNR} \geqslant \text{RSNR}_{\text{min}}\}, \tag{11}$$

this estimates the probability that the RSNR exceeds a minimum value.

3.2 TEST IN SYNTHETIC SIGNALS

To prove the performance of the NOR-CS and compare it with the other approaches mentioned above, the methods are tested with synthetic low-pass signals. Basically, n-dimensional instances x, localized and k-sparse in a certain basis S are generated starting from an instance of a random vector x' with zero mean and correlation ma-

trix $\mathcal{X}'$. Computing $\xi' = S^{-1}x'$, we obtain ξ by keeping only the k higher absolute values in ξ'. Finally, the synthetic signal is obtained as $x = S\xi$. The correlation matrix $\mathcal{X}'$ is expressed as a Toepliz matrix $\mathcal{X}'_{i,j} = r^{\frac{\beta}{n}|(i-j)|}\ \forall i,j \in \{0,1,\ldots,n-1\}$, with $r \in [0,1]$, x' is a chunk of a stationary stochastic process with low pass profile. The factor β is empirically imposed to prevent an abrupt decay of the profile when n is large. In this work, the Discrete Cosine Transform (DCT) is used as the orthonormal basis S and the setting of the parameters to generate the synthetic signals are reported in the Table 1 including the corresponding values of $\mathcal{L}_x$.

Table 1: Parameter settings to generate the synthetic signals.

Sparsity Basis	n	r	k	β	$\mathcal{L}_x$
DCT	128	0.7	12	150	0.026
DCT	256	0.7	25	150	0.024

For the Rak-CS approach we use the typical value $l = 0.5$ [7], with NeO-CS, for each value of c, we have limited the number of iterations to $Z_{max} = 10^5$ to generate a new row $\hat{m}$. As p_{min} for NOR-CS we adopt 10^{-3}, i.e, the expected average value of trials to be used for each row generation is 10^3. Once the parameters are established, for each method, 100 instances of the matrix A were drawn, each of them encoding 20 different instances of x. Also, non-idealities were modeled adding white Gaussian noise to each input vector x adapted to an Intrinsic Signal to Noise Ratio (ISNR) equivalent to 60 dB. For the estimation of the PCR in (11), we consider a value $\text{RSNR}_{min} = 55$ dB. Finally, the instances are reconstructed solving (1) with the SPGL1 toolbox[1]. Tuning of c for both NeO-CS and NOR-CS is done in order to reduce as much as possible the number of measurements required to obtain a PCR $= 0.95$. We refer to each of them value as m_{min}.

1 This tool is available online in https://www.cs.ubc.ca/$\sim$mpf/spgl1/download.html

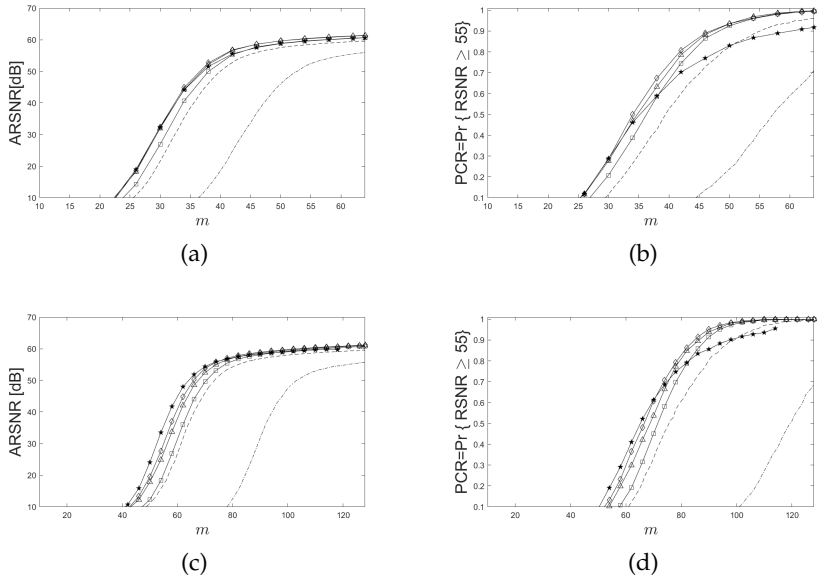


Figure 9: Performances in terms of ARSNR for a) n = 128 and c) n = 256 and PCR for b) n = 128 and d) n = 256 between the Std-CS and the optimized CS approaches. NeO-CS and NOR-CS are evaluated with the best values of c reported in the Table 2.

Performances are shown in the Figure 9, where we can observe that the ARSNR and the PCR of all the optimized methods mentioned above have a better performance than the Std-CS approach, that is why from now only the optimized methods will be considered. In Figures 9 c) and 9 a) we can observe a similar performance in ARSNR for NeO-CS and NOR-CS, both were evaluated with the best value of c reported in the Table 2. Also, we can observe in the figure that NeO-CS and NOR-CS slightly outperform the Rak-CS. Interestingly, the Figures 9 b) and 9 d) shown a notably better performance in terms of PCR for NOR-CS compared with the other methods. Same results are in the Table 2, where NOR-CS shows a very positive impact in the reduction of m_{min}.

In the Figure 10 a) we can observe how the localization is reduced by the parameter l in the process to generate the matrix A by three different versions of NOR-CS. This localization profile corresponds to the setting that provides the better performances reported in the

Table 2: Best performances for NeO-CS and NOR-CS.

Approach	l_0	$n = 128$		$n = 256$	
		c	m_{min}	c	m_{min}
NeO-CS	1.0	0.2031	51	0.2031	104
NOR-CS	0.5	0.1250	39	0.1328	85
	1.0	0.1406	38	0.1484	81
	1.5	0.1406	39	0.1328	83

Table 2 with $n = 256$, the constant values of l for the NeO-CS and Rak-CS are included only as a reference. In addition, we can observe in the Figure 10 b) the average of the number of iterations $\mathbf{E}[Z]$ as a function of m according with the localization profile showed in the Figure 10 a). As was expected, the computational effort to generate a new row with NOR-CS is approximately constant after a certain number of row generation.

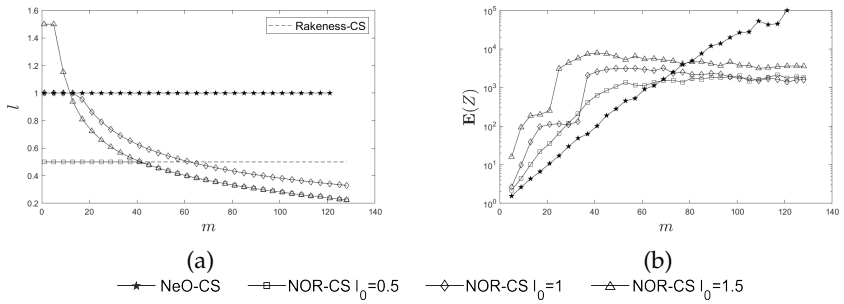


Figure 10: a) Values of l along the matrix A generation, b) Average of the number of iterations $\mathbf{E}[Z]$ of the optimized CS methods with geometric constraints.

According with the Table 2 the NOR-CS with $l_0 = 1$ presents a slight outperformance from the other versions (at least dealing with synthetic signals). For this reason, from now we refer generally to NOR-CS implying $l_0 = 1$. In Table 3 we also report values of c that maximize the PCR for same reference values of m. This table includes the corresponding final l values as well as performance for the Rak-CS where the observed average values of c are included. This is also

the case of the last proposed comparison in Figure 10. Here, c is the value that maximize the performance for each considered number of rows in A. The Figure 11 a) shows that the NOR-CS approach presents a performance not different from NeO-CS in terms of ARSNR and both NeO-CS and NOR-CS have an outperforming with respect to the Rak-CS. However, the Figure 11 b) provides a considerable better performance of the NOR-CS in terms of PCR than NeO-CS and Rak-CS.

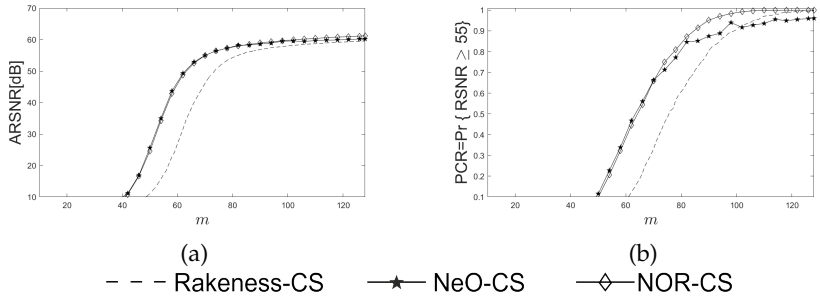


Figure 11: ARSNR for and PCR with the optimum values of c that maximize the performance in each row.

Table 3: Best value of c to obtain the best PCR in each row where: i) l values for NOR-CS indicate the observed final values with $l_0 = 1$; ii) performance for Rak-CS include observed average value of c, $\mu(c)$.

	NOR-CS			**NeO-CS**			**Rak-CS**		
m	c	l	PCR	c	l	PCR	$\mu(c)$	l	PCR
80	0.1328	0.3621	0.8455	0.1875	1	0.7981	0.3699	0.5	0.6075
90	0.1484	0.4319	0.9520	0.1953	1	0.8745	0.3705	0.5	0.8050
100	0.1406	0.3439	0.9870	0.2031	1	0.9094	0.3710	0.5	0.9065
110	0.1484	0.3806	0.9995	0.2031	1	0.9359	0.3713	0.5	0.9705

3.3 TEST IN SYNTHETIC ECG SIGNALS

The approaches are tested also with electrocardiographic (ECG) signals. This real-life signals comply with the requirements of highly localization and sparsity; at the same time, they are the center of in-

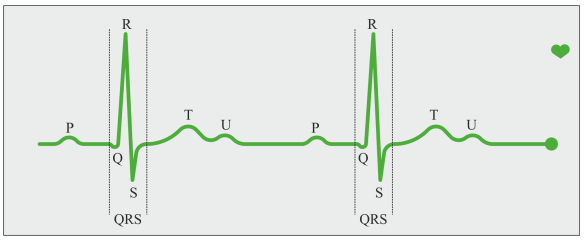


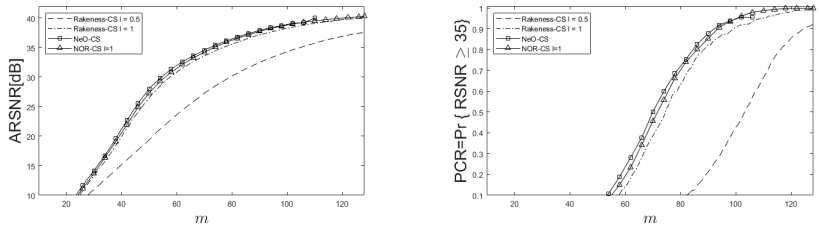
Figure 12: CS based system.

terest for the improvements in the bio-medical prototypes. An ECG signal is a graphical representation of the potential generated by the activity of the heart. This signal is obtained by electrodes attached to the surface of the skin in specific places. A cycle of ECG signal is composed by the P wave, which represents the depolarization of the atria, the QRS complex, which represents the ventricular depolarization, and the T wave, which represents the repolarization of the ventricles [46] as is shown in the Figure 12.

The ECG signal is obtained by a synthetic generator discussed in [52]. This generator provides signals without noise, the Gaussian noise is added empirically, and it is chosen considering an environment realistic. The Gaussian noise is added with 40 dB SNR. The heart rate is randomly selected in the range of 50 and 70 Hz and a sampling rate = 256 Hz. The sparse basis is obtained with the Symlet-6 Wavelet. For each value of c, the number of iterations is limited to $Z_{max} = 10^5$, for the Rak-CS the typical value $l = 0.5$ is used

For the estimation of the PCR in (11), we consider a value $RSNR_{min} = 35$ dB. Also, c is tuned for both NeO-CS and NOR-CS to reduce as much as possible the number of measurements m_{min} required to obtain a PCR $= 0.95$. Performances in terms of ARSNR and the PCR of the optimized methods are shown in the Figure 13. NeO-CS and NOR-CS are evaluated with the best values of c reported in the Table 2. Particularly, in this kind of signal we can observe some interesting things. There is no considerable difference of m_{min} and the best value of c is relatively large between NeO-CS and NOR-CS, which means

geometric constraints do not have a big impact to improve the performance both in PCR and ARSNR. This can be proved by evaluating the Rak-CS with $l = 1$. The explanation is because this type of ECG is highly localized, that is, most of the signal energy is concentrated in a specific region in the signal space.



(a) (b)

Figure 13: Performances in terms of a) ARSNR and b) PCR for $n = 256$ between the optimized CS approaches.

It is necessary to evaluate if the slight increase in performance is worth the computational effort required by the geometric constraints based CS adaptations. At this point we have had a successful result using synthetic low-pass signals and not very blunt result with ECG synthetic signals. Therefore, bellow a test with a less localized signal will be proposed.

3.4 TEST IN EEG SIGNALS

In addition, the reconstruction of electroencephalograph (EEG) signals is tested to demonstrate the effectiveness of the NOR-CS. An EEG is a set of signals recorded from several electrodes on the scalp to analyze the brain activity. These signals provides information to identify different brain conditions and is useful to monitor the patient's health and diagnosis in the neuroscience, cognitive science and cognitive psychology areas. In particular we focus on Evoked Potentials (EPs), that consist in recordings of the electrical activity of the brain

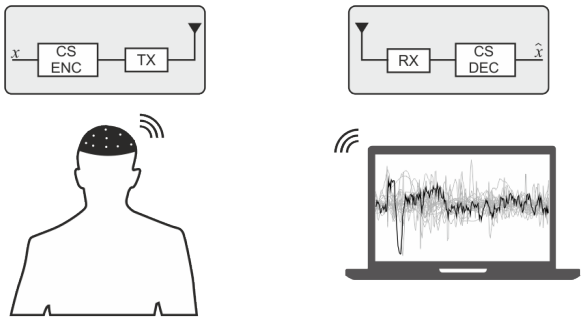


Figure 14: CS based system.

following an repetitive auditory stimulus with a time interval of 1 s between them to evaluate the auditory perceptual threshold [67, 68]. However, the individual responses of a spontaneous EEG are visually indistinguishable, for this reason the analysis of EPs suggests that the ensemble of the signals post-stimuli (epochs) has to be averaged to detect the response of the auditory stimulus. A possible scenario is showed in the Figure 14, where the EEG of the patient is collected by sensors of a battery-powered system. Due to the amount of information that each channel generates, the compression stage plays an important role for energy saving. It is proved that systems based in CS techniques are characterized by a low energy consumption in portable devices because the rate of the output data is reduced considerably [51, 57]. A properly designed sensing matrix A could greatly reduce the data to be transmitted or stored without compromising the quality of the input signal (in this case without compromise the correct EEG interpretation).

The data set consists in EP recordings from a normal-hearing patient subjected to an auditory task. Basically, the test consists of listening speech syllables in one second intervals. The EEG acquisition was collected by 23 channels according the International 10/20 system of electrode placement and two channels used as noise reference to reject ocular artifacts. Each channel is organized in 700 epochs divided in two parts, the first 350 epochs are used as Training Set (TS), while

the second half named Data Set (DS) is used to test the CS-based approaches. According to [68] the channel Cz gives the strongest auditory response. However, also the reconstructions of the channel FC6 are compared . The associated correlation matrix is calculated by (6) from the average denoised signal of the TS as in [4] using [83–85] to remove the artifacts.

To estimate the reconstructions quality, the Mean Squared Error (MSE) is used $MSE = \frac{1}{n}\sum_{i=0}^{n}(\mathbf{x_i} - \hat{\mathbf{x}}_\mathbf{i})^2$ that is, the average squared difference between the original signal $\mathbf{x}$ and the reconstructed signal $\hat{\mathbf{x}}$, where $\mathbf{x}$ represents the average of the raw signals and $\hat{\mathbf{x}}$ is the average of the reconstructed signals. The MSE is estimated for three values of $CR = \{4, 8, 16\}$, that is, $m = \{64, 32, 16\}$. Results in the Tables 4 and **??**confirm that all optimized CS approaches have a better performance than the Standard CS. Remarkably, approaches based on geometric constrains present a better performance in terms of MSE than the Rak-CS. However, the proposed NOR-CS approach with $l_0 = 1.5$ presents the best performance among the optimized CS approaches. Note that for MSE, lower values correspond to a higher quality.

Table 4: Performance of the CS methods on EEG signals (Channel Cz).

	Std-CS	**Rak-CS**	**NeO-CS**		**NOR-CS**	
m	MSE (μV^2)	MSE (μV^2)	MSE (μV^2)	c	MSE (μV^2)	c
16	1.3065	0.8274	0.6954	0.1875	0.5859	0.1641
32	1.2055	0.6793	0.4246	0.1953	0.3827	0.1875
64	0.7580	0.3175	0.1852	0.1641	0.1435	0.2266

To visualize the performance related to the acquisition and reconstruction of the EPs, consider the case with $m = 16$. Figures 15 a) and 15 a) show that the matching between the average reconstructed EPs using the Std-CS is very poor. The matching is improved using the Rak-CS as we can observe in the Figures 15 b) and 16 b); however, Figures 15 c) and 16 c) show that the quality of the reconstruction

Table 5: Performance of the CS methods on EEG signals (Channel FC6).

	Std-CS	**Rak-CS**	**NeO-CS**		**NOR-CS**	
m	MSE (μV^2)	MSE (μV^2)	MSE (μV^2)	c	MSE (μV^2)	c
16	1.0953	0.7290	0.6281	0.2344	0.6043	0.1484
32	0.9114	0.6326	0.5394	0.2344	0.5146	0.1484
64	0.7299	0.4732	0.4215	0.2109	0.3450	0.2109

using the Rak-CS is outperformed by the NeO-CS. As highlighted by the Figures 15 d) and 16 d), the best performance is given by the NOR-CS, where the peaks of the auditory stimulus are easier to distinguish than those given by the other approaches.

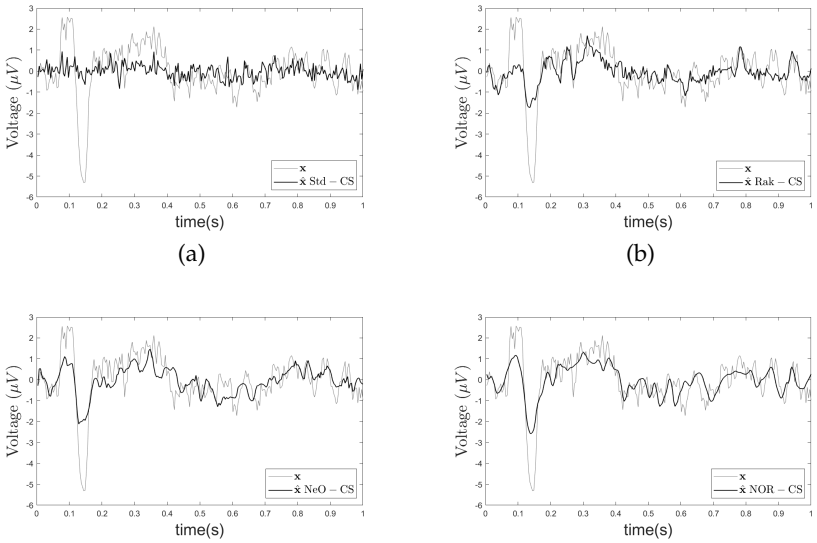


Figure 15: Comparison between the Average of the EEG raw signals $\mathbf{x}$ and the average of reconstructed signals $\hat{\mathbf{x}}$ with m = 16 for a) Std-CS, b) Rak-CS, c) NeO-CS and d) NOR-CS for the channel Cz.

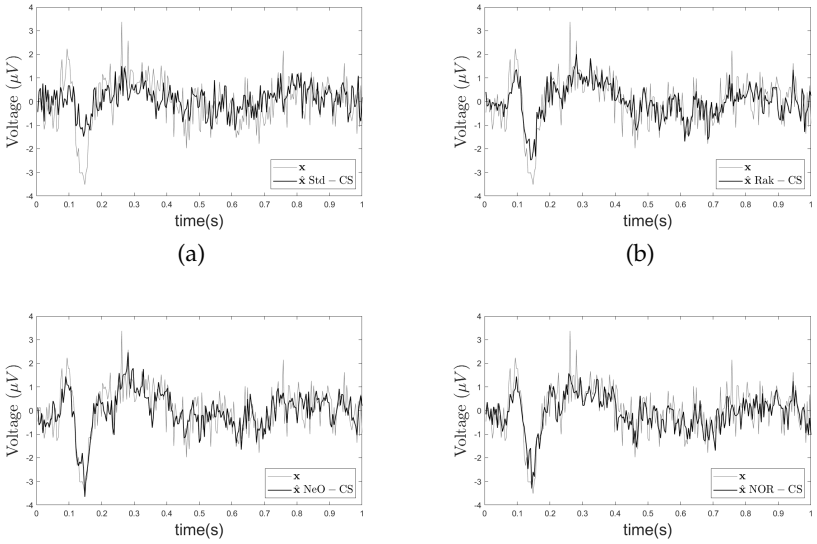


Figure 16: Comparison between the Average of the EEG raw signals $\mathbf{x}$ and the average of reconstructed signals $\hat{\mathbf{x}}$ with m = 16 for a) Std-CS, b) Rak-CS, c) NeO-CS and d) NOR-CS for the channel FC6.

4

CONCLUSION

In this first part, a critical review of the state of the art of the optimized CS methods in the sensing stage (Rakeness-based CS and Nearly Orthogonal-based CS). Advantages and limitations using these methods to generate the sensing matrix are discussed. In addition a new algorithmic solution named Nearly Orthogonal Rakeness-based CS is proposed with the aim to overcome the limitations founded in the methods reviewed. This technique allows to exploit the geometric constraint as much as possible characterizing the sensing sequences of A with an adapted localization. After intensive simulations with synthetic low pass signals and electroencephalographic signals, the performances of all the methods discussed in this paper were compared. Results shows a remarkably better performance in therms of PCR of the proposed approach using synthetic low pass signals. Also, results demonstrate that electroencephalographic signals reconstructed with the proposed method present the best quality regardless the compression ratio. For this reason, using a high compression ratio, NOR-CS it can be considered as a suitable method to identify Evoked Potentials.

Part II

DNA ANALYSIS

This part provides general concepts of biology and current developments in bioinformatics. Questions like what a sequencer does?, What kind of data it produces? and How is analyzed this data? are intended to be answered. An important fact about human genomes is that, if two genomes corresponding of two unrelated human beings are compared, the sequences are very similar (about 99.8% - 99.9% similar), that is about one or two differences every 1000 bases, for this reason the someone else's genome can be used as kind of template. The Human Genome Project was an international scientific research project to identify and map the human genome, this can help us to understand how different are two human genomes? and how is the genetic predisposition linked to different diseases?. This developments in bioinformatics use powerful computational techniques to align the assemble the data with a reference sequence to find mutations, gene expression, or single nucleotide polymorphisms (SPNs), that are the most common genetic variations. Another reliable technique used in the analysis of the DNA without sequencing is the High Resolution Melting (HRM) analysis, this technique is used to find differences between two strands of DNA. HRM is also discussed to finally design a HRM analysis software.

5

INTRODUCTION

Deoxyribonucleic acid (DNA) is the biological molecule that stores genetic information in the cell of each organism. The DNA is a *polymer*, which means that it consists of repeating units called *nucleotides*. Each nucleotide consists in three parts: a *deoxyribose sugar*, a *nitrogenous base* and a *phosphate group*. There are four different types of nitrogenous bases: Adenine, Guanine, Cytosine and Thymine (A, G, C, T). Adenine and Guanine are *purines* while cytosine and thymine are *pyrimidines*. Two nucleotides of the same strand are connected via *phosphodiester bond* between the 3rd carbon of the deoxyribose sugar of one of them and the 5th carbon of the deoxyribose sugar of the next nucleotide. In a like manner, a single strand of DNA (ssDNA) can combine with another complementary (hybridization) to form the double helix structure typical of DNA. This connection is possible with *hydrogen bonding* between the nitrogenous of each base forming the double-stranded DNA (dsDNA). Guanine-Cytosine (GC) pairs are bounded by three hydrogen bonds, while Adenine-Thymine (AT) pairs are bounded by only two hydrogen bonds. The structure of de DNA is illustrated in the Figure 17.

The human genome is organized into a 23 pairs of chromosomes (maternal and paternal set) that encode in a kind of recipe the functional proteins. Amazingly, the human genome contains an approximate of six billion nucleotides, and all the ~trillion cells of the human body of different types (blood, skin, hair, brain. etc.) contain the same *molecular blueprint*. Due to the importance of DNA to life has spurred developments to obtain new methods for the analysis

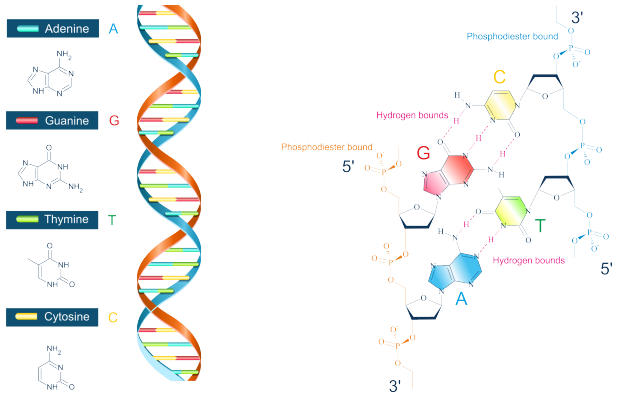


Figure 17: Structure of the DNA.

and interpretation of this molecule. Sequencing Technologies used to read the genome can be classified in two generations. First generation sequencing whit the Sanger (Chain termination sequencing) [30] and the second generation (2007) with the NGS (Next-Generation Sequencing).

5.1 DNA SEQUENCING

In order to explain the DNA sequencing consider the representation of DNA as a chain of lego pieces as is shown is the Figure 19.

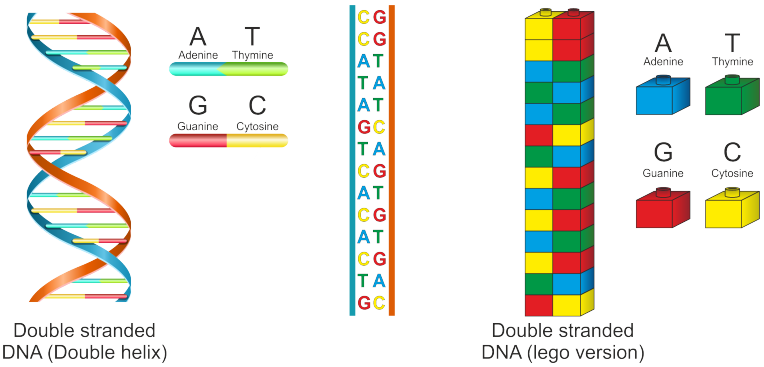


Figure 18: DNA representation.

In the DNA replication process two identical replicas of DNA are produced from one original DNA molecule. First, the double-stranded DNA splits into two complementary strands (Figure 19).

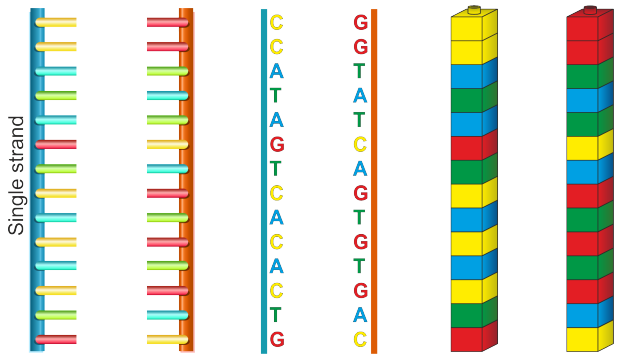


Figure 19: Representation of DNA denaturation.

Each single-strand acts as kind of template for the production of its counterpart. There are many proteins associated with the DNA replication, but the most prominently is the DNA polymerase. This enzime synthesizes the new strands by adding complementary nucleotides. That is, given one of these single-stranded templates DNA, polymerase biulds the complementary strand piece by piece resulting in a new double-stranded. If this process is done this for both of the template strands, two double stranded copies of the original DNA are obtained. Following the lego example, this process is shown in the Figure 20.

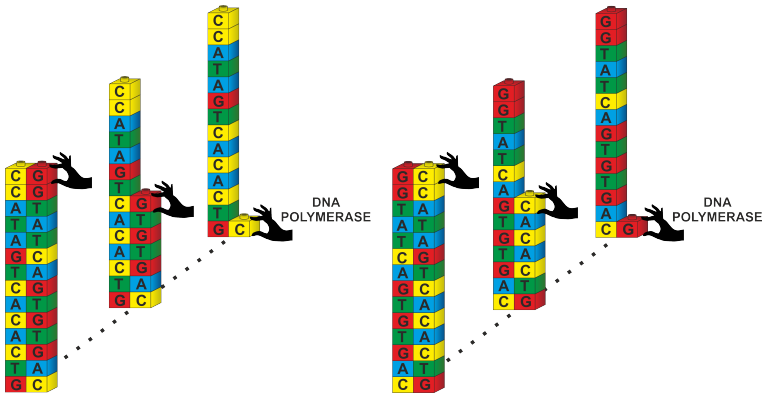
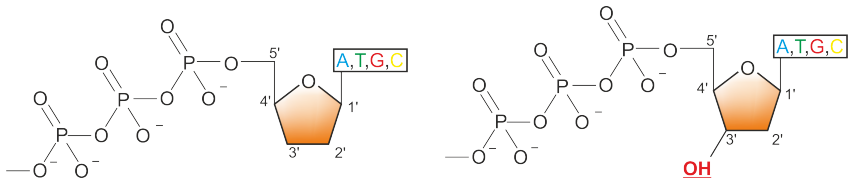


Figure 20: Representation of DNA replication.

5.1.1 *Sanger sequencing*

Sanger sequencing is a method for determining nucleotide sequences in DNA developed by Frederick Sanger in 1977 [30]. This technique incorporates dideoxynucleotides which act as specific chain-terminating inhibitors of DNA polymerase. This method requires a single-stranded DNA template mixed with a DNA primer, DNA polymerase, normal deoxynucleotidetriphosphates (dNTPs) and modified dideoxynucleotidetriphosphates (ddNTPs). The dNTPs differ from ddNTPs because the latter lack a 3'-OH group required to form the phosphodiester bond between two nucleotides as is illustrated in the Figure 21



(a) (b)

Figure 21: a) Dideoxynucleoside trhyphosphate (ddNTP), b) Deoxinucleotide 5' triphosphate.

The ddNTPs are radioactively or fluorescently labelled for detection in automated sequencing machines. In general terms, Sanger sequencing can be summarized as follows:

- Figure 22 shows the first step, where the double helix structure of the DNA is separated (denaturized) with sodium Hydroxide NaOH and a single strand is chose.

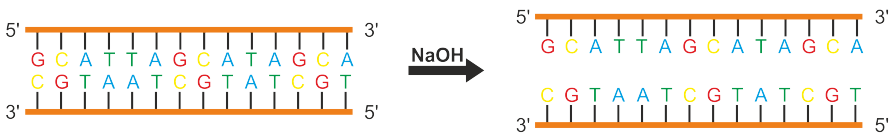


Figure 22: Denaturation of the DNA.

- Then, the couple of single strands of DNA are mixed with a labeled DNA primer, DNA Polimerase, all four types of deoxinucleotide 5′ triphosphates (dATP, dGTP, dCTP, dTTP) and a tiny quantity of one specific Dideoxynucleoside trhyphosphate (ddATP, ddTTP, ddCTP or ddGTP) (approximately 1% respect to the deoxinucleotide 5′ triphosphates).

- Previous step is repeated for the other three remaining ddNTPs.

- Once the four reactions are completed, the resulting DNA fragments are heat denatured and separated by size using gel electrophoresis. Each reaction mixture is placed into a lane to produce a total of four lanes. The results are transferred onto a polymer sheet, which is then exposed to x-ray autoradiography. The relative positions of the different bands among the four lanes, from bottom to top, are then used to read the DNA sequence. Figure 23 shows the Sanger sequencing process.

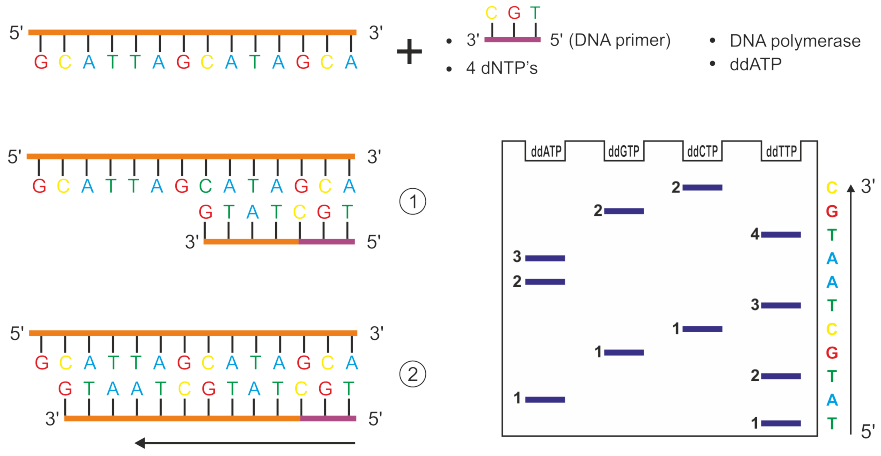


Figure 23: Sanger sequencing process.

The Human Genome Project (HGP) was an international scientific research project with the aim of read the entire sequence of nucleotide base pairs of all the human genome, as well as identify and mapping all the genes (20,000-25,000), it took over 13 years with

a cost nearly $3 billion dollars using capillary electrophoresis-based Sanger sequencing. The HGP can help us understand diseases including genotyping of specific viruses to direct appropriate treatment, identification of mutations, the design of medication, advancement in forensic applied sciences, agriculture, animal husbandry, bioprocessing, bioarcheology, anthropology and evolution.

5.1.2 *Next-Generation Sequencing*

The principle of the Next-Generation Sequencing (NGS) is similar to that of the Sanger sequencing method, where DNA polymerase adds fluorescent nucleotides one by one onto a DNA template strand identified by its fluorescent tag. The difference between the both methods is the sequencing volume. The advantage of NGS method, known as shotgun sequencing, is that the genome can be fragmented and sequenced in parallel. Massive amounts of data in less time and lower cost it can be produced whit NGS, while Sanger method only sequences a single DNA fragment at a time.

At the beginning, the cost of the NGS method was very high; however after the HGP the cost of the NGS has decreased incredibly, outpacing Moore's Law. In 2014, over 45 human genomes were sequenced in a single day for approximately $1000 each. When is evaluated NGS costs, the sample volume for the study has to be considered. In general, for analyzing only a few (< 20) targets on a few samples, Sanger sequencing is more useful. For sequencing more than 20 target regions or high sample volumes, NGS is better. In the Figure 24, the area above the line represents higher cost-effectiveness with targeted DNA sequencing compared to Sanger sequencing.

In NGS methods, the DNA polymerase adds of fluorescently labeled deoxyribonucleotide triphosphates (dNTPs) into a DNA tem-

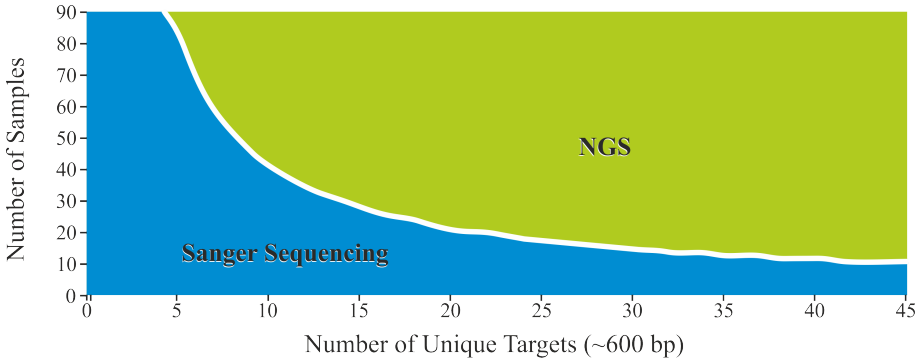


Figure 24: Sanger sequencing vs NGS.

plate strand during sequential cycles in the DNA synthesis. During each cycle the nucleotides are identified by fluorophore excitation. Sequencing Approaches for NGS:

- Pyrosequencing.
- Sequencing by synthesis.
- Sequencing by Ligation.
- Ion Semiconductor Sequencing.

More than 90% of the world's sequencing data are generated by Illumina sequencing by synthesis(SBS) chemistry. SBS can be summarized in four steps:

- *Library Preparation:* NGS library is prepared by fragmenting a DNA sample and ligating specialized adapters to both ends.
- *Cluster Generation:* Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification. For example, consider a sequence of DNA obtained with a blood sample. In this blood sample there are a lot of cells each with a copy of the genome. Now, imagine that this DNA string is formed by the sequence: CCATAGTATATCTCGGCTCTAGGCCCTCATTT (commonly the sequences are much

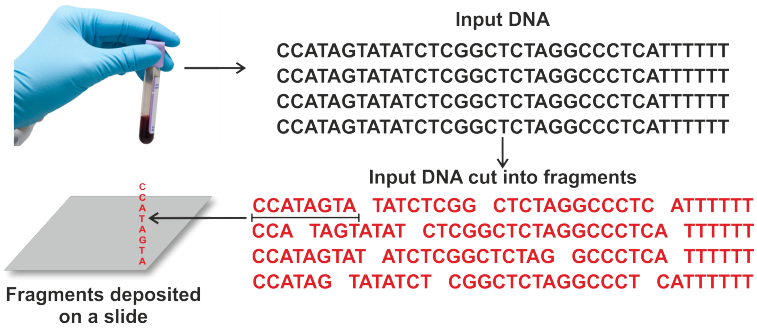


Figure 25: Sequencing step.

longer). In the laboratory the DNA is extracted and fragmented into little pieces and they taken those short single-stranded templates and deposit them randomly across a sort of flat surface like a slide. The Figure 25 shows only one template; however the slide goes way out in every direction and there are many template strands attached to the slide.

- *Sequencing:* Labeled nucleotides are added one by one in a repeated cycles, where the cluster is imaged recorded. The emission wavelength is used to identify the base. In the Figure 26 there is a LEGO version (slice in gray and three template strands) to explain this step.

  Each read is random and is probabilistically expected to overlap another fragment such that, in theory, the entire genome can be assembled by algorithms to reconstruct the genomic sequence of an organism comparing the similarity of the overlapping reads and pasting these together into increasingly larger, contiguous sequences called contigs. There are three paradigms for genome assembly: *i)* Greedy constructions of contigs, *ii)* Overlap layout consensus (OLC), *iii)* De Brujin graphs.

- *Data Analysis:* The reads are aligned to a reference sequence by software. After the alignment, differences between the reference genome and the sequenced reads are identified. These reads are

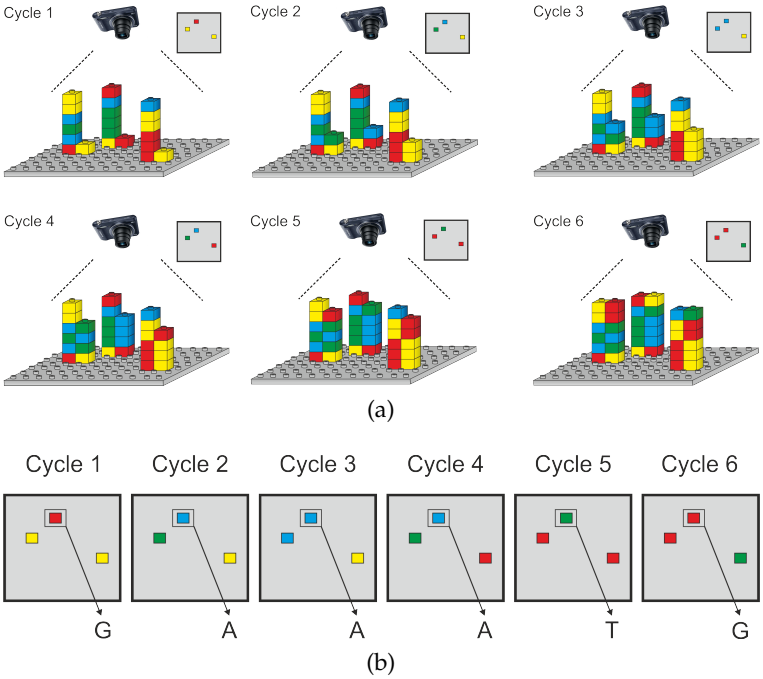


Figure 26: a) Sequencing process, b) Images produced by six cycles.

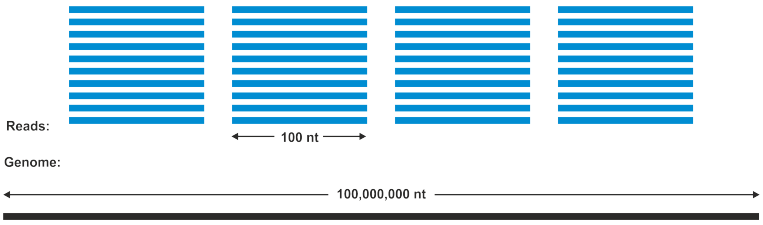


Figure 27: Reads.

very short compared to the length of the input DNA (Figure 27). For example massively parallel sequencers produce about 100 or 250 base pairs (bp) long, one human chromosome is on the order of hundred millions bp long. However, this technology provides millions and millions of these reads, enough to cover the whole genome many times over, in other words a lot of redundant information.

Differences between genomes or the predisposition to disease can be analyzed by sequence alignment. Nevertheless, it is no easy. This is analogous to putting back together a puzzle, but with a guide of the picture of the completed puzzle as is repre-

Figure 28: Puzzle analogy of sequence alignment

sented in the Figure 28. An important fact about human genomes is that two unrelated persons have very similar genome sequences (about 99.8% - 99.9%), that is about one or two differences every 1000 bases. For this reason, is used the genome sequence of some person as kind of template or a guide.

Despite the fact that Next Generation Sequencing (NGS) is a powerful tool to read DNA, there is a more accessible, affordable and faster method for genomics analysis called High Resolution Melting (HRM) curves capable to find DNA variations without sequencing. Melting analysis by fluorescence was introduced in 1997 as a method to differentiate Polymerase Chain Reaction (PCR) products [63] (Appendix B). Nowadays, thanks to the advances on instruments, fluorescent dyes and software for DNA melting analysis, HRM analysis has become in a stable and reliable method used for fast identification of variants in regions of interest without sequencing [34, 44, 55, 56, 61, 74]. Clinical applications of HRM analysis includes single base change genotyping, mutation, zygocity, transplant compatibility and gene espresion [28, 29, 31, 81]. The advantages that HRM offers (faster and most economic than sequencing) increase its application in plant research, in particular in food sector e. g. authenticity procedures in food products and seeds [60, 70, 73].

6

HIGH RESOLUTION MELTING CURVES ANALYSIS

6.1 HRM ANALYSIS

HRM is based in a fundamental propriety of the DNA: *melting* i.e. the propriety of thermal separation (denaturation) and annealing (hybridization) of its double-stranded helical state (dsDNA) and its single-stranded random coil state (ssDNA). First, Polymerase Chain Reaction (PCR) is used prior to HRM analysis in order to amplify the DNA region of interest (amplicon). After that, specialized dsDNA binding dye is added and the DNA section is gradually denatured by increasing temperature, typically in a range from 55 to 95 °C in steps of 0.01 to 0.2 °C, where high fluorescence is observed in presence of dsDNA and poorly fluorescent in the unbound state. At the end of the experiment, a file with the values of the fluorescence in function of the temperature of each amplicon is generated and HRM curves analysis begins.

The principal aim to the HRM analysis software is extract the significant clinical information extracting the melting curve (background subtraction and normalization), melting temperature identification T_m (commonly obtained with the negative first derivative) of these raw signals and improve methods for clustering and classifying the results [42, 43, 55, 56]. Also, in order to design a reliable assay in applications in which the sequence is known (e. g. designing primers), run the amplicon through software to predict melting curves with thermodynamic models to know the number of peaks in the deriva-

tive curve could be useful to discard a possible contamination and not repeat the experiment. [24, 25].

In general the melting curve profile depends of the length of the strand, GC content, methylation in the CpG islands, sequence and heterozygosity of the target.

The general flowchart of the HRM analysis is showed in the Figure 29

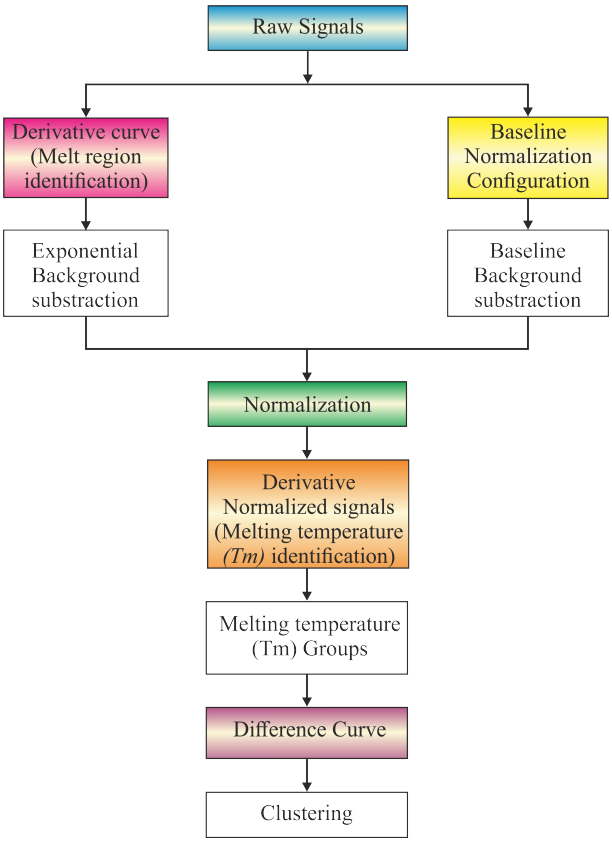
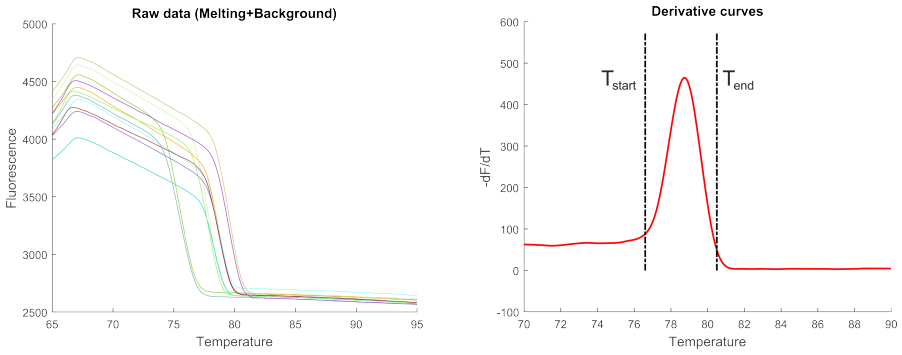


Figure 29: Flowchart of the HRM analysis.

6.1.1 *Melting region identification*

Figure 30 a) shows the raw data, this signal has to be processed to obtain the significant information. Melting region temperature curves represents the proportion of DNA that is denatured. This section of the curve is characterized by a significantly increase of the slope and

finish when the slopes is nearly flat. The task is the location of the melt start temperature T_{start} and the melt end temperature T_{end}, this is showed in the Figure b). Identification of the melting section is absolutely necessary in order to subtract the exponential background of the raw signal. Commonly, the method used to find the melting region is based in the angle formed by the curve of the negative first derivative curve of each curve, this identification can be manually or with algorithms help. Application of smooth and filtering functions are recommended in order to decrease the noise produced by digital differentiation, Savinsky-Golay smoothing filter is used by [55, 56]. In this work, the angle corresponding to the T_{start} is $\theta_1 = 50°$ and for T_{end} is $\theta_2 = 20°$. T_{start} and T_{end} are founded by fitting a line in certain number of points (by default the number of points contained in 1 °C) in the negative first derivative curve of the raw data, the angle formed by the inverse tangent between the points and the baseline of the x-axis is evaluated repeatedly until $\theta_1 = 50°$ and $\theta_2 = 20°$. In the program, the user can modify θ_1, θ_2 and the window of points used in the fitting.



(a) (b)

Figure 30: a) Raw curve, b) Melting region identification.

6.1.2 *Background subtraction and normalization*

As is shown in the figure 30 a), Differences between raw curves belonging to different genotypes are difficult to observe. For this reason, Background subtraction is required in order to extract the melting curves of the raw curves [55, 56, 66]. After that, normalization is necessary to get a better comparison between the probes. Baseline method and exponential Background subtraction (EBS) are the two major methods to extract melting curves from raw data.

6.1.2.1 *Baseline Background subtraction*

Baseline method consists in the linear approximation of the range above and below the melting transition $L_0(T)$ and $L_1(T)$ respectively. The baseline method uses the following approximation:

$$M(T) = \frac{F(T) - L_0(T)}{L_1(T) - L_0(T)} \tag{12}$$

where $M(T)$ estimates the melting curve.

Sometimes Base-linear methods fails when L_0 and L_1 intersect below the graph of $F(T)$ due to its concavity, and the denominator in (12) goes to zero, leading divergence of M(T). In these case an exponential model to subtract the background produces better results.

6.1.2.2 *Exponential Background subtraction*

In the exponential background removal method, the raw melting signal $F(T)$ is modeled as the sum of the Melting curve $M(T)$ plus an exponential decaying background $B(T)$

$$F(T) = M(T) + B(T). \tag{13}$$

Basically, the exponential background subtraction is resumed in the following steps:

- First, two temperatures $T_L = T_{start}$ and $T_R = T_{end}$ have to be identified. Fitting the slope of $F(T)$ at these temperatures, where T_L sufficiently below and T_R sufficiently above from the melting transition temperatures so that the slope is not significantly affected. This is quantified in therms of a radio of derivatives of F(T) constant for a pure exponential plus a constant, such that at T_L and T_R the melting curve component $M(T)$ are effectively zero

$$\frac{dM}{dT}(T_L) = \frac{dM}{dT}(T_R) = 0 \tag{14}$$

  and the slope of the raw fluorescence is completely attributable to the slope of the background. Differentiating (13)

$$\frac{dF}{dT} = \frac{dM}{dT} + \frac{dB}{dT}, \tag{15}$$

  and joined with (14)

$$\frac{dF}{dT}(T_L) = \frac{dB}{dT}(T_L) \quad \text{and} \quad \frac{dF}{dT}(T_R) = \frac{dB}{dT}(T_R). \tag{16}$$

- These slopes are used to fit the background by the exponential function

$$B(T) = B_0 + Ce^{a(T-T_L)} \tag{17}$$

where the argument of the exponential is shifted to T_L for numerical stability. The parameters a and C are obtained first differentiating (17)

$$\frac{dB}{dT} = aCe^{a(T-T_L)} \tag{18}$$

and then, evaluating (18) at T_L

$$aC = \frac{dF}{dT}(T_L) \tag{19}$$

and T_R

$$aCe^{a(T_R-T_L)} = \frac{dF}{dT}(T_R). \tag{20}$$

Dividing (19) and (20) and taking the logarithm to obtain

$$a = \frac{\ln\frac{dF}{dT}(T_R) - \ln\frac{dF}{dT}(T_L)}{T_R - T_L} \tag{21}$$

and substituting a in (19)

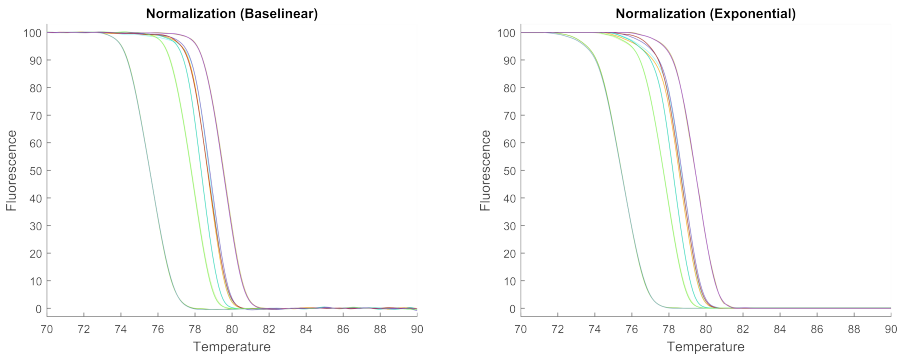
$$C = \frac{\frac{dF}{dT}(T_L)}{a} \tag{22}$$

- The melting signal is obtained with the background subtraction as

$$M(T) = F(T) - Ce^{a(T-T_L)}. \tag{23}$$

Finally,the extracted melting curve is normalized as is showed in the Figure 31. Commonly, the melting curve is re-scaled by 100 to visualize as a percentage with the equation

$$M(T)_{normal} = \frac{100(M(T) - \min(M(T))}{(\mathrm{Max}(M(T) - \min(M(T))}. \tag{24}$$



(a) (b)

Figure 31: a) Baselinear background subtraction, b) Exponential background subtraction.

6.1.3 *Melting temperature* (T_m) *identification*

The melting temperature T_m is the temperature where the 50% of the sample is dsDNA and 50% is ssDNA. Melting temperature T_m is basically the highest slope of the melting curve, and is determined as the peak of the negative derivative curve as is showed in the Figure 32.

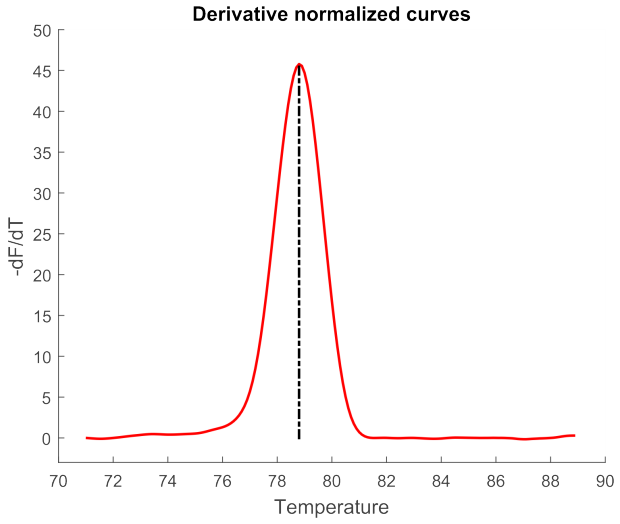
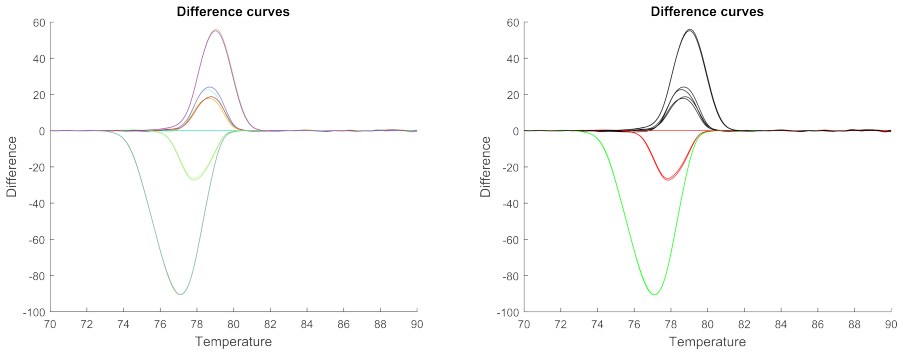


Figure 32: Melting temperature T_m.

6.1.4 *Difference curve*

Other analytical graphic is represented by the Figure 33, it is useful to make comparisons between the melting curves is the difference curve. This graph helps to distinguish between genotypes subtracting a known genotype considered as the reference (also in this program is supported two or more reference curves) from the another curves.



(a) (b)

Figure 33: a) Difference curve, b) Clustering.

6.1.5 *Clustering and classifying melting curves*

Although it is true that the experience of the researcher is very important in order to identify the differences between the probes with the previous graphs, a flexible software for grouping and classify the results is necessary. In general clustering is based in computing the distances between the curves [43, 56]. Automated methods for classification of the genes are presented in [1, 37, 62]. In this work the clustering is done from the difference curve using the k-means function of Matlab as is shown in the Figure 33 b).

7

CONCLUSION

The design of an High resolution Melting Analysis software with the AppDesigner of Matlab was presented in this second part. To do that, a brief overview in DNA sequencing techniques is provided in this work. In the software developed, Baselinear and Exponential background subtraction is supported, grouping based in the melting temperature identification T_m, average of targets in the difference graphic, Bio-Rad CFX platforms supported and at the same time is very easy to use.

Part III

APPENDIX

A

GLOSARY

- *Primer:* Is a strand of short nucleic acid sequences that serves as starting point for DNA syntesis
- *Nucleotide:* One of the structural units of DNA composed by a sugar, phosphate group and a base.
- *Bases:* Nucleotides (A (Adenine),G (Guanine),C (Cytocine), and T (Thymine).
- *Base pairs (bp):* A measure of length of DNA using the number of nucleotide pairs.
- *Hydrogen bond:* In a Hydrogen bond, a hydrogen atom is shared by two electronegative atoms. These are the strongest type of intermolecular bonds. The group that contains the H-atom is the H-bond donor while the group that accept it is the H-bond accepter.
- *Contigs:* A stretch of continuous sequence, in silico, generated by aligning overlapping sequencing reads.
- *Oligonucleotide:* A short DNA or RNA sequence.
- *Genotype:* The set of genes of an organism or individual, which determines one of its characteristics.
- *Phenotype:* They are the observable physical properties of an organism.
- *Zygosity:* Is the degree of similarity of the alleles for a trait in an organism.

- *Heterozygote:* Two different alleles in the two chromosomes of a gen in a diploid organism.
- *Homozygote:* Neither allele contains a mutation in the two chromosomes.

B

POLYMERASE CHAIN REACTION

Polymerase chain reaction (PCR) is a method to make amplify DNA segments (Genes duplications). This method was developed by Kary Mullis in 1983 [53] and, nowadays is extremely useful because millions or billions of copies of a particular DNA segment can be generated in a short period of time. This DNA segment of interest to amplify can be very long (i.e. 10,000 nucleotides) and commonly it is not necessary to know the sequence of DNA segment to be amplified (target). However, PCR absolutely requires the knowledge of the flanking sequences of the segment of interest. PCR requires several components and reagents, where the most remarkable are: The primers (Two different primers are included in the solution one for each flanking sequence), DNA polymerase and the deoxynucleoside triphosphates or dNTPs

The flanking sequence is the segment of DNA that we do not want to replicate. It flanks the target DNA.

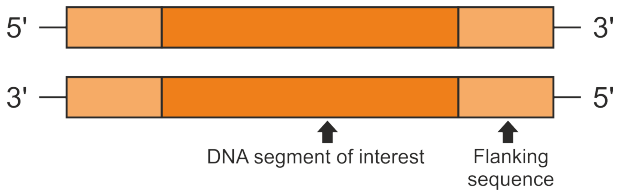


Figure 34: DNA sequence conformed by a DNA segment of interest and a flanking sequence.

Almost all the PCR methods rely on thermal cycling, specifically DNA melting and enzyme-driven DNA replication. Basically, the PCR methods can be summarized as follows:

- *Denaturation* of the double helix. This step consists of heating the dsNDA molecule in to a range of $95 - 98°$ C for about 15 seconds. This gives enough time to separate double-stranded

DNA template by breaking the hydrogen bonds between complementary bases (DNA melting or denaturation), yielding two individual strands.

- *Annealing* of DNA primers. The heated DNA solution is cooled to $50 - 65°$ C, this allows the annealing (hybridization) of the primers to each of the single DNA template.

A DNA primer anneals beginning in the 3′ end of each strand. This because DNA polymerase synthesizes in the direction 5′ to 3′.

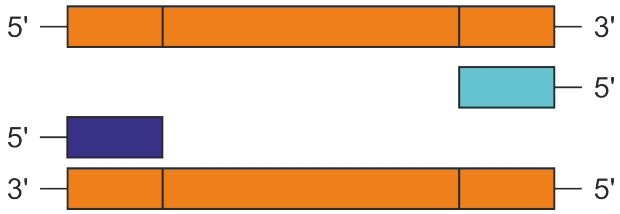


Figure 35: Annealing of the primers.

- *Replication* of DNA with heat-resistant DNA polymerase, a special DNA polymerase that is heat resistant, such as Taq DNA polymerase [17], an enzyme from thermophilic bacterium. After annealing step, the cooled solution is heated to a temperature of 72° C and it begins DNA synthesis of a new DNA strand complementary to the template by adding free dNTPS from the solution in the 5′ to 3′ direction on both ends. After one cycle of PCR, the number of DNA molecules is doubled.

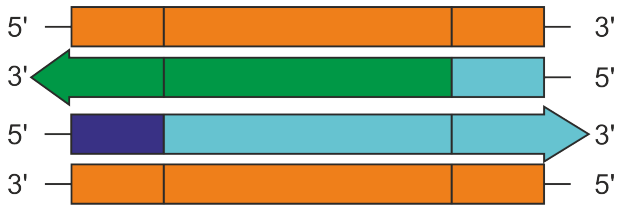


Figure 36: First cycle of PCR.

After two cycles of PCR, four DNA molecules are obtained, in general: 2^n = Number of copies of the original double-stranded DNA target region (assuming 100% reaction efficiency), where n is the number of cycles. For example, a reaction for $n = 30$ results in $2^{30} = 1,073,741,824$ DNA copies.

- *Repeat* the cycle again by changing temperature.

- Select **Menu ▶ Open file** (Ctrl+O) and select the *csv file. Automatically the program split the data contained in the file detecting the amount of probes that were used in the experiment as well as the temperature resolution. Then, the raw data is plotted. In this program is possible to visualize two graphics selecting the **Graph 1** or the **Graphic 2** with the switch (green circle 38). By default the cut-off temperatures are established on both the left and right sides (TL=70 and TR=90) as is illustrated in the Figure 37. This values can be edited with the Spinner TL (circle blue) and TR (circle red). Erroneous probes can be discarded of

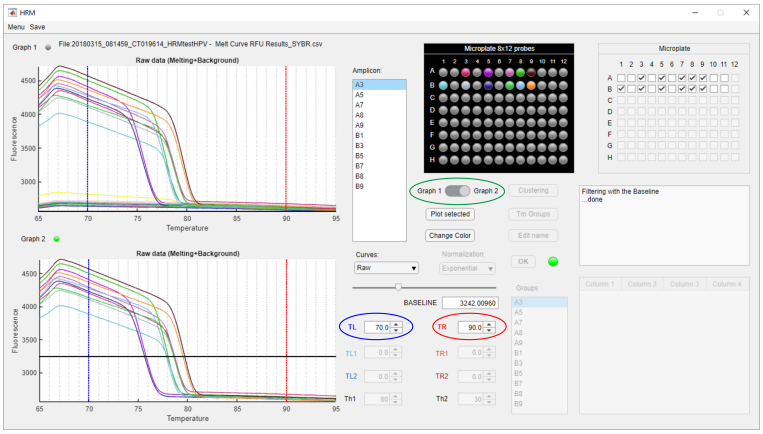


Figure 37: Graph 1: Raw signals, Graph2: Raw data filtered by the Baseline value. TL blue vertical line, TR red vertical line.

the analysis with the help of the Microplate check boxes array or establishing a value of fluorescence in the Baseline line edit field such that all the curves that are below this value are discarded (also the check boxes are modified). This value in the

Baseline also can be edited by a slider and is represented by a black and horizontal line in the Graphic 2 of the figure 37.

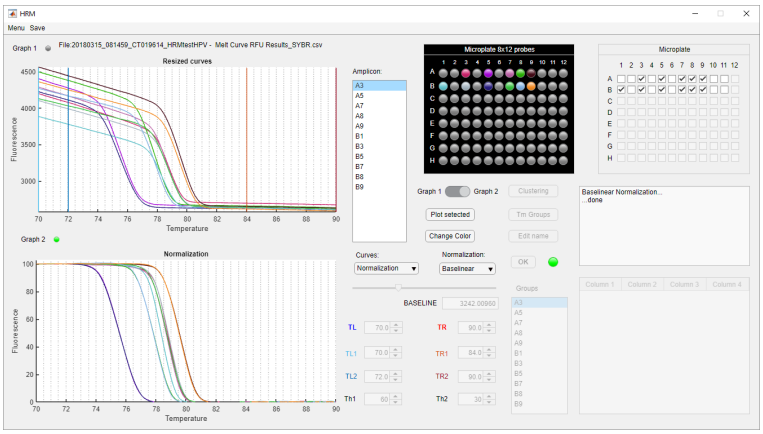


Figure 38: Graph 1: Resize , Graph2: Normalization.

- Select **Resize** in the drop down button of the curves, the raw signals are visualized in the temperature range established by TL and TR. Note that the Spinners TL1, TL2, TR1 and TR2 are now enabled. TL2 represents the melt start temperature, while TR1 the melt end temperature. The region in the middle of TL2 and TR1 is the melting region temperature of the curves, this section of the curve is characterized by a significantly increase of the slope and finish when the slopes is nearly flat. TL2 and TR2 are used to establish the section to fit the section of the signal to be removed. In this program a Base-linear Background subtraction and Exponential Background subtraction are available.

- Now we can visualize the normalized graphic by selecting **Normalization** in the drop down button of the curves.

- The Melting temperature T_m is defined as the highest slope of the melting curve, and is determined as the peak of the negative derivative curve. Select **Derivative** to visualize the first negative derivative of the raw data (Savinsky-Golay smoothing

filter is used to decrease the noise produced by digital differentiation), selecting **Derivative Norm** the derivative of the normalized data is observed in the graphic. T_m of each probe is shown by selecting the amplicon in the List boxes. Select **Tm Groups** to obtain a table with the probes that have the same T_m. It is possible change the name of the groups with **Edit name** and **OK**.

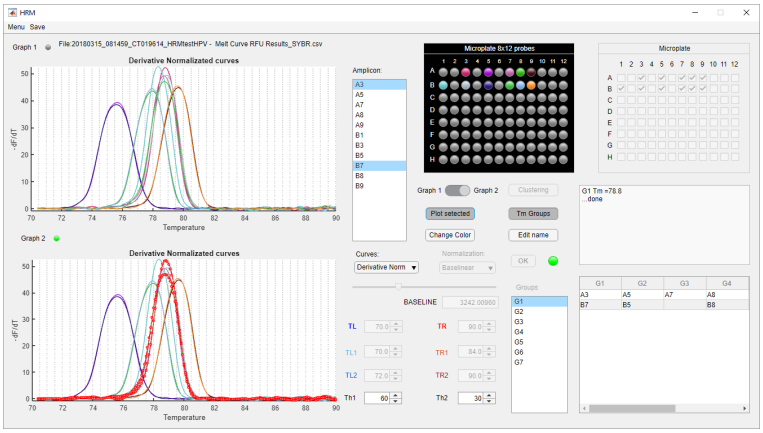


Figure 39: Graph 1: Derivative , Graph2: Highlighted selected probes.

- Other analytical graphic to make comparisons between the melting curves is the difference curve. This graph helps to distinguish between homozygotes and heterozygotes subtracting a known genotype considered as the reference (target) from the another curves. Also, the reference curve could be an average between two or more reference curves. Selecting **Difference**, the difference between the curves and one or more targets is plotted. Targets can be selected through both List boxes. Finally, with the button **Clustering** groups are created (with the k-means function). **Edit name** and **OK**.

- Select **Save ▶ Figure** (Ctrl+S) to save the last plot. This figure will be saved automatically in the same path of the *.csv file with the name of the file +/(Raw, Derivative,Difference,...etc.). In addition, selected curves (List Boxes) can be highlighted with

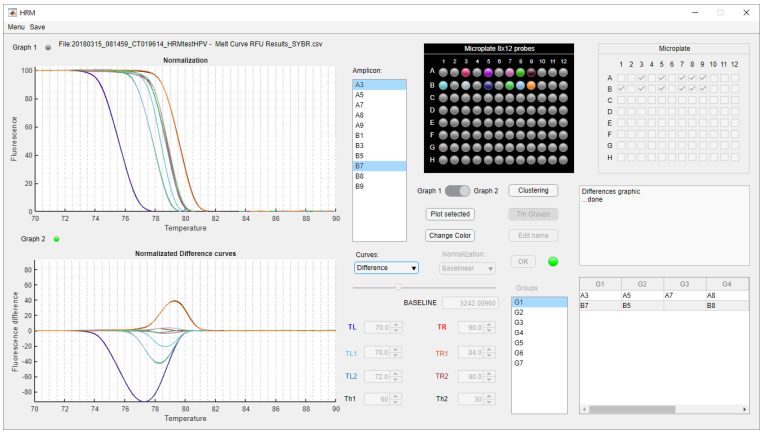


Figure 40: Graph 1: Normalization , Graph2: Difference curve.

the button **Plot selected** as well change their color with **Change Color**.

BIBLIOGRAPHY

[1] Pornpat Athamanolap, Vishwa Parekh, Stephanie I Fraley, Vatsal Agarwal, Dong J Shin, Michael A Jacobs, Tza-Huei Wang, and Samuel Yang. "Trainable high resolution melt curve machine learning classifier for large-scale reliable genotyping of sequence variants." In: *PloS one* 9.10 (2014), e109094. DOI: 10.1371/journal.pone.0109094.

[2] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. "Model-Based Compressive Sensing." In: *IEEE Transactions on Information Theory* 56.4 (2010), pp. 1982–2001. DOI: 10.1109/TIT.2010.2040894.

[3] E. van den Berg and M. P. Friedlander. *SPGL1: A solver for large-scale sparse reconstruction*. http://www.cs.ubc.ca/labs/scl/spgl1. June 2007.

[4] N. Bertoni, B. Senevirathna, F. Pareschi, M. Mangia, R. Rovatti, P. Abshire, J. Z. Simon, and G. Setti. "Low-power EEG monitor based on compressed sensing with compressed domain noise rejection." In: *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2016, pp. 522–525. DOI: 10.1109/ISCAS.2016.7527292.

[5] Thomas Blumensath and Mike E. Davies. "Iterative hard thresholding for compressed sensing." In: *Applied and Computational Harmonic Analysis* 27.3 (2009), pp. 265 –274. ISSN: 1063-5203. DOI: http://dx.doi.org/10.1016/j.acha.2009.04.002. URL: http://www.sciencedirect.com/science/article/pii/S1063520309000384.

[6] M. Borgerding, P. Schniter, J. Vila, and S. Rangan. "Generalized approximate message passing for cosparse analysis compressive sensing." In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 3756–3760. DOI: 10.1109/ICASSP.2015.7178673.

[7] Valerio Cambareri, Mauro Mangia, Fabio Pareschi, Riccardo Rovatti, and Gianluca Setti. "A rakeness-based design flow for analog-to-information conversion by compressive sensing." In: *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*. IEEE. May 2013, pp. 1360–1363.

[8] E. J. Candes, J. Romberg, and T. Tao. "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information." In: *IEEE Transactions on Information Theory* 52.2 (2006), pp. 489–509. ISSN: 0018-9448. DOI: 10.1109/TIT.2005.862083.

[9] E. J. Candes and T. Tao. "Decoding by linear programming." In: *IEEE Transactions on Information Theory* 51.12 (Dec. 2005), pp. 4203–4215. ISSN: 0018-9448. DOI: 10.1109/TIT.2005.858979.

[10] Emmanuel J. Candès and Michael B. Wakin. "An introduction to compressive sampling." In: *Signal Processing Magazine, IEEE* 25.2 (Feb. 2008), pp. 21–30.

[11] Emmanuel J. Candès, Michael B. Wakin, and Stephen P. Boyd. "Enhancing Sparsity by Reweighted ℓ_1 Minimization." In: *Journal of Fourier Analysis and Applications* 14.5 (2008), pp. 877–905. ISSN: 1531-5851. DOI: 10.1007/s00041-008-9045-x. URL: https://doi.org/10.1007/s00041-008-9045-x.

[12] Emmanuel Candes and Justin Romberg. *ℓ_1-MAGIC: Recovery of Sparse Signals via Convex Programming*. https://statweb.stanford.edu/ candes/l1magic. Oct. 2005.

[13] A. Caprara, F. Furini, A. Lodi, M. Mangia, R. Rovatti, and G. Setti. "Generation of Antipodal Random Vectors With Prescribed Non-Stationary 2-nd Order Statistics." In: *Signal Processing, IEEE Transactions on* 62.6 (Mar. 2014), pp. 1603–1612. ISSN: 1053-587X. DOI: 10.1109/TSP.2014.2302737.

[14] P. Charalampidis, A. G. Fragkiadakis, and E. Z. Tragos. "Rate-Adaptive Compressive Sensing for IoT Applications." In: *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*. 2015, pp. 1–5. DOI: 10.1109/VTCSpring.2015.7146042.

[15] R. Chartrand and Wotao Yin. "Iteratively reweighted algorithms for compressive sensing." In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2008, pp. 3869–3872. DOI: 10.1109/ICASSP.2008.4518498.

[16] Xi Chen, Ehab A. Sobhy, Zhuizhuan Yu, Sebastian Hoyos, Jose Silva-Martinez, Samuel Palermo, and Brian M. Sadler. "A Sub-Nyquist Rate Compressive Sensing Data Acquisition Front-End." In: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 2.3 (Sept. 2012), pp. 542–551. ISSN: 2156-3357. DOI: 10.1109/JETCAS.2012.2221531.

[17] A. Chien, Edgar D. B., and J. M. Trela. In: *Journal of bacteriology* 127.3 (1976), 1550–1557. ISSN: 0021-9193.

[18] George M Church, Yuan Gao, and Sriram Kosuri. "Next-generation digital information storage in DNA." In: *Science* 337.6102 (2012), pp. 1628–1628.

[19] M. A. Davenport, A. K. Massimino, D. Needell, and T. Woolf. "Constrained Adaptive Sensing." In: *IEEE Transactions on Signal Processing* 64.20 (2016), pp. 5437–5449. ISSN: 1053-587X. DOI: 10.1109/TSP.2016.2597130.

[20] D. L. Donoho. "Compressed Sensing." In: *IEEE Transactions on Information Theory* 52.4 (Apr. 2006), pp. 1289–1306. ISSN: 0018-9448. DOI: 10.1109/TIT.2006.871582.

[21] D. L. Donoho, A. Javanmard, and A. Montanari. "Information-Theoretically Optimal Compressed Sensing via Spatial Coupling and Approximate Message Passing." In: *IEEE Transactions on Information Theory* 59.11 (2013), pp. 7434–7464. ISSN: 0018-9448. DOI: 10.1109/TIT.2013.2274513.

[22] David L. Donoho, Arian Maleki, and Andrea Montanari. "Message-passing algorithms for compressed sensing." In: *Proceedings of the National Academy of Sciences* 106.45 (Nov. 2009), pp. 18914–18919.

[23] Julio Martin Duarte-Carvajalino and Guillermo Sapiro. "Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization." In: *Image Processing, IEEE Transactions on* 18.7 (2009), pp. 1395–1408.

[24] Z. L. Dwight, R. Palais, and C. T. Wittwer. "uAnalyze: Web-Based High-Resolution DNA Melting Analysis with Comparison to Thermodynamic Predictions." In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9.6 (2012), pp. 1805–1811. ISSN: 1545-5963. DOI: 10.1109/TCBB.2012.112.

[25] Zachary Dwight, Robert Palais, and Carl T. Wittwer. "uMELT: prediction of high-resolution melting curves and dynamic melting profiles of PCR products in a rich web application." In: *Bioinformatics* 27.7 (2011), pp. 1019–1020. DOI: 10.1093/bioinformatics/btr065. URL: http://dx.doi.org/10.1093/bioinformatics/btr065.

[26] M. Elad. "Optimized Projections for Compressed Sensing." In: *IEEE Transactions on Signal Processing* 55.12 (2007), pp. 5695–5702. DOI: 10.1109/TSP.2007.900760.

[27] Y. C. Eldar, P. Kuppinger, and H. Bolcskei. "Block-Sparse Signals: Uncertainty Relations and Efficient Recovery." In: *IEEE Transactions on Signal Processing* 58.6 (2010), pp. 3042–3054. DOI: 10.1109/TSP.2010.2044837.

[28] Tze-Kiong Er and Jan-Gowth Chang. "High-resolution melting: Applications in genetic disorders." In: *Clinica Chimica Acta* 414 (2012), pp. 197 –201. ISSN: 0009-8981. DOI: https://doi.org/10.1016/j.cca.2012.09.012. URL: http://www.sciencedirect.com/science/article/pii/S0009898112004470.

[29] Maria Erali and Carl T. Wittwer. "High resolution melting analysis for gene scanning." In: *Methods* 50.4 (2010). The ongoing Evolution of qPCR, pp. 250 –261. ISSN: 1046-2023. DOI: https://doi.org/10.1016/j.ymeth.2010.01.013. URL: http://www.sciencedirect.com/science/article/pii/S1046202310000289.

[30] Sanger F, Nicklen S, and Coulson AR. "DNA sequencing with chain-terminating inhibitors." In: *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (1977), 5463–5467. DOI: 10.1073/pnas.74.12.5463. URL: https://www.ncbi.nlm.nih.gov/pubmed/271968.

[31] Marco Galarza, Manuel Fasabi, Kelly S Levano, Edith Castillo, Nadia Barreda, Mitzi Rodriguez, and Heinner Guio. "High-resolution melting analysis for molecular detection of multidrug resistance tuberculosis in Peruvian isolates." In: *BMC infectious diseases* 16.1 (2016), p. 260. DOI: 10.1186/s12879-016-1615-y.

[32] D. Gangopadhyay, E. G. Allstot, A. M. R. Dixon, K. Natarajan, S. Gupta, and D. J. Allstot. "Compressed Sensing Analog Front-

End for Bio-Sensor Applications." In: *IEEE Journal of Solid-State Circuits* 49.2 (2014), pp. 426–438. ISSN: 0018-9200. DOI: 10.1109/JSSC.2013.2284673.

[33] Nick Goldman, Paul Bertone, Siyuan Chen, Christophe Dessimoz, Emily M LeProust, Botond Sipos, and Ewan Birney. "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA." In: *Nature* 494.7435 (2013), p. 77.

[34] Cameron N Gundry, Joshua G Vandersteen, Gudrun H Reed, Robert J Pryor, Jian Chen, and Carl T Wittwer. "Amplicon melting analysis with labeled primers: a closed-tube method for differentiating homozygotes and heterozygotes." In: *Clinical chemistry* 49.3 (2003), pp. 396–406.

[35] C. Guo and M. E. Davies. "Near Optimal Compressed Sensing Without Priors: Parametric SURE Approximate Message Passing." In: *IEEE Transactions on Signal Processing* 63.8 (2015), pp. 2130–2141. ISSN: 1053-587X. DOI: 10.1109/TSP.2015.2408569.

[36] Giovanni Jacovitti, Alessandro Neri, and Gaetano Scarano. "Texture synthesis-by-analysis with hard-limited Gaussian processes." In: *Image Processing, IEEE Transactions on* 7.11 (1998), pp. 1615–1621.

[37] Sami Kanderian, Lingxia Jiang, and Ivor Knight. "Automated classification and cluster visualization of genotypes derived from high resolution melt curves." In: *PloS one* 10.11 (2015), e0143295. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0143295.

[38] F. Krahmer and R. Ward. "Stable and Robust Sampling Strategies for Compressive Imaging." In: *IEEE Transactions on Image Processing* 23.2 (2014), pp. 612–622. ISSN: 1057-7149. DOI: 10.1109/TIP.2013.2288004.

[39] Naveen Kumar, Fatemeh Fazel, Milica Stojanovic, and Shrikanth S. Naryanan. "Online rate adjustment for adaptive random access compressed sensing of time-varying fields." In: *EURASIP Journal on Advances in Signal Processing* 2016.1 (2016), p. 48. ISSN: 1687-6180. DOI: 10.1186/s13634-016-0348-9. URL: https://doi.org/10.1186/s13634-016-0348-9.

[40] M. Leinonen, M. Codreanu, and M. Juntti. "Sequential Compressed Sensing With Progressive Signal Reconstruction in Wireless Sensor Networks." In: *IEEE Transactions on Wireless Communications* 14.3 (2015), pp. 1622–1635. DOI: 10.1109/TWC.2014.2371017.

[41] Bo Li, Liang Zhang, Thia Kirubarajan, and Sreeraman Rajan. "A projection matrix design method for MSE reduction in adaptive compressive sensing." In: *Signal Processing* 141.Supplement C (2017), pp. 16 –27. ISSN: 0165-1684. DOI: https://doi.org/10.1016/j.sigpro.2017.05.019. URL: http://www.sciencedirect.com/science/article/pii/S0165168417301901.

[42] Huaizhong Li, Ruiting Lan, Niancai Peng, Jing Sun, and Yong Zhu. "High resolution melting curve analysis with MATLAB-based program." In: *Measurement* 90 (2016), pp. 178 –186. ISSN: 0263-2241. DOI: https://doi.org/10.1016/j.measurement.2016.04.057. URL: http://www.sciencedirect.com/science/article/pii/S0263224116301233.

[43] M. Li, R.A. Palais, L. Zhou, and C.T. Wittwer. "Quantifying variant differences in DNA melting curves: Effects of length, melting rate, and curve overlay." In: *Analytical Biochemistry* 539 (2017), pp. 90 –95. ISSN: 0003-2697. DOI: https://doi.org/10.1016/j.ab.2017.10.015. URL: http://www.sciencedirect.com/science/article/pii/S0003269717304001.

[44] Mei Li, Luming Zhou, Robert A Palais, and Carl T Wittwer. "Genotyping accuracy of high-resolution DNA melting instruments." In: *Clinical chemistry* 60.6 (2014), pp. 864–872.

[45] S. Li, L. D. Xu, and X. Wang. "Compressed Sensing Signal and Data Acquisition in Wireless Sensor Networks and Internet of Things." In: *IEEE Transactions on Industrial Informatics* 9.4 (2013), pp. 2177–2186. DOI: 10.1109/TII.2012.2189222.

[46] L.S. Lilly and H.M. School. *Pathophysiology of Heart Disease: A Collaborative Project of Medical Students and Faculty*. Wolters Kluwer/Lippincott Williams & Wilkins, 2011.

[47] Sujuan Liu, Meihui Zhang, Wenshu Jiang, Junshan Wang, and Peipei Qi. "Theory and hardware implementation of an analog-to-Information Converter based on Compressive Sensing." In: *2013 IEEE 10th International Conference on ASIC*. 2013, pp. 1–4. DOI: 10.1109/ASICON.2013.6812033.

[48] H. Mamaghanian, N. Khaled, D. Atienza, and P. Vandergheynst. "Design and Exploration of Low-Power Analog to Information Conversion Based on Compressed Sensing." In: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 2.3 (2012), pp. 493–501. ISSN: 2156-3357. DOI: 10.1109/JETCAS.2012.2220253.

[49] M. Mangia, F. Pareschi, R. Rovatti, and G. Setti. "Adaptive Matrix Design for Boosting Compressed Sensing." In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 65.3 (2018), pp. 1016–1027. ISSN: 1549-8328. DOI: 10.1109/TCSI.2017.2766247.

[50] Mauro Mangia, Riccardo Rovatti, and Gianluca Setti. "Rakeness in the design of Analog-to-Information Conversion of Sparse and Localized Signals." In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 59.5 (May 2012), pp. 1001–1014.

[51] A. Marchioni, M. Mangia, F. Pareschi, R. Rovatti, and G. Setti. "Sparse sensing matrix based compressed sensing in low-power ECG sensor nodes." In: *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. 2017, pp. 1–4. DOI: 10.1109/BIOCAS.2017.8325155.

[52] P. E. McSharry, G. D. Clifford, L. Tarassenko, and L. A. Smith. "A dynamical model for generating synthetic electrocardiogram signals." In: *IEEE Transactions on Biomedical Engineering* 50.3 (2003), pp. 289–294. ISSN: 0018-9294. DOI: 10.1109/TBME.2003.808805.

[53] CA) Mullis Kary B. (Kensington, CA) Erlich Henry A. (Oakland, CA) Arnheim Norman (Woodland Hills, CA) Horn Glenn T. (Emeryville CA)and Saiki Randall K. (Richmond, and CA) Scharf Stephen J. (Berkeley. "Process for amplifying, detecting, and/or-cloning nucleic acid sequences." In: 4683195 (1987). URL: http://www.freepatentsonline.com/4683195.html.

[54] Deanna Needell and Joel A Tropp. "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples." In: *Applied and Computational Harmonic Analysis* 26.3 (2009), pp. 301–321.

[55] R. A. Palais and C. T. Wittwer. "Melting Curve Analysis with exponential Background Substraction." In: (Nov. 2011).

[56] Robert Palais and Carl T. Wittwer. "Chapter 13 Mathematical Algorithms for High-Resolution DNA Melting Analysis." In: *Computer Methods, Part A*. Vol. 454. Methods in Enzymology. Academic Press, 2009, pp. 323 –343. DOI: https://doi.org/10.1016/S0076-6879(08)03813-5. URL: http://www.sciencedirect.com/science/article/pii/S0076687908038135.

[57] F. Pareschi, M. Mangia, D. Bortolotti, A. Bartolini, L. Benini, R. Rovatti, and G. Setti. "Energy Analysis of Decoders for Rakeness-Based Compressed Sensing of ECG Signals." In: *IEEE Transac-*

tions on Biomedical Circuits and Systems 11.6 (2017), pp. 1278–1289. ISSN: 1932-4545. DOI: 10.1109/TBCAS.2017.2740059.

[58] Fabio Pareschi, Pierluigi Albertini, Giovanni Frattini, Mauro Mangia, Riccardo Rovatti, and Gianluca Setti. "Hardware-Algorithms Co-Design and Implementation of an Analog-to-Information Converter for Biosignals Based on Compressed Sensing." In: *IEEE Transactions on Biomedical Circuits and Systems* 10.1 (Feb. 2016), pp. 149–162. ISSN: 1932-4545. DOI: 10.1109/TBCAS.2015.2444276.

[59] J. T. Parker, V. Cevher, and P. Schniter. "Compressive sensing under matrix uncertainties: An Approximate Message Passing approach." In: *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*. 2011, pp. 804–808. DOI: 10.1109/ACSSC.2011.6190118.

[60] Leonor Pereira, Sónia Gomes, Sara Barrias, José Ramiro Fernandes, and Paula Martins-Lopes. "Applying high-resolution melting (HRM) technology to olive oil and wine authenticity." In: *Food Research International* 103 (2018), pp. 170 –181. ISSN: 0963-9969. DOI: https://doi.org/10.1016/j.foodres.2017.10.026. URL: http://www.sciencedirect.com/science/article/pii/S0963996917307135.

[61] Jan Radvanszky, Milan Surovy, Emilia Nagyova, Gabriel Minarik, and Ludevit Kadasi. "Comparison of different DNA binding fluorescent dyes for applications of high-resolution melting analysis." In: *Clinical Biochemistry* 48.9 (2015), pp. 609 –616. ISSN: 0009-9120. DOI: https://doi.org/10.1016/j.clinbiochem.2015.01.010. URL: http://www.sciencedirect.com/science/article/pii/S0009912015000326.

[62] Valin Reja, Alister Kwok, Glenn Stone, Linsong Yang, Andreas Missel, Christoph Menzel, and Brant Bassam. "ScreenClust: Ad-

vanced statistical software for supervised and unsupervised high resolution melting (HRM) analysis." In: *Methods* 50.4 (2010), S10 –S14. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2010.02.006.

[63] Kirk M. Ririe, Randy P. Rasmussen, and Carl T. Wittwer. "Product Differentiation by Analysis of DNA Melting Curves during the Polymerase Chain Reaction." In: *Analytical Biochemistry* 245.2 (1997), pp. 154 –160. ISSN: 0003-2697. DOI: https://doi.org/10.1006/abio.1996.9916. URL: http://www.sciencedirect.com/science/article/pii/S0003269796999169.

[64] Riccardo Rovatti, Gianluca Mazzini, and Gianluca Setti. "Memorym antipodal processes: spectral analysis and synthesis." In: *Circuits and Systems I: Regular Papers, IEEE Transactions on* 56.1 (2009).

[65] Riccardo Rovatti, Gianluca Mazzini, Gianluca Setti, and Stefano Vitali. "Linear probability feedback processes." In: *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*. IEEE. 2008, pp. 548–551.

[66] Lindsay N. Sanford, Jana O. Kent, and Carl T. Wittwer. "Quantum Method for Fluorescence Background Removal in DNA Melting Analysis." In: *Analytical Chemistry* 85.20 (2013). PMID: 24070125, pp. 9907–9915. DOI: 10.1021/ac4024928. URL: https://doi.org/10.1021/ac4024928.

[67] B. Mca. Savers, H. A. Beagley, and W. R. Henshall. "The mechanism of auditory evoked EEG responses." In: *Nature* 247.5441 (1974), pp. 481–483.

[68] Michael Scherg, Jiri Vajsar, and Terence W. Picton. "A Source Analysis of the Late Human Auditory Evoked Potentials." In: *Journal of Cognitive Neuroscience* 1.4 (1989). PMID: 23971985, pp. 336–355. DOI: 10.1162/jocn.1989.1.4.336. eprint: https://doi.

org/10.1162/jocn.1989.1.4.336. URL: https://doi.org/10.1162/jocn.1989.1.4.336.

[69] M. Shoaran, M. H. Kamal, C. Pollo, P. Vandergheynst, and A. Schmid. "Compact Low-Power Cortical Recording Architecture for Compressive Multichannel Data Acquisition." In: *IEEE Transactions on Biomedical Circuits and Systems* 8.6 (2014), pp. 857–870. ISSN: 1932-4545. DOI: 10.1109/TBCAS.2014.2304582.

[70] Ivan Simko. "High-Resolution DNA Melting Analysis in Plant Research." In: *Trends in Plant Science* 21.6 (2016), pp. 528 –537. ISSN: 1360-1385. DOI: https://doi.org/10.1016/j.tplants.2016.01.004. URL: http://www.sciencedirect.com/science/article/pii/S1360138516000054.

[71] J. A. Tropp and A. C. Gilbert. "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit." In: *IEEE Transactions on Information Theory* 53.12 (2007), pp. 4655–4666. DOI: 10.1109/TIT.2007.909108.

[72] John Hasbrouck Van Vleck and David Middleton. "The spectrum of clipped noise." In: *Proceedings of the IEEE* 54.1 (1966), pp. 2–19.

[73] Carl T Wittwer. "High-resolution DNA melting analysis: advancements and limitations." In: *Human mutation* 30.6 (2009), pp. 857–859.

[74] Carl T Wittwer, Gudrun H Reed, Cameron N Gundry, Joshua G Vandersteen, and Robert J Pryor. "High-resolution genotyping by amplicon melting analysis using LCGreen." In: *Clinical chemistry* 49.6 (2003), pp. 853–860.

[75] S. M. H. T. Yazdi, H. M. Kiah, E. Garcia-Ruiz, J. Ma, H. Zhao, and O. Milenkovic. "DNA-Based Storage: Trends and Methods." In: *IEEE Transactions on Molecular, Biological and Multi-*

Scale Communications 1.3 (2015), pp. 230–248. ISSN: 2332-7804. DOI: 10.1109/TMBMC.2016.2537305.

[76] Juhwan Yoo, Stephen Becker, Matthew Loh, Manuel Monge, Emmanuel Candès, and Azita Emami-Neyestanak. "A 100MHz-2GHz 12.5x sub-Nyquist rate receiver in 90nm CMOS." In: *2012 IEEE Radio Frequency Integrated Circuits Symposium*. June 2012, pp. 31–34. DOI: 10.1109/RFIC.2012.6242225.

[77] J. Zhang, Z. Gu, Z. L. Yu, and Y. Li. "Energy-Efficient ECG Compression on Wireless Biosensors via Minimal Coherence Sensing and Weighted ℓ_1 Minimization Reconstruction." In: *IEEE Journal of Biomedical and Health Informatics* 19.2 (2015), pp. 520–528. ISSN: 2168-2194. DOI: 10.1109/JBHI.2014.2312374.

[78] J. Zhang, Y. Suo, S. Mitra, S. (Peter) Chin, S. Hsiao, R. F. Yazicioglu, T. D. Tran, and R. Etienne-Cummings. "An Efficient and Compact Compressed Sensing Microsystem for Implantable Neural Recordings." In: *IEEE Transactions on Biomedical Circuits and Systems* 8.4 (2014), pp. 485–496. ISSN: 1932-4545. DOI: 10.1109/TBCAS.2013.2284254.

[79] Z. Zhang, T. Jung, S. Makeig, and B. D. Rao. "Compressed Sensing for Energy-Efficient Wireless Telemonitoring of Noninvasive Fetal ECG Via Block Sparse Bayesian Learning." In: *IEEE Transactions on Biomedical Engineering* 60.2 (2013), pp. 300–309. DOI: 10.1109/TBME.2012.2226175.

[80] Z. Zhang and B. D. Rao. "Sparse Signal Recovery With Temporally Correlated Source Vectors Using Sparse Bayesian Learning." In: *IEEE Journal of Selected Topics in Signal Processing* 5.5 (2011), pp. 912–926. DOI: 10.1109/JSTSP.2011.2159773.

[81] Luming Zhou, Lesi Wang, Robert Palais, Robert Pryor, and Carl T Wittwer. "High-resolution DNA melting analysis for simul-

taneous mutation scanning and genotyping in solution." In: *Clinical Chemistry* 51.10 (2005), pp. 1770–1777. ISSN: 0009-9147. DOI: 10.1373/clinchem.2005.054924. URL: http://clinchem.aaccjnls.org/content/51/10/1770.

[82] H. Zhu, G. Leus, and G. B. Giannakis. "Sparsity-Cognizant Total Least-Squares for Perturbed Compressive Sampling." In: *IEEE Transactions on Signal Processing* 59.5 (2011), pp. 2002–2016. ISSN: 1053-587X. DOI: 10.1109/TSP.2011.2109956.

[83] Alain de Cheveigné and Jonathan Z. Simon. "Denoising based on time-shift PCA." In: *Journal of Neuroscience Methods* 165.2 (2007), pp. 297–305. ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2007.06.003.

[84] Alain de Cheveigné and Jonathan Z. Simon. "Denoising based on spatial filtering." In: *Journal of Neuroscience Methods* 171.2 (2008), pp. 331–339. ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2008.03.015.

[85] Alain de Cheveigné and Jonathan Z. Simon. "Sensor noise suppression." In: *Journal of Neuroscience Methods* 168.1 (2008), pp. 195–202. ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2007.09.012.

DECLARATION

Italy, July 2020

César Hugo Pimentel

Romero