Theoretical and Applied Informatics
ISSN 1896–5334
Vol.23 (2011), no. 3-4
pp. 177–192
DOI: 10.2478/v10179-011-0012-x

# On the Application of Compressive Sampling Techniques to High Throughput Data in Computational Genomics

ENRIQUE HERNÁNDEZ-LEMUS

Computational Genomics Department
National Institute of Genomic Medicine
Periférico Sur 4124, Piso 5, 01900, México City, México
ehernandez@inmegen.gob.mx,

Center for Complexity Sciences
National Autonomous University of México (UNAM)
Torre de Ingeniería, Circuito Escolar s/n, 04510, México City, México

*Received 3 September 2011, Revised 7 November 2011, Accepted 12 November 2011*

**Abstract:** With the advent of high throughput experiments in genomics and proteomics, the researcher in computational data analysis is faced with new challenges, both with regards to the computational capacities and also in the probabilistic/statistical methodology fields; in order to handle such massive amounts of data in a systematic coherent way. In this paper we describe the basic aspects of the mathematical theory and the computational implications of a recently developed technique called *Compressive Sampling*, as well as some possible applications within the scope of Computational Genomics, and Computational Biology in general. The central idea behind this work is that most of the information sampled from the experiments turns out to be discarded (for being non-useful) in the final stages of biological analysis, hence it would be better if we could find an algorithm to remove *selectively* such information in order to get rid of the computational burden associated with processing and analyzing such huge amounts of data. Here we show that a possible algorithm for doing so it is precisely Compressive Sampling. As a working example, we will consider the data-analysis of whole-genome microarray gene expression for 1191 individuals within a breast cancer project.

**Keywords:** Compressive Sampling, Compressed Sensing, Computational Genomics

## 1. Introduction

Within the context of computational analysis of data generated in high throughput experiments in genomics (e.g. massive DNA sequencing, whole genome gene-expression, genome-wide genotyping or proteomics) one is usually confronted with the

acquisition and reconstruction of large data-vectors or matrices. Suppose our data is in the form of a vector X in $\mathcal{R}^m$. According to the usual tenets in signal processing (i.e. Nyquist-Shannon Sampling Theory) [10, 11] this would require $m$ samples. However, if we know *a priori* that X is compressible by means of transform coding with a known coding function, of course we can choose to acquire X by sampling instead $n$ general linear functionals. In the case that the collection of functionals is well-chosen and by considering also that we allow a certain degree of Sampling/Reconstruction error, the size $n$ of the sampling set could be drastically reduced (i.e. $n << m$). For example, in some instances in digital signal processing a sampling space of size $n = \mathcal{O}(m^{\frac{1}{4}} \log^{\frac{5}{2}} m)$ is required to fully recover a m-pixel image.

This is of course in stark contrast with common wisdom assuring us that we have to account for at least the number of Shannon-Nyquist (SN) [15, 16] modes to reconstruct such image, since according to SN Theorem, in order to reconstruct a signal (without error) we need to comply with the signal's *bandwidth*, i.e. the length of the shortest interval which contains the support of the signal's spectrum, which correspond to its dimension $m$ [8, 5]. SN sampling theorem and most of its extensions were stated primarily for band-limited functions (signals), and not for general random processes. However, these are more relevant to the information theorist and computational scientist, as opposed to the electric/electronic engineering applications of band-limited signals. Nevertheless, most of these sampling expansions can be extended easily to random processes by means of data-transformation. The original sampling theorem states that ...*If a function f (t) contains no frequencies higher than W cps [i.e. Hz] it is completely determined by giving its ordinates at a series of points spaced (1/2W) s apart...* [16]. It is pertinent to mention that due to the symmetry of the Fourier transform pairs (in which the proof is based), the sampling theorem is also valid for time-limited functions.

Even if the SN sampling theorem sets a minimum sampling rate, it only refers to standard (independent signal) measurements. However, *sampling with derivatives*, for example, increases the sample spacing required, or in other words it allows the reconstruction of the band-limited signal with a sampling rate less than the Nyquist rate. Another approach aiming at the same goal was established [12] and it is related with the problem of the representation and construction of wide-sense stationary stochastic signals, *not from one set of data but from several sets of sampled values* obtained by using a multiple channel sampling scheme. It was showed that with the optimum combination of prefilters and post-filters, in the case where two sets of sample values are taken, the frequency range of the input signal is limited by the prefilters to a total width twice the SN prediction. It has also been proposed [17] the use of multiple channels to reconstruct deterministic band-limited signals with a sampling rate less than the Nyquist rate. The sample rate needed is inversely related to the number of the channels used and directly proportional to the Nyquist rate.

These examples (illustrative but by no means comprehensive) shown that by changing the sampling strategy it is possible to overcome the sampling rate established by SN theory. Let us begin the consideration of how sampling scheme modifications (and in particular ideas from compressive sampling) could be applied in the case of biosignals. One interesting point with regards to biological signal processing (that also applies to other kinds of signals such as digital images) is that: a) there is a certain degree of noise involved in the related measurement processes b) there is a also a tolerated error bound associated with reconstruction. Since biosignals are usually noisy, there is no need for complete accuracy, since any reconstruction error below the noise bound would be unnoticed in further analysis. Of course, in order to attain acceptable signal-to-noise ratios, *careful* reconstruction techniques are needed. In the present case, 'careful' reconstruction is related with taking into account issues such as data sparsity and compressibility. Sparsity leads to dimensionality reduction and effective modeling strategies; whereas compressibility enables optimal data handling and processing. All of this issues have been considered in the past and as such they constitute the cornerstone of signal processing techniques. However, the present paradigm of acquiring the full signal, then calculating the entire set of transform coefficient to *compress* the signal, encode the largest coefficients (e.g. the Principal Components) and discard all others has become extremely cumbersome, specially in the case of high throughput experiments (for an example, see Figure 2 and related discussion below).

One possible way-out for this situation is based on considering data-objects $X$ that possess a sparse representation in some basis (in general orthonormal) such as wavelets, Fourier modes, principal components, or even Shannon-Weaver, Kullback-Liebler or Gabor optimals. If X has an sparse representation, then all of its coefficients belong to a certain $l^p$ ball with $0 < p < 1$ and also that the $N$ *most-important* coefficients in the expansion allow a reconstruction with $l^2$ error of order $\mathcal{O}(N^{\frac{1}{2}} - \frac{1}{p})$ [8]. Nevertheless if some error is allowed (below the noise bounds) then it is possible to design a small dimension space by means of *Basis Pursuit* a nonadaptive sampling technique, thus reducing the gap between sampling and processing solving at least partially the so-called *dimensionality problem*.

Having these facts in mind, the rest of the paper will be organized as follows: Section 2 will present the mathematical basis of Compressing Sampling in terms of sparse data matrixes and their relation with sampling below the Shannon-Nyquist limit, in particular with regards to the possibility of performing *a priori* data compression under optimal recovery conditions. Then in Section 3 an approximate setting termed *near optimal recovery* is discussed. Section 4 establish these ideas in the context of noisy data and robustness of the reconstruction. After this, Section 5 refers to the application of the Compressed Sampling algorithm to gene expression data for a Computational Genomics analysis. Section 6 deals with the Materials and Methods used in such analysis. Finally,

Section 7 presents a discussion of this proof-of-concept and gives some perspectives on it.

## 2. Undersampling and Sparsity

Let us consider a general *signal reconstruction problem* for a vector $x$ in $\mathcal{R}^N$ from a set of $y$ linear measurements of the form:

$$y_k = \langle x, \phi_k \rangle; \;\; k = 1, 2, \ldots K \;\; \text{or} \;\; \overrightarrow{y} = \Phi \overrightarrow{x} \tag{1}$$

If we take a look at the underdetermined case $K << N$ we are facing the *dimensionality problem* in which we have many fewer measurements than unknown signal values. This is one of many challenges during the analysis of data for genome-wide gene expression experiments [1]. The usual setting here is the consideration of hundreds of thousands of gene-probes characterizing the expression of thousand genes (i.e. different mRNA transcripts) out of a set that, commonly, is in the tens to, at most few hundreds of samples (microarrays) [1]. Underdetermined reconstruction problems are ill-fated unless they satisfy two conditions: 1) there is a finite noise-bound and 2) they are compressible. Compressibility implies that the signal depends on a number of degrees of freedom smaller than N. If our target signal is sparse, then it can be written either exactly or at least accurately (i.e. with an error below the noise bound) as a superposition of a number $N_s < N$ of vectors in some fixed basis. As we shall see $N_s$-reconstruction turns out to be nothing but a simple convex optimization problem [4].

If we consider the so-called principle of *transform sparsity* (that we have just stated) then it is satisfied that for some $0 < p < 2$ and for some $Z > 0$:

$$\|\theta\|_p = \left( \sum_i \; |\theta_i|^p \right)^{\frac{1}{p}} \leq Z \tag{2}$$

Here the $\theta_i$ are the **sparse transform** (compression) coefficients from an orthonormal basis $\phi_i$ defined so that $\theta_i = \langle x_i, \phi_i \rangle$ [7]. Equations (1) and (2) thus define the usual procedures of sampling and compressing a signal. Within this setting we have a so-called $l^p$ -norm sparsity constraint. It can be shown that for $p \geq 2$ no sparsity is present. However, specially for the cases $0 < p \leq 1$ (highly compressive data-objects) it is possible to construct a near-optimal algorithm (NOA) which can be based on linear programming (essentially by a combination of $l^1$ and even $l^0$ minimization). This NOA basically computes the coefficients to reconstruct (and decompress) the object $X$ with the smallest $l^1$-norm that is consistent with the information $y_n$. For the details of the reconstruction see [4, 5].

### 2.1. A relation between $l^1$ and $l^0$ minimization in the context of Optimal Recovery

The reason for the NOA to actually work is based on a recent finding relating the minimization problems constrained by $l^1$ and $l^0$ -norms, respectively. As we have just said, the NOA is based on $l^p$ minimization ($0 < p \leq 1$) that is a highly non-trivial, non-convex optimization problem. For p=1 however, solving $l^1$ minimization (LOm) is related with solving $l^0$ minimization (LZm). Let us examine this relation in detail.

$$LZm : \min \|\theta\|_0 \quad \text{subject to} \quad \Phi\,\theta = X \tag{3}$$

Here the zero-norm of $\theta$ , is of course just the number of non-zero entrances in $\theta$, i.e. the sparsity measure. Ordinarily solving LZm would require combinatorial optimization. However when LZm has an sparse solution, LOm could find it. In fact, when $\theta$ has at most $\mathcal{O}(n/\log m)$ non-zeros (for $n << m$), the LZm and LOm have the same unique solution [5] (see Theorem 8). This fact takes relevance in view of the works of Candes, Romberg and Tao [4] that studied the design matrices built by taking $n$ rows at random from an $m$ by $m$ Fourier matrix (i.e. a sparse orthonormal representation of a compressible data source) and proved an $\mathcal{O}(n/\log m)$ bound. They were thus proving that the NOA was feasible. In other words, even if the system of equations is massively underdetermined, $l^1$ minimization and its sparse solution coincide- when the result is sufficiently sparse.

## 3. Near-Optimal Recovery

In general, signals that result interesting in practice may not have complete support in space or even within a transformation domain. Instead, they very often are concentrated near a sparse set. For example, within the context of harmonic analysis one usually assumes that the coefficients of elements taken for a signal class decay very fast, usually with a power-law decay. Of course, a wide-variety of signals, both smooth and piecewise smooth are susceptible of harmonic modeling. Under this scenario, a common question would be: *how well can one recover a nearly-sparse signal?*. For an arbitrary vector $x$ in $\mathcal{R}^N$, let us call $x_S$ its best $S$-sparse approximation, i.e. the approximation obtained by keeping the $S$ largest entries of $x$ setting all other entries equal to zero. It turns out that if the sensing matrix obeys a restricted isometry hypothesis called the *Uniform Uncertainty Principle* (UUP) [4] [i.e. that every set of columns with cardinality less than $S$ approximately behaves like an orthonormal system], then the recovery error is not-much-worse than $\|x - x_s\|\, l^2$. Of course, for sparse systems this error turns out to be asymptotically small.

It is noticeable that the very fact that makes high dimensionality undersampled problems complex to solve in a traditional way (high dimension of the support space and

small sampling numbers) make the design matrices and sensing vectors very sparse (in general, but specially under orthonormality), thus obeying UUP and being natural candidates for CS reconstruction techniques.

## 4. Robustness and Compressive Sampling

As we have stated, any realistic signal reconstruction technique has to take into account that signals are noisy and recovery algorithms generate errors. It is thus necessary to examine the issue of robustness of compressive sampling against measurement errors. Of course, in order for CS to be widely applicable, the algorithm should be stable (i.e. small fluctuations in the data should give rise to small reconstruction errors). Let us consider a simple reconstruction model like the one in equation (1) but including an error term:

$$\overrightarrow{y} = \Phi \overrightarrow{x} + \overrightarrow{\epsilon} \tag{4}$$

Here $\overrightarrow{\epsilon}$ is a stochastic or deterministic error term with bounded energy (mean variance) $E = \|\overrightarrow{\epsilon}\|l^2$ , by doing so, the reconstruction program will be of the form:

$$\min \|\widetilde{x}\|_{l^1} \ \text{ such that } \ \|\Phi\widetilde{x} - \overrightarrow{y}\|_{l^2} \tag{5}$$

Here $\widetilde{x}$ is an *approximant* to $\overrightarrow{x}$. The program defined by equation (5) is a convex second order cone program [3], these methods are extremely reliable and could be reduced, either to quadratically constrained quadratic programs, or to linear matrix inequalities in the constrains and can be solved with great efficiency by interior point methods. These problems are so ubiquitous that there exist generic solving platforms for them such as `Mosek` (http://www.mosek.com/) or `OpenOpt` (http://openopt.org/Welcome). In the particular case of the problem in equation (5) one can notice that the reconstruction error is the sum of a term proportional to the size of the measurement error and a term corresponding to the noiseless case. Due to this linear additivity it is possible to prove that there is no recovery method that outperforms the program in equation (5) for arbitrary perturbations (of size $\epsilon$) [4]. Thus CS programs obtained by following equation (5) are robust against bounded perturbations.

Having studied the main properties, scope and limitations of the Compressive Sampling algorithm as a means to attain a sub-Shannon-Nyquist sampling limit, making use of sparsity, near-optimal recovery and the relation between $l^1$ and $l^0$ minimization; and once it is proved that this procedure is linear-robust against noise-levels, and it is thus a tractable convex optimization problem, even implemented in both, commercial and open-source packages; we are now in a position to sketch its possible application within the realms of Computational Genomics.

## 5. A plausible framework for the application of Compressive Sampling in Computational Genomics

In recent times, high density oligonucleotide arrays have become widely used both in basic and biomedical research in genomics. The system made use of oligonucleotides, usually of 25 base-pairs in longitude that are used to probe genes. Each gene is generally represented by a set of 16-20 pairs of those oligonucleotides known as probe sets. One of each pair of these oligos is known as the perfect match (PM) probe and correspond to an exact segment of the complementary sequence of the associated gene, whereas the other one, known as the mismatch probe (MM) is made by changing the middle (13th) base in order to look up for the effects of non-specific binding [1].

A question arise as how to *combine the data* for the set of 16-20 PM-MM pairs to define a measure of expression that represent in an optimal way the amount of the associated mRNA species [9]. This is not a trivial issue since there are a lot of variables involved in the analysis (several probes for a probe-set, tens of thousands of probe-sets for whole genome approaches in humans and other mammals, for instance; versus some dozens of samples, a few hundreds at most), and the resulting signals are very noisy. These facts imply that the usual frequentist approach to probability and statistics has to be modified to deal with whole-genome gene expression data. As we already have described in the previous sections, this kind of high dimensional data calls for the implementation of techniques such as compressive sampling.

Let us consider this issue in a greater detail. Statistical analyses of the PM and the MM probes under controlled experimental conditions have revealed that for large values of genetic abundance the differences between PM and MM probes have a bi-modal distribution with the second mode occurring for negative differences. This effect has been related with heteroscedasticity (unequal variances in the distributions). Commonly, hybridization noise characteristics at the high expression regime are Poisson-like, whereas its characteristics for the small expression levels are more complex. Hence to assess the statistical validity of gene expression differences between two experiments we must characterize the fluctuation caused purely by experimental measurement. It is known that noise depends strongly on the expression level. Therefore, an expression-dependent distribution function is needed to characterize the variability between replicates. In order to correct (or at least take into account) for these issues an optimization of the signal-to-noise-ratio (SNR) has been proposed in the form of the so-called *background correction* [9, 1]. In the other hand, observed intensity levels also depend on sample preparation, manufacture of the arrays, and lab processing of such arrays (dye labeling, hybridization and scanning), for this reason, unless arrays are correctly *normalized* comparing data from different arrays can lead to misleading results. In short, from very large data-sets we have to define a method for combining probe-level samples of noisy-data, measured across different scales and filter them in

a clean cut way to make them ready for statistical analysis and probabilistic modeling.

A plethora of computational and mathematical approaches ranging from support vector machines, to bayesian predictors, neural networks and machine learning techniques have been developed to this end. However, a *simple* and reliable algorithm designed to meet all the aforementioned requirements of *normalization*, *background correction*, and *probe-summarization* has been proposed [2]. The algorithm called `RMA` consists of three steps: a background adjustment, quantile normalization and finally summarization. Mathematical details of the algorithm have been analyzed [1, 2]. With regards to them, let us just state that background adjustment is made by means of a (linear) conditional expectation transformation, and that quantile normalization and probe summarization are attained also as linear applications; so that the whole RMA processing algorithm could be visualized as a *signal reconstruction* and *data-compression* (yet incomplete) problem.

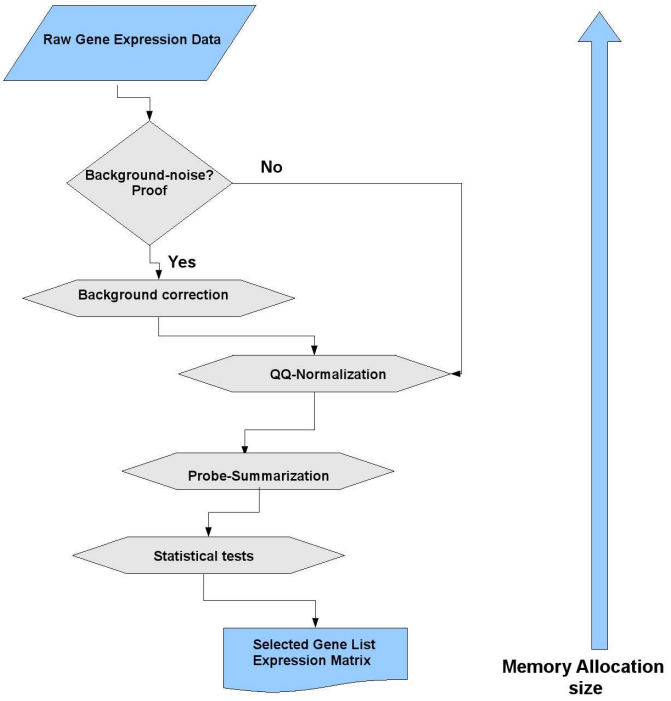


Fig. 1. Processing stages of Whole Genome Gene Expression data from Microarray Experiments

The resulting gene expression (GE) design matrix (a $N_1$ by $N_2$ matrix representing $N_1$ different mRNA expression values in $N_2$ different experimental conditions or samples) is thus a Compressive-Sampling Matrix (CS-Matrix) i.e. one that can be *acquired* by means of compressive sampling techniques. This is so, because GE matrices possess a certain degree of linear independence (but not complete independence) among all small groups of columns (e.g. mRNA samples belonging to different individuals or to a same individual in purposely different conditions). This is nothing but condition CS1 of Candes, Romberg and Tao CS theory [8]. With regards to condition CS2, linear combinations of small groups of columns give vectors that *look* much like random noise, this condition is attained because quantile-normalized columns may belong to different quantile levels and so their linear combinations are to some degree *incoherent*. Finally, with regards to condition CS3, it states that for every vector constructed from a submatrix of the GE matrix, the quotient norm is asymptotically the $l^1$ norm. This condition is also fulfilled since gene expression vectors (after RMA algorithm has been applied) are normalized. The mathematical proof that conditions CS1-CS3 completely determine that a data-matrix as a CS matrix (too complex to be stated here), it is based on the abstract theory of *Gel'fand n-widths*, the interested reader could check it on [4].

If we look at Figure 2 we will be able to understand the need for optimized sampling strategies in the case of computational analysis of biological data. The problem corresponds to Genome-Wide Gene Expression studies (GWGE) in a series of 1191 human mRNA samples for a combined breast cancer study that is being carried out in the Computational Genomics department at the National Institute of Genomic Medicine (INMEGEN) in Mexico. These experiments correspond to the so-called GPL96 protocol which is based on the Affymetrix HGU-133A microarray GeneChip platform. The set includes over 1,000,000 unique oligonucleotide features covering more than 39,000 transcript variants, which in turn represent greater than 33,000 of the best characterized human genes. The data, as it comes from Affymetrix scanning data-acquisition server, is stored in raw binary files (.CEL files), the amount of memory allocated by the 1191 CEL files corresponds to about 21 Gb, whereas the over and under expressed gene list (which is the final outcome of data-processing), that is, the file examined by the molecular biologists to make their analysis (for the very same dataset) corresponds to about 1.4 Mb. This implies a **15313-fold** memory allocation reduction, i.e. we sampled fifteen-thousand times the amount of data we ended-up using. This fold-change is expected to be highly-increased with the advent of Next Generation Sequencing (NGS) techniques, since NGS experiments could generate up to 2 Terabyte of raw data in a single experimental run.

Let us consider a hypothetical instance in which we can actually attain the Candes-Romberg-Tao (CRT) sampling space size limit, i.e. $n = \mathcal{O}(m^{\frac{1}{4}} \log^{\frac{5}{2}} m)$. Since we are trying to recompose a 39,000 variable gene-data set (for the Affymetrix HGU133A

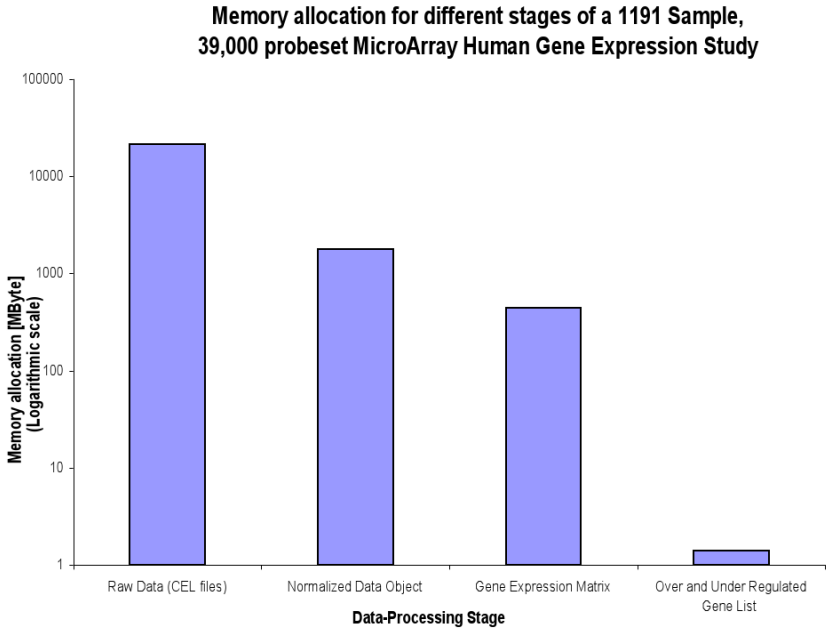


Fig. 2: Memory allocation reduction through different stages of computational processing of gene expression data (Notice logarithmic scale in the y-axis)

array), at the gene level, we will just have to sample $\mathcal{O}(162.1294655)$ or, about 163 samples to have full coverage at this level. Even if we wish to consider the whole gene-probe level data, in order to cover the full one-million oligonucleotide sampling space (something that is rarely made in practice) it would be theoretically feasible to make that with just $\mathcal{O}(476.797144)$ or 477 samples, again in the case of the HGU133A GeneChip. Of course the CRT sample limit is a theoretical lower bound, however, if we consider the actual nature of the data and the desired results in the usual GWGE analyses we will see that compressive sampling could be implemented in an efficient manner in practice. In the Breast Cancer project that we are using here, several important issues could be raised. First, if we consider figure 3, we could see that the vast majority of genes are expressed on a similar level in both the cancer and normal tissue mRNA samples (i.e. they are distributed along, or very close to the identity line), hence the data of all these genes is uninformative with regards to the actual mechanisms behind tumorigenesis. This means that a large amount of the gene expression information that is actually being sampled is not used in further analysis.

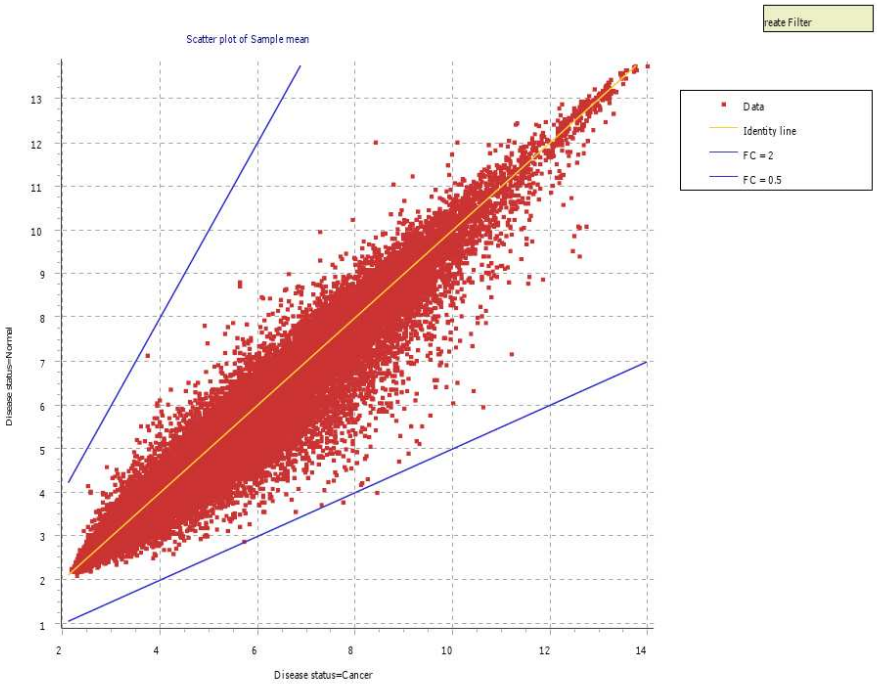


Fig. 3. Scatter-Plot of gene expression

Let us look at this in greater detail, in figure 4, we plotted an histogram of the log-ratios or log-fold changes of gene expression intensity, i.e. base-two logarithms of the ratio of the expression of a gene in the cancer samples (on average) to the expression of the same gene in the normal samples. Within the biomedical community a usual rule of thumb for considering a gene "interesting" or "important", is that its log2-ratio is greater than 0.5 or 1 (corresponding to about 1.5-fold to twice the expression in a normal condition) or lesser than -0.5 or -1 (i.e. 1.5-fold to twice the expression in a diseased condition). Now, if we look at figure 4 we could notice that a vast majority of genes sampled are considered *uninteresting*, since just the, say $> 0.5$ and $< -0.5$ right and left tails of the distribution are further considered in the analyses. Expression fold-change is not the only issue related to the actual *compressibility* of GWGE data, we have to consider also the question of repeatability and coherence of the data between different samples, an issue that is strongly related with noise and statistical significance. If we define a way to know if a given log-fold change is statistically significant, a set of statistical proofs need to be performed in the normalized gene expression matrix (see Figure 1).

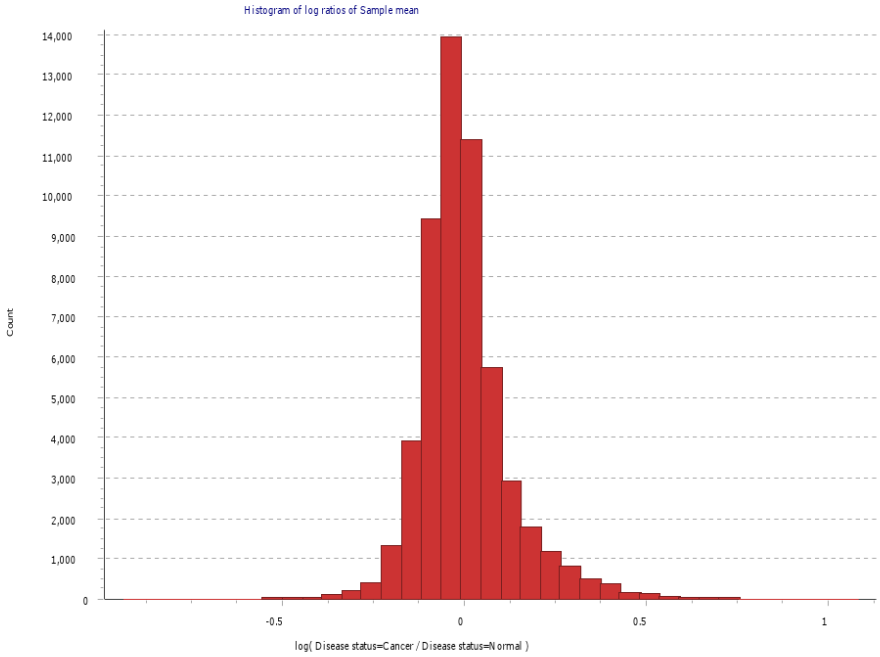


Fig. 4. Histogram of differential gene expression

In the particular case of the 1191 sample study we are carrying-out, we have performed common analysis of variance ANOVA, an improved version of the t-test (CyberT by Baldi & Long), two-sample Bayes t-test, Local Pooled Error, as well as an Empirical Bayes Algorithm. In our particular case the best performing algorithm was Baldi & Long's CyberT although all of the methods (even the ANOVA) performed well. After carrying out these statistical analysis we adjusted for multiple testing by means of Bonferroni, False Discovery Rate (FDR) and The Family Wise Error Rate (FWER) proofs. After all these calculations are done, it is possible to generate an *Overabundance plot* that, in some sense will work as a further measure of compression-rate. In Figure 5 it is displayed the overabundance plot (number of statistically significant genes versus confidence level) for the CyberT tested, FDR-corrected analysis.

It is noticeable that just about 6-7% of the genes (that is to say, some 2,350 genes) comply with the $p - value < 0.05$ significance. This means that, in the end most of the sampled data ends up not-being used. However, the very fact that this data was sampled (and foremost was acquired, stored and processed) generated computational and logistic costs that could be significantly reduced if a CS technique would be implemented in the data-acquisition stage.

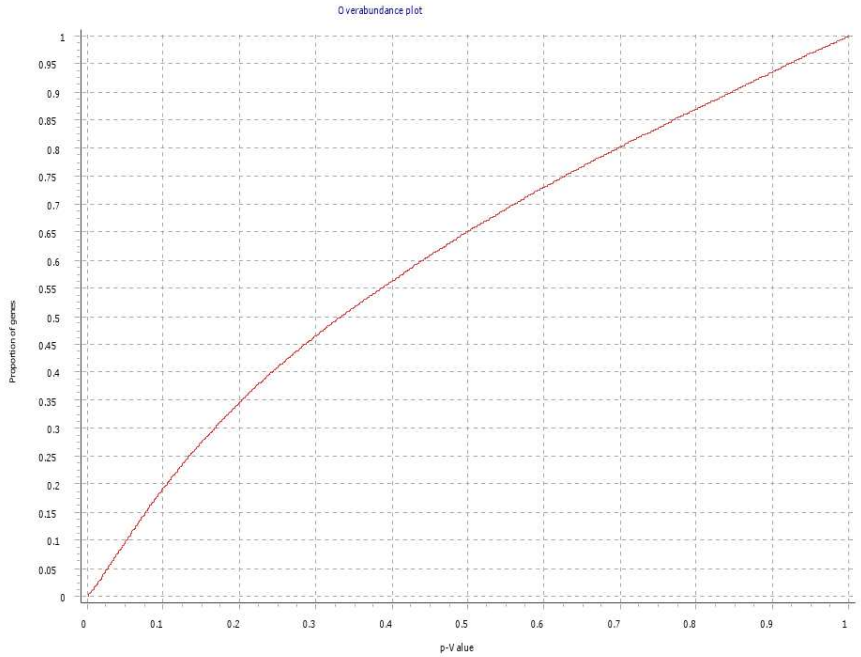


Fig. 5. Overabundance plot

We have just sketched some hints so as to how compressed sensing or compressing sampling techniques could be used as auxiliary tools to facilitate massive data handling and analysis in the context of computational genomics. The level presented here is that of a *proof-of-concept*, however, in order to outline a detailed algorithm or implementation, further insight in biological data structure need to be attained by means of simulation [13] or bootstrapping calculations [14]. Important limitations in this regard are due to the reluctance of genomic analysis equipment producers to make public the data structure of the binary files that constitute the output of the experimental procedures.

## 6. Methods and Materials

For the reported study we used a curated set of 1191 publicly available GWGE Microarrays corresponding to several independent experiments, all of them processed according with the GPL96 protocol over Affymetrix HGU133A arrays. This set is part of a computational analysis project that is being carried out in the Computational Genomics department at INMEGEN. RMA preprocessing (background correction, normalization and probe-summarization) was done by using `[R] / Bioconductor`, Statistical tests

were performed on [R] and data-visualization was done using the FlexArray suite developed by Genome Quebec. All pre-processing was done on a 128 Gb RAM 8-Power5+ dual core-processor, symmetric multiprocessing (SMP) unit by IBM. Whereas all Statistical tests were performed on a Dell Precision Series 8 Gb RAM QuadCore Workstation. Both SMP unit and workstation running under Linux (RedHat and Gentoo, respectively) and visualization was done in a Dell Optiplex Dual-Core, 4 Gb RAM running under Windows XP.

## 7. Discussion

In the section above we have discussed at the *proof-of-concept* level the possibility of application of compressive sampling techniques within the settings of real GWGE data. We have shown by means of benchmarks and statistical tests, that a great amount of the huge data sampled in this kind of experiments results discarded in some of the pre-processing and/or statistical significance stages, i.e. before any biological hypothesis could be tested on it. We also discussed the conditions that make CS plausible within the actual settings. No-actual CS algorithm, however has been possible to apply since data-acquisition for these experiments is made by proprietary software and algorithms belonging to the technology vendors (in this particular case Affymetrix, Inc.) and no open-formats are available.

Nevertheless by showing the plausibility of this strategy and the technical and economical convenience of CS techniques we are in a position to approach genomic technology vendors (such as Affymetrix) to suggest them that their Bioinformatics and Information Technology units consider the possibility of implementing computational tools that allow CS to be an option for real-time (well, almost real-time in reality) experimental data acquisition (doing this seem easier to do than suggest them to open the formats of their proprietary software data-management suits). In fact, some two years ago in the user's meeting for a genomic technology company called *Illumina*, it was presented a prototype system of real-time data-acquisition and pre-processing for genotyping/sequencing data.

We believe that it is still early time to implement such CS techniques, specially in view of the up-coming super-high throughput NGS experiments and also the massive-imaging proteomic spectrometric data that will in the near-future call out for optimized data-acquisition and pre-processing methodologies. As a means of contrast, pre-processing of the 1191 MicroArray dataset from raw data (.CEL files) involved the use of 1766 processor-hours for the SMP unit (i.e. about nine and a half days in 8 processors) and it was *just* about 21 Gb, just imagine processing times for 2 Tb sequencing experiments or for long electrophysiological (about 1 Gb of data per hour per experiment as a rule of thumb) or proteomic imaging studies (hundreds of thousands of high-throughput

of spectral peak files) (Note: the pre-processing for sequencing data and for GWGE, electrophysiological or proteomic data is different, so it is not just a matter of scaling these figures). CS methodologies or its analogs will thus, very surely, be involved in modern data acquisition / processing in computational biology and genomics.

In this paper we have discussed some instances for the applicability of Compressive Sampling techniques for the near-optimal *data reconstruction* of high throughput samples within the context of Computational Genomics, and in particular we outlined an application to whole-genome gene expression analyses. The issue of compressive sampling is extraordinarily rich and vast, and its applicability in signal processing and data reconstruction has only been glimpsed up to this day. However, this techniques seem very promising in the future to cope with problems of undersampling, complex multidimensional signal-processing and also in problems like data-storing limitations, for example in next generation sequencing and whole-proteome imaging settings.

Much is still to be explored with regards to reliability of the algorithm's implementation, benchmarking against other data-compression/data-analysis techniques, etc., before CS could be established as a data-processing gold standard but the future seem promissory.

## References


1. K. Baca-López, E. Hernández-Lemus, M. Mayorga: *Information-theoretical analysis of gene expression data to infer transcriptional interactions*, Rev. Mex. Fis. 55, 6, 456-466, (2009).
2. B.M. Bolstad, R.A. Irizarry, M. Astrand, T.P. Speed: *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias*, Bioinformatics 19, 2, 185-193, (2003).
3. S. Boyd, L. Vandenberghe: *Convex Optimization*, Cambridge University Press, (2004). Available online at http://www.stanford.edu/ boyd/cvxbook/.
4. E.J. Candes, J. Romberg, T. Tao: *Signal Recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math. 59, 8, 1207-1223, (2005).
5. E.J. Candes: *Compressive Sampling, Proceedings of the International Congress of Mathematicians*, European Mathematical Society, Madrid, Spain, 1-20, (2006).
6. E.J. Candes, J. Romberg, T. Tao: *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Information Theory 52, 2, 489-509, (2006).
7. D.L. Donoho, M. Vetterli, R.A. DeVore, I.C. Daubechies: *Data Compression and Harmonic Analysis*, IEEE Trans. Information Theory 44, 6, 2435-2476, (1998).
8. D.L. Donoho: *Compressed Sensing*, Technical report, Stanford University, (2004). A revised version appears in Donoho, D.L., Compressed Sensing, IEEE Trans, Information Theory 52, 4, 1289-1306 (2006).

9. R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, T.P. Speed: *Exploration, normalization, and summaries of high density oligonucleotide array probe level data*, Biostatistics 4, 2, 249-64, (2003).

10. A. Jerri: *The Shannon Sampling Theorem-Its Various Extensions and Applications: A Tutorial Review*, Proceedings of the IEEE, 65, 11, 1565-1595, (1977).

11. A. Jerri: Correction to The Shannon sampling theorem-Its various extensions and applications: A tutorial review, Proceedings of the IEEE, 67, 4, 695, (1979).

12. R.E. Kahn, B. Liu: *Sampling representation and optimum reconstruction of signals*, IEEE Trans. Inform. Theory, 11, 3, 339-347, (1965).

13. P. Langfelder, S. Horvath: *Eigengene networks for studying the relationships between co-expression modules*, BMC Systems Biology, 1:54, (2007) doi:10.1186/1752-0509-1-54 (see particularly the supplementary material available at http://www.biomedcentral.com/content/supplementary/1752-0509-1-54-s7.pdf and the Code for the simulations available at: http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/EigengeneNetwork/).

14. T.H.E. Meuwissen, M.E. Goddard: *Bootstrapping of gene-expression data improves and controls the false discovery rate of differentially expressed genes*, Genetics Selection Evolution 36, 2, 191-205, (2004) doi:10.1186/1297-9686-36-2-191.

15. H. Nyquist: *Certain topics in telegraph transmission theory*, Trans. AIEE, 47, 2, 280-305, (1928). Reprinted as a classic paper in: Proc. IEEE, 90, 2, 276-279,(2002).

16. C.E. Shannon: *Communication in the presence of noise*, Proc. Institute of Radio Engineers 37, 1, 10-21, (1949). Reprinted as a classic paper in: Proc. IEEE, 86, 2, 447-457, (1998).

17. D.E. Todd: *Sampled data reconstruction of deterministic band limited signals*, IEEE Trans. Inform. Theory, 19, 6, 809-811, (1973).