Identifying the Impact of G-Quadruplexes on Affymetrix Exon Arrays using Cloud Computing

Farhat N. Memon, Olivia Sanchez-Graillet, Graham J. G. Upton, Anne M. Owen and Andrew P. Harrison*

Departments of Mathematical Sciences and Biological Sciences, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, United Kingdom, Tel: +44 (0) 1206 873040, http://bioinformatics.essex.ac.uk/users/harry/

Summary

A tetramer quadruplex structure is formed by four parallel strands of DNA/ RNA containing runs of guanine. These quadruplexes are able to form because guanine can Hoogsteen hydrogen bond to other guanines, and a tetrad of guanines can form a stable arrangement. Recently we have discovered that probes on Affymetrix GeneChips that contain runs of guanine do not measure gene expression reliably. We associate this finding with the likelihood that quadruplexes are forming on the surface of GeneChips.

Our original discovery was made using 3’ arrays. We have now extended our analysis to look at Affymetrix Exon arrays. In order to cope with the rapidly expanding size of Exon array datasets in the public domain, we have explored the use of cloud computing. This is a recently introduced high-performance solution that takes advantage of the computational infrastructure of large organisations such as Amazon and Google.

We expect that cloud computing will become widely adopted because it enables bioinformaticians to avoid capital expenditure on expensive computing resources and to only pay a cloud computing provider for what is used. Moreover, as well as financial efficiency, cloud computing is an ecologically-friendly technology, it enables efficient data-sharing and we expect it to be faster for development purposes. Here we describe the use of cloud computing to perform a large data-mining analysis of public domain Exon arrays.

1 Introduction

1.1 G-Quadruplex

The binding of guanine to cytosine and adenine to thymine usually occurs through the famous Watson-Crick interactions in double stranded DNA. However, in a single-strand of DNA sequence, a guanine can bind to another guanine through a Hoogsteen hydrogen bond. A tetrad of guanines can then form a loop, in which each guanine can bind to two other guanines at 90 degrees (similar to the edges of a square). Indeed, this occurs throughout a genome because single-strands of DNA sequence that have frequent occurrences of guanine runs are capable of forming four-stranded structures, known as G-Quadruplexes, G-tetrads, or G4 DNA [1].

*Corresponding author: harry@essex.ac.uk
In a single strand of DNA, a G-quadruplex consists of four runs of guanines (called the stems of G-quadruplex) with three loops in between the four stems. \textit{GGAGGGTTGACGGGAAGGG}, a segment of single strand of DNA sequence for instance, can form a G-quadruplex in which the four sets of underlined Gs represent four stems of guanine and the nucleotides in between these stems create loops. Both the stem size and loop size have biological significance. As the stem size increases, a G-quadruplex becomes more stable; whereas an increase in loop size weakens the stability of quadruplexes.

[2] demonstrated that G-rich nucleic acid sequences can adopt quadruplex structures that are stabilised by the presence of G-quartets(Figure- 1). A G-quadruplex may not necessarily form through a single nucleic acid sequence; sometimes two or four parallel nucleic acid sequences may form a G-quadruplex collectively.

Figure 1: Schematic presentation of G-quartet structures. (A) G-quartet. (B) Different layouts/topologies and loop orientation of quadruplexes (Source: http://nar.oxfordjournals.org/cgi/content/full/31/8/2097)

Figure 1(B) illustrates a number of different topologies for G-quadruplexes. For example, Monomer Chair and Monomer Basket showing G-quadruplexes that are formed in a single nucleic acid sequence, whilst Dimer Chair and Dimer Basket illustrate that two G-rich nucleic acid strands are capable of forming a G-quadruplex. Indeed a tetramer can result from four parallel strands forming a G-quadruplex. A quadruplex form through more than one sequences falls in the category of Intermolecular Quadruplex structures. Thus, Dimer and Tetrimer are both the examples of Intermolecular Quadruplex structures [3]. By keeping tetramer quadruplex structure in mind, we are investigating the tools (Microarray) that are used to analyse genomic data.
1.2 Affymetrix GeneChips enable whole-transcriptome studies of the Genome

The production of messenger RNA reflects the activity level of a gene, and many biologists are interested in the conditions in which a specific gene is turned on or turned off. Microarray technology allows the simultaneous study of many genes in parallel, providing a snapshot of how a genome is operating. A microarray usually consists of a glass slide, containing a 2D array of an orderly arrangement of fragments of single-stranded DNA, referred to as probes, that represent the genes of an organism. Each DNA fragment representing a gene is assigned a specific location on the array and a fluorescently labelled DNA or RNA (target sequence) will stick through hybridisation to their complementary probe. The genes that are active are detected through measuring the light from the excited fluorescence of the labelled DNA or RNA.

There are many types of microarray that are available commercially. However, in this study we focus on the Affymetrix GeneChip, a high density oligonucleotide array. An Affymetrix GeneChips consists of 25-mer oligonucleotide probes which have been synthesised in-situ through photolithographical methods. Each gene is represented by several probes, collectively called a probe set. The size of a GeneChip covered by an array of probes is 1.28cm × 1.28cm. Due to improvements in array manufacturing technology, the number of distinct probe sequences within this area has increased over time, with some of the latest designs having over 5 million different cells, each containing many thousands of copies of a distinctive probe sequences. Figure 2 shows the basic construction of an Affymetrix GeneChip.

Affymetrix have released GeneChips for most major model organisms. One of their most widely used designs is known as the 3' Arrays, because most probes are selected towards the 3' region of a gene. Although the probes are selected to ideally avoid non-specific hybridisation to other transcripts, such cross-hybridisation is unavoidable. The Affymetrix 3' design attempts to compensate for this by including probes that are identical to the gene-specific probes except for a mismatched base at its centre (13th base). These mismatch (MM) probes are placed immediately adjacent to their perfect match (PM) probes. In this way, each gene is represented by 22 different probes (11 Perfect match probes and 11 Mismatch probes). The design philosophy is that 11 signal intensities measure a particular gene fragment plus a sequence-specific
background; while 11 mismatch probes report a close approximation to the sequence-specific background; the MM signal is subtracted from the PM signal resulting in a measure of a gene’s expression; the multiple measurements of gene expression are collated into one expression measure.

Affymetrix has introduced another chip design, the **Exon array**, which is designed to investigate exon-level expression. They have smaller probesets, typically 4 probes, and these probesets detect exons across the gene, not just towards the 3’ end. Mismatch probes do not exist in Exon arrays. There are approximately four probes per exon and roughly 40 probes per gene. Exon arrays enable “exon-level” analysis, which allows us to distinguish between different isoforms of a gene, and to detect specific alterations in exon usage, some of which may play a central role in disease mechanism and etiology. The Exon arrays also allow "gene-level" expression analysis, in which multiple probes on different exons are summarised into an expression value of all transcripts from the same gene.

### 1.3 Identifying problems in GeneChip data

Affymetrix report that over 10,000 published papers have used or described their technology. As each typical study comprises multiple GeneChips, there are now many tens of thousands of GeneChips in the public domain that are now available for meta-analysis. Although the power of GeneChip technology is widely recognised, many open questions remain about the appropriate analysis of GeneChip data. This is particularly true now that we have the opportunity to mine large GeneChip datasets in order to discover novel signatures associated with diseases.

It is expected that if a particular gene is highly expressed then all the probes in a probe set representing that gene will be consistent in showing the presence of that particular gene. However, [4] found that probes containing runs of guanine show abnormal affinities; they tend to have increased cross hybridisation signals and reduced target-specific hybridisation signals, presumably due to multiplex binding forming G-quartet structures. We [5] recently confirmed that probes having a sequence of four or more guanines, which we termed G-spots, are poorly correlated with other probes in their probeset. We suggested that the intensities reported by these probes should not be used in the calculation of gene expression values, due to their poor correlation with the other probes in their probeset. We also suggested that G-spot probes should not be included within future array designs.

[5] suggested that structures closely resembling G-quadruplexes are forming on GeneChips, and this is why probes containing runs of guanine are not fit for purpose. Neighbouring probes with the same sequence can come into physical contact on a GeneChip. For most sequences which lack complementary sections they will not be expected to hybridise to each other. But for probes containing runs of guanine, it is possible that a stack of Hoogsteen hydrogen bonds can occur [5]. A grouping of four probes can then form a stable tetrad at each guanine, and the resulting stack of tetrads forms a G-quadruplex. In such a G-quadruplex the guanines face inwards and are not available to hybridise to target sequences. But in the interpretation of [5], the formation of a G-quadruplex frees up space in the immediate surroundings of the four probes. This reduction in probe density increases the rate, and strength, of hybridisation between target RNA sequences containing runs of cytosines and the neighbouring probes, all of which contain runs of guanine. This results in cross-hybridisation dominating for these probes, and the G-spot
probes not detecting the target RNA for which they were chosen. This accounts for why the G-spot sequences are poorly correlated with other probes that are able to measure target RNA reliably.

We aim to focus on Exon arrays in order to examine whether the problems found in 3’ array, specially the misbehaviour of G-spot probes, also affect Exon arrays. Although we have only used Human Exon 1.0 ST arrays throughout our study, our results should apply to any Affymetrix Exon array.

2 Method

Section 2.1 explains our approach to analyse Affymetrix Exon arrays and section 2.2 describes cloud computing, a high-performance technology we have adopted for this study.

2.1 Our approach

We have designed a pipeline to analyse Affymetrix exon arrays, downloaded from NCBI’s Gene Expression Omnibus (GEO). Our pipeline processes CEL files, the data files that contain average fluorescence intensity of each probe in the array. The pipeline includes unique mapping of probes to exons, calibration process for quality control analysis, and creation of heatmaps for all Ensembl-defined exons. We have adopted a similar methodology as our previous work on 3’ arrays [6], but necessarily ignore the contribution from MM probes (which are missing on Exon arrays).

2.1.1 Unique probe mappings

We avoid multiple targeting probes, probes that align exactly to more than one places in the genome. We only consider those probes which are uniquely mapped to an exon, in order to reduce the effect of cross-hybridisation. Although we end up throwing away lots of probes, the remaining set should provide reliable measurements. We have described previously [6] that we consider a probe to be uniquely mapping to an exon if it aligns 25 bases (100% identification) to only one exon and to any synonymous exons i.e. the probes map to the same genomic region). Moreover, we insist that the 100% identification should be to only one place on the exon and the probe should not map to any exon-exon junction. Furthermore, the probes should not map partially (20 or more bases) to any other exon. [6] provides more details about our way of establishing unique mappings.

2.1.2 Calibration process

[7] reported that by introducing some amount of spatial noise, they found that many gene expression measures more than doubled. Thus an important issue in the analysis of microarray is quality control. For better analysis we therefore applied calibration process that includes normalisation and detection of spatial flaws at probe level in all CEL files.
2.1.3 Heatmaps for all Ensembl exons

The last phase is to generate heatmaps of all Ensembl exons. If an exon is within transcripts in the sample then all the probes detecting that exon should, ideally, respond in the same way, i.e. the fluorescent signals from these probes should be correlated. We examine the correlation coefficient value between the pairs of probes using the processed CEL files and then generate the heatmap for quick visualisation and easy analysis. We use a heatmap as a graphical representation of the correlations between the levels of expression of different probes across a number of samples. Each cell in a heatmap is colour coded according to a colour scale. In our case, this colour scale is chosen according to the cell’s correlation coefficient value. A bright cell usually indicates highly correlated pairs of measurements while a darker cell represents a poor/low correlation. All the cells in diagonal always contain 10 because correlation coefficient value for a probe with itself is always 1.

These heatmaps can show a wide range of different behaviours of G-spot probes, some of them are expected and some may be new. Figure 3 illustrates a heatmap that is representing correlation between pairs of 7 probes uniquely mapping to Ensembl exon ENSE00000330846. Different colour shades are representing different correlation coefficient between probe pairs. Probes 2, 3, 4, and 5 are represented by bright cells as these probes have high correlation among them, whilst probes 1 and 6 are behaving as outliers and are poorly correlated with other probes, but partially correlated with each other.

<table>
<thead>
<tr>
<th>N. Probes</th>
<th>Xy</th>
<th>ProbesSetEmb</th>
<th>Position</th>
<th>Sequence</th>
<th>G.mean</th>
<th>S.Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>5108919</td>
<td>1718,1995</td>
<td>3927112</td>
<td>231</td>
<td>dgcagtctcgtgcgcggtagc 120</td>
<td>1.47</td>
</tr>
<tr>
<td>6</td>
<td>22701137</td>
<td>1876,886</td>
<td>3937112</td>
<td>227</td>
<td>agcagctctgcggcggcgcggc 136</td>
<td>1.38</td>
</tr>
<tr>
<td>5</td>
<td>3153467</td>
<td>2106,1231</td>
<td>3937111</td>
<td>125</td>
<td>agcagcgcgcggcggcgggagggg 290</td>
<td>1.97</td>
</tr>
<tr>
<td>4</td>
<td>12315114</td>
<td>1837,486</td>
<td>3937111</td>
<td>124</td>
<td>tgcatctcgctcgtgcgcggc 307</td>
<td>1.98</td>
</tr>
<tr>
<td>3</td>
<td>10395631</td>
<td>2160,1942</td>
<td>3937111</td>
<td>123</td>
<td>ggcgcgcgcgcgcggcgcgcgc 318</td>
<td>2.02</td>
</tr>
<tr>
<td>2</td>
<td>17926483</td>
<td>2306,2341</td>
<td>3937111</td>
<td>121</td>
<td>ggcgcgcgcgcgcgcggcgcgc 326</td>
<td>1.92</td>
</tr>
<tr>
<td>1</td>
<td>17090866</td>
<td>1942,66</td>
<td>3937109</td>
<td>25</td>
<td>agcagcgcgcggtggtggtgcggc 252</td>
<td>1.29</td>
</tr>
</tbody>
</table>

Figure 3: Collection of all probes that represent Ensembl exon ENSE00000330846. First column contains sequence number in which these probes appear in heatmap. Next column have probe IDs followed by X and Y positions, probe set ID, their position in Ensembl exon, sequence of nucleotide, geometric mean, and standard deviation. The values inside each cell of heatmap represent the rounded value of (correlation x 10).

We have confirmed in our previous work that in GeneChip arrays G-spot probes are poorly correlated with other members of their probesets but they are highly correlated with each other [5]. Thus, we are expecting same behaviour of G-spot probes in exon arrays.
2.2 Cloud Computing

There was a time when companies used their own generators to produce electricity for running their factories or plants. This usually required a large capital expenditure when purchasing dynamos, and also required maintenance costs. But this business model was quickly dropped and companies started to buy electricity from a utility supplier of electricity. This was because it proved to be cheaper to buy electricity as a commodity, and more efficient to not have to worry about maintenance and updating equipment.

It is becoming increasingly apparent that computing is performing a similar transition at present, with computing, and other information technologies, being sold as a commodity which can be purchased from utility suppliers. The availability of significant computational opportunities is being made available by several companies, such as Amazon, Google, and Microsoft. They provide high-performance solutions that enable users to utilise their computational infrastructure, and to only pay for the resources used. The web services platform of these organisations are suitable for user groups of any size, including individuals. The “Cloud computing” concept is very simple: the computing resources are located somewhere (not in your office/computer room) and you will connect to them and use them according to your requirement.

Cloud computing enables bioinformaticians to avoid setting-up of expensive computing resources. This makes life easier for bioinformaticians, as they will avoid capital expenditure on computers which rapidly devalue in value, minimise the time and effort required to maintain infrastructure and remove the requirement of space and cooling systems needed to house the computers. We expect that cloud computing will be widely adopted by bioinformaticians in the near future. Furthermore, cloud computing is a green technology, as the carbon footprint of one large datacentre is much less than that of many groups housing the own inefficient computational infrastructure. Moreover, many users can easily gain access to shared data on the cloud, and don’t have to worry about the inconvenience of managing, and paying for, lots of data transfer.

We have begun to explore the use of cloud computing through Amazon’s platform. However, when using Amazon Web services, we do not have any long term commitments. They provide us with the flexibility to choose any development platform or programming model that is most appropriate for the problems to be solved. AWS are providing different services which includes Amazon Elastic Compute Cloud (Amazon EC2), Amazon Simple DB, Amazon Simple Storage Service (Amazon S3), Amazon CloudFront, Amazon Simple Queue Service (SQS), Amazon Elastic MapReduce, AWS Premium Support.

AWS is already hosting some public data sets, including Ensembl and some of the NCBI databases [8]. We expect that Ensembl and NCBI will continue their practice of uploading all their data, as it grows beyond the petabyte scale[9]. This is beneficial to our work, as we already use several of these databases, and we do not need to cover the costs of uploading this data.

We use Amazon Elastic Compute Cloud (EC2) to use high computing power and Amazon Simple Storage Service to store our own data; whilst Amazon’s public data sets enable us to use some of the Ensembl and NCBI data freely. To use Amazon EC2 service, an Amazon
Machine Image (AMI) is required. An AMI is an encrypted machine image that contains all the information required to boot instance of our software and it stores in Amazon Simple Storage Service (S3). In our case, we created our own AMI but one can use public AMIs as well. Bundling an AMI performs certain tasks related to confidentiality and authentication which include compression of AMI in order to minimise bandwidth usage and storage requirements, encryption of the AMI, breaking down the encrypted AMI into smaller chunks to upload, and creation of a file that contains the details about image’s small chunks with their checksum values. Then one or more instances can be launched for that AMI and finally we administer these instances as we do on our server. The block diagram to show the flow of EC2 is depicted in the figure 4.

![Figure 4: Amazon Elastic Cloud Compute (EC2) Flow (Source: http://aws.amazon.com/)](image)

3 Results

We have generated heatmaps in order to analysis Human Exon 1.0 ST arrays. These heatmaps are created according to each Ensembl exon and include only the probes that are uniquely mapping to these exons. The probes are ordered in antisense (initial position < end position) and sense (initial position > end position) direction for each exon.

References

  Online Journal: [http://nar.oxfordjournals.org/cgi/content/abstract/33/9/2908](http://nar.oxfordjournals.org/cgi/content/abstract/33/9/2908)


*Online Journal*: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/19/2566?ck=nck


