Identifying the impact of G-Quadruplexes on Affymetrix $3'$ Arrays using Cloud Computing

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

Departments of Mathematical Sciences and Biological Sciences, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, United Kingdom
http://bioinformatics.essex.ac.uk/
Email: fnmemo@essex.ac.uk; owena@essex.ac.uk; osanch@essex.ac.uk; gupton@essex.ac.uk; harry@essex.ac.uk

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.

In order to cope with the rapidly expanding size of GeneChip array datasets in the public domain, we are exploring the use of cloud computing to replicate our experiments on $3'$ arrays to look at the effect of the location of G-spots (runs of guanines). Cloud computing 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 propose the advantageous use of cloud computing to perform a large data-mining analysis of public domain $3'$ 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 single-stranded DNA sequences, 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-stranded DNA sequences that have frequent occurrences of guanine runs are

*Corresponding author: harry@essex.ac.uk

capable of forming four-stranded structures, known as G-Quadruplexes, G-tetrads, or G4 DNA [1].

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. GGGAGCGGGTTGACGGGAAGGG, a segment of single stranded 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 [1].

[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.

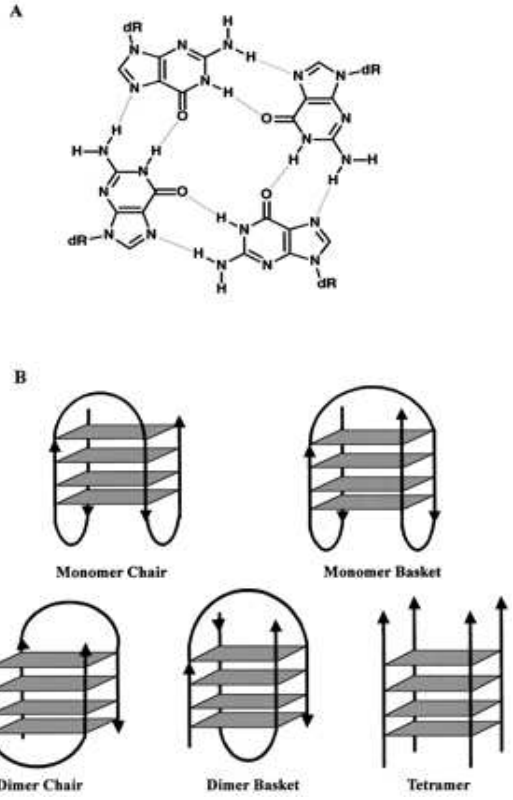

Figure 1: Schematic presentation of G-quartet structures. (A) G-quartet. (B) Different layouts/topologies and loop orientation of quadruplexes (Source: http://nar.oxfordjournals.org/cgi/content/full/31/8/2097)

Figure 1(B) illustrates a number of different topologies for G-quadruplexes. For example, the Monomer Chair and Monomer Basket show G-quadruplexes that are formed in a single nucleic acid sequence, whilst the 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 that forms through more than one sequence falls into the category of Intermolecular Quadruplex structures. Thus, the Dimer and Tetramer are both examples of Intermolecular Quadruplex structures [3]. Keeping tetramer quadruplex

structures in mind, we are investigating the implications for microarrays 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. A fluorescently labelled DNA or RNA (target sequence) will stick through hybridisation to its 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 commercially available. However, in this study we focus on the Affymetrix GeneChip, a high density oligonucleotide array. An Affymetrix GeneChip 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$\times$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 sequence. Figure 2 shows the basic construction of an Affymetrix GeneChip.

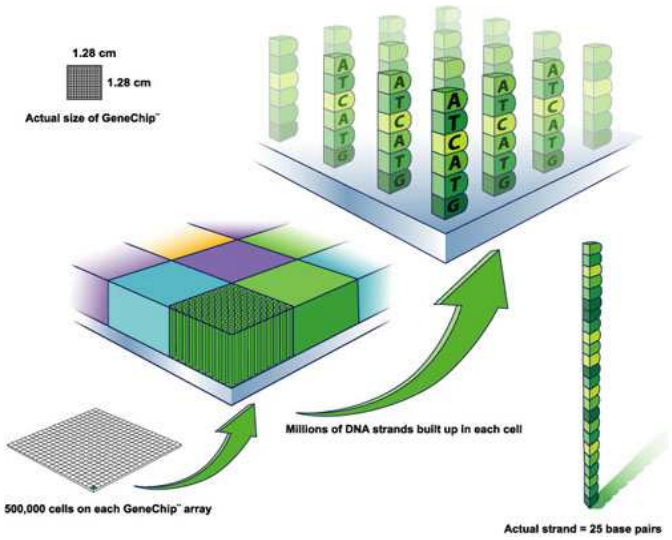

Figure 2: Basic Structure of an Affymetrix GeneChip (Source: http://electronicdesign.com/Files/29/10603/Figure_06.jpg)

Affymetrix has released GeneChips for most major model organisms. One of their most widely used designs is known as a $3'$ array, because most probes are selected towards the $3'$ region of a gene. Some cross-hybridization to other transcripts can occur even though the probes are selected to ideally avoid such cross-hybridisation. This led to the Affymetrix $3'$ design including, for each gene-specific probe, a probe that is identical in sequence except for a complementary

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 intention is that subtraction of the MM signal from the PM signal will result in a measure of a genes expression, though strategies are required to deal with the cases where it is the MM signal that is the greater. The multiple measurements of gene expression are collated into one composite expression measure.

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 recently confirmed that probes having a sequence of four or more guanines, which we termed G-spots, typically have poor correlation with other probes in their probeset [5]. However, we went further in discovering that the intensities reported from these G-spot probes are correlated with each other. We suggested that the intensities reported by these probes should not be used in the calculation of gene expression values and these G-spot probes should not be included within future array designs.

We have proposed that structures closely resembling G-quadruplexes are forming on GeneChips, and this is why probes containing runs of guanine are not fit for purpose [5]. 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.

2 Method

Section 2.1 explains our approach to analyse Affymetrix GeneChip HG_U133A 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 data from Affymetrix HG_U133A arrays, downloaded from NCBI's Gene Expression Omnibus (GEO). Although we only used HG_U133A arrays in this study, our analysis can be applied to any set of Affymetrix $3'$ arrays. 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 processes for quality control analysis, and the creation of a contour map/plot in order to find the effect of the G-run's positions within a group of probes.

2.1.1 Unique probe mappings

Rather than using information from all the probes on an array, we are selective and only use probes which are uniquely mapping to an exon, in order to reduce the effects of cross-hybridisation. We have described previously [6] that we consider a probe to be uniquely mapping to an exon if it completely aligns with 25 bases to only one exon and to any of its synonymous exons (i.e. exons located on the same genomic region, although they have different Ensembl identifiers). Moreover, we insist that the alignment of completely 25 bases should only be at one place on the exon. Furthermore, the probes should not map partially or totally (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 many gene expression measures more than doubled when they introduced typical levels of spatial noise seen in raw GeneChip data. Thus an important issue in the analysis of microarray data is quality control, and we apply a calibration process that includes normalisation of all the CEL files [8].

2.1.3 Generation of Contour map/plot

We used data from 352 randomly choosen CEL files relating to 179 Gene Expression Series (GSEs). We have focussed on looking at the effect of the location of a G-run within the probe. We first filter out the probes that have only a single occurance of exactly four guanines and are mapping uniquely to an exon. Suppose, k represents to the location of G-run within probes; the possible values of k can be 1,2,3,...,22. A 22 by 22 matrix is generated in order to create a contour plot. Each element of the matrix represents to the average correlation coefficient of two groups of probes. For example, matrix[1,2] stores an avergare correlation coefficient value

between group of probes in which k=1 and group of probes in which k=2. The contour plot shows an overview of the entire correlation surface. We have used a different dataset to that used by [5] but we find a very similar correlation contour plot (Figure 3) to that found by [5]. In particular, we confirm the unexpectedly high correlations between 4G-probes taken from different genes. We believe that this is the first published result for using the Amazon Web Services cloud for analysing GeneChip experiments.

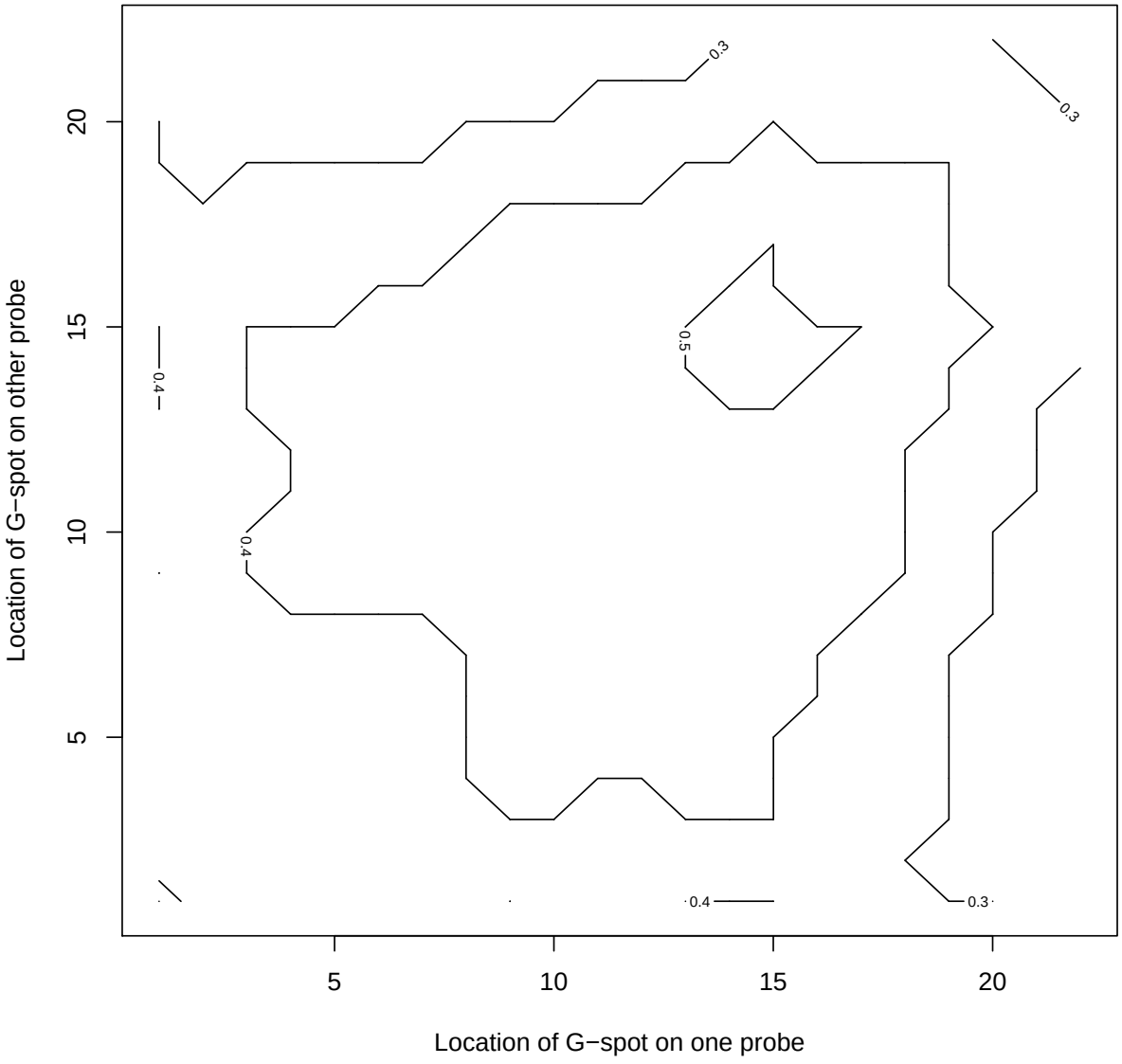

Figure 3: A contour plot showing the variation of the average correlation coefficient for pairs of probes containing 4G runs at different positions

2.2 Cloud Computing

Many companies are providing their computational resources as a utility which can be purchased from these companies. Amazon, Google, Microsoft are some examples of such com-

panies. Users can use computational infrastructure of these companies and only pay for the resources that are used. Any size of user groups and the individuals can easily use the web services platform of these organisations. Simply, we can say that the computing resources of large companies are located somewhere in the world; you have to connect to them through internet and use the required resources and pay for them.

Cloud computing enables bioinformaticians to avoid capital expenditure on computers which rapidly decrease in value. It also minimises the time and effort required to maintain large clusters and removes the requirement for 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 their 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 website, though Amazon does not require any long term commitment of its users. They provide us with the flexibility to choose any development platform or programming model to solve the problems. Amazon Elastic Compute Cloud (Amazon EC2), Amazon Simple DB, Amazon Simple Storage Service (Amazon S3) are some of the services that Amazon Web Services (AWS) provide.

AWS is already hosting some public data sets, including Ensembl and some of the NCBI databases [9]. We expect that Ensembl and NCBI will continue their practice of uploading all their data, as it grows beyond the petabyte scale[10]. 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.

To get high computing power, we use Amazon Elastic Compute Cloud (EC2) that provides an environment to run virtual servers on demand and Amazon Simple Storage Service (S3) is used 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 a file that contains all the necessary information that are required to boot an instance of our software. These AMIs are stored in Amazon Simple Storage Service (S3). Users can either create their own AMI or use public AMIs (either as it is or with some changes). The next step, bundling an AMI, performs certain tasks related to confidentiality and authentication which include the 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 the image's small chunks with their checksum values. Then one or more instances can be launched for that AMI. We can administer these instances as we would on our own server. The block diagram to show the flow of EC2 is depicted in Figure 4.

We have developed our own AMI that is based on Ubuntu Linux 8.04 with R and Bioconductor installed. In order to save time, we created our AMI by using a public AMI. We first selected an AMI (ami-b55dbbdc) that is based on Ubuntu Linux 8.04 with some bioinformatics tools and then launched an instance of this AMI to install R and Bioconductor. Finally, we uploaded

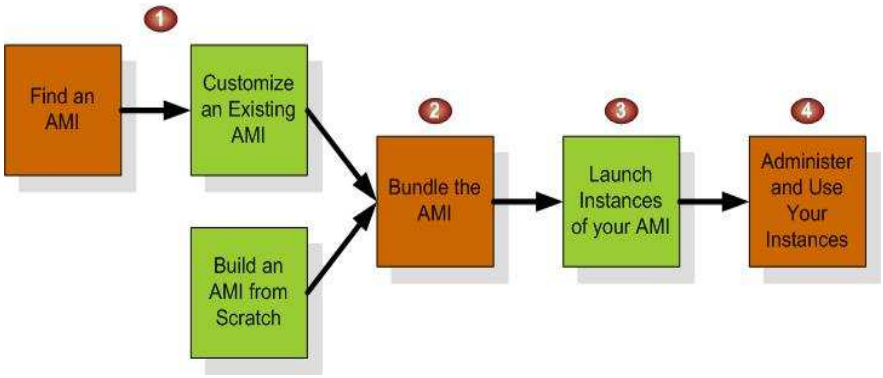

Figure 4: Amazon Elastic Cloud Compute (EC2) Flow (Source: http://aws.amazon.com/)

the new modified version of AMI to Amazon S3. We registered the AMI which assigned the AMI an ID to enable us to run instances of the AMI later on.

References

[1] J.L. Huppert and S. Balasubramanian. Prevalence of quadruplexes in the human genome. *Nucleic Acids Research*, 33(9):2908-2916, 2005
Online Journal: `http://nar.oxfordjournals.org/cgi/content/abstract/33/9/2908`

[2] V. Dapic, V. Abdomerovic, R. Marrington, J. Peberdy, A. Rodger, J.O. Trent and P.J. Bates. Biophysical and biological properties of quadruplex oligodeoxyribonucleotides. *Nucleic Acids Research*, 31(8):2097-2107, 2003
Online Journal: `http://nar.oxfordjournals.org/cgi/content/abstract/31/8/2097`

[3] V. K. Yadav, J. K. Abraham, P. Mani, R. Kulshrestha and S. chowdhury. QuadBase: genome-wide database of G4 DNA - occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Research*, 36(Database issue):D381-D385, 2008
Online Journal: `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2238983`

[4] C. Wu, H. Zhao, K. Baggerly, R. Carta and L. Zhang. Short Oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays. *Bioinformatics*, 23(19):2566-2572, 2007
Online Journal: `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/19/2566?ck=nck`

[5] G. J. G. Upton, W. B. Langdon and A. P. Harrison. G-spots cause incorrect expression measurement in Affymetrix microarrays. *BMC Genomics*, 9:613, 2008
Online Journal: `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2628396`

[6] O. Sanchez-Graillet, J. Rowsell, W. B. Langdon, M. Stalteri, J. M. Arteaga-Salas, G. J. G. Upton and A. P. Harrison. Widespread existence of uncorrelated probe intensities from

within the same probeset on Affymetrix GeneChips. *Journal of Integrative Bioinformatics*, 5(2):98, 2008
Online Journal: http://journal.imbio.de/index.php?paper_id=98

[7] M. Reimers and J. N. Weinstein. Quality assessment of microarrays: Visualization of spatial artifacts and quantitation of regional biases. *BMC Bioinformatics*, 6:166, 2005
Online Journal: http://www.ncbi.nlm.nih.gov/pubmed/15992406

[8] W. B. Langdon, G. J. G. Upton, R. S. Camargo and A. P. Harrison. A Survey of Spatial Defects in Homo Sapiens Affymetrix GeneChips. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2008
Online Journal: http://www2.computer.org/portal/web/csdl/doi/10.1109/TCBB.2008.108

[9] Amazon Web Services. http://aws.amazon.com/publicdatasets/

[10] A. Bateman and M. Wood. Cloud Computing. *Bioinformatics*, 25(12):1475, 2009
Online Journal: http://bioinformatics.oxfordjournals.org/cgi/reprint/25/12/1475