The Bustle of Bioinformatics: Cloudy with a Chance for Big Data

Samuel Chapman[1], Scott H. Harrison[2], Marwan Bikdash[1], Dukka B. KC[1]
[1] Dept. of Computational Science and Engineering
[2] Dept. of Biology
North Carolina A&T State University, Greensboro, North Carolina, USA
sdchapma@aggies.ncat.edu, scotth@ncat.edu, bikdash@ncat.edu, dbkc@ncat.edu

Abstract

The rapid pace of data production in scientific research has required new methods for analytical processing and interpretation. In the area of biology, large volumes of data have been challenging to navigate as they are generated from rapidly developing technologies and systems-level investigations. The discipline of bioinformatics has arisen to manage and analyze complex varieties of biological data. In addition to new hardware and software technologies, productivity within the cross-disciplinary arena of bioinformatics has required strategic considerations of personnel, data management architecture, and organizational support. The bustling range of methods and challenges of bioinformatics serves as a powerful example of how computational analyses of scientific phenomena are undergoing revolutionary transformations.

Keywords: Data; Bioinformatics; Parallel; Cloud; Genome; Database; Human.

1. Introduction

The era of so-called "Big Data" is upon us, where large volumes of data ranging from terabytes to exabytes are affecting many areas of research, across both public and private contexts, including physics [1], chemistry [2], astronomy [3], business [4], and biology [5]. In the area of biology, collecting and analyzing large data sets has become essential for multilevel analyses ranging from cells and genes to whole organisms and the evolution of life's diversity. These opportunities are being accelerated by radical reductions in the costs for data production, most notably for genomic sequencing [6]. Here we review big data challenges and solutions impacting the overall area of bioinformatics.

2. Biological Data Avalanche

2.1 New Technologies

More than twenty years ago, a seminal project in bioinformatics began to sequence the entire human genome of several billion base pairs. The first draft was completed in 2000, at a cost of roughly one dollar per base pair [7]. At that time, less than a hundred organisms had been fully sequenced, most of them bacteria and archaea, which have smaller genome sizes than humans and other eukaryotes. Currently, there are more than 3,000 fully-sequenced organisms, with 95% being bacterial and archaeal [8]. It is now possible to sequence an entire human genome for only a few thousand dollars [9]. In 2008, a major wave of improvement in sequencing capacity was "Next-Generation Sequencing" (NGS) where genomic material is first broken into small fragments that are sequenced [10, 11, 9]. Based on overlapping regions, sequenced fragments are then assembled into larger pieces of the original genome [10]. The aggressive pace of new efficiencies in genomic data production is eclipsing another common scenario of exponential growth, Moore's Law. In its original form, Moore's Law stated that the number of transistors on a chip doubles roughly every 24 months [12]. This is often generalized to the speed of computation, or more particularly, the speed at which data can be processed. Between 2001 and 2013, the cost to sequence a megabase of DNA dropped from approximately $5,200 to six cents; the price to sequence an entire human-sized genome dropped from approximately $95,000,000 to $5,000. These numbers represent decreases in costs by factors of 90,000 and 18,000, respectively (a more than fivefold change every 24 months). Concomitant with these reduced costs, there has been more than a million-fold improvement in the rate of DNA sequencing over the past three decades [13], leading to an estimated 15 petabytes per year capacity of genomic sequencing worldwide [5].

Genomic sequencing is only an initial stage of investigation that is often followed by alignment and annotation with respect to reference sequence data and functional inferences from the sequencing pattern. For instance, DNA sequences of protein-coding regions have been used to model the structure of the inferred protein, and help predict dynamics of intermolecular interaction and regulatory effects. Genomic sequencing technologies have been used to sequence mRNA, an intermediate product expressed from protein-coding DNA regions prior to their translation into protein [14]. In an overall methodology referred to as RNA-Seq, NGS has helped quantify levels of both coding and non-coding RNA varieties to reveal changes in transcriptional expression over time [15]. Finally, genomic sequencing data are used to detect genetic variation, including splice variants and single nucleotide polymorphisms [14]. Detection of genetic variation has been integral for multiple initiatives in biological research, including for the analysis of heritable disease [16].

2.2 Bias, Complexity and Sparsity

As databases of biological systems grow, gaps in our knowledge are becoming increasingly apparent. Some of this is due to data that have remained missing from the massive production of biological data. Absence of data can be attributed to differing factors including the extraordinary scope of life's diversity, research agenda, and cost. The level of additional inquiry that is needed is considerable. It is estimated that only 14% of species have been discovered on Earth and 9% in the ocean [17]. Among those that are known, only three thousand have had their genomes fully sequenced [8]. There have been hints and speculation about entire branches of life that are undiscovered. In 2012, a team in

Norway discovered a single-celled eukaryote that did not seem to belong to any known classification of life [18]. In 2011, versions of highly-conserved genes from environmental samples led to the conjecture of a fourth domain of life [19]. Biological data production and collection has been overall biased heavily in favor of a small number of model organisms. Model organisms are often selected based on whether they can be used to answer specific scientific questions and their ease of use in the laboratory. Answers that are obtained from model organisms however do not necessarily hold true for others, even ones that are similar [20].

Despite improved efficiencoies for the production of biological data, gaps in experimental analyses remain. For instance, gene expression over time requires samples to be taken across desired intervals, and this can be costly [21]. Even as data sets grow to cover multiple temporal states or other factors of a biological system, major problems have arisen from the Curse of Dimensionality and the Curse of Data Sparsity. The Curse of Dimensionality refers to the fact that large numbers of classes, such as genes, can be present in data sets. A large number of classes often associates with the Curse of Data Sparsity where some of these classes do not have enough sample measurements to provide for conclusive modeling [22]. Another common challenge that follows from sparsity includes the difficulty of profiling the underlying, true distribution of values expected from a class.

2.3 Data Repositories

In addition to raw data available from major public repositories which include genomic, proteomic and other biological data [23, 24], there are over a thousand other biological databases many of which have refined and curated information [25]. In terms of repositories that provide basic annotations and reflect the overall amount of data for gene and protein sequences, the RefSeq and Pfam databases are common starting points for bioinformaticists, and continue to grow dramatically. The RefSeq database [26] contains curated non-redundant gene and protein sequence information. At the time of the first release in June 2003, there were 4.3 billion DNA base pairs representing 65,000 genes, along with 263 million amino acids in the database representing 785,000 proteins. As of March 2014, there were 395 billion base pairs representing 5.5 million genes, along with 13.1 billion amino acids representing 37.8 million proteins. In summary, gene and protein sequence information stored in RefSeq has expanded by more than a magnitude over the past decade [27]. The Pfam database is a catalogue of proteins arranged into protein families [28]. The Pfam database has exploded in size. The first release in 1996 contained 100 protein families comprised of ten thousand proteins and 2.2 million amino acids. The most recent release from 2013 contains 14,800 families comprised of 18.5 million proteins and 4.4 billion amino acids [29].

3. Cloud Computing

Cloud computing refers to the access to software and computing hardware from a provider such that an individual can use these resources irrespective of location [30]. Cloud computing offers on-demand service and an enormous amount of resources to individuals and small organizations, resulting in a high level of scalability [30]. The installation and configuration of software applications within the cloud is expected to drive Software as a Service (SaaS) business models that may be especially fruitful for enabling analyses that would otherwise require applications that are hard to install, maintain and configure. Another aspect of cloud computing that may be very helpful to the bioinformatics community is Data as a service (DaaS). The transfer of large amounts of data can take a long time, and databases may not be well-maintained. A cloud-based service can keep large volumes of data situated in one physical location and the costs of database maintenance can be distributed across the set of users. Here, we list some cloud-based solutions for bioinformatics analyses.

- **NGS sequence analysis workflow tools.** Galaxy is a web-based platform for coordinating workflows across bioinformatics applications, and for which cloud-based implementations have been developed [31, 32, 33]. Eoulsan is a tool that supports multiple analysis tasks on distributed computers, and can be used with Galaxy [34].
- **Sequence similarity comparison.** BLAST searches for sequences that are close to a query sequence [35]. A cloud-based implementation of BLAST is called CloudBLAST [36].
- **Scalable searching of large sequence data.** BioPig is a Hadoop-based toolkit that breaks down and process large datasets across multiple computers such as in a cloud [37]. BioPig is designed for scalable sequence analysis (up to 500 gigabytes of sequence data) and for ease in programming. SeqPig is another library and toolkit based on Hadoop [38].
- **Read mapping to reference genomes.** CloudAligner [39] and CloudBurst [40].
- **Assembly of reads without a reference genome.** Contrail [41].
- **SNP calling.** Crossbow [42].
- **Analysis of differential RNA expression.** Myrna [43].
- **Conversions of BAM files.** Hadoop-BAM processes NGS alignments that are stored in BAM (Binary Alignment/Map) format [44].
- **Protein-ligand binding sites.** Cloud-PLBS (Cloud Protein-Ligand Binding Site) [45].
- **Digital subtraction.** Human sequencing data may contain sequencing reads from host-associated microorganisms [46]. Pathseq is a cloud-based tool for subtractive identification of microbial sequences in human sequence data [47].

4. Parallel Processing

4.1 Parallel Computing Hardware

Parallel computing, as opposed to regular serial computing, refers to the use of more than one processing unit, such as a chip core or separate computer, to complete a task. By separating a problem into smaller pieces, many computers can

work on the pieces and solve the problem faster. There are two general categories of parallel computing: distributed computing and cluster computing.

Distributed or grid computing is a software system in which components are located on a network of computers and actions are coordinated through message passing without a central server [48]. Cloud computing is similar to distributed computing; the primary difference is for cloud computing to operate in a business model that outsources equipment maintenance overhead to an external organization [49]. Folding@Home [50] is a distributed computing project that simulates protein folding, computational drug design, and other types of medical dynamics. This project uses idle processing resources of thousands of personal computers owned by volunteers. As of March 2014, all of the Folding@home computers taken together have a speed of nearly 43,000 teraflops [51], faster than the fastest supercomputer on Earth, the Tianhe-2 [52]. Another distributed computing project is Rosetta@home for protein structure prediction [53].

Cluster computing is an architecture of parallel computing where many nodes, each having several processors and disks, are linked together by high-speed interconnections [54]. For large enterprises, this allows for optimizations across hardware architecture, since there is full control over the hardware, and cluster computing does not have the same privacy concern as with cloud computing, because data remain within the network. Major research institutions and universities often have a cluster computing infrastructure dedicated for bioinformatics research.

2.4 Parallel Programming

In order for multiple pieces of hardware to work together in parallel, the programs themselves must be written to work in a parallel fashion so that the data can be processed. A number of tools are available for parallel programming, such as OpenMP [55], Message-Passing Interface (MPI) [56] and CUDA [57], which makes use of graphics cards. The programmer must also take into account the type of problem, which affects the approach used.

There are two broad categories of parallel programming problems. The easier and faster type is the so-called "embarrassingly-parallel" problem. This type of problem can be broken up into different parts so that each computing unit can work on its own without the need for communication between units. The final result can then be combined when needed. An example in bioinformatics is the BLAST algorithm, mentioned earlier [35]. One way to parallelize BLAST is to divide query sequences among the different processors, and then pool together the overall set of results from each list of queries [58].

Another type of parallelism is referred to as distributed. In this case, the different computing units must communicate with one another in order to solve the problem. BLAST can be parallelized in this way by splitting up the sequence database into smaller databases. The full list of queries is searched against each sub-database; intermediate results from each sub-database are then merged and the search is refined [59, 60]. Another example of distributed parallelism is the Folding@home project that seeks to determine the correct three-dimensional structures of proteins from their sequence of amino acids [50].

5. Parallel Algorithmic Strategies

Several primary algorithmic strategies in the analysis of large biological data sets include parallel processing, transformation of data structures, and data set reduction. These strategies are interrelated. While parallel processing offers great advantages, depending on the problem, it may not always increase computational throughput and reduce memory usage to adequate levels. One solution is to use different data structures. This has been an essential approach for metagenomics. Metagenomics is the process of sampling genomic sequences from the environment, especially for scenarios involving multiple, diverse lineages. The multitude of sequences are broken up into smaller reads and assembled to comparatively profile the diversity of organismal lineages and genomic content within that environment. A common method is to use connectivity graphs called de Bruijn graphs [61]. However, the assembly of de Bruijn graphs requires a great deal of memory, because many sequences must be compared at once. Depending on the number of reads in the metagenomic sample, terabytes of memory may be required [61]. A recent innovation for de Bruijn-based approaches has been the use of a data structure called a Bloom filter [61]. Bloom filters are based on an indexing structure that works within a fixed memory limit by implementing membership queries through a probabilistic structure, and the memory needed for graph traversal decreases by 95.0% to 97.5% compared to other de Bruijn-based approaches.

Another interesting strategy for dealing with the ever-increasing amount of raw data is to only utilize a subset for analysis. While this does not provide for a fully exact treatment, in many cases it can provide an answer that is essentially the same as an exact one. A very common type of subsampling is called bootstrapping, which repeatedly subsamples items in a data set in order to estimate a statistic from the overall distribution [62]. In this way, only some data need to be examined. Bootstrapping is classified as an embarrassingly-parallel problem, so parallel programming methods can be used with it [63]. A recent development has made use of a variation of bootstrapping, called Bag of Little Bootstraps [63] which averages the results of several groups of bootstrapped subsamples. This method has been shown to provide accurate confidence intervals of data, including biological data, in a much faster time period than other bootstrapping methods.

Other methods for subset selection are based on reducing the level of data associated with each specimen or item being analyzed across a group. One such method is feature selection, which only looks at a restricted set of features in high-dimensional data sets [64]. Another strategy has been for reducing sequence assembly data sets in a method called "Digital Normalization." During sequence assembly, multiple overlapping reads can map to the same place in the genome, providing a depth of coverage. Without knowing the underlying sequence of an organism, it is possible to estimate coverage by examining the numbers of different k-mers, or possible "words" of size k, within the reads. Digital

normalization recognizes that only a particular level of median coverage may be needed across the genome to get a satisfactory result. Therefore, if it determines that a new read has coverage above a certain threshold, it simply discards it. In practice, digital normalization requires a much lower amount of memory and time compared to other assembly methods and has comparable error rates [65].

6. The Human Enterprise

As it spans from human activity to endpoints of productivity, there are economic implications and disciplinary factors to the navigation of large-scale biological data. Increased time and resources for upstream and downstream analyses are becoming major costs for projects with large biological data sets [66]. Efficiency to these investments relies upon decisions associated with hardware and software infrastructure, comprehensiveness of data collection and analysis, and the recruitment, training, and selection of personnel. Inefficiencies may be expected to have profound implications for important areas of application including human health, and we suggest that there may be additional instances beyond those travails that have been well-publicized such as caBIG and HL7 [67, 68]. Strategic investment into the next generation of qualified cross-disciplinary personnel for big data biology is essential for scientific advancement [69].

The nuts-and-bolts aspect of being cross-disciplinary has increasingly required researchers in the life sciences to adopt scripting environments such as python, R, and the command-line to be their 21st century "scientific calculator." Scripting environments have an ideal sandbox-type framework for inventiveness and emulation. These environments help surmount the lack of prior art upon the changing landscape of biotechnologies and research questions and to emulate analyses that may not be directly reproducible due to bioinformatics tools and algorithms whose guidelines for usage are complex or cryptic, whose overall testing and usage is limited, or whose maintenance has atrophied [70]. Scripting environments expedite the revisiting of basic assumptions upon which an infrastructure of bioinformatics software may be built. For instance, significant debate and discussion has been ongoing with categorizations of species [71, 72], noisiness of coverage and annotation [73], measurement of evolutionary pressure and change [74, 75], and multiple classification schemes on genes and pathways [76, 77]. In general for bioinformatics research, sharpened quantitative skills built jointly with formative experiential training are needed for analyses that are both powerful and accurate. Beyond the elegance of math, which has a commercial benefit for members of the workforce [78], the need for computational thinking [79] and software carpentry [80, 81] to work with the heterogeneity of biological data continues to rise.

Interactive workflows represent the non-automated aspect of human-mediated computational analysis upon which pipelines can be selectively invoked or synthesized into larger constructs. Several trends indicate emerging directions for bioinformatics workflow tools. Web-based systems continue to increase [82, 83, 84, 85]. For the navigation and assembly of data sets across different spans of data, data-mart type approaches are a common technique of many web-based systems [86, 87, 88, 89]. The new wave of cloud-based tools is expected to further reduce maintenance overhead and invigorate the availability of bioinformatics resources [90]. Many of the client-side graphical user interface tools that remain most popular are built with cross-platform languages (e.g., java) and have a user and development community active enough to support predictably high levels of usage and reliability across different operating system platforms [91, 92, 93, 94].

The interplay between non-automated workflows versus pipelines, which are fully automated computations, relates in part to how one person's science is another person's technology. A classic exposition of this science versus technology divide is on the topic of statistics [95, 96], but this divide carries forward into protocols and algorithms involved with separation science and annotation of biomolecular data sets such as DNA, RNA, proteins, and metabolites. Within institutions, a common approach for bringing together an overall ensemble of expertise has been with service-oriented cores which act as community knowledge bases. Across institutions, consortia have been playing a leading role in shepherding larger-scale efforts, typically with an agenda for greater comprehensiveness of data collection and annotation. These consortia have had greatest activity for model organisms such humans [97] and mice [98]. Other consortia like the Human Microbiome Project have recently emerged to chart the combinatoric complexity of communities and coevolutionary change [99]. Industry is becoming increasingly involved in establishing workflow tools that build upon applications released by commercial and non-commercial developers [100]. Recent developments from non-commercial funding agencies have sought to advance integrative data analysis approaches with, for instance,the Department of Energy Systems Biology Knowledgebase [101] and the increasing commitment of the National Institutes of Health for translational medicine [102]. In summary, productive outcomes in big data biology require an in-depth commitment to training across the life science and computational science disciplines, attentiveness to the architectural options for workflows and pipelines, and organizational means for support of the enterprise.

7. Discussion and Conclusion

The degree by which new methods may keep pace with large biological data sets remains to be seen. Considering the explosion and variety of new bioinformatics data resources and technologies that have been produced in the past ten years, it is difficult to predict the next ten years. Fifteen years ago, it was a multiyear, multibillion dollar triumph to sequence a single human's genome, but new technologies may soon allow for the sequencing of a human genome for $1,000 within two hours [103]. Beyond the goal for obtaining a comprehensive range of genomic sequencing data across life's diversity, other challenges include the analysis of whole populations within their ecosystems, the accurate annotation of gene functions, and the modeling gene product interactions with other cellular components. The near future may further involve fantastical scenarios of bioinformatics resources driving the accelerated development of artificial changes to life in the area of synthetic biology [104]. A sobering reminder of current limitations in bioinformatics is that, despite the large amount of

bioinformatics research performed on humans, analysis and interpretation remains incomplete. Genome-wide association studies (GWAS) have been able to account for only a fraction of the heritability of various human traits [105]. For both now and as may be expected for the future, the increasing volume of large biological data sets is running parallel to an ongoing upheaval in laboratory and computational technologies.

Acknowledgment

This material is based in part upon work supported by the National Science Foundation under Cooperative Agreement No. DBI-0939454. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors would also like to acknowledge an AWS in Education Research Grant from Amazon.

References

[1] Brumfiel, Geoff: Down the petabyte highway. Nature, **469.20**, 282--283 (2011)

[2] Lusher, Scott J., McGuire, Ross, van Schaik, René C., Nicholson, C. David, de Vlieg, Jacob: Data-driven medicinal chemistry in the era of big data. Drug discovery today, (2013) DOI: [http://dx.doi.org/10.1016/j.drudis.2013.12.004]

[3] Stephens, Matt: Mapping the universe at 30 terabytes a night: Jeff Kantor, on building and managing a 150 Petabyte database. [Internet]. The Register (October 3 2008) [cited April 1 2014]. Available from: [http://www.theregister.co.uk/Print/2008/10/03/lsst_jeff_kantor/]

[4] McAfee, Andrew and Brynjolfsson, Erik: Big data: the management revolution. Harvard business review, **90.10**, 60--68 (2012)

[5] Schatz, M. C. and Langmead, B.: The DNA data deluge. IEEE Spectrum, **50.7**, 28--33 (2013)

[6] Mardis, E.: A decade's perspective on DNA sequencing technology. Nature, **470**, 198--203 (2011)

[7] Office of the Press Secretary, White House: President Clinton announces the completion of the first survey of the entire human genome [Internet]. Human Genome Project Information Archive (2000) [cited March 29 2014]. Available from: [http://web.ornl.gov/sci/techresources/Human_Genome/project/clinton1.shtml]

[8] Genomes Online Database (GOLD) statistics [Internet]. [updated March 31 2014; cited April 1 2014]. Available from: [http://genomesonline.org/cgi-bin/GOLD/index.cgi]

[9] Wetterstrand, Kris: DNA sequencing costs: data from the NHGRI large-scale genome sequencing program [Internet]. Large-Scale Genome Sequencing Program. National Human Genome Research Institute (2012) [cited March 29 2014]. Available from: [http://www.genome.gov/sequencingcosts/]

[10] Hall, Neil: Advanced sequencing technologies and their wider impact in microbiology. Journal of Experimental Biology, **210.9**, 1518--1525 (2007)

[11] Quail, Michael A., et al: A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC genomics, **13.1**, 341 (2012)

[12] Moore's law [Internet]. Encyclopedia Britannica Online (2014) [cited March 29 2014]. Available from: [http://http://www.britannica.com/EBchecked/topic/705881/Moores-law]

[13] Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. Nature, **458(7239),** 719-724.

[14] Nagalakshmi, Ugrappa, et al: The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, **320.5881**, 1344--1349 (2008)

[15] Nie, Lei, Wu, Gang, and Zhang, Weiwen: Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in Desulfovibrio vulgaris: a quantitative analysis. Genetics, **174.4**, 2229--2243 (2006)

[16] Amberger, J., Bocchini, C. and Hamosh, A.: A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). Human Mutation, **32.5**, 564-567 (2011)

[17] Camilo, Mora, Tittensor, Derek P., Adl, Sina, Simpson, Alastair G.B. and Worm, Boris: How many species are there on Earth and in the ocean? PLoS biology, **9.8**, e1001127 (2011)

[18] University of Oslo: Rare protozoan from sludge in Norwegian lake does not fit on main branches of tree of life [Internet]. ScienceDaily; (April 26 2012) [cited March 29 2014]. Available from: [http://www.sciencedaily.com/releases/2012/04/120426104853.htm]

[19] Wu, Dongying, et al: Stalking the fourth domain in metagenomic data: searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees. PLoS One, **6.3**, e18011 (2011)

[20] Gilbert, Scott: The adequacy of model systems for evo-devo: modeling the formation of organisms/ modeling the formation of society. Mapping the Future of Biology, Boston Studies in the Philosophy of Science, **266**, 57-68 (2009)

[21] Falin, Lee J. and Tyler, Brett M.: Using interpolation to estimate system uncertainty in gene expression experiments. PLoS One, **6.7**, e22071 (2011)

[22] Somorjai, Ray L., Dolenko, B. and Baumgartner, Richard: Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics, **19.12**, 1484--1491 (2003)

[23] Sayers, E. W., Barrett, T., Benson, D. A. et al.: Database resources of the national center for biotechnology information. Nucleic acids research, **38.Suppl 1**, D5--D16 (2010)

[24] Vizcaíno, J. A., Côté, R., Reisinger, F. et al: The Proteomics Identifications Database: 2010 update. Nucleic Acids Research, **38.Suppl 1**, D736--D742 (2010)

[25] Fernández-Suárez, X. M. and Galperin, M. Y.: The 2013 Nucleic Acids Research Database Issue and the Online Molecular Biology Database Collection. Nucleic acids research, **41.D1**, D1--D7 (2013)

[26] Pruitt, Kim D., Tatusova, Tatiana and Maglott, Donna R.: NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research, **35.suppl 1**, D61--D65 (2007)

[27] RefSeq ftp available at: [ftp.ncbi.nlm.nih.gov]

[28] Bateman, Alex, et al: The Pfam protein families database. Nucleic acids research, **32.suppl 1**, D138--D141 (2004)

[29] Pfam ftp available at: [ftp.sanger.ac.uk]

[30] Armbrust, Michael et al: A view of cloud computing. Communications of the ACM, **53.4**, 50--58 (2010)

[31] Giardine, Belinda et al: Galaxy: a platform for interactive large-scale genome analysis. Genome research, **15.10**, 1451--1455 (2005)

[32] Afgan, Enis et al: Galaxy CloudMan: delivering cloud compute clusters. BMC bioinformatics, **11.Suppl 12** S4 (2010)

[33] Liu, Bo, Li, Jianqiang, and Liu, Chunchen: Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses. Journal of biomedical informatics (2014)

[34] Jourdren, Laurent, Bernard, Maria, Dillies, Marie-Agnès and Le Crom, Stéphen: Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. Bioinformatics, **28.11**, 1542--1543 (2012)

[35] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology, **215.3**, 403--410 (1990)

[36] Matsunaga, Andréa, Tsugawa, Maurício and Fortes, José: Cloudblast: combining mapreduce and virtualization on distributed resources for

bioinformatics applications. eScience, 2008. IEEE Fourth International Conference on. IEEE, (2008)

[37] Nordberg, Henrik, Bhatia, Karan, Wang, Kai and Wang, Zhong: BioPig: a Hadoop-based analytic toolkit for large-scale sequence data. Bioinformatics, **229.23**, 3014--3019 (2013)

[38] Schumacher, André et al: SeqPig: simple and scalable scripting for large sequencing data sets in Hadoop. Bioinformatics, **30.1**, 119--120 (2014)

[39] Nguyen, Tung, Shi, Weisong, and Ruden, Douglas: CloudAligner: a fast and full-featured MapReduce based tool for sequence mapping. BMC research notes, **4.1**, 171 (2011)

[40] Schatz, Michael C.: CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics, **25.11**, 1363--1369 (2009)

[41] Schatz, Michael C., Sommer, Dan, Kelley, David and Pop, Mihai: Contrail: Assembly of large genomes using cloud computing. CSHL Biology of Genomes Conference (2010)

[42] Langmead, Ben, Schatz, Michael C., Lin, Jimmy, Pop, Mihai and Salzberg, Steven L.: Searching for SNPs with cloud computing. Genome Biol, **10.11**, R134 (2009)

[43] Langmead, Ben, Hansen, Kasper D., and Leek, Jeffrey T.: Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol, **11.8**, R83 (2010)

[44] Niemenmaa, Matti et al: Hadoop-BAM: directly manipulating next generation sequencing data in the cloud. Bioinformatics, **28.6**, 876--877 (2012)

[45] Lei, Xie and Bourne, Philip E.: A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. BMC bioinformatics, **8.Suppl 4**, S9 (2007)

[46] Schmieder, Robert and Edwards, Robert: Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PloS One **6.3**, e17288 (2011)

[47] Kostic, Aleksandar D. et al: PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nature biotechnology, **29.5**, 393--396 (2011)

[48] Foster, Ian: What is the grid? a three point checklist [Internet]. (July 20, 2002) [cited March 29 2014]. Available from: [http://dlib.cs.odu.edu/WhatIsTheGrid.pdf]

[49] Kondo, Derrick, Javadi, Bahman, Malecot, Paul, Cappello, Franck and Anderson, David P.: Cost-benefit analysis of cloud computing versus desktop grids. IEEE International Symposium on Parallel & Distributed Processing (2009)

[50] Larson, Stefan M., Snow, Christopher D., Shirts, Michael and Pande, Vijay S: Folding@ Home and Genome@ Home: Using distributed computing to tackle previously intractable problems in computational biology (2002)

[51] Folding@home client statistics by OS [Internet]. [updated March 29 2014; cited March 29 2014]. Available from: [http://fah-web.stanford.edu/cgi-bin/main.py?qtype=osstats2]

[52] Meuer, Hans, Strohmaier, Erich, Simon, Horst, and Dongarra, Jack: The TOP500 List [Internet]. November 2013 [cited March 29 2014]. Available from: [http://www.top500.org/list/2013/11/]

[53] Rosetta@home [Internet]. [cited March 29 2014]. Available from: [https://boinc.bakerlab.org/rosetta/]

[54] Buyya, Rajkumar: High performance cluster computing. F'rentice, New Jersey (1999)

[55] Dagum, Leonardo, and Menon, Ramesh: OpenMP: an industry standard API for shared-memory programming. Computational Science & Engineering, IEEE **5.1**, 46--55 (1998)

[56] Gropp, William, Lusk, Ewing, and Skjellum, Anthony: Using MPI: portable parallel programming with the message-passing interface. Bioinformatics, **1**, MIT Press (1999)

[57] Nvidia, C. U. D. A.: Compute unified device architecture programming guide. (2007)

[58] Braun, R., Pedretti, K., Casavant, T. et al: Parallelization of local blast service on workstation clusters. Future Gener. Comput. Syst., **17.6** (2001)

[59] Darling, A., Carey, L., and Feng, W.: The Design, implementation, and evaluation of mpiBLAST. Proceedings of the ClusterWorld Conference and Expo, in conjunction with the 4th International Conference on Linux Clusters: The HPC Revolution (2003)

[60] Lin, H., Ma, X., Feng, W., and Samatova, N.: Coordinating computation and I/O in massively parallel sequence search. IEEE Transactions on Parallel and Distributed Systems (May 2010)

[61] Pell, Jason et al: Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. Proceedings of the National Academy of Sciences, **109.33**, 13272--13277 (2012)

[62] Efron, Bradley and Tibshirani, Robert: An introduction to the bootstrap. **109.33**, CRC press (1994)

[63] Kleiner, Ariel, Talwalkar, Ameet, Sarkar, Purnamrita and Jordan, Michael: The big data bootstrap. arXiv preprint arXiv:1206.6415 (2012)

[64] Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga: A review of feature selection techniques in bioinformatics. Bioinformatics, **23.19**, 2507--2517 (2007)

[65] Brown, C. Titus, Howe, Adina, Zhang, Qingpeng, Pyrkosz, Alexis B. and Brom, Timothy H.: A reference-free algorithm for computational normalization of shotgun sequencing data. Proceedings of the National Academy of Sciences, arXiv preprint arXiv:1203.4802 (2012)

[66] Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K. and Gerstein, M. B.: The real cost of sequencing: higher than you think! Genome biology, **12.8**, 125 (2011)

[67] Califano, A., Chinnaiyan, A. M., Duyk, G. M. et al: An assessment of the impact of the NCI Cancer Biomedical Informatics GRID (caBIG). (2011)

[68] Hasman, A.: HL7 RIM: an incoherent standard. Proceedings of MIE2006, **124**, 133 (2006)

[69] NAS 2010: Rising above the gathering storm, revisited: rapidly approaching category 5 [Internet]. By Members of the 2005 "Rising Above the Gathering Storm" Committee; Prepared for the Presidents of the National Academy of Sciences, National Academy of Engineering, and Institute of Medicine [cited March 29, 2014]. Available from: [http://www.nap.edu/catalog.php?record_id=12999]

[70] Veretnik, S., Fink, J. L. and Bourne, P. E.: Computational biology resources lack persistence and usability. PLoS computational biology, **4.7**, (2008)

[71] Cotterill, F. P., Taylor, P. J., Gippoliti, S., Bishop, J. M. and Groves, C. P.: Why one century of phenetics is rnough: tesponse to ‘Are there really twice as many Bovid species as we thought?. Systematic biology, syu003 (2014)

[72] Doolittle, W. F. and Papke, R. T.: Genomics and the bacterial species problem. Genome biology, **7.9**, 116 (2006)

[73] Chain, P. S. G., Grafham, D. V., Fulton, R. S. et al: Genome project standards in a new era of sequencing. Science, **326**, 5950 (2009)

[74] Pond, S. L. K. and Frost, S. D.: Not so different after all: a comparison of methods for detecting amino acid sites under selection. Molecular biology and evolution, **22.5**, 1208--1222 (2005)

[75] Boussau, B., Szöllösi, G. J., Duret, L. et al: Genome-scale coestimation of species and gene trees. Genome research, **23.2**, 323--330 (2013)

[76] Karp, P. D.: Pathway databases: a case study in computational symbolic theories. Science, **293.5537**, 2040--2044 (2001)

[77] Caspi, R., Dreher, K. and Karp, P. D.: The challenge of constructing, classifying, and representing metabolic pathways. FEMS microbiology letters, **345.2**, 85--93 (2013)

[78] Cipra, B. A.: More math means more money. Science, **243**, 314 (1989)

[79] Settle, A., Goldberg, D. S. and Barr, V.: Beyond computer science: computational thinking across disciplines. Proceedings of the 18th ACM conference on Innovation and technology in computer science education, 311--312 (July 2013)

[80] Via, Allegra et al: Best practices in bioinformatics training for life scientists. Briefings in bioinformatics, **14.5**, 528--537 (2013)

[81] Perkel, J. M.: Coding your way out of a problem. Nature methods, **8.7**, 541 (2011)

[82] Glaab, E., Garibaldi, J. M. and Krasnogor, N.: ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization. Bmc Bioinformatics, **10.1**, 358 (2009)

[83] Dereeper, A., Guignon, V., Blanc, G. et al: Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic acids research, **36.Suppl 2**, W465--W469 (2008)

[84] Suyama, M., Torrents, D. and Bork, P.: PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic acids research, **34.Suppl 2**, W609--W612 (2006)

[85] Delport, W., Poon, A. F., Frost, S. D. and Pond, S. L. K.: Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics **26.19**, 2455--2457 (2010)

[86] Markowitz, V. M., Chen, I. M. A., Palaniappan, K. et al: IMG: the integrated microbial genomes database and comparative analysis system. Nucleic acids research, **40.D1**, D115--D122 (2012)

[87] Kasprzyk, A.: BioMart: driving a paradigm change in biological data management. Database, bar049, (2011)

[88] Cole, J. R., Wang, Q., Fish, J. A. et al: Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic acids research, **42.D1**, D633--D642 (2014)

[89] DeSantis, T. Z., Hugenholtz, P., Larsen, N. et al: Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiology, **72.7**, 5069-5072 (2006)

[90] Tan, T. W., Xie, C., De Silva, M. et al: Simple re-instantiation of small databases using cloud computing. BMC Genomics, **14.Suppl 5**, S13 (2013)

[91] Thorvaldsdóttir, Helga, Robinson, James T. and Mesirov, Jill P.: Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics, **14.2**, 178--192 (2013)

[92] Robinson, James T., Thorvaldsdóttir, Helga, Winckler, Wendy et al: Integrative Genomics Viewer. Nature Biotechnology, **29**, 24--26 (2011)

[93] Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, Michèle and Barton, Geoffrey J.: Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics, **25**, 1189--1191 (2009)

[94] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S. et al: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, **13.11**, 2498--2504 (2003)

[95] Guy, W. A.: On the original and acquired meaning of the term "statistics"' and on the proper functions of a statistical society: also on the question whether there be a science of statistics; and, if so, what are its nature and "social science". Journal of the Statistical Society of London, 478--493 (1865)

[96] Fernholz, L. T. and Morgenthaler, S.: A conversation with John W. Tukey. Statistical Science, 346--356 (2003)

[97] ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome. Nature, **489.7414**, 57--74 (2012)

[98] Bono, H., Kasukawa, T., Furuno, M., Hayashizaki, Y. and Okazaki, Y.: FANTOM DB: database of functional annotation of RIKEN mouse cDNA clones. Nucleic acids research, **30.1**, 116--118 (2002)

[99] Turnbaugh, P. J., Ley, R. E., Hamady, M. et al: The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature, **449.7164**, 804 (2007)

[100] Marusina, Kate: Genetic engineering & biotechnology news. **32.15**: 1, 34--40. (September 1 2012) DOI: doi:10.1089/gen.32.15.12

[101] Nordberg, H., Cantor, M., Dusheyko, S. et al: The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic acids research, **42.D1**, D26--D31 (2014)

[102] Zerhouni, E. A.: Translational and clinical science—-time for a new vision. New England Journal of Medicine, **353.15**, 1621--1623 (2005)

[103] DeFrancesco, Laura.: Life Technologies promises $1,000 genome. Nature Biotechnology **30.2**, 126 (2012)

[104] Blakes, J., Twycross, J., Romero, F. J. and Krasnogor, N.: The Infobiotics Workbench: an integrated in silico modelling platform for Systems and Synthetic Biology. Bioinformatics, **27.23**, 3323--3324 (2011)

[105] Maher, Brendan: The case of the missing heritability. Nature, **456.7218**, 18--21 (2008)