Self-reflection I

Siddharth Tomar

May 18, 2017

Topic: Future software development challenges in cryo-EM
By: Jose Miguel de la Rosa Trevin on 8 – May – 2017 / Seminar: Exploring the Complexity of Life by Cryo-EM

Introduction

Bioinformatics is associated with creation of tools to analyze biological data and deals with development of computational methods. The presentation by Jose Miguel de la Rosa Trevin concentrated on problems encountered within this process, with a more generalized approach towards bioinformatics ending with his implementation, a possible solution in field of Cryo-EM to mitigate the problems[1]. The reason why I chose this presentation for self-reflection is because the master's programme primarily deals with creation and utilization of tools for biological data analysis, and thus software development process is closely related to the programme.

Background

Algorithms and software in bioinformatics range from essential tools like database search algorithms[2] and visualization software[3] to specialized tools for molecular dynamics[4] and medical imaging[5]. Most of the tools fall under certain categories, like "visualization" tools to "sequence" processing tools for alignment[6] and gene-ontology. Most of these tools employ common libraries or shared pieces of codes which can be used to quickly perform a particular task. For example, tools which use GPU(graphical processing unit) acceleration to speedup tasks[7] use CUDA libraries which enable processing on GPU. More often, the tools themselves are built around these common set of libraries which perform the basic functions. Jose Miguel gave example pertaining to Cryo-EM, with three commonly used packages Relion[8], Spider[9] and Xmipp[10]. These three packages share common core components, with modifications towards the end of processing pipeline. These tools for the purpose of processing during intermediate stages create files in different format which are not interchangeable. This also introduces another problem, file formats: Bioinformatics has considerable formats for the same category. For example, multiple sequence alignment programs have different formats to represent same type of data, and these different formats more often don't lead to any substantial gain in information.

Reflection

Since most of the tools in Bioinformatics are built on preexisting tools and libraries, Jose Miguel proposed a more pragmatic approach to software development, whereby news tools are created as an extension of the previous versions. This prevents code duplication and enables better tracking of development and code revision. For example, tool A which works on sequences is superseded by tool B, which is an optimized version of A. Instead of duplicating the code from tool A, it is better to use the code directly and layer modifications on top of the code. Similarly, there is a common trend in bioinformatics to create new file formats with new tools, which are either incompatible with other software or don't add any significant information. This practice not only prevents a "mix and match" pipeline (i.e. to use any software for any stage in pipeline) but also makes the data available in public domain incompatible. This can be prevented by using a more standardized approach equivalent to one used in other areas of computing; formats like zip[11] and JPEG[12] are governed by international bodies and have a standard set of rules. The method[1] developed by Jose Miguel handle different formats and parameters for different software in Cryo-EM analysis pipeline by creating an interchangeable format which can be converted to support any tool required for the given stage. This enables a more portable and reproducible pipeline, which gives flexibility to researcher so he/she can use tools they prefer for specific stages. This effort, which aims to end the gap in different implementations reflect how seemingly dissimilar tools are related and can potentially share a more open framework, only if the developers/researchers allow so in the design.

In conclusion, I would like to point out the disparity in software development practices in field of bioinformatics and life sciences in general, which are detrimental to research and dissemination of data. This has led to experimental pipelines which essentially perform identical tasks but are incompatible with each other. A possible solution to this is creation of a set of rules which define how data is handled in Bioinformatics, and use of better software development practices. But the fundamental underlying problem is the direction which research groups take, i.e. to have more impact by creating a propriety set of tools and formats, even when the tools themselves are improvement over previously existing tools; instead of having a more open approach which is beneficial for scientific community in general. Ultimately it is the mindset (whether governed by ego or "grants") which has led to this problem, and changing it would require a paradigm shift to accommodate a more universal model of development.

References

[1] J.M. de la Rosa-Trevin, A. Quintana, L. del Cano, A. Zaldivar, I. Foche, J. Gutierrez, J. Gomez-Blanco, J. Burguet-Castell, J. Cuenca-Alba, V. Abrishami, J. Vargas, J. Oton, G. Sharov, J.L. Vilas, J. Navas, P. Conesa, M. Kazemi, R. Marabini, C.O.S. Sorzano, and J.M. Carazo. Scipion: A software framework toward integration, reproducibility and validation in 3d electron microscopy. *Journal of Structural Biology*, 195(1):93 – 99, 2016.

[2] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, Sep 1997.

[3] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[4] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilard Pall, Jeremy C. Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19 – 25, 2015.

[5] R. Cheng, S. Xu, A. Bokinsky, E. McCreedy, W. Gandler, B. J. Wood, and M. J. McAuliffe. GPU based multi-histogram volume navigation for virtual bronchoscopy. *Conf Proc IEEE Eng Med Biol Soc*, 2014:3308–3312, 2014.

[6] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1):195–197, Mar 1981.

[7] M. Mielczarek and J. Szyda. Review of alignment and snp calling algorithms for next-generation sequencing data. *Journal of Applied Genetics*, 57(1):71–79, 2016.

[8] Dari Kimanius, Bjorn Forsberg, and Erik Lindahl. Accelerated cryo-em structure determination with parallelisation using {GPUs} in relion-2. *Biophysical Journal*, 112(3, Supplement 1):575a –, 2017.

[9] Tanvir R. Shaikh, Haixiao Gao, William T. Baxter, Francisco J. Asturias, Nicolas Boisset, Ardean Leith, and Joachim Frank. Spider image processing for single-particle reconstruction of biological macromolecules from electron micrographs. *Nat. Protocols*, 3(12):1941–1974, Dec 2008.

[10] C. O. Sorzano, R. Marabini, J. Velazquez-Muriel, J. R. Bilbao-Castro, S. H. Scheres, J. M. Carazo, and A. Pascual-Montano. XMIPP: a new generation of an open-source image processing package for electron microscopy. *J. Struct. Biol.*, 148(2):194–204, Nov 2004.

[11] Information technology – Document Container File – Part 1: Core. Standard, International Organization for Standardization, Geneva, CH, October 2015.

[12] William B. Pennebaker and Joan L. Mitchell. *JPEG: Still Image Data Compression Standard (Digital Multimedia Standards) (Digital Multimedia Standards S)*. Springer, 1992.