Statika: managing cloud resources, bioinformatics tools and data

Alexey Alekhin, Evdokim Kovach, Pablo Pareja-Tobes, Marina Manrique, Eduardo Pareja, Raquel Tobes and Eduardo Pareja-Tobes[*]

Oh no sequences! Research Group. Era7 bioinformatics

*eparejatobes@ohnosequences.com

Abstract. Next Generation Sequencing (NGS) has brought a revolution to the bioinformatics landscape, definitely reshaping fields such as genomics and transcriptomics, by offering sheer amounts of data about previously inaccessible domains in a cheap and scalable way. Thus biological data analysis demands, more than ever, high performance computing architectures; in particular, Cloud Computing, a comparable breakthrough in the IT world, holds promise for being the foundation on which a solution could be built (as already demonstrated by pioneering efforts such as Galaxy or CloudBioLinux). It provides a perfect framework for high throughput data analysis: deploying architectures with as much computing capacity as needed, scaling in an horizontal way, being also able to scale down adjusting to the computing needs real time, or the pay-as-you-go model make for a strong case.

However, fast, reproducible, and cost-effective data analysis in the cloud at such scale remains elusive. Certainly, one fundamental prerequisite for achieving this is having the ability to manage both the tools and data to be used in a robust, reproducible, and automated way. High throughput analysis, where a lot of resources are to be used and paid for, needs to have a robust configuration system to rely on. In the cloud computing world, due to its on-demand nature, automated resource configuration is a critical factor. This is even more so in the case of bioinformatics analysis where pretty often a pretty intricate and unstable chain of dependencies underlies tools and data; knowing beforehand that all the resources to be used are properly configured is invaluable.

Statika (http://ohnosequences.com/statika) aims to be a basic tool for the declaration and deployment of composable, versioned and reproducible cloud infrastructures for the bioinformatics space.

Data, tools and infrastructure are treated on an equal footing, and a expressive domain specific language allows the user to express complex dependency relationships, check for possible version conflicts and automatically choose a safe resource creation order.

By making use of advanced features of the Scala programming language such as dependent types and type-level computations a great deal of structure can be expressed abstractly, and checked at compile time before any cost is incurred. A strong versioning system where both data and tools are included makes reproducibility not only possible but actually enforced.

Statika has been put to work on scenarios as different as a cloud-based system for scaling inherently parallel computations in the bioinformatics domain: Nispero, or by providing versioned and modular automated deployments of Bio4j, a graph database integrating all data from key resources in the bioinformatics data space, including: UniProt, Gene Ontology, the NCBI Taxonomy or UniRef. We use it internally for the integration and automated deployment of all sort of bioinformatics tools and data.

Statika is open source, available under the AGPLv3 license.