ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich
Institute of Molecular Systems Biology

CloudBroker

IBM

SyBIT
SystemsX.ch
Biology IT

Accelerating 3D Protein Modeling Using Cloud Computing

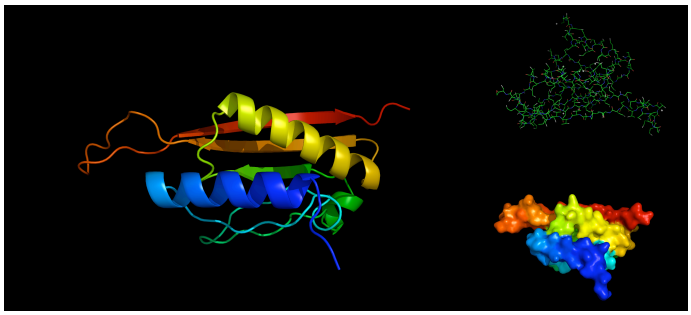
Lars Malmström, Ruedi Aebersold – ETH Zürich, IMSB http://www.imsb.ethz.ch
Wibke Sudholt, Nicola Fantini – CloudBroker GmbH http://www.cloudbroker.com
Roland Reifler, Marcel Lautenschlager, Stefan Ruckstuhl – IBM Schweiz http://www.ibm.com/ch
Peter Kunszt – SystemsX.ch, SyBIT http://www.sybit.net

Cloud Computing and Rosetta

Biology as a scientific domain needs a growing amount of computational power. However, not every researcher has access to high performance computing resources locally. Today, it is easy to buy computing resources on demand from public cloud providers like IBM, **paying only for the amount of computing that is really being used**. However, the difficulty of setting up the simulation and operating the virtual infrastructure is also often a showstopper for scientists to use cloud resources. CloudBroker fills this gap by providing interfaces to scientific software **to be used immediately** on the cloud as a service.

The **Rosetta** software suite focuses on the prediction and design of protein structures, protein folding mechanisms, and protein-protein interactions. It is one of the tools that are provided as a service by CloudBroker on the IBM SmartCloud, ready to be used by the scientists.

Modeling Streptococcus Proteins



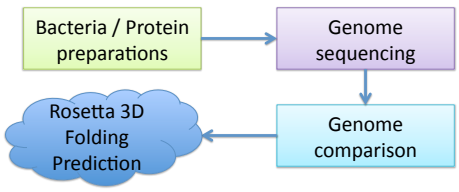
Detailed knowledge about the protein structures in highly virulent pathogens is essential in the fight against antibiotics-resistant bacteria. In an ongoing experiment, the genome of a mutated and highly virulent strain of streptococcus (AP1) was sequenced and compared to a low virulent strain (sf370), identifying proteins that have higher mutation rates than expected. The question is how the mutated proteins are **structurally different** to be able to investigate differences in protein-protein affinities to the host.

Predicting the protein structure from the amino acid sequence involves a lot of simulation and is very compute intensive. Each structural domain is modeled independently. Some models are very compute intensive and others are relatively quickly done. On average there are 10'000 models for each structural domain. Finally, during scientific post-analysis, where many additional factors are taken into account, the best or most accurate models are selected and evaluated.

Number of proteins	1697
Structural domains	832 de-novo, 1440 homology
Number of models	22'720'000
Total CPU hours needed	800'000

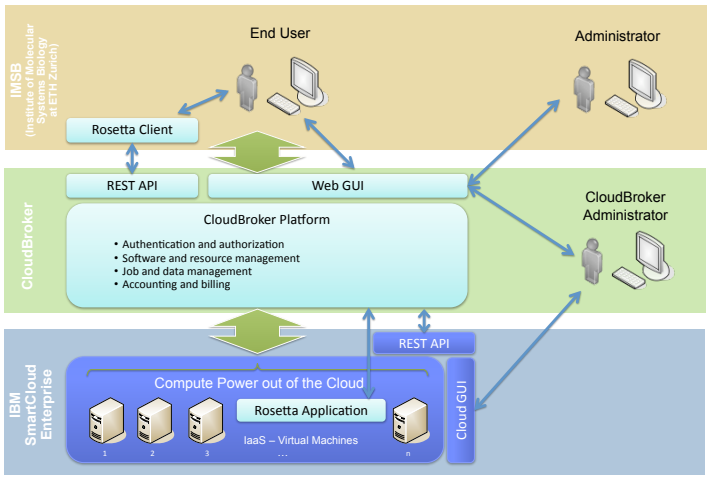
Approach

Identification of mutations in the protein structure between low virulent strains (sf370) and high virulent strains (AP1) is done by sequencing and comparing both genomes. The 3D protein structures are then predicted and compared as well.



In just two weeks, the most significant proteins could already be modeled, allowing for scientific data post-analysis. On the available university shared cluster resources this calculation would have taken several months.

Cloud Usage Overview

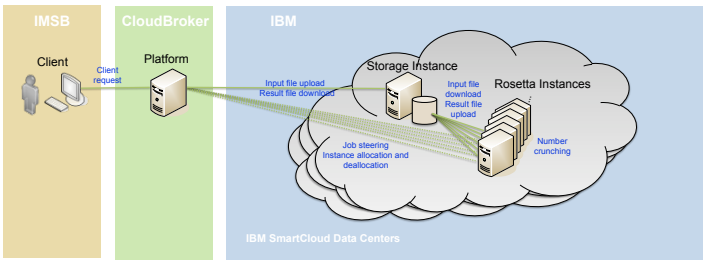


The IBM SmartCloud Enterprise infrastructure provides an API and a GUI to the users. This is being used by the CloudBroker Platform to control the automatic provisioning of the infrastructure and launches the instances with the Rosetta Application.

The CloudBroker Platform manages the Rosetta jobs automatically and monitors the execution. Fail-safes are built into the system with automatic restarts and shutdowns.

The end-user is given a CloudBroker Rosetta Client that can be used on the command line very similarly to the Rosetta software itself. It is a smart client that automatically creates the jobs from the input files as necessary. The smart client was used at IMSB to submit the parameters for the protein structure modeling. The jobs can be monitored using the web GUI, which also allows to manually steer the execution, to submit or to cancel jobs and to upload and download data. Administrators can also use the interfaces directly to monitor the execution or to intervene in case of problems.

Control and Data Flow



Data is stored on a dedicated in-cloud storage instance. The CloudBroker Platform manages the data placement and steers Rosetta instance allocations and de-allocations. The CloudBroker Platform also provides interfaces to the client for data upload and download and smart job submission. In the Cloud, the data are staged to the instances running the calculations dynamically.

Structural domains modeled on the Cloud	249 out of 832 de-novo
Number parallel Cloud CPU resources	up to 1008
Net CPU hours processed	~ 249'000
Number of jobs	~ 36'000
Number of models	~ 2'300'000