ExaScience Life Lab helps pharma companies dig for drug candidates in existing data

High-performance computers process and mine the enormous amounts of data from molecule testing for drug discovery to get new insights.

Pharma companies test millions of molecules to find potential drugs for specific diseases. But they could use the tests for a much wider screening, predicting the effect of the molecules for many more biological processes and drug targets. One of the bottlenecks, however, is processing and mining the enormous amounts of available data. Roel Wuyts, a senior scientist at the imec-based ExaScience Life Lab, explains how he and his colleagues work to solve such computational bottlenecks for life sciences applications. Starting from a recent project to repurpose the results of high-throughput cell imaging, he shows how the expertise center unleashes the power of today's (and tomorrow's) high-performance computers to help improve people's quality of life.

Sifting through millions of pictures

To fill their research pipeline with new drug candidates, pharmaceutical companies regularly screen the effects that hundreds of thousands of candidate molecules have on cells, the basic building blocks of our body. To do so, they use plates that have hundreds of little wells. In each well, they deposit a culture of the cells they want to look at. Then, to each well separately, they add one of the molecules for which they want to see the effect. They give it some time to react, add contrast and coloring liquids and make one or more high-resolution microscopy pictures of the wells.

These pictures are then automatically processed, looking at a number of morphological characteristics of the cells and their organelles that reveal what the effect is of the molecules that were added. Has the cell grown, or has it shrunk? Is the cell wall still intact, or does it show damage? And what about the cell's nucleus, where the genetic material is sitting?

"The process is called high-throughput cell imaging (HTI)," says Roel Wuyts, "and the tool that is often used to sift through the pictures is the open-source CellProfiler software developed by the Broad Institute. CellProfiler was designed to enable biologists without training in computer vision or programming to automatically process large amounts of pictures, quantifying specific characteristics of cells. The CellProfiler script sets up a pipeline, calling a number of subsequent programs, each taking as input the output of the previous program."

These are elaborate tests, assays in pharma parlance. Usually they are set up to look at the effects of a library of molecules on only one specific biological process, e.g. where a specific receptor molecule is lodged in the cell and how its distribution is influenced by adding the molecule. “That makes it easy for the statistical methods used,” says Wuyts. “It only has to look at a handful of all the available cell characteristics. But the downside is that it mines only a small portion of the information that is potentially available in those expensive pictures.”

Unlocking all the information

In a recent project, we collaborated with researchers from Janssen Pharmaceutica, the Broad Institute, and a number of research partners to unlock the information content in such images by using high-performance computing.

Each cell hosts thousands of biochemical processes and potential drug targets, all of which are exposed to the chemical compounds used in a specific assay. And apart from the specific process for which the assay was set up, many other processes and targets also have an impact on the cell morphology and can thus be studied from the images.

So the researchers wanted to see if it was possible to make an unbiased image-based fingerprint of each well, a fingerprint that could then be used to predict the activity of all tested molecules on a great many processes.

One important difference between a single-purpose assay and an unbiased fingerprint is that each high-resolution picture, and there may be millions in one assay, has to be scanned for hundreds or even thousands of characteristics instead of only a handful. And that turns out to be a computational bottleneck, a bottleneck that Roel Wuyts and his colleagues helped solve.

“In this study,” says Wuyts, “we had a second look at the pictures that were taken to study the influence of half a million different molecules on H4 neuroglioma cells, a specific type of brain tumer cell. The original goal was to see how the molecules impacted the move of glucocorticoid receptors from the cell cytoplasm into the nucleus. This is a process that is called nuclear translocation and that can be studied visually.”

“We received imaging data from close to 2,000 plates, each with 384 wells. So there were millions of high-resolution pictures to process, totaling over 10 TByte of data. The goal was to use CellProfiler to extract quantitative data for about 1,400 features per image. But the CellProfiler script originally developed at the Broad Institute was not optimized to analyze this quantity of data. More specifically, it was not able to make good use of a state-of-the-art high-performance computer infrastructure with an array of multiprocessor computers each hosting multiple multicore processors. So running it took a prohibitively amount of expensive computer resources, which is one of the reasons why this kind of comprehensive analysis was never done before.”

Solving the computational bottleneck

The ExaScience Life Lab has a high-performance compute cluster with 32 processing nodes, each consisting of 36 cores. The lab’s data center is also certified by Janssen Pharmaceutica, a partner in both the ExaScience Lab and this project. Such certification is a security prerequisite to process this type of sensitive biological data.

“The challenge is to distribute the computation as efficiently as possible over all processors and nodes of the compute cluster”, says Wuyts. “One way to do that is to rewrite the programs, making them maximally suitable for parallel processing. We’ve done that with great success for a number of other projects, but here we decided to stick to CellProfiler and adapt the way it is run. So our experts used a bag of tricks to run and distribute all processes over the available cores as efficiently as possible, e.g. by setting up scripts that ran CellProfiler over all images in headless mode, without interaction from a user, and splitting the total work in small chunks that can be executed on individual processor cores. With these interventions we managed to cut the processing time on the Lab’s supercomputer by almost two thirds, and we see potential for further gains.”

Of course, the image processing was only one step, albeit an important one, in the repurposing process. In their conclusions, the project’s researchers stated that their “results indicate that images from HTI screening projects that are conducted in many institutions can be repurposed for increasing hit rates in other projects, even those that seem unrelated to the primary purpose of the HTI screen. Consequently, it might be possible to replace particular assays with the potentially more cost-efficient imaging technology together with machine learning models.”

“Our lab was able to remove a computational bottleneck, making a sizable difference for the practical feasibility of projects like these,” says Wuyts. “We have proven that with some careful and targeted interventions, it is possible to slash the computer time for processing this amount of images substantially. And as computer time and the associated cost is often a prohibitive factor in life sciences projects such as these, we believe we have enabled future projects that will improve the health and wellbeing of many.”

Want to know more?

The Exascience Life Lab is an expertise center for high-performance and big-data computation in life sciences. The lab is housed at imec and was started in 2013 as a joint initiative by Intel, Janssen Pharmaceutica, imec, and all Flemish universities. Its core mission is to remove computational bottlenecks in software applications, allowing them to be used to solve real-world problems in the life sciences industry. This specific project was supported by research grants IWT130405 ExaScience Life Pharma, IWT130406 ExaScience Life HPC, and IWT150865 Exaptation from VLAIO, the Flanders Innovation and Entrepreneurship agency.

Biography Roel Wuyts

Roel Wuyts is principal scientist at imec and part-time professor at imec- DistriNet – KU Leuven. His main research interest is in the runtime management layer of future high-performance computing hardware. Before joining imec, he was Associate Professor at the ULB (Université Libre de Bruxelles). He obtained his PhD in computer science from the VUB (Vrije Universiteit Brussel). Roel Wuyts has served as member in various conference program committees such as ECOOP, OOPSLA, SC, Net.ObjectDays, or ESUG, organized workshops such as the DATE'08 Workshop on Software Engineering for Embedded Systems, and reviewed papers for TOPLAS or TOSEM.