Roundtable: The Continuing Evolution of Bioinformatics in the Pharma Industry

In life sciences research, advances in instrumentation, new types of measurements that are now possible, and improved computational tools, are generating not only significantly more data than ever before, but also more diverse modalities and higher resolution data. The positive effect is that we now have evidence to inform ever more comprehensive models of disease, disease progression, and health, at both the individual level and for the population as a whole.

Moreover, pharma and biopharma organisations can leverage the new 'mountain' of available data to open up the potential for them to expedite the discovery and delivery of new, more effective, safer therapies for patients. As far back as 2013, McKinsey & Company[1] spoke about the importance of having data that are consistent, reliable, and well linked. An ability to manage and integrate data, from discovery to real-world use after regulatory approval, is now seen by most as a fundamental requirement that allows companies to:

- Derive maximum benefit from technology and computational advances
- Implement infrastructure that is fit for purpose, cost effective, and easy to use across the enterprise
- Support clear industry goals faster, de-risking routes to the clinic and market approval using informed innovation, data to drive decision making and improved efficiency, for example.

In this roundtable, we speak with leaders in bioinformatics about the challenges facing the pharma industry today as they look to turn data rapidly and efficiently into knowledge, and knowledge into understanding and commercial success.

Data is the Currency of Computational Biology

The revolution in DNA-sequencing technologies in the last 25 years means that there is now an abundance of data about the human genome and how it varies between individuals. John Quackenbush, Professor at the Harvard T. H. Chan School of Public Health, who specialises in using massive data to probe how many small effects combine to influence human health and disease and has also founded a number of influential companies such as Genospace, explains: "It's been incredibly exciting to have been working in a field that has been transformed so fundamentally – and which is still undergoing massive change. Technology is advancing at such a pace that today we can develop datasets that give us a foothold in addressing questions that were unanswerable even two or three years ago."

The bioinformatics field is continuing to expand, according to Professor Quackenbush, and in a different direction to what scientists would have predicted. "What's fascinating is that things haven't evolved in the way some thought they might. When the first human genome was sequenced 20 years ago, people were saying 'now we've identified all the genes we'll be able to find the root cause of all human diseases.' But it didn't work out like that, because biological systems are more complex than that, which in turn requires enormous datasets."

With progressively more sparse data at hand, the challenges in connecting, searching, and analysing within and across multiple datasets continues to grow. As digital transformation is fast-tracked by many, pharma companies must ask how they can make sure they have data organised and rapidly accessible in a scalable, easy to use, computational platform that is future proofed to meet requirements that we cannot predict today.

The Biobank Revolution

Biobanks typically aggregate genetic information such as WGS (whole genome sequencing), WES (whole exome sequencing), and SNP (single nucleotide polymorphisms) with a range of other data on the same individuals: health records like GP data, hospitalisations, diagnoses, prescriptions, MRI (magnetic resonance imaging), lab

results from biochemistry and haematology, and patient-reported information such as family history, behavioural history, and socio-demographics are all included.

For example, the UK Biobank, which stores data from around 500,000 patients in the UK is a repository that presents a great source of public data. Martin Hemberg, Assistant Professor of Neurology, Brigham and Women's Hospital and Member of the Faculty, Harvard Medical School, located at The Evergrande Center for Immunologic Diseases, explains: "I see biobanks as simply increasing the number of opportunities available to us. Perhaps more exciting still is the opportunity to combine data sets and data types and see more than you could see with just one modality." Hemberg's work focusses on developing methods for analysing single-cell RNA-seq data and he's primarily interested in computational genomics and developing methods and models to help understand gene regulation.

Thanks to computational biology, we've got a lot better at defining diseases in terms of what's going on at a molecular level, says David De Graaf, CEO of Abcuro, a biotechnology company developing immunotherapies for autoimmune diseases and cancer. "And that's important because patients don't come into the clinic and complain that their PI3 [kinase] hurts, for example. We need to continue to progress findings at the molecular level, with these new datasets available, and make connections through the whole chain – from identifying symptoms, to understanding the molecular basis of disease, through to drug discovery. That's the grand aim – there's a lot we still don't know, for example, why patients don't always benefit as expected from gene therapy. There's a lot of questions that still need answering, and computational biology could support with that."

Despite advancements in genomic research, there are still barriers that organisations face to support their ongoing disease studies and many inefficiencies with current computational approaches. De Graaf comments: "a huge inefficiency is that it has typically been harder to find old data – I think that computational tools that annotate relevant data, and allow you to search across it, could really pay off." He continues, "Another opportunity is being able to take externally curated content and understand it in the context of your own experiments. Then there's the concept of making a link between patients with a similar molecular phenotype. And finally, where computational tools could make a big difference is being able to extract value from studies that haven't worked. People don't always think about the fact that 99% of the money in the pharma industry is spent on things that end up going wrong. But I don't believe it's right to simply forget about that work – we should bring it together and gather insights from it. I think all of these things have the potential to be huge drivers in accelerating drug discovery and development."

If the industry can expand its computational toolbox to deliver rapid insights more efficiently, with technology platforms that can handle large-scale heterogeneous data, then further advances in drug discovery can be made.

Single-cell Changes the Landscape Again

For many in the field, the bioinformatic landscape has been altered significantly by the ability to collect and analyse data

from single-cells rather than aggregated cell populations, or organ systems. Single-cell RNA sequencing (scRNA-seq) allows 'omics analysis, notably genomics, transcriptomics, epigenomics and proteomics at the single-cell level and enables, for example, the identification of minor subpopulations of cells that may play a critical role in a biological process. It can also provide an ultra-sensitive tool to clarify specific molecular responses to therapy and pathways and thus reveal the nature of cell heterogeneity. The Sanger Institute commented in a 2020 blog post[2] that:

"The volume of single cell data that will be generated will exceed the volume of genotype sequencing data by orders of magnitude. While the human genome has 20K genes, there are 300 different cell types in the human body comprising 37 trillion cells. Previously, scientists would take billions of cells together and measure an average of gene activity. Now, it is feasible to measure each cell's individual gene expression profile."

Martin Hemberg notes the change he has seen in his research where this technology has helped to transform results: "Specifically, in terms of single-cell analysis, new technology development is really driving the field – I have to keep my ear to the ground to understand these new technologies and how they can help us solve day-to-day problems we have. For example, I recall that when the costs of cell isolation and sequencing RNA started to fall, researchers took advantage, new protocols emerged where pooled samples were sequenced 'in bulk', and the results were deconvoluted to identify individuals. The amount and complexity of the data increased dramatically. In 2016, the problem did not exist, then around 2018 the first publications came out relating to these new experiments, and we started looking at enhancing analysis methodology."

As such, scientists must remain aware of how quickly the landscape can change as data evolves. Hemberg continues: "What's become obvious over the years is that a difficult computational challenge can be totally solved by a better assay – or, conversely, a new assay can throw up a new, interesting and challenging bioinformatics and computational problem. The moral is clear: you need to be nimble."

The 'Grand Challenges' – Can We Solve Them?

The abundance of data now available is unprecedented but, without the right tools, its potential value will remain unrealised.

Combining genomic data, imaging data and patient metadata with computational methods to create multi-tiered models that encompass genetic variants, transcription, gene expression, methylation, microRNAs, and more will be important going forward, says Professor Quackenbush. "Bringing all these tools together gives us a better chance of extracting insights from the fantastically complex datasets that are now available."

To prepare for changes within the field, scientists and organisations should ensure their informatics systems can support new methods and approaches. An integrative analytics platform that streamlines translational research by reducing or eliminating extract transform and load processes, is a computational approach that many organisations should look towards if they want to transform the way they organise, store, compute, and integrate their multi-modal scientific data without an army of data janitors.

As we look towards the 'next big thing' in pharma, accelerating scientific discoveries through the production of higher-resolution data and new computing paradigms through quantum data is set to soar. With a digital transformation already underway, quantum methods will create an increased requirement for scientists to have their data organised and accessible in scalable analytics platforms future-proofed to meet emerging requirements. It will be more important than ever that organisations are ready to adapt to change in the coming years to enable further bioinformatics-driven breakthroughs.

With thanks to John Quackenbush Ph.D., Martin Hemberg, Ph.D., and David De Graaf, Ph.D., for their valuable insights on some of the bioinformatics challenges faced by the pharmaceutical industry today.

REFERENCES

1. Mckinsey&Company. 2021. How big data can revolutionize pharmaceutical R&D. [online] Available at: <https://www.mckinsey.com/industries/life-sciences/our-insights/how-big-data-can-revolutionize-pharmaceutical-r-and-d> [Accessed 14 December 2021].
2. Mapping the Human Cell Atlas – charting the body's cellular world, Sanger Institute (2020/04/08) https://sangerinstitute.blog/2020/04/08/mapping-the-human-cell-atlas-charting-the-bodys-cellular-world/#:~:text=20%2C000%20dimensions,may%20use%20to%20varying%20extents

Zachary Pitluk

Zachary Pitluk, PhD., VP of Life Sciences at Paradigm4, has worked in sales and marketing for 23 years, from being a pharmaceutical representative for BMS to management roles in Life Science technology companies. Since 2003, his positions have included VP of Business Development at Gene Network Sciences and Chief Commercial officer at Proveris Scientific. Zach has held academic positions at Yale University Department of Molecular Biophysics and Biochemistry: Assistant Research Scientist, NIH Postdoctoral Fellow and Graduate Student, and has been named as co-inventor on numerous patents.

Email: zpitluk@paradigm4.com
Web: www.paradigm4.com

Marilyn Matz

Marilyn Matz is CEO and co-founder, along with Turing laureate Michael Stonebreaker, of Paradigm4. The scientific analytics solutions company enables scientists and data scientists to transform their research with an integrative analytics platform that powers massively scalable analytics and machine-learning. Prior to Paradigm4, after completing a MS degree at the MIT AI lab, she was one of three co-founders of Cognex Corporation, now a publicly traded, global industrial machine vision company, where she was Senior Vice President and Business Unit Manager of its Vision Software Products Group. Marilyn was the recipient of the sixth annual Women Entrepreneurs in Science and Technology (WEST) Leadership Award; a co-recipient of the SEMI industry award for outstanding technical contributions to the semiconductor industry; and a 2020 NACD Directorship 100. She also serves on the Board of Directors of Teradyne, a leading supplier of automation equipment for test and industrial applications.