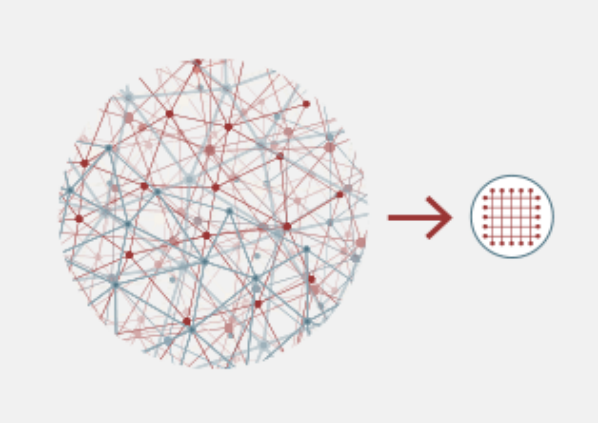
2016 Canadian Bioinformatics and Computational Biology Strategic Framework Report

February 16, 2016

Preamble

Data-intensive scientific research lays the groundwork for the development of the revolutionary approaches to problem solving and decision-making processes that are the primary drivers of innovation. Major technological advances during the last decade have accelerated the production of data of an unprecedented volume and complexity that in turn have driven the development of the new approaches and infrastructures needed to collect, store, manage, mine and integrate this information. In this context, many, including all of us in Canada, are struggling to understand all these data (big and small), to derive maximum value from the billions of dollars invested in research and to realize the promised benefits to Canadian society.



Big Data requires special expertise to analyze large data sets and convert them into user-friendly formats.

In the life sciences, data are generated from a variety of sources, including biological research on all organisms, data-intensive technologies such as quantitative imaging, large population cohorts, multinational clinical trials, international genomic and epigenomic sequencing initiatives. In recent years, the increase in data production has been particularly dramatic in the "omics" technologies, specifically in the area of genome sequencing. The Human Genome Project, completed in 2003, required hundreds of sequencing machines and cost over $1 billion over a 10-15 year period. In 2016, it is possible to sequence an individual's genome in 2-3 days for little more than $1000 (the cost of a day at the hospital). These numbers are not static and some estimates suggest that, by 2020, data will be

generated at up to one million times the current rate, which is orders of magnitude faster than the growth of computational power as predicted by Moore's law (i.e., doubling of computing power every two years). The analysis of human genomes, transcriptomes, epigenomes, proteomes, interactomes, metabolomes and microbiomes will provide the basic knowledge necessary to diagnose, understand and cure many diseases leading directly to reductions in health costs for society. Similar research efforts in the agriculture, energy, environment, fisheries, forestry, and mining sectors are providing important knowledge to guide pest management strategies, sustainable farming practices, natural resource management, crop developments and environmental monitoring in the face of climate change.

When analyzed, interpreted and applied correctly, 'big data' are of significant value to Canada. However, for raw data to be useful to end-user groups — whether they be economists, environmentalists, natural resource managers, or health-care professionals — special expertise must be applied to its analysis and interpretation.

To analyze, interpret and integrate such large datasets, they must be made available in a community-supported standardized format that is readily accessible to the widest possible array of new tools and innovative approaches, so as to derive the maximum possible socio-economic benefit. Intersecting disciplines, which we collectively refer to as Bioinformatics and Computational Biology (B/CB), address this need. As a multidisciplinary science that combines aspects of life sciences, computer science, chemistry, biochemistry, statistics, mathematics, engineering, physics and medical sciences, B/CB plays the crucial role in the analysis of complex biological data, processes and mechanisms. B/CB integrates biological themes together with the help of computer tools and biological databases, providing new knowledge of the systems under study. New innovative and cutting-edge algorithms and computational and statistical techniques generated by B/CB scientists are vital in efforts to effectively mine, rapidly access, and efficiently analyze vast quantities of data as well as integrating across related datasets to harvest and apply the information contained within them.

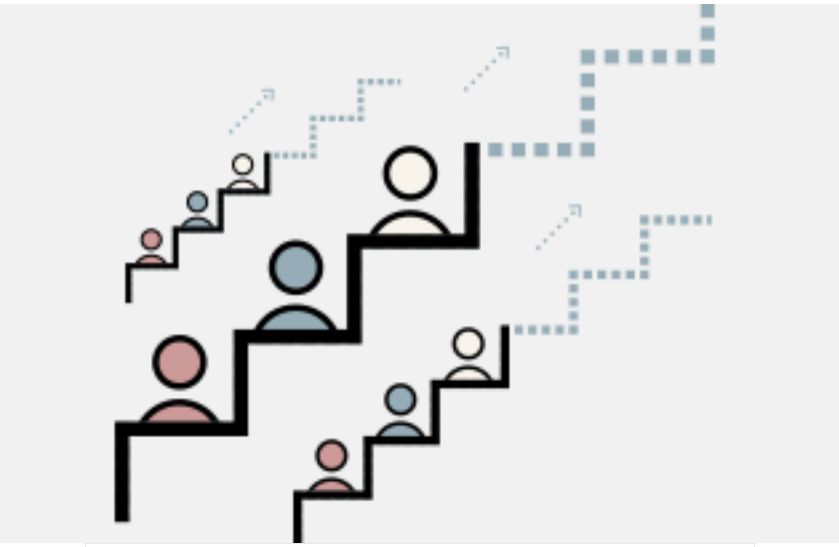
CANADIAN B/CB STRATEGIC FRAMEWORK

The vision underscoring the Canadian B/CB Strategic Framework is to build fully integrated B/CB capacity across the life sciences to ensure that Canadians derive maximum economic, health and social benefits from biological data. It will also further position Canada as a leader in international initiatives to derive full value from the billions of dollars invested in research globally. To achieve this vision, concerted and coordinated efforts among all stakeholders are required in the following areas:

Strengthening the B/CB Research Enterprise:

Resourcing to meet current and future demands

Due to extensive changes in research, technologies and methods, the bottleneck in scientific productivity has shifted from data production to data management, communication and interpretation. Without strong B/CB capabilities, crucial insights and discoveries would stay buried in the data, providing minimal return on the substantial investments in the life sciences and generating little benefit. The new tools and algorithms generated by the B/CB research community are, and will continue to be, the key to data interpretation and integration, bridging the gap between knowledge generation, its application, and generating new knowledge wealth from the initial work. Without completing this circle we cannot benefit from the original investments.



Resourcing to meet current and future demands.

Research funding. There is significant commonality in the B/CB needs of different sectors, with B/CB tools and algorithms developed in health, for example, being valuable in agriculture or the environment. This illustrates the integrating role of B/CB across the range of life sciences. While this situation helps foster research collaboration across sectors, securing research funding through traditional/investigator-initiated mechanisms has been challenging as this field straddles the mandates of federally funded agencies and does not fall squarely within the core purview of any.

Canadian research funding agencies have launched several strategic research initiatives that either specifically target B/CB (e.g., Genome Canada/CIHR Bioinformatics and Computational Biology competition) or promote the integration of B/CB in large-scale genomics projects (e.g., Genome Canada/CIHR Large-Scale Applied Research Project Competition in Genomics and Personalized Health), but more is needed. Of special note is the Discovery Frontiers Program: Advancing Big Data Science in Genomics Research (NSERC, Genome Canada, CIHR and CFI), which is supporting the flagship "Cancer Genome Collaboratory" at the OICR. However, core funding for B/CB research programs cannot be sustained through one-time strategic funding, which is currently the situation for many such programs. Tapping the potential of B/CB research to achieve scientific

excellence and societal impact will require Canada's research funding agencies to evolve strategic, episodic programs into ones that provide sustainable, predictable funding.

State-of-the-art computing. Although B/CB technologies hold the key to data analysis, benefits can only be achieved through the availability and effective management of the appropriate digital infrastructure that makes this feasible. To address this necessity, a computing needs-assessment and corresponding technology plan will be required to ensure that the required infrastructure resources are made available to support planned B/CB research initiatives. Ideally funding mechanisms should require coordination between B/CB professionals and infrastructure providers in the project design stage, as well as confirmation of adequate technology support at the award stage, in order to support a strong and sustainable B/CB research community. At a higher level, there should be close coordination between funding mechanisms designed to support the B/CB community and both service providers such as Compute Canada and CANARIE and the funders of such digital infrastructure resources such as CFI.

Immediate Actions

Coordinated action plan among life science research and infrastructure funders to promote alignment of B/CB funding opportunities, including:

- Research initiatives designed to harness the potential of big data recognizing the importance of funding for both short-term demonstration projects as well as longer-term, large-scale projects and their sustainability.
- Open programs to provide sustained and predictable support for investigator-initiated B/CB research: This research aspect and the need for new discovery is not going away.
- Tools to promote the integration of B/CB and computing infrastructure into the planning of small and large-scale life sciences projects: B/CB deals with data at all scales, and all may require custom solutions.

For the B/CB research community to thrive and respond to the demand for innovative tools and algorithms, a coordinated action plan among Canadian life sciences research and infrastructure funders is required to ensure the appropriate alignment of funding opportunities. Through this increased coordination, Canada will be better positioned to maximize and leverage the impact of previous federal investments in research.

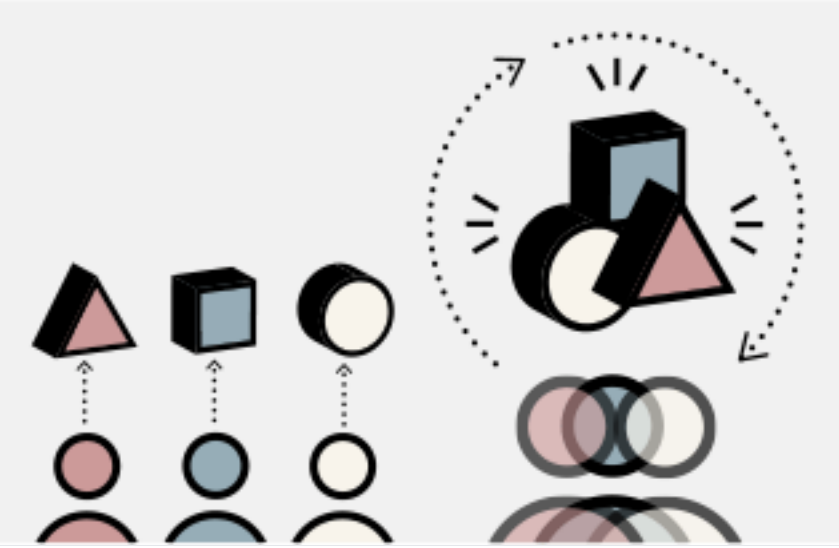
Transdisciplinary Capacity Building:

The key to sustaining and growing the B/CB enterprise

Core data analysis and computational modeling skills, coupled with experience working with data-intensive biological information (bioinformatics), are essential for the effective use and translation of high throughput data by the scientific community. As a result, B/CB

expertise has become a skill set in high demand within both the private and public sectors, and B/CB professionals are highly sought after across the life sciences.

Transdisciplinary approaches to training are required to provide the necessary skills to tackle the challenge of increasingly complex large data sets and the rapid pace of technological progress. B/CB crosses many academic departments, including molecular biology, biochemistry and genetics and requires integration with statistics, mathematics, computer sciences and engineering. Training in both biological and computational methodologies is made possible by integrating academic centres in computer science, statistics, molecular biology, and biotechnology, as well as with translational research groups at hospitals and at the clinical interface. This approach needs to also include a strategy to attract and retain B/CB leaders to oversee and be involved in these training programs. In addition, to keep up with the pace of leading technologies used in computational biology, short-term focused courses (such as the bioinformatics.ca workshop series) are required to ensure the continuous training/skills development of experts and users. If we are to achieve the broader vision of the national B/CB strategy and address the data challenges of tomorrow, all stakeholders with a vested interest in training will see the merit in supporting this activity.



Transdisciplinary approaches to training are required to provide the necessary skills to tackle the challenges of complex data sets.

Immediate Actions

Coordinated action plan championed by academic institutions and research funders, in partnership with other stakeholders requiring professional B/CB expertise that will:

- Determine approaches and mechanisms required to foster the creation of sustainable transdisciplinary B/CB graduate training programs,

- Develop continuous, training approaches and processes to address challenges of tomorrow.

Academic institutions and funding agencies, working with infrastructure providers and other stakeholders (such as industry), need to develop a coordinated approach to inspire the evolution of transdisciplinary B/CB graduate training programs across Canada that are also linked nationally and internationally.

Formation of a National B/CB Network:

National coordination for international impact

Coordination and collaboration have become the norm in data production efforts where Canadian B/CB researchers are playing a prominent role such as the International Cancer Genome Consortium, the International Rare Diseases Research Consortium, the Global Microbial Identifier, the International Human Epigenome Consortium, the International Wheat Sequencing Consortium and the International Cooperation to Sequence the Atlantic Salmon Genome to name a few. In response to the enormous amounts of data, hardware, and software produced by these efforts, new international alliances have recently been formed to facilitate data sharing and promote the development of standard strategies and procedures for data management, such as the Global Alliance for Genomics and Health. Similarly, coordinated, collaborative approaches are emerging among the infrastructure funders and providers who are a vital resource for the B/CB community, such as the Canada Foundation for Innovation, Compute Canada and CANARIE. Stronger linkages among these organizations, data scientists and software and tool developers will better align computing infrastructure and their management with the needs of the B/CB professionals, reducing duplications and optimizing data storage and analysis. Furthermore, CANARIE, which manages an ultrahigh-speed network and associated funding programs is developing and implementing next-generation technologies in collaboration with the private sector that will ensure Canada has access to the information technology tools necessary to effectively and efficiently support the modern research enterprise. All these interactions are essential to better align computing infrastructure with the needs of the research community, reducing duplication and optimizing data storage and analysis capacity.



National coordination is required for international impact.

Moreover, the changing dynamic and sheer scale of data production has also created new ethical, legal and social challenges that impact the timely development and implementation of policies and guidelines. In response, several international, federal and provincial initiatives have recently been launched to provide sound data management and stewardship policies for Canada and promote open access to information and data, maximizing the availability of research data to researchers and the private sector. A mechanism is required to continuously inform policy and guideline development that involves the B/CB research and user communities, which in turn will better position these communities in fulfilling their advocacy role for the field. This is generally what is done in many Genome Canada projects where a GE3LS component is required. This kind of socio-economic perspective needs to be a feature of B/CB activities when appropriate and called for.

Immediate Actions

- Establish a Canadian B/CB conference to connect the research community and stakeholders, to build an integrated community that is linked internationally.
- Bring all stakeholders together to develop an action plan for the formation of a national network, including how it needs be structured and funded to ensure its sustainability.
- Funders and related communities (e.g. the genomics ethics and policy communities) are strongly encouraged to attend and/or contribute to the key conferences and other meetings attended by the B/CB community, to perceive and optimally align their individual mandates and vision with that of the research and application community.
- Implement programming and associated funding to ensure that the ethical/legal/social context of the B/CB framework is well-positioned to deliver the benefits of large-scale data analysis.

Keeping up with the demands from the life sciences community for innovative, and increasingly complex algorithms, tools and databases is surpassing existing researcher's capacity. Along with a robust transdisciplinary research capacity action plan, we need to coordinate and integrate what is currently being done in Canada to be able to meet the demands of the user community. The establishment of a national conference will be a first step in coordinating efforts within the B/CB community, involving research funders, infrastructure providers, users and other stakeholders. The inaugural conference will be held in Toronto in May 2016, providing a forum for discussions as to how best to bring all stakeholders together as part of a sustainable national network. The conference is being established in coordination with the International Society for Computational Biology regional conference, and represents a key first step in the unification and integration of the Canadian B/CB community in support of the B/CB strategic framework.

The Way Forward: Connect, Coordinate and Train

The rate of data and new knowledge acquisition continues to accelerate, in turn escalating data storage and analysis challenges in the face of opportunities for the bio-economy and health sector. Scientists in Canada and around the world are struggling to understand and manage this vast amount of new data in order to ensure that deriving full value from past investments are fully realized and that critical information is not lost or forgotten.

This strategic framework is a national effort to rally all B/CB stakeholders as we strive to build fully integrated B/CB capacity across the life sciences. By connecting, coordinating and training highly skilled personnel, Canada will derive maximum economic, health and social benefits from big data. The several funding agencies, digital infrastructure providers, together with the B/CB community, must coordinate the development of any new initiatives and their associated activities. To this end, we recommend the formation of a pan-Canadian body representative of all of the various B/CB stakeholders. We assert that independent institution-based committees cannot accomplish what is needed, but rather a national committee is required to oversee, drive and coordinate new initiatives as they arise. Such a body would assume responsibility for ensuring that coordinated efforts are of high impact and are positioned to deliver what is needed for this important scientific endeavor.

This document was jointly authored by the B/CB Advisory Committee in 2014-2015 and subsequently shared with the community for comments and feedback. 159 individuals responded. Their comments were then consolidated and used for clarification, extension and adjustment of many of the ideas presented or declared in the framework document. The B/CB Advisory Committee supports the text presented here, and assumes responsibility for any persisting errors, omissions or inconsistencies.

Francis Ouellette & William Crosby
Co-Chairs, B/CB Advisory Committee

B/CB Advisory Committee:

Gary Bader, *University of Toronto*
Robert Beiko, *Dalhousie University*
Guillaume Bourque, *McGill University*
Fiona Brinkman, *Simon Fraser University*
Michael Brudno, *University of Toronto*
Liz Conibear, *University of British Columbia*
Bill Crosby, *University of Windsor*
Mark Dietrich, *Compute Canada*
Francis Ouellette, *Ontario Institute for Cancer Research*
Peter Wilenius, *Canarie*

The advisory group acknowledges support of staff from Genome Canada (GC) and the Canadian Institute for Health Research, Institute of Genetics (CIHR, IG):

Naveed Aziz, *GC*
Cindy Bell, *GC*
Paul Lasko, *CIHR, IG*
Eric Marcotte, *CIHR, IG*
Stephanie Robertson, *CIHR, IG*