novo nordisk fonden

Bioinformatics Summit

Thinking Big for Big Data

Contents

Bioinformatics Summit photographs taken by Christian Als

Preface

The Novo Nordisk Foundation is an independent Danish foundation with corporate interests. It has two objectives: to provide a stable basis for the commercial and research activities conducted by the companies in the Novo Group* and to support scientific, humanitarian and social causes.

The history of the Foundation dates back to 1922, when Nobel Laureate August Krogh returned home to Denmark from Canada with permission to manufacture insulin in Scandinavia. This was the starting point for developing world-class diabetes medicine and a Danish business and export venture. It also led to the establishment of several foundations that, many years later, merged into today's Novo Nordisk Foundation.

The Novo Nordisk Foundation's vision is to contribute significantly to research and development that improves the lives of people and the sustainability of society.

With Denmark as its centre of gravity, the Foundation's focus is to improve the lives of people by improving health, education and developing a knowledge-based sustainable society.

Major grant areas include biomedical and health science; patient-centred and research-based care; life science research and industrial applications promoting sustainability; and natural and technical science. Since 2010, the Foundation has awarded grants totalling more than DKK 16 billion (€2.2 billion), primarily for research at public institutions and hospitals in Denmark and the other Nordic countries, as well as research-based treatment and prevention of diabetes. The Foundation supports the entire research chain – from education to innovation.

The Foundation is always on the lookout for new ideas and innovative initiatives. To help develop strategic initiatives within the area of bioinformatics, the Foundation hosted a Bioinformatics Summit on 5–6 November, 2018, during which key stakeholders and national and international experts were gathered to provide their valuable advice on potential initiatives the Novo Nordisk Foundation and other private/public funders can support and develop to strengthen bioinformatics research and education in Denmark. This report recapitulates the major points from the summit.

Read more about the Novo Nordisk Foundation at novonordiskfonden.dk/en.

*The Novo Group companies comprise Novo Nordisk A/S, Novozymes A/S and Novo Holdings A/S.

Executive Summary

Given the huge potential gains in health and wellbeing from utilising 'big data' in the life sciences, the Novo Nordisk Foundation wishes to develop strategic initiatives in bioinformatics to help ensure Denmark is able to fully capitalise on this potential. To help inform these initiatives, the Foundation commissioned a landscape report assessing current and near-future supply and demand estimates of competencies within the field. This was followed by a Bioinformatics Summit bringing together 36 stakeholders and leading experts in the field.

The landscape report concluded that the number of individuals involved in bioinformatics research in Denmark has grown significantly over the past 5 years and is expected to increase further over the coming 5 years, particularly within industry and academia. Such individuals not only include those with specific bioinformatics backgrounds but also those such as computer scientists and statisticians. Despite the expected increase in the number of such individuals in the coming years, Denmark will likely struggle to meet future demand, particularly for those with 'core bioinformatics' skills profiles. While the landscape report was based on individual stakeholder surveys and interviews, the Bioinformatics Summit was organised to thoroughly explore the challenges in realising the potential of bioinformatics/biological data science in Denmark and develop prototype solutions to meet these challenges.

The following overall challenges were described by the group of experts at the Summit:

- **Lack of talent**: too few individuals with the right skillsets are attracted into bioinformatics and/or retained
- **Lack of a (defined) career path**: bioinformatics is still seen by many as a 'support' function
- **Lack of role models**: talent attracts talent, and there are too few 'superstars' in Denmark
- **Poor compensation**: potential talent is lured away by other players such as financial technology (fintech) companies and Google
- **Conservatism in educational programmes**: bioinformatics is not incorporated into enough courses and programmes
- **Traditional, hierarchical structure in academia**: bioinformatics requires a marriage of disciplines and cooperative working, which is not fostered within academia
- **Data infrastructure**: there is a wealth of health data in Denmark, but it is currently not pooled or shared effectively
- **Engagement with the public**: there is not enough engagement with young students or the public
- **Branding**: there is a more general lack of 'branding' and positioning around bioinformatics at all levels
- **High-performance computing (HPC)**: without appropriate computing power, it is difficult to attract talent and deliver sustainable solutions

The following, potentially overlapping, prototype solutions were outlined by the expert group:

1. Bioinformatics Research Institute
An interdisciplinary, matrix-structured research institute – its researchers maintaining their affiliation with their home institutions – with strong leadership overseeing challenge-driven research.

2. Centres of Excellence
Alternatively, the 'Bioinformatics Research Institute' solution could be a more virtual, umbrella organisation, working with existing research centres across Denmark to strengthen their positions.

3. 'Blue-sky' Professorship
Long-term funding for pure research with complete freedom to think about challenges and projects (high risk but with potentially ground-breaking reward).

4. Education Academy
Provide PhD programmes and postdoctoral training along with oversight and organisation of the outreach and vocational training activities outlined below. Also working with the universities to bring bioinformatics to more courses.

5. Vocational Training
Educational initiatives bringing data sciences to those with a relevant degree in the life sciences, bringing life sciences to data scientists and keeping existing researchers and medical doctors up to date.

6. Outreach Activities
Activities targeting school-age students to generate awareness of and excitement around bioinformatics.

7. Data Standardisation and Sharing
Establish a formal body to facilitate inter-entity data exchange and pooling, leading to country-wide (or global) data format standardisation. A flagship project, or projects, based on these data could demonstrate the benefits of such research and position Denmark as a world leader in bioinformatics.

8. HPC Ecosystem
While there were differences of opinion on the state of HPC in Denmark, appropriate computing power needs to be available and may need investment.

In summary, the experts recommended a variety of solutions with an underlying strategy of establishing and fostering rewarding, interdisciplinary, challenge-driven research utilising FAIR (Findable, Accessible, Interoperable and Reusable) data while strengthening educational initiatives to encourage both high school students and university students in other disciplines to consider a career in bioinformatics. Differences of opinion on the current and future state of HPC in Denmark warrant further exploration. While not listed as a recommended solution above, there is a clear need for a 'marketing' campaign to highlight the potential in bioinformatics and the role of those involved in making a difference to global health and wellbeing, as well as business opportunities. Describing this critically important field as 'biological data science' rather than 'bioinformatics' was also recommended to further help improve attractiveness and bolster its own deserved right as a forward-thinking, independent yet cross-disciplinary field.

1. Background

A wealth of advances in research and development can be gained from utilising big data within the life sciences, but the ability to process and apply insights from these datasets is often rate limiting. With bioinformatics positioned between the traditionally separate disciplines of life science, computer science and mathematics, individuals with the required skillsets may not exist in sufficient numbers to fully capitalise on the potential of big data.

The Novo Nordisk Foundation wishes to develop strategic initiatives in bioinformatics to help ensure that current and future needs for talented bioinformaticians in research and development are met across industry, academia and healthcare in Denmark. To inform these initiatives, the Foundation has now completed two key activities: (1) the commissioning of a landscape report assessing the current and near-future bioinformatics position in Denmark; and (2) bringing together key stakeholders and experts in bioinformatics and associated disciplines for a Bioinformatics Summit.

What is ‘Big Data’?

‘Big data’ is a term used to describe sets of data that are too large or complex (with too many variables) to be analysed adequately by traditional data-processing software. With high-performance computing, big data can be analysed to reveal patterns, trends and associations. Currently, the term ‘big data’ may also be used to refer to the analytics involved.

The applications and implications include predicting human behaviour (business trends, purchasing patterns, crime prevention), healthcare (disease prevention, personalised medicine) and scientific research (particle physics, astronomy, genomics).

Capturing and analysing such data has been made possible by advances in computing power and capacity (bandwidth and storage). When the human genome was originally sequenced, it took 10 years to complete – it can now be achieved in less than a day.

2. The Novo Nordisk Foundation Bioinformatics Landscape Report

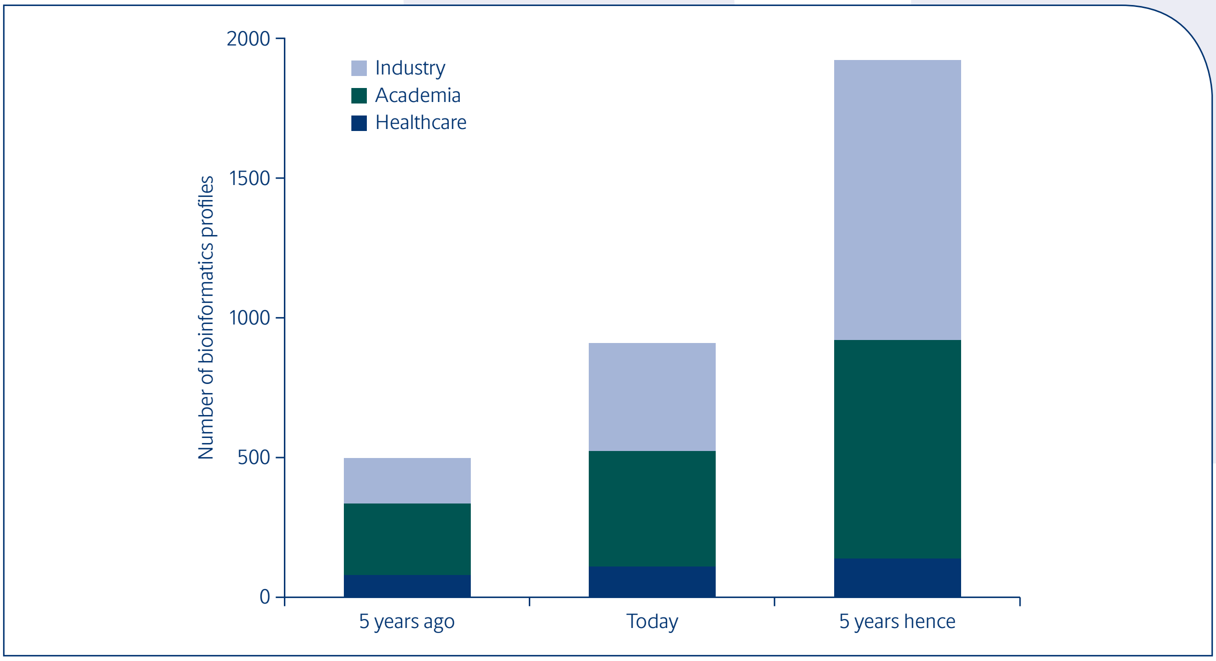
To assess the current situation and gain insight into the near-future situation in Denmark, an analysis was conducted combining desk research with interviews and surveys with key stakeholders currently using bioinformatics across industry, academia and healthcare in Denmark. Life science research is strong in Denmark, with bioinformatics research activity largely concentrated in four geographic areas: the Copenhagen region, Aarhus, Aalborg and Odense. The Copenhagen and Aarhus areas account for over two-thirds of bioinformatics activity in Denmark, and their research environments are typically organised as independent centres with an explicit focus on bioinformatics. Large pharmaceutical companies and specialist consulting and software development companies also operate within this region.

The mapping revealed a range of skills that are needed in order to gain actionable knowledge from large biological datasets and capitalise on these insights, which were grouped into four key competence profiles:

- **Core Bioinformatics profiles:** experts in translating and interpreting large amounts of biological data into insights, typically holding Master's degrees in bioinformatics
- **Application profiles:** experts with deep knowledge in biology or medicine with a basic understanding of programming and algorithm design to provide support and solve problems relevant to their biological science and medical fields
- **Data Science & Software Design profiles:** experts with an educational background in computer science
- **Profiles from Other Domains:** experts from different fields who possess skills applicable to bioinformatics research and development (such as physics, statistics or mathematics)

The number of such individuals involved in bioinformatics research in Denmark has grown significantly over the past 5 years across the life science industry, academia and healthcare, and demand is expected to increase further (Fig. 1). Over the next 5 years, anticipated acceleration of growth within academia and industry translates to a requirement for 87% and 160% more individuals, respectively. This increased demand is driven by the growing importance of bioinformatics and an expanding scope – particularly as more teams in academia work across disciplines and industry becomes aware of the potential for bioinformatics to drive innovation and accelerate solution development within healthcare and life sciences.

Fig. 1. Overall increasing demand for individuals with competences applicable to bioinformatics across industry, academic and healthcare sectors, seen during the previous 5 years and expected over the next 5 years (total of all four identified competence profiles).



A key element to meet this growing demand is to ensure sufficient educational throughput in Denmark. Across institutions, the number of students admitted to relevant identified Master's degree programmes each year increased steadily over the 2012–2016 period, and this increase is expected to continue over the next few years. In particular, the number of individuals within the Core Bioinformatics profile graduating each year is expected to increase by 60% within the next 5 years, again reflecting an increased awareness of the subject area and its growing importance and potential. However, it is very likely the supply of Core Bioinformatics profiles will struggle to meet future demand. For the three other profile types, there will be more than enough graduates to potentially satisfy demand. However, these profiles are integral to a wide range of educational programmes (for example, Application profiles include biochemistry and molecular biology graduates). Thus, while appropriate graduate profiles exist, competition for these among employers will continue to exist outside of bioinformatics. In the case of Data Science & Software Design profiles, competition is even larger as it extends beyond life science into business sectors such as banking, marketing and social media.

The landscape report defined key challenges to ensure an adequate supply of individuals with the required skillsets based on individual stakeholder surveys and interviews. To further explore the challenges involved in meeting future demand for bioinformatics and realising the potential of the field, as well as advance prototype solutions to these challenges, the Foundation organised a Bioinformatics Summit: a 1.5-day, workshop-based meeting facilitating group discussion and collaborative planning among key stakeholders and experts.

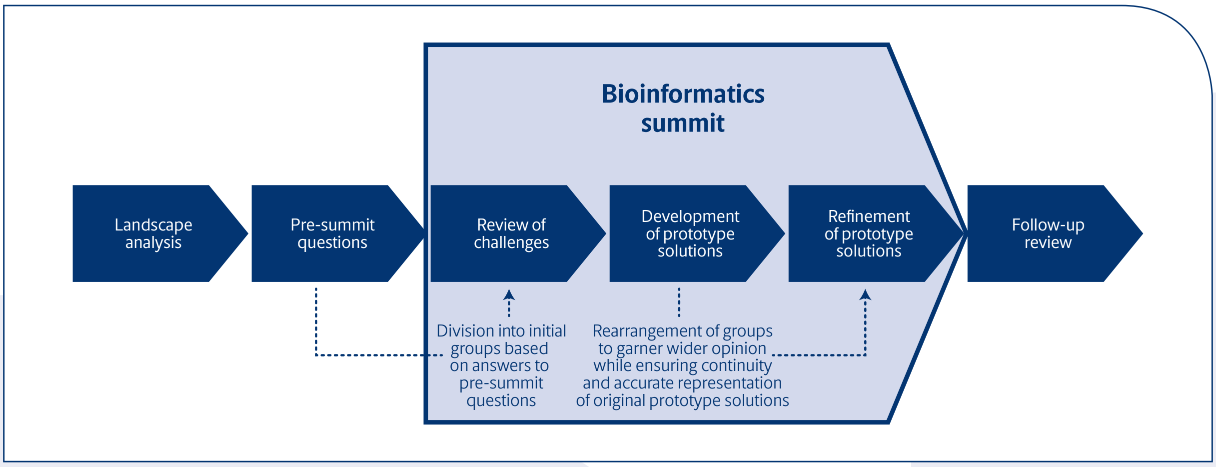
A Brief History of Danish Bioinformatics

In Denmark, bioinformatics has evolved as an interdisciplinary field of research over the last 25 years, anchored at several Danish universities. The first Danish research environment dedicated to bioinformatics took shape at the Technical University of Denmark in Copenhagen in 1993. It was based on a grant from the Danish National Research Foundation to the Centre for Biological Sequence Analysis. Some years later, in 2001, the Bioinformatics Research Centre at Aarhus University was established as a collaboration between the Faculty of Science and the Faculty of Health. In 2002, the Bioinformatics Centre was founded at the University of Copenhagen (partly by a grant from the Novo Nordisk Foundation). The centre played a key role in kick-starting bioinformatics research and education at the university. Around the same time, research groups with a focus on specific areas of bioinformatics and computational biology were formed at the University of Southern Denmark and Aalborg University. In 2007, the Centre for Protein Research was established at the University of Copenhagen through a grant from the Novo Nordisk Foundation. The Centre has a strong focus on computational biology represented by the Translational Disease Systems Biology Group. More recently, other research environments have been established, often as a collaboration between the established centres at the universities and the hospital units. Examples include the Department of Molecular Medicine in Aarhus and the Unit for Genomic Medicine at Rigshospitalet in Copenhagen.

3. The Bioinformatics Summit

The Bioinformatics Summit brought together 36 key stakeholders, including academic leadership and leading experts with extensive expertise in bioinformatics, data science (including physicists), genomics and computational biology. The participants were primarily from academia, healthcare and industry in Denmark, complemented by international experts from the US, UK, Sweden and Spain (see Appendix 1). Representatives from the Novo Nordisk Foundation Board of Directors, advisory panel and secretariat also participated in the meeting.

Fig. 2. Overview of the processes involved leading up to and during the Bioinformatics Summit.



Prior to the Summit, the invited delegates had been asked to list their three top challenges regarding meeting future needs in bioinformatics along with one piece of advice for overcoming them. Based on these responses, the delegates were divided into groups to collectively define the most important challenges to strengthening Denmark's position in bioinformatics and then develop prototype solutions with regards to the research environment, education, infrastructure and healthcare. On the second day, groups were changed such that each new group contained a representative from each of the original groups, ensuring discussions and outputs were not developed by each group in isolation. Feedback sessions and 'gallery walks' further enhanced the group validation of the identified challenges and prototype solutions. The final session of the Summit was a collective review of the prioritised and refined solution suggestions. This report represents a consolidated review of the discussions, and the outputs from the meeting described herein should be viewed as an initial draft of recommended options for further exploration and validation.

4. 'Faster Horses' and the Evolution of Bioinformatics

Shomit Ghose, Managing Director of the US-based venture capital firm Onset Ventures, set the scene for the Summit with a lecture on the evolution of bioinformatics. Importantly, he grounded his presentation by highlighting that, just as healthcare needs to be scalable in order to provide benefit for humanity, bioinformatics must become scalable; the benefits cannot be confined to a specific population or group.

The question of what constitutes 'biological data' was put rhetorically to the delegates; the answer being that *all* data are biological data. To illustrate this concept, Shomit Ghose drew upon the example of how financial technology (fintech) is changing banking. Traditionally, consumer 'financial data' used by banks to judge whether to provide credit would include elements such as bank balance, salary history, credit history, monthly expenses: all clearly financial in nature. However, a number of fintech players, working with traditional banks or directly with consumers, have broadened the definition of 'financial data' and are making credit decisions based on new and unconventional markers such as whether the applicant capitalised their name correctly, their typing speed on a mobile device, whether they read the user agreement correctly and even the battery charge status of their mobile phone at the time of application. While the mechanistic relationship with creditworthiness may be unclear, companies using these markers have been profitable and successful, including in scaling the technology to bring appropriate credit to the emerging markets. Thus, *all data are financial data*.

These advances in predicting human financial behaviour have been made possible by combining two raw elements: data and machine learning. We are generating continuous, individualised data trails in modern daily life, resulting in massive volumes of data and allowing for high degrees of statistical correlation on analysis. Machine learning (with corresponding computing power) allows us to elucidate previously invisible correlations and associations. However, limited data lead to limited possibilities; so, we must be 'data hungry' in the areas in which we wish to develop insights. As with the fintech markers, the mechanisms underlying a correlation may not be immediately clear, but the more biological data that are gathered, the more 'hidden insights' machine learning will uncover.

This important point was emphasised with the apocryphal Henry Ford quote: "If I had asked people what they wanted, they would have said 'faster horses.'" While Ford did not actually say this, it eloquently underlines the principles behind the mission: 'thinking big for big data'. Those working within the field need to continually push the definition of what data are, and think differently about what can be achieved, to fully realise the potential – just as the definition of 'personal transport' was redefined by the affordable automobile.

Shomit Ghose reviewed some of the advances already being seen from big data in healthcare, including an article in *Cell Research* showing that facial morphology provides a more reliable ageing biomarker than blood profiles and can better reflect general health status than chronological age. 'Facial morphology' biomarkers are being collected by the millions every day in the form of selfies. Other examples include diagnosis of melanoma using mobile phone cameras, diagnosing genetic disorders from facial-recognition software and predicting both cardiovascular risk and cognitive impairment from voice signals. The adoption of wearables means even more data are now being collected. The Apple Watch's heart sensors can identify signs of cardiac dysfunction, but based on the combined data such wearables collect, there may be correlations beyond heart function that we have yet to even contemplate. While we do not know what these correlations may be, some will be revolutionary and we must be open to finding them; in short, we must not be in the 'faster horses' business.

Importantly, Shomit Ghose highlighted that Google, Facebook and Amazon are all moving rapidly in the health arena and, by the nature of their activities, have access to a tremendous amount of data. Denmark will not be able to compete with Google but, by becoming a significant player and capitalising on the data available, has the potential to become a large enough player to drive scalable health and wellness solutions for the benefit of humanity.

5. Crucial Challenges

Delegates were divided into thematic discussion groups focusing on research environment, education, infrastructure and healthcare. Within each group, delegates suggested challenges to realising the potential from bioinformatics individually before these challenges were reviewed and discussed among the group. Each group was ultimately tasked with agreeing and detailing their top three major challenges.

Consolidating feedback across groups, the following key challenges can be outlined:

1. **Lack of talent**
   As described in the landscape report and confirmed across all groups at the Summit, there are simply too few individuals with the right skillsets. There are not enough individuals being attracted to bioinformatics and there are issues with retaining individuals, for the reasons explored below. A lack of diversity was also raised as a challenge, in terms of low numbers of women and ethnic minorities, but also in diversity of skills.

2. **Lack of (defined) career path**
   The attraction and retention issues were linked to a lack of clear and/or attractive career paths within bioinformatics, particularly as bioinformatics is still seen as a 'support' function by many. There should be more opportunities and flexibility with respect to career progression for those who want it.

3. **Lack of role models**
   Compounding the lack of a defined career path is a lack of role models within the discipline. Talent attracts talent and there are too few 'superstars' to act as role models in Denmark.

4. **Poor compensation**
   Most groups discussed the relatively poor compensation packages for bioinformatics-related careers compared with other industry sectors, such as fintech and Google. However, it was agreed that it would be difficult to match such salaries. Rather, the challenge is in creating an attractive package beyond remuneration, particularly exciting research opportunities and the rewarding purpose of solving impactful problems.

5. **Conservatism in educational programmes**
   Bioinformatics needs to be integrated into more courses at earlier stages to generate a critical mass of bioinformatics-aware teachers and scientists at universities. This includes exposing computer science students to the life sciences and vice versa. Providing a grounding in bioinformatics in other programmes may inspire individuals who would otherwise not have considered a graduate degree or job in bioinformatics to pursue these paths.

6. **Traditional, hierarchical structure in academia**
   The current structure in academia is not built to foster bioinformatics progress, which requires a marriage of disciplines and cooperative working. Moreover, the current funding system does not promote 'freedom', 'risk taking' and 'blue-sky' thinking. By contrast, these elements are increasingly found in industry. This compounds the lack of strong bioinformatics leadership and role models (as discussed above).

7. **Data accessibility**
   While there is a wealth of data in Denmark, they cannot currently be pooled owing to differing methodologies in capturing and storing the data and legal frameworks that limit sharing.

8. **Engagement with students and the public**
   There is not enough engagement with young students and the public, which is not just an issue for bioinformatics but also for STEM careers in general. This also needs to counteract the negative perceptions of personal data (mis)use.

9. **Branding**
   In addition to engagement with young students and the public, there is a more general lack of 'branding' around bioinformatics. The visibility and value among current academics and graduate students are low, and bioinformatics needs to be given an identity and positioned as an area where researchers can indeed make a difference to society. One element of this should be renaming bioinformatics to 'biological data science', a specialised part of the data sciences.

10. **High-performance computing (HPC)**
    Without appropriate computing power and capacity, it is difficult to attract serious talent and deliver sustainable solutions. The experts disagreed on the current state of HPC in Denmark, but this may be a key challenge as the discipline advances.

The Birth of Big Data in Biology and the Need for Infrastructure

Dr Niklas Blomberg, Director of ELIXIR

In an inspirational talk at the Summit, Dr Niklas Blomberg, Director of ELIXIR, traced the birth of Big Data in Biology to 1493: the year Columbus brought back new biology from the New World. The new technology of ocean-going ships ushered in the first wave of globalisation, allowed explorers to criss-cross the Atlantic and revealed a New World of flora and fauna. More importantly, a 'systematic recording' of the data began as explorers brought things back to Europe. The benefits were not limited to curiosity and cuisine; the plants of the New World led to vital new medicines. Spain appointed a Master Curator, a 'bioinformatician' tasked with letting go of the old, challenging Aristotelian classifications, and starting to classify and analyse the new 'data'.

Niklas Blomberg drew parallels between the discussion at the Summit around establishing a Danish Data Lake with the 15th-century Botanical Gardens built to house the plants from the New World. These gardens were carefully curated repositories for the data, with robust data governance built within secure walls. While each garden was secure, seeds were exchanged between gardens, just as the Summit highlighted the benefits of pooling and sharing 21st-century biological data. Indeed, Niklas Blomberg stated that the greatest opportunity for Denmark (given its size) is to work towards linking data and establishing collaborations with the other Nordic countries and the rest of Europe.

For both the first and current explosion of Big Data in Biology, research programmes and infrastructure need to be developed in conjunction. In addition to an ability to bring the data together, there must be strong governance and, while scientific research is mostly funded at the national level, it needs to be conducted at the global level. This should be easier to achieve now that transport technology has advanced since Columbus set sail.

6. Potential Solutions

In response to the defined challenges, the groups developed prototype solutions, which were then refined based on peer feedback. On the second day, the groups were shuffled, and the prototype solutions were reviewed by each of the new groups, who selected and refined their priority initiatives. Given the degree of overlap between the recommendations made by each group, the outputs from each are not listed in detail; rather, the following solutions have been distilled from the discussions.

6.1 Bioinformatics Research Institute

A number of working groups presented ideas for either a 'research institute' or the qualities that would be incorporated within such an institute (such as leadership, role models, interdisciplinary matrix structure) to address the identified challenges of rigid hierarchical structure in academia, limited interdisciplinary collaboration, and a lack of role models and clear and varied career options. Opinions on the nature of this institute differed across discussion groups. Some groups suggested a purely virtual, umbrella organisation linking existing researchers anchored in their home affiliations, perhaps utilising temporary 'make a space' departments to establish a physical core of bioinformaticians for the duration of a project. Other groups believed such an institute should be a standalone, 'bricks and mortar' establishment directly employing its own faculty.

Following discussions, the majority of experts agreed that there needs to be a physical institute to bring researchers from different disciplines together and create a critical mass. However, researchers could have dual affiliation and remain linked to their home institutions. Some experts also suggested that the Bioinformatics Research Institute could be an evolution of an already-established institute within existing university space to reduce budget requirements.

Such an initiative addresses many of the challenges identified by the groups as it provides a critical mass for ideation and a natural setup for 'blue-sky', challenge-driven research. It also creates a beacon for Nordic research, serves as an example of an interdisciplinary, matrix-structured institution, exemplifies career opportunities and provides better opportunities for young researchers to mix with role models. To ensure that individuals within the institution work collaboratively, they would be subject to ongoing reviews, overseen by an advisory board. There was some discussion on how much such a review process would be based on output, as there was a concern that output-driven reviews would not solve the expressed challenge of short termism within academia. Those not exhibiting the expected collaborative behaviours (or meeting minimum output requirements if such metrics are included in the assessments) could be removed from the institute (and its associated benefits) and simply remain attached to their home institutions.

Inspiration can be gained from a number of international examples, including Health Data UK,[1] a national body where scientists are employed by their associated universities but funded by Health Data UK. The UK's Turing Institute[2] was also cited as an example, along with Bell Labs[3] and the Isaac Newton Institute.[4]

An outstanding question from the meeting is whether such a research institute would have PhD and postdoctoral programmes entirely within its walls. Most groups included new cross-functional bioinformatics PhD programmes in their suggested solutions as there is a clear need for ambitious PhD programmes and open scholarships to attract the very best young research talent; the research institute could provide this.

Other prototype solutions that could be included as part of the institute are international visiting professors. In discussions on how to promote challenge-driven research, one group raised the possibility of establishing an international competition, combining research funding with a prize element to be worked towards as a relatively low-cost strategy. The key aim would be to promote international, cross-discipline collaboration by necessitating that different disciplines 'self-assemble' to solve a problem. The institute should be 'Danish without walls', acting as a global example for bioinformatics. Some also included the 'blue-sky professorship' as part of the institute. Collectively, these components would address the lack of openness and collaboration between disciplines that is borne out of a lack of funding flexibility.

Estimates of timelines for the initiative varied among the groups with suggestions of between 2 and 5 years to be fully established and up to 10 years to achieve a high impact. Although the institute would require significant investment, it was felt that such an initiative would provide the greatest impact of all the suggested options in the long term.

6.2 Centres of Excellence

Some groups addressed the challenge of promoting interdisciplinary collaboration by suggesting that existing centres be strengthened and supported. This would also raise the profile of bioinformatics in Denmark to help attract and retain talent. This could be pursued in the absence of a 'bricks-and-mortar' initiative such as the bioinformatics institute.

The centres of excellence would instead form a network across Denmark (with funding spread across the country) to promote cutting-edge postgraduate research and education, with strong links to industry and healthcare centres. It was suggested that existing centres in Copenhagen, Odense, Aalborg and Aarhus be designated centres of excellence, and strengthened and networked as such. Such an initiative would not only address the challenge of attracting and retaining talent, but would also provide a framework for a robust mentorship programme and the development of visionary leaders of the future.

1. https://www.hdruk.ac.uk; 2. https://www.turing.ac.uk; 3. https://www.bell-labs.com; 4. https://www.newton.ac.uk

6.3 'Blue-sky' Professorship

During discussion of the challenges, some individuals criticised current funding systems for being too focused on short-term project outputs. To be able to expand the definition of bioinformatics and fully realise its potential, long-term, hands-off funding, focused on people, not projects, is required. The suggested 'blue-sky professorship' solution was very highly regarded among the various groups. Specifically outlined by an international expert as requiring 10-year funding purely for research (no teaching), such a role would have complete freedom to think about challenges and projects in bioinformatics with the aim of implementing novel methodologies over the long term. While many noted the idea as high risk, with impact clearly in the long term, the potential outcomes could be ground-breaking and would certainly serve as a beacon and role model to others, providing additional impact in the more immediate term by attracting the very best research talent. Such an initiative would promote the innovation that is so often lacking in research as a result of the customary focus on shorter-term project funding rather than truly ground-breaking, longer research endeavours.

6.4 Education Academy

Educational activities are clearly needed to address the challenge of there being too few people within the field with the required competence profiles. As such, while there was some disagreement about the precise nature of an 'academy', all workshop groups selected such an initiative and/or educational programmes as a priority to address these challenges. Some groups combined research activities within or under the academy (as opposed to a bioinformatics institute) but the key activity to be addressed by this initiative is education. Such education would target all levels from high school students, university educational programmes, PhD and postdoctoral programmes through to vocational training of existing researchers and medical doctors. Therefore, a critical success factor for the academy is strong leadership.

There was some discussion over such an academy's role in educating those at university level, as this should be something the universities themselves are responsible for. Therefore, it was highlighted that this initiative would have to work very closely with the universities to be successful and take advantage of synergies. Some suggested the academy might fund PhDs and postdocs within universities. It was also highlighted that, in order to be successful, courses in bioinformatics need to be nationally accredited.

Parallels were drawn with the Danish Diabetes Academy[1] for this initiative. Similarly, an analogy was made with the Manufacturing Academy of Denmark[2] with regards to its university and industry ties, together with its PhD research programme. Whatever the precise nature of the academy, it would be expected to overcome the challenge of integrating education and research to ensure that flexible career paths are available for all.

6.5 Vocational Training

To address the immediate need for breaking down barriers of siloed working and overcoming the challenges of biologists working with computer scientists, vocational training activities were selected as a priority. As some advisors noted, biologists today should know how to program. From the reverse angle, there is a need to generate interest in the biological sciences among data scientists and machine learning specialists. Moreover, there is a clear need to educate teachers in bioinformatics if they are expected to educate children.

Vocational training initiatives should be designed to address both sides of the equation: bringing data science to life scientists and bringing life science to data scientists. As the field is fast evolving, there are benefits to be had from providing lifelong education to existing researchers, particularly as this may expand the pool of researchers able to supervise PhD students who use bioinformatics tools and enable the industry as a whole to accelerate research. Furthermore, the landscape report suggested that the demand for bioinformatics profiles was weakest in the healthcare sector and will remain so, perhaps because of limited awareness of the field. Some found this surprising due to our entering into the era of personalised medicine. Vocational training initiatives developed specifically for healthcare could stimulate activity within the sector and also encourage collaboration with industry and academia. Such initiatives would provide good impact within a short timeframe.

6.6 Outreach Activities

There is a clear challenge resulting from lack of awareness of bioinformatics and career paths within the field impacting the supply of the relevant profiles in sufficient numbers. There are simply too few people in the field, even though the landscape report predicts an increase in the number of students. To address this, and ensure that young people are motivated and attracted to careers in bioscience in Denmark, high-school students need to be aware of what bioinformatics is (and how the scope continues to widen), how exciting it can be to develop meaningful solutions with the discipline and what opportunities exist within Denmark. Delegates also highlighted information regarding the decreasing number of young people in Denmark (the number of 21-year-olds will decrease by 15% in the coming years) and a general backdrop of all STEM areas being undersubscribed.

To address this, outreach activities need to be developed to inspire students and ensure they are aware, from school age, what bioinformatics is and what they can achieve if they develop the necessary skills. Such activities could be materials outlining the area and career options, materials for teachers to incorporate into their own courses and

1. www.danishdiabetesacademy.dk; 2. https://www.made.dk

'demo days'. They should not only address bioinformatics, but also 'humanise the STEM subjects', highlighting the chance to do good for humanity. It was suggested that the Foundation take inspiration from US science fairs, perhaps organising a national project on meaningful data for children and teachers to present to parents. Similarly, others described this as letting students dive into the data lake to learn how to swim. Indeed, while high school students may be limited in what they can do in the lab, everyone is able to use their computer or mobile phone, meaning they may be more amenable to such a project, particularly when they can see the potential impact. It was also highlighted that social media should be used when promoting to young would-be bioinformaticians. Initiatives like this that overcome the challenge of engaging with young students and the wider public will be instrumental to reframe the perceptions around the use of personal data as something that has the potential to have huge benefits for healthcare and wider humanitarian issues.

6.7 Data Standardisation and Sharing

Denmark is in a strong position with its centralised health system and associated data but, while the data are there, the meta-data layer that makes them usable together is not well established. Regional variations in methodology and a 'closed-down' approach make it very difficult to capitalise on these data. The state and the five regions have a say in how data are organised but a top-down approach to data management is required. Therefore, it was suggested that a formal organisation is established to facilitate inter-entity data exchange and a country-wide standardisation of data collection and sharing. The political obstacles were acknowledged by the experts but, if the drive is supported by educational and branding activities, the potential benefits will help bring about changes in healthcare and beyond.

A high-profile flagship project, or projects, based on these data would serve to demonstrate the potentially huge benefits of such research. A clear healthcare benefit resulting from such a project could also help lessen public and hospital concern over the collection and usage of such data. Such projects could also be run in conjunction with the Institute or Academy to reinforce their positions and provide raw material education and training. EarthCube – a community of geoscientists, geoinformaticians and data scientists focused on related issues for geoscience data – was raised as an example.

Discussions took place over whether the formal body to facilitate data exchange should be part of the Academy or the Institute, but many felt this should be an entirely separate initiative with a remit independent of any other body or sponsor. It should be clearly established as an independent body working to standardise data formats for all parties to benefit from. There were also suggestions that this initiative should aim to create a global healthcare data format from the outset, rather than a Danish standard only.

6.8 HPC Ecosystem

Access to a fast, flexible and secure infrastructure and the ability to combine different types of sensitive data and perform analyses are key prerequisites for conducting bioinformatic research and participating in collaborative research with international scientists. Up-to-date supercomputers can also attract and retain talented individuals.

There was some disagreement between the experts on the state of HPC in Denmark. Some believe a good HPC ecosystem already exists, but that it just needs to be maintained and strengthened; others believe a current lack of computing power is a big issue for Denmark. Denmark certainly has recent success in strengthening its life science HPC infrastructure in the form of The Danish National Life Science Supercomputing Center, named Computerome: an HPC facility specialised for life science. Computerome is physically installed at the Technical University of Denmark's campus in Risø, funded by grants from the Technical University of Denmark, University of Copenhagen and the Danish e-infrastructure Cooperation. Current users include research groups from all Danish universities and large international research consortia, as well as from industry and the healthcare sector. A second-generation super computer (Computerome 2.0) is currently under construction.

It was suggested that computing power issues will be solved through cloud computing and international collaborations, and should not be a priority (rather, the data are a priority). Without a change in the legal framework in Denmark (currently mandating that Danish data stay within Denmark), cloud computing outside of Denmark may not be an option. However, establishing a 'private cloud solution' within Denmark may be possible to provide further capabilities beyond those of infrastructures like the Computerome.

Building Institutions

Professor Chris Sander, Harvard Medical School, Dana Farber Cancer Institute

Chris Sander is a computational biologist, and is currently the Director of the cBio Center at Dana Farber Cancer Institute within the Department of Biostatistics and Computational Biology. He is also a Professor of Cell Biology at Harvard Medical School and an Associate Member of the Broad Institute. He drew on his vast experience, in particular with the European Bioinformatics Institute and the European Molecular Biology Laboratory (EMBL), to discuss the key elements involved in building a successful institution. The EMBL is of particular note because of its great impact in becoming a turntable for post-docs and young scientists to be spun out across various European countries, something very relevant to the discussions around how to increase the numbers of bioinformatics skills profiles.

Based on Chris Sander's experience, the key elements in building a successful institution are:

- Ambitious vision
- Long-term commitment
- Collaborative and open environment (culture and buildings)
- Nice buildings in which to work
- Funding

Chris Sander also highlighted the importance of collaboration and bringing computational biology and quantitative sciences to appropriate biological problems, as he does at Dana Farber and has done in the past while chair of the Computational Biology Programme at the Memorial Sloan-Kettering Cancer Center (MSKCC) in New York. Research needs to focus on turning data into knowledge but, more importantly, this knowledge needs to lead to action; action that provides benefit to society and contributes to what he sees as a continued evolution of our species. He regards evolution as not just DNA-based life but also all that DNA-based life creates. As we continue to evolve in this way, the definition of, and potential 'action' from, bioinformatics will be refined and expanded.

On the subject of definitions, Chris Sander also suggested that bioinformatics should be renamed to 'biological data science' to clearly place it as a biological speciality within the data sciences rather than a separate discipline, which might limit its scope and potential.

In summary, to build an institution that can successfully bring about meaningful beneficial action, one needs to define a good problem, then define projects to address the problem, define funding, build a nice open collaborative space, bring in leadership with the vision to support long-term, open collaboration, then leverage global resources to support the work.

7. Summary

In summary, the expert delegates recommended a variety of solutions with an underlying strategy of establishing and fostering rewarding, interdisciplinary, challenge-driven research utilising FAIR (Findable, Accessible, Interoperable and Reusable) data, while strengthening educational initiatives to encourage both high school students and university students in other disciplines to consider a career in bioinformatics. Differences of opinion on the current and future state of HPC in Denmark warrant further exploration.

While not listed as a defined recommended solution, there is a clear need for a 'branding' campaign across all activities. Most groups at the Summit raised compensation as a challenge to attracting and retaining talent, but it was also agreed that it will not be possible to compete with salaries available in fintech and from companies like Google and Facebook. Rather, careers in bioinformatics (in industry, academia and healthcare) need to be made sufficiently attractive in other ways such that salary becomes a secondary issue. Five parameters to retaining high-quality bioinformaticians were outlined: "interesting data, smart colleagues, a nice location in which to live and work, a quality computer and a good salary". If Danish biosciences cannot compete on the last element, the first four need to be nurtured and clearly communicated so that potential recruits are excited by the potential outcomes from bioinformatics and see a clear role and career path for those involved in using bioinformatics to make a difference to global health and wellbeing. Describing this critically important field as 'biological data science' rather than 'bioinformatics' was also recommended by some experts to further help improve attractiveness and bolster its own deserved right as a forward-thinking, independent yet cross-disciplinary field.

8. Appendix: Summit Participants

Summit Participant	Affiliation
Anders Krogh	University of Copenhagen Dept of Biology, Bioinformatics Center
Bengt Persson	SciLifeLab, Sweden
Bernhard Palsson	Technical University of Denmark, NNF Center for Biosustainability
Chris Sander	Harvard Medical School, Dana Farber Cancer Institute, US
David T. Jones	University College London, Francis Crick Institute, UK
Finn Cilius Nielsen	Rigshospitalet
Henrik Bindslev	University of Southern Denmark, Faculty of Engineering
Henrik Geertz-Hansen	Novozymes
Henrik Semb	University of Copenhagen, NNF Center for Stem Cell Biology
Ivo Glynne Gut	Centro Nacional de Análisis Genómico, Barcelona, Spain
Jakob Skou Pedersen	Aarhus University, Dept of Molecular Medicine
Jens Christian Godskesen	IT University of Copenhagen
Jens Kringelum	Evaxion Biotech
Jens Lundgren	Rigshospitalet
John Renner Hansen	University of Copenhagen, Faculty of Science
Katrine Krogh Andersen	Technical University of Denmark
Lars Bo Nielsen	Aarhus University, Faculty of Health
Lars Fogh Iversen	Novo Nordisk
Martin Bøgsted	Aalborg University, Dept of Clinical Medicine, Hospital
Martin Lange	Novo Nordisk
Martin Zachariasen	University of Southern Denmark, Faculty of Science
Mikkel Heide Schierup	Aarhus University, Bioinformatics Research Center
Niels Christian Nielsen	Aarhus University, Faculty of Science
Niklas Blomberg	ELIXIR, UK
Ole Nørregaard Jensen	University of Southern Denmark, Dept of Biochemistry and Molecular Biology
Phillip E. Bourne	University of Virginia, US
Rasmus Larsen	Technical University of Denmark
Roald Forsberg	Raven Biosciences
Rolf Apweiler	European Bioinformatics Institute, EMBL, UK
Shomit Ghose	Onset Ventures, US
Steen Pedersen	Technical University of Denmark
Søren Brunak	University of Copenhagen, NNF Center for Protein Research
Søren Egested	Novozymes
Torben Hansen	University of Copenhagen, NNF Center for Basic Metabolic Research
Ulla Wewer	University of Copenhagen, Faculty of Health and Medical Sciences
Wouter Krogh Boomsma	University of Copenhagen, Dept of Computer Science

NNF: Novo Nordisk Foundation