Department of Computational Biology at Cornell University
A Vision Statement

Contributors (alphabetical): Andrew Clark*, Alon Keinan, Susan McCouch, Philipp Messer, Jason Mezey*, Amy Williams, Haiyuan Yu*

September 5, 2017, revised January 24, 2018

Executive Summary

- Strength in basic and applied research in Computational Biology is essential for the health of biological research at Cornell and the broader Cornell mission.
- There are compelling reasons to form a new Department of Computational Biology at Cornell, including:
    - Its mission-critical function encourages joint ownership by multiple Colleges and Schools that contribute faculty lines.
    - Cornell is already a magnet for top talent in this area, making it easily in our grasp to form a department with a world-class reputation composed of faculty with collaborative research connections throughout Cornell.
    - Faculty mandate of research, consulting, and teaching in computational biology will have broad impact throughout the life sciences.
    - The department will focus on areas aligning with current and future Cornell research strengths: methodology development, genomic/ proteomic data analysis, and modeling of underlying processes.
- The new CB department will be unusually outward looking and collaborative.
- The document closes with an Appendix that identifies leading peer institutions in computational biology, showing that we have some catching up to do to match their level of commitment and engagement.

Strong Computational Biology at Cornell is essential for the life sciences

Every discipline of the biological and life sciences now incorporates high dimensional data. At present, the most prolific high-dimensional biological data are genomic sequences, but proteomics, imaging, electronic records, and electronic monitoring data are also emerging as essential components of basic and applied biological research. Using high-dimensional data for research requires skill sets that are beyond the training of the majority of biologists educated just a decade ago. For example, many of these data sets require specialized management and manipulation techniques (*e.g.*, they are

too large to be opened in text editors or Excel), they require database and other tools for storage facilitating integration with large public data repositories (*e.g*., processes that would be inefficient or impractical by hand); for analysis they often rely on specialized modeling techniques or new computational analysis methods that require high-performance computational resource (*e.g*., they cannot be run on a desktop computer); the output of the analyses of large-scale genomic studies include long lists of likely candidate genes, requiring high-throughput computational-experimental integrated assays to identify a small number of functionally relevant targets to deliver actionable hypotheses. Every academic institution that plans to be a leader in biological and life science research will need to have a strong computational biology presence, where such a presence includes: 1. Professors developing new research that pushes modeling and methodology development, 2. A world-class community of undergraduate, graduate, and postdoctoral researchers being trained in many facets of the computational life sciences, 3. Resources and leadership initiatives that enable computational biologists and experimental biologists working throughout the basic and applied life science disciplines. This is why every major academic institution and every major peer of Cornell is making a significant investment in computational biology (see Appendix).

The current state of Computational Biology at Cornell

Cornell is a world leader in life sciences research thanks to its fantastic breadth and depth. However because of structural arrangements, we lag behind our peer institutions in Computational Biology. On Cornell's main campus, we have a cluster of professors working in computational biology situated in the Department of Biological Statistics and Computational Biology (BSCB) and many others in half a dozen other departments. Over the past decade, the Computational Biology faculty in BSCB have developed world class reputations through their lab research and highly regarded collaborations that have generated publications in the highest impact journals, as well as impressive funding portfolios. BSCB has also been responsible for the bulk of undergraduate and graduate students trained in Computational Biology, where many of these students entered or continued their careers in Computational Biology or related fields. These successes are all the more impressive because BSCB has been kept well below critical mass (currently 3 mid-career tenured faculty, 2 junior pre-tenure faculty, and 2 senior faculty with joint appointments), in part because of the success of mid-career faculty who left for offers at other institutions (e.g. Rasmus Nielsen, Carlos Bustamante, Adam Siepel) and because of inadequate recruiting. There have been recent individual hires of professors working in Computational Biology in other departments and schools at Cornell (*e.g.*, Adam Boyko and Praveen Sethupathy in Veterinary Medicine, Charles Danko in the Baker Institute, Jeffrey Varner in Chemical and Biomolecular Engineering,

Ilano Brito, Ben Cosgrove, and Iwijn De Vlaminck in BME, Zhenglong Gu in Nutritional Sciences, Amnon Koren in MGB, Adam Bogdanove and Kelly Robbins in Integrative Plant Sciences; where it should be noted that 10 of these 11 hires have a significant experimental component to their research programs with assigned wet lab spaces in their start-up packages). The Department of Computer Science (and more broadly the program of Information Science) at Cornell have not hired any Computational Biology faculty in over a decade. Weill Cornell Medical College has been increasing their investment in Computational Biology with 3 junior and mid-career pre-tenure professors in the department of Physiology and Biophysics (Olivier Elemento, Chris Mason, Ekta Khurana). The Cornell Tech campus has not hired any computational biologists beyond engaging a few postdocs trained in Computational Biology. Overall, while Cornell has had a successful footprint in Computational Biology, this success has been driven by a relatively small number of faculty compared to peer institutions (see Appendix). A renewed investment in Computational Biology is necessary if Cornell is to continue as a leader in the biological and life sciences.

The Vision: A new Department of Computational Biology

In this section we lay out a plan for moving from the current departmental arrangement to a new structure that we feel will be more forward looking and allow Cornell to recruit and retain talent at all levels in the area of Computational Biology. We eagerly seek input from Deans and the Provost to develop this departmental plan. It should include aspirations for location, resources, linkages with other departments, rules for joint appointments, and be accommodating to lab-oriented faculty. Here is a step-by-step plan to develop a truly outstanding department of Computational Biology for Cornell.

1. Developing University-wide support. The current BSCB has been precluded from adequate expansion in both faculty recruiting as well as training outreach by being housed entirely within CALS. There are compelling reasons to broaden this support. The new CB department will be a trend-setting unit, both in teaching and in research, with a mission that could be critical to several other colleges. We envision a department structured somewhat like MBG, with individual faculty lines coming from each of several colleges. As faculty lines open, negotiation for startup packages, salaries, etc. would be done with the Dean of the relevant college. Indirects would be partitioned as they are in MBG, as would teaching credits. CALS would continue to benefit from the mission of this unit, and its teaching and graduate training of many of the faculty would continue to match the mission of CALS. But on top of this, many faculty in CB will have Computer Science degrees, and will focus on methodology and algorithm development for solving problems in computational genomics and biology. The training mission in this area is closely aligned with that of CIS, and the linkages with high dimensional statisticians in

the current BSCB department will continue to be encouraged. These linkages will best be fostered by having CIS get part ownership of CB.

The College of Arts and Sciences already has faculty in departments like Molecular Biology and Genetics who are computational biologists (*e.g*. Amnon Koren, Jeff Pleiss, Andrew Clark) and the basic science mission of CAS is also well served to have a part of CB. Similarly, the College of Veterinary Medicine is increasingly experiencing the demand for computational approaches in research, education and its medical services. Recent recruits like Charles Danko (Baker Institute), Adam Boyko and Praveen Sethupathy (Biomed Sciences) would be at home in a Computational Biology department in any major university. If the Vet College had part ownership (and faculty lines) in Computational Biology, they would gain the cache of being one of the few vet colleges with a CB department, and they would gain access to the research resources and training expertise of the group. The College of Engineering may also have a part of CB. Even though the missions of CB and Biomed Engineering may seem different as overall units, individual faculty within these two units may have considerable overlap, and so the College of Engineering may find the same reasons for buy-in to CB to be compelling. Finally, there will be enormously important opportunities for the Weill Medical College to realize the power of a unified and cohesive CB department at the Ithaca campus, as is seen at key competitors like Stanford, Harvard and the University of Washington. This document will not go into detail articulating exactly how to achieve this balance between the Weill and Ithaca campuses, but the role of computers in medicine is expanding at a dizzying rate, and the Weill campus will need to be poised to capture these opportunities.

2. Core areas for faculty recruitment in CB. Given the small size of the current computational biology faculty, we have little choice other than to build on current strengths. The three areas that align with current Cornell research strengths, future research needs, and intellectual investments of Cornell peer institutions include computational methodology development, genomic/proteomic data analysis, and systems biology modeling. Below we expand on each of these in turn:

Methodology development: The current BSCB department had a good vision, namely to bring together people focused on the biological questions of computational biology with statisticians working on high-dimension problems. This has given BSCB a unique slant, and has helped build the reputation of past and current members of the department. In the future, we need to retain this strength by maintaining strong ties with high-dimensional statisticians while building strength in analysis of high-dimension functional genomics data. The current department, while small, has some highly visible stars

among its faculty in the area of methodology development, and these will serve as a seed to nucleate faculty expansion that retains and develops excellence.

Genomic/proteomic Data Analysis: The area of genomic and proteomic data analysis is the one with the greatest demand in training and in person-hours of research demand. This is in part driven by the flood of data, but also by the terrific research opportunities brought about by the easy of generating impressively informative (and large) data sets.

Systems Biology: Systems biology is an expanding field that is inherently interdisciplinary, incorporating ideas and methods from biology, statistics, computer science, engineering, and other physical sciences. It is built upon the understanding that "the whole is greater than the sum of the parts." The main goal of systems biology is to develop technologies and models to systematically and accurately measure and predict physiochemical properties and behaviors of biomolecules within cells, tissues and whole organisms, and to learn how such complex systems change over time and under different conditions. Great progress has been made by integrating systems biology approaches with genomics information to discover new biomarkers for disease and to stratify patients for personalized treatment. Cornell is in a unique position to broaden the impact of systems biology into non-human organisms, such as systems agriculture. The center of mass of Systems Biology currently lies in Engineering, but with Haiyuan Yu in CB, there will be further opportunities to hire in this area within CB.

3. Joint appointments. An important means of buttressing the department in its next phase of recruitment is to identify key faculty who are computational and/or systems biologists but who are not in the CB department, to agree to play a role in the growth of CB through commitments in the form of joint appointments. This will help seed all subsequent recruitment efforts. While it seems that the fused department (with CB and BS faculty) never quite caught fire, it is appreciated that high-dimensional statistics is a critical aspect of genomic analysis and so every effort needs to be taken to retain close ties with selected statistical faculty, and joint appointments in CB would be encouraged. These joint appointments would initially be structured like the current joint appointments of Drs. Clark and McCouch, where the primary department remains the sole tenure home, and even teaching commitments, but joint hires would be tapped for their expertise to help with faculty recruiting, curriculum and other departmental decisions. Similarly, the computational biologists whose primary affiliation is in other units would be encouraged to seek joint appointments in CB. This would include people like Adam Boyko, Ilana Brito, Charles Danko, Iwijn De Vlaminck, and Praveen Sethupathy.

***4. Overlaps with current departments*.** It is inevitable that diverse department across campus will seek to hire computational biologists. There will be candidates like Amnon

Koren, to single out one, whose primary questions (origins of DNA replication) fit in a department like MBG, but whose methods make novel and creative use of computation. It is important that CB not be territorial, but that it celebrate advance of the field by the distributed hiring. It may come to pass that CB would evolve to be more focused on methodology development, but in any case its focus would be less likely to have faculty that drill narrowly into a single biological problem. Hires in other departments would be sought out for joint appointments, and the existence of CB would become an added magnet to assist in recruiting top talent into many other departments. There will be excellent opportunities for cooperation with BioMed Engineering in recruiting efforts, as well as formal collaboration, and through open communication there will emerge a natural partitioning of areas that are the focus of one or the other. Similarly, at many universities, Computer Science has a strong contingent of faculty in Computational Biology, and while this is not currently so at Cornell, algorithmic and other fundamental CS problems that arise in Computational Biology should continue to be an attractive possibility for recruitment.

5. ***Curriculum*****.** One of the ways to secure broader buy-in across the university is for a broader group of individuals to see that CB has something to offer to them. Training in CB is clearly important for many disciplines, and so accessibility of that training for a broad group of undergraduates and graduate students (and postdocs) is important. Careful design of an engaging curriculum that targets these needs will go far to buy success of CB at Cornell. Benefits of the undergraduate and graduate education programs generated by faculty in CB will arguably be accrued campus-wide, but each of the Colleges mentioned above will have its own faculty and students also get benefit from the teaching mission of the new CB department.

6. Computational Biology consulting**.** Cornell's unique strength in the life sciences is its diversity, with its faculty probing a vast array of biological questions. Many of these faculty are not up to speed with the latest computational and genomics approaches, and they are or soon will be held back by this. This gives the new CB department a core mission that is different from many departments. If they could embrace a service mission, encouraging faculty to give some small portion of their time to consulting services, similar to the Cornell Statistical Consulting unit, this could serve to raise the standards of computational analysis university wide. This will have to be phased in carefully. Not all statisticians are consultants, and similarly there could be individuals hired to specialize in this service. We currently have a Bioinformatics Core which provides much of this form of consulting, and we see advantages to pushing this important service to a new level.

How a new department of Computational Biology will amplify the strengths of Cornell life sciences

A new department of Computational Biology, with the structure and mandates described in the previous section and supported as described in the following section, will strengthen life sciences Cornell in five interconnected dimensions:

1. Solidifying Cornell as a world leader in biological and life sciences research. A new department will ensure Cornell's position as a leader in the development of field-leading research in Computational Biology. Being an institutional leader in any field is desirable, but these areas will also continue to expand and grow in importance with the rapid expansion of biological data sets. What's more, tools developed in the field of Computational Biology are fast becoming essential components of biological and life sciences research. An appropriately designed new department will ensure Cornell-wide collaborations that will enhance the unique breadth and depth of life science research spread across 23 departments on the Cornell main campus. These collaborations will be enabled by the faculty with primary appointments in the new department with collaboration mandates and the joint-appointed top faculty with a strong computational component of their research programs appointed in departments where they have a natural collaborative fit. In short, a new department with a University-wide organization and collaboration mandate bringing together faculty that have expertise in system-level analysis, simulation research, computational expertise and new technologies for generating high throughput, comprehensive and quantitative experimental data will help nucleate access to the very best purely computational colleagues, and allow people with a primarily theoretical or computational background novel opportunities for lab research. Cornell could easily situate itself to be recognized as being particularly good at fostering this cross-disciplinary work.

2. Cornell will offer the highest quality education and training in Computational Biology for a diversity of students at all levels. A new appropriately mandated department will provide education and training in Computational Biology that is essential for performing research in the life sciences, and for obtaining any job that requires computational biology expertise. The next generation interested in careers involving basic or applied research, whether directly in the areas of Computational Biology or in the life sciences more broadly, will soon require training in Computational Biology. More broadly, the number of jobs requiring skills or a working knowledge of Computational Biology in industry, medicine, academia, the non-profit sector, or government will be dramatically increasing. A new department would lead the Cornell education and training effort in Computational Biology. This would include a spread of courses having strong enrollment that would pull in students from across the university. Beyond courses

developed and organized by the faculty of a new department that will be aimed directly at undergraduates and graduate students specializing in Computational and System Biology, the department will organize a broader curriculum of courses serving the needs of the life science majors and departments across Cornell. Additionally, the new department will have a consulting mandate for faculty that will include training programs providing a diversity of needed skill sets for Cornell students at all levels, development of an MPS program for training in Computational Biology, and continued / expanded leadership in organizing cross-Cornell campus education programs (*e.g.*, including courses taught by video-conference on multiple Cornell campuses or graduate programs that span the campuses, two areas where Cornell Computational Biology faculty currently have a strong track record). In short, a new department would serve as a hub for exciting and innovative educational and training opportunities, showing students powerful and creative approaches to use computers to advance life science research at the undergraduate, graduate, and post-graduate level.

3. Augmenting Cornell's reputation as a top academic institution. The academic institutions with reputations as being among the best in computational biology will all be leaders in the life sciences. This is a consequence of the increasingly rapid pace of discoveries and technology advances in the life sciences that are touching every aspect of people's lives, from medicine to the food they eat. To be a leader in the life sciences over the foreseeable future, an institution will need to be strong in Computational Biology. The value to life science research and education that a new department of Computational Biology will bring, will be a critical pillar in augmenting and building Cornell's reputation as a world-leading academic institution. A new department will enhance Cornell's reputation, which in turn will enhance recruitment to the Cornell community, influence on thought leadership, and the ability to achieve Cornell's core missions

4. Contributions to Cornell's financial health. A new department of Computational Biology will contribute to the financial health of Cornell in at least five ways: 1. It would increase the number of successful grants in the Cornell life sciences portfolio, 2. It could host an MPS program in Computational Biology, 3. It would serve as a donation target for philanthropy and companies, 4. It would generate Intellectual Property (IP), and 5. It would serve as a foundation for start-ups built by Cornell faculty. For the first, governmental and institutional grants in the life sciences will include Computational Biology components. Funding success for Cornell faculty in the life sciences will be improved with a new department. For the second, MPS programs focused on Computational Biology are increasingly in demand. Masters students are willing to pay the cost of such education and given the increased employability and salaries of students who complete top programs in these areas, asking for the investment is ethical and justifiable. For the third, Cornell peer institutions have demonstrated that

philanthropy is increasingly being aimed at enhancing Computational Biology. Companies are also investing in University Computational Biology initiatives to ensure a workforce for the future. To share in these initiatives, Cornell will need a strong department of Computational Biology. For the fourth, the national percentage of income generating IP being produced by basic and applied science is increasing rapidly. A new department will provide the foundation of critical expertise that Cornell life sciences faculty will need to make such licensable discoveries. For the fifth, while traditionally University faculty in computer science and engineering fields produced the largest number of startups, life science faculty are now joining this trend. A new department will enable the already excellent track record of Cornell life science startups (*e.g*., Embark started by Adam Boyko). These life science startups will generate an increasingly large amount of revenue for Cornell.

5. Attaining the "One Cornell" vision. Compared to peers, Cornell University has a number of exceptional challenges that make a unified University difficult, including a combined private / public structure and geographically separated campuses. While there are many components that need to be address as outlined in the "One Cornell" vision, having a collaborative research environment that spans departments, schools / programs, and campuses is a clearly important aspect. Research engagement not only promotes intellectual exchange but also encourages other unifying activities (*e.g*., joint grant writing, travel for collaborative purposes, education, administrative connections and initiatives). Collaborative research across the different cultures of departments is already challenging, where geography only adds to the challenge. Unlike most research in the life sciences, Computational Biology research is particularly well-suited for spanning these divides, both because these areas are becoming essential for research and because the critical components are mobile (e.g., concepts, methods, and data). The current small number of Computational Biology faculty in BSCB have long-term research collaborations that span departments, schools, and campuses (see Appendix 6). A new department with a collaboration mandate would quickly become a conduit and foundation for an increasingly large number of collaborations among faculty throughout Cornell.

Critical Components and Investments

In this section, we list eight critical areas requiring the influence and investment by Cornell leadership to assure success of a new department of Computational Biology:

1. Departmental Structure, Mandates, and Autonomy. To achieve the goal of building a new field-leading department that is a research, service, and education hub

for Computational Biology at Cornell, it will be critical that Cornell leaderships put necessary structure, mandates, and autonomy controls in place. These include a core faculty with primary appointments in the department, as well as faculty from other departments with a joint appointments, that the primary faculty have a clear mandate to build their own research program and collaborate with Cornell faculty with whom research interests naturally intersect (which can be encouraged by departmental / school ownership of faculty lines), that they have an contractual service mandate (similar to Statistics), and that they have a contractual teaching mandate. It is also essential that the department has autonomy on critical administrative decisions including tenure, raises, teaching and committee assignments, and selection of a departmental chair, despite its multiple college foundation.

2. Vigorous Faculty Recruitment. The core of the new department will be the current Computational Biology in BSCB. However, we need to immediately recruit a visionary chairperson for the new department without delay. This is will provide essential leadership (where we have seen the impacts of not having such leadership). We are also need to immediately begin recruiting new junior faculty (mid-career or senior if appropriate targeted hires become available), since the faculty roles are dangerously close to being below critical mass. Successful recruitment of five additional faculty plus the chair over the next three to five years will be the minimum needed to achieve the critical mass to achieve the goal of building a new field-leading department that is a research, service, and education hub for Computational Biology at Cornell.

3. Space. The new department needs a space where all of the faculty with primary appointments can be co-located. This will require dry lab space for faculty and lab groups of the five current Computational Biology faculty in BSCB, plus space for the five new faculty plus the chair, with additional space for growth (5-6 positions) over the subsequent five years. This space will also require access to flex wet lab space in the building for faculty who have experimental components to their research program (currently just Haiyuan Yu, but some of the new recruits will undoubtedly have wet lab space needs.

4. Teaching Support. To remain in step with peer institutions, the teaching load for the primary faculty of the new department will need to be similar to that of molecular biologists, roughly one full semester lecture course for majors and one additional symposium course per year. TAships will need to be allocated for these courses (we note that currently the TA situation for BSCB is rather ad hoc, and terms need to be solidified).

5. Support for the Graduate Field in Computational Biology. The primary source of students for Computational Biology faculty in BSCB is the graduate field of Computational Biology. This field is undergoing a re-building effort but the financial health of the field has not yet been assured by obtaining a training grant. Until this occurs, it will be essential that the field is able to continue supporting the first year of incoming students to be competitive with peer programs

6. Computational Resources. Research in Computational Biology is compute intensive, requiring high performance compute (HPC) cluster resources. At present, the Computational Biology faculty at Cornell have access to an excellent HPC resource that is subsidized by Cornell leadership. Support of this resource will need to continue and increase as the growth of the department puts increasing pressure on this resource. Supporting an HPC resource is well in line with the investment peer institutions are making to support their Computational Biology faculty. For example, with an $11 million investment in computer hardware, the Johns Hopkins Center for Computational Biology has a cluster with 18,000 CPUs and 22 petabytes of storage, roughly 20 times the computing resource available at Cornell. We are starting discussions with the Data Science Initiative Task Force about resource issues.

7. Control of the Bioinformatics Core. An important mandate for the new department is a strong service and consulting component, similar to the model for Statistics faculty. However, unlike statistical consulting, the analogous consulting in Computational Biology is not just conceptual and methodological advice, but also requires additional programming and other support that will be beyond the capacity of a department faculty member. The current Bioinformatics Core operates with autonomy from the research faculty working in Computational Biology and is guided by a faculty advisory panel. It is highly likely that a closer tie to a department like CB would leverage these consulting services, pull in more faculty for the grey area between consulting and collaborating, and provide all with better service and greater collaborative opportunities.

8. Financial management and office staff. As the new department grows, there will be a growing administrative burden. While the current administration support for the Computational Biology faculty of BSCB is outstanding, this will need to be increased to support the aggressive grant writing, research, and other activities of existing and new faculty, to support the Computational Biology graduate field and the new MPS program, as well as the administrative needs of a department. Increase in administrative support will be offset by the direct (i.e., the MPS) and indirect (i.e., grants) financial successes of the new department.

Appendix 1 – Organization of Computational Biology at peer institutions

Here is a sampling of other programs at peer institutions, showing the high level of commitment to computational genomics and biology:

Carnegie Mellon has a Computational Biology Department of 20 primary faculty in its School of Computer Science.
https://ccb.jhu.edu/

Harvard has a Department of Biostatistics with a section of 41 primary faculty in Bioinformatics and Computational Biology.
https://www.hsph.harvard.edu/biostatistics/bioinformatics-computational-biology/

Johns Hopkins has a Center for Computational Biology with 28 primary faculty.
https://ccb.jhu.edu/

MIT has more than 70 faculty in its program in Computational and Systems Biology, with 15 in Biological Engineering, 27 in Biology, 8 in Chemical Engineering and 23 in Electrical Engineering and Computer Science.
http://csbi.mit.edu/#

Stanford has a Computational Biology group in Computer Science,
http://compbio.stanford.edu/
and a Center for Computational Evolutionary and Human Genomics,
https://cehg.stanford.edu/
and the Department of Biology has a Computational Biology group,
https://biology.stanford.edu/research/research-areas/computational-biology
as well as a Department of Biomedical Data Science.
http://med.stanford.edu/dbds.html

University of California at Berkeley has a Center for Computational Biology with 51 faculty across 5 colleges.
http://ccb.berkeley.edu/

UCLA has a Center for Biomedical Informatics and a Department of Biomathematics. Their faculty are tied together by a graduate program in Bioinformatics that has 50 active faculty members. Computational Biology is well represented in their Computer Science department as well.
http://bioinformatics.ucla.edu/faculty/

University of Michigan has a Department of Computational Medicine and Bioinformatics with 15 primary faculty.
https://medicine.umich.edu/dept/computational-medicine-bioinformatics

University of Washington has 38 faculty in its Program in Computational Molecular Biology,
http://cmb.washington.edu/
and 31 faculty in its Department of Genome Sciences.
http://www.gs.washington.edu/

Washington University has a Division of Biology and Biomedical Sciences with a program in Computational Biology.
http://dbbs.wustl.edu/divprograms/compbio/Pages/default.aspx