The NIH Bioinformatics and Computational Biology Roadmap

Eric Jakobsson

Chair, NIH Bioinformation Science and Technology Initiative Consortium

Director, NIGMS Center for Bioinformatics and Computational Biology

For the President's Information Technology Advisory Committee, June 17, 2004

Why Computational Biology at the NIH

- Because computation and information technology is an invaluable tool for understanding biological complexity, which is at the heart of advance in biomedical knowledge and medical practice.
- “You can’t translate what you don’t understand”---Elias Zerhouni, Director of the National Institutes of Health, commenting on the relationship between basic research and translational research, that transforms the results of basic research into a foundation for clinical research and medical practice.

Some important problems with biomedical computing tools are:

- They are difficult to use.
- They are fragile.
- They lack interoperability of different components
- They suffer limitations on dissemination
- They often work in one program/one function mode as opposed to being part of an integrated computational environment.
- There are not sufficient personnel to meet the needs for creating better biological computing tools and user environments.

Why the problems with biological computing tools must be fixed.

- Today computation is at the heart of all leading edge biomedical science. For leading examples, consider this past year’s Nobel prizes:
- Structure of voltage-gated channels—required sophisticated computation for image reconstruction for x-ray diffraction data, the mathematical techniques for which were the subject of a previous Nobel prize.
- Discovery of water channels—The experimental work required augmentation by bioinformatics for identification of water channel genes by sequence homology.
- Magnetic resonance imaging—A large share of the prize work was for the mathematical and computational techniques for inferring structure and image from nmr spectra.

The Paradox of Computational Biology--Its successes are the flip side of its deficiencies.

- The success of computational biology is shown by the fact that computation has become integral and critical to modern biomedical research.
- Because computation is integral to biomedical research, its deficiencies have become significant rate limiting factors in the rate of progress of biomedical research.

Mission Statement

- In ten years, we want every person involved in the biomedical enterprise---basic researcher, clinical researcher, practitioner, student, teacher, policy maker---to have at their fingertips through their keyboard instant access to all the data sources, analysis tools, modeling tools, visualization tools, and interpretative materials necessary to do their jobs with no inefficiencies in computation or information technology being a rate-limiting step.

Historical Highlights

- 1999—Botstein-Smarr Committee Recommends Establishment of BISTI and of national biomedical computing centers.
- 2001—BISTI established—search for Chair launched
- 2003—Chair hired, Funding Announcement Issued for National Centers for Biomedical Computing, Digital Biology Week is held in Collaboration with NSF and NIST

In 2004

- First Set of National Centers for Biomedical Computing will be established
- Funding Announcement will be issued for smaller projects to collaborate with the centers in creating the national biomedical computing infrastructure.

In 2005

- Establish second set of National Centers for Biomedical Computing
- Establish first set of collaborating projects with the National Centers for Biomedical Computing.
- Ramp up to full functionality oversight and coordination of the national network of collaborating projects to create the National Program of Excellence in Biomedical Computing for the creation of a excellent national biomedical computing environment.
- Expand coordinated efforts with other agencies where there is synergy. (Currently NSF-NIH Biology-Mathematics Initiative is in its third year, multi-agency multiscale modeling initiative will be announce later this year, formal and informal planning meetings are underway with other agencies.)

2006 and Beyond

- Continued Implementation of the Ten-year NIH Roadmap in Bioinformatics and Computational Biology, with on-the-fly adjustments as appropriate.

Computational Biology at the NIH—why, whence, what, whither. ---Whither: The NIH Bioinformatics and Computational Biology Roadmap:

- Was submitted to NIH Director Dr. Elias Zerhouni on May 28, 2003
- Is the outline of an 8-10 year plan to create an excellent biomedical computing environment for the nation.
- Has as its explicit most ambitious goal “Deploy a rigorous biomedical computing environment to analyze, model, understand, and predict dynamic and complex biomedical systems across scales and to integrate data and knowledge at all levels of organization.

1-3 year roadmap goals: relatively low difficulty

- 1. Develop vocabularies, ontologies, and data schema for defined domains and develop prototype databases based on those vocabularies, ontologies, and data schema
- 2. Require that NIH-supported software development be open source.
- 3. Require that data generated in NIH-supported projects be shared in a timely way.
- 4. Create a high-prestige grant award to encourage research in biomedical computing.
- 5. Provide support for innovative curriculum development in biomedical computing
- 6. Support workshops to test different methods or algorithms to analyze the same data or solve the same problem.
- 7. Identify existing best practice/gold standard bioinformatics and computational biology products and projects that should be sustained and enhanced.
- 8. Enhance training opportunities in bioinformatics and biomedical computing.

1-3 year roadmap goals: moderate difficulty

- 1. Support Center infrastructure grants that include key building blocks of the ultimate biomedical computing environment, such as: integration of data and models across domains, scalability, algorithm development and enhancement, incorporation of best software engineering practices, usability for biology researchers and educators, and integration of data, simulations, and validation.
- 2. Develop biomedical computing as a discipline at academic institutions.
- 3. Develop methods by which NIH sets priorities and funding options for supporting and maintaining databases.
- 4. Develop a prototype high-throughput global search and analysis system that integrates genomic and other biomedical databases.

4-7 year roadmap goals: relatively low difficulty

- 1. Supplement existing national or regional high-performance computing facilities to enable biomedical researchers to make optimal use of them.
- 2. Develop and make accessible databases based on domain-specific vocabularies, ontologies, and data schema.
- 3. Harden, build user interfaces for, and deploy on the national grid, high-throughput global search and analysis systems integrating genomic and other biomedical databases.

4-7 year roadmap goals: moderate difficulty

- 1. Develop robust computational tools and methods for interoperation between biomedical databases and tools across platforms and for collection, modeling, and analyzing of data, and for distributing models, data, and other information.
- 2. Rebuild languages and representations (such as Systems Biology Markup Language) for higher level function.

4-7 year roadmap goals: high difficulty

- 1. Ensure productive use of GRID computing through participation of biologists to shape the development of the GRID.
- 2. Develop user-friendly software for biologists to benefit from appropriate applications that utilize the GRID.
- 3. Integrate key building blocks into a framework for the ultimate biomedical computing environment.

8-10 year roadmap goals: relatively low difficulty

- 1. Employ the skills of a new generation of multi-disciplinary biomedical computing scientists

8-10 year roadmap goals: moderate difficulty

- Produce and disseminate professional-grade, state-of-the art, interoperable informatics and computational tools to biomedical communities. As a corollary, provide extensive training and feedback opportunities in the use of the tools to the members of those communities.

8-10 year roadmap goals: high difficulty

- Deploy a rigorous biomedical computing environment to analyze, model, understand, and predict dynamic and complex biomedical systems across scales and to integrate data and knowledge at all levels of organization.

Initial Steps on the Roadmap Plan I

- We have released a funding announcement, and received proposals, for the creation of four NIH National Centers for Biomedical Computing. Each Center is to serve as the node of activity for developing, curating, disseminating, and providing relevant training for, computational tools and user environments in an area of biomedical computing. We hope ultimately to establish eight centers.

Initial Steps on the Roadmap Plan II

- We are preparing a funding announcement for investigator-initiated grants to collaborate with the National Centers. **Instead of having big science and small science compete with each other, we will create an environment in which they will work hand in hand for the benefit of *all* science.**

Initial Steps on the Roadmap Plan III

- We are preparing a funding announcement for work on creating and disseminating curricular materials that will embed the learning and use of quantitative tools in undergraduate biology education for future biomedical researchers. **We are committed to pressing a reform movement in undergraduate biology education to ensure an adequate number of quantitatively trained and able biomedical researchers in the future.**

Initial Steps on the Roadmap Plan IV

- We are in the initial stages of establishing a formal assessment and evaluation process. A possible form is that an external group of scientists will establish criteria by which to evaluate the program, and a professional survey research group will work with the scientists to implement the ongoing assessment and evaluation plan, so that prompt and appropriate mid-course corrections and tuning will take place.

Key Features of the NIH Bioinformatics and Computational Biology Roadmap Process

- Every component goes through NIH peer review system.
- Larger components are by cooperative agreement rather than grant, with active continued participation by NIH program staff.
- There is complete transparency about the rules and the process (except for the confidentiality necessary for peer review).
- Assessment and Evaluation are built in from the start.
- Program, review, and evaluation are independent of each other.

Possible areas of productive interaction with other agencies

a. with DOE on microbial science and nanoscience and biotechnology
b. with DARPA on microbial science and on nanoscience and biotechnology
c. with USDA on nutrition and agricultural science
d. with NIST on data and software standards and on nanoscience
e. with NSF on biology at all levels, on integrating biomedical computational science with the cyberinfrastructure initiative, on fostering interdisciplinary collaborative science, on nanoscience, and on biology education
f. with NASA and NOAA on environmental issues related to health

A Possible Division of Responsibility for High-performance Computing in Biology

- For NIH, domain-specific software development to efficiently and maximally utilize powerful computers to speed the pace and expand the scope of biomedical research.
- For NSF, building the national computing grid.
- For NSF and DOE, providing the hardware and middleware for “heroic” computing.

Some Scientific Challenges Significant for the Mission of the NIH that Potentially Involve High-Performance Computing

- Near-term: Being able to do accurate in silico screening of lead compounds for drugs.
- Near-term: Being able to do high quality homology modeling of proteins.
- Intermediate term: Being able to do accurate computer-aided design of self-assembled nanodevices.
- Intermediate term: Being able to do reliable computer-aided design of biomaterials.
- Longer term: Accurate dynamical modeling at higher levels of biological organization. (Cells, tissues, populations)