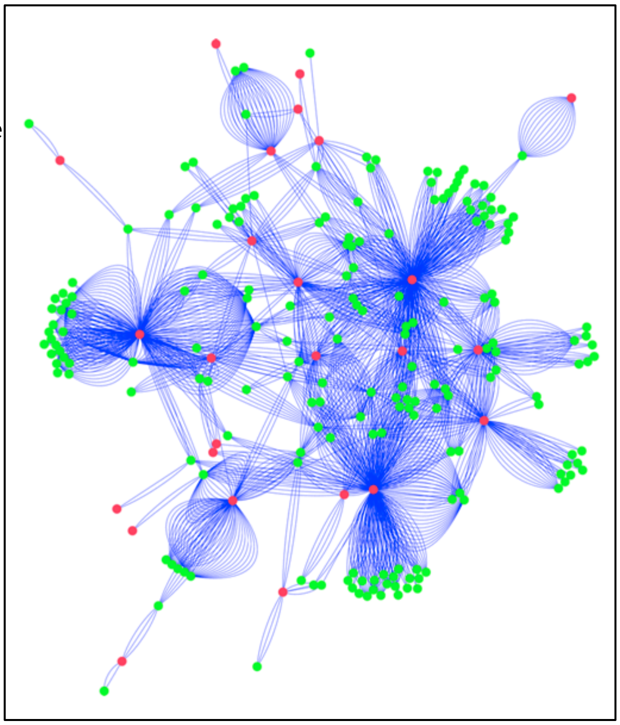
Chain Reactions: Information Technology and Biomedical Discovery

Russ Altman
Stanford University



The field of biomedical informatics (BMI) has emerged in the last 20 years as a direct result of the dual revolutions in information technology and biological measurement. BMI is focused on the creation of novel computational methods to represent, store, annotate, search, and analyze biomedical data — everything from molecular data to human medical records and population-based data. BMI has depended fundamentally on the Networking and Information Technology Research and Development (NITRD) investment.

The NIH has supported the growth of biomedical informatics (particularly the individual institutes NLM, NIGMS, NHGRI, NCRR and NCI, among others). The informatics problems in biomedicine are myriad and incredibly challenging, attracting young scientists with dual facility in informatics and biology. In 1994, Stanford first offered an introductory course in biomedical informatics, and each year a mixture of computer science, biology, medicine, and other students gain an entrée into the field. (In 1995, a Stanford student named Sergei Brin sat in on the course, wondering if biology was a good area for his interests; Brin went on to found Google). Further demonstrating the interdisciplinary nature of BMI, DOE funded this area very early on as it took a leadership role in the Human Genome Project. NSF has supported basic algorithmic research and applications in genetics, genomics, and systems biology. And most recently, many pure IT companies, including Microsoft, IBM, HP, Oracle and others, have become engaged in this space.

We are now faced with a curious and unusual situation: our ability to generate DNA sequence data for the last five years has outpaced Moore's law. This has had real consequences — all of the weaknesses in our algorithms are exposed as our ability to analyze and store the data deteriorates, and requires major innovation. But it also creates great opportunities. The key networking and information technologies used today include:

I. The entire biomedical published literature (20×10^6 article abstracts) is available in a single resource (Pubmed) that serves the entire world using high speed networks, and optimized search. It hosts almost 100,000,000 searches per month.

II. Very large stores of DNA data, three-dimensional biological macromolecular data, cellular gene expression data, and small molecule databases are available for search and analysis freely.

III. High precision physical simulation of biological molecules is enabled by supercomputers, clusters, and now graphical processing units (GPUs) that allow these numerically intensive simulations to achieve time scales relevant to biological phenomena.

IV. Biomedical imaging data are generated and transmitted around and between hospitals every day, with increasing resolution (both time and space) and requiring substantial compute and networking capabilities.

V. New models for biomedical research are emerging, including crowd-sourced evaluation of drug side effects, drug interactions, disease symptoms, and other health-related information.

The trends for the future are simple: (1) basic molecular and genetic measurements will continue to become cheaper and more accurate. At the same time, (2) our ability to instrument the health care delivery system (collecting clinical data such as clinical notes, lab tests, diagnostic imaging) will also grow. The chief challenge will be to capture, store and analyze the emerging information so that we can discover new ways to improve health.