BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Bonnie Berger

eRA COMMONS USER NAME (credential, e.g., agency login): BABERGER

POSITION TITLE: Professor of Mathematics and Electrical Engineering & Computer Science

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Completion Date MM/YYYY	FIELD OF STUDY
Brandeis University, Waltham, MA	AB	06/1983	Computer Science
Massachusetts Institute of Technology	SM	01/1986	Computer Science
Massachusetts Institute of Technology	Ph.D.	06/1990	Computer Science
Massachusetts Institute of Technology	Postdoc	06/1992	Applied Mathematics

A. Personal Statement

Many of the advances in modern biology revolve around recent advances in automated data collection and the subsequent large data sets drawn from them. I am considered a pioneer in the area of bringing computer algorithms to the study of biological data, and I am one of the founders of the community that has grown up in this area over the last 20 years. Largely via algorithmic insights, I have contributed to many areas of computational biology and biomedicine, as evidenced by the thousands of citations to my papers and widely used software. My group works on a diverse set of challenges, including Large-scale Genomics, Network Inference, Population Genomics, Structural Bioinformatics, and Comparative Genomics. Moreover, we collaborate closely with biologists, MDs, and software engineers, implementing these new techniques in order to design experiments to maximally leverage the power of computation for biological explorations. Over the past five years I have been particularly active in the analysis of large and complex biological data sets; for example, my lab has played integral roles in ModEncode (non-coding RNA annotation), MPEG (biological data compression standard), and the Broad Institute's sequence analysis efforts.

I have trained more than 100 students and postdoctoral fellows, with many of them holding top academic positions. Among my PhD students are: Drs. Serafim Batzoglou – Stanford University tenured; Phil Bradley – Fred Hutchinson and U. of Washington tenured; Manolis Kellis – MIT tenured; Lior Pachter – UC Berkeley and Caltech tenured; Nathan Palmer – Harvard Biomedical Informatics group leader; Mona Singh – Princeton University tenured; and Russell Schwartz – CMU tenured. Some of my more recent PhD students are: Michael Baym – Harvard Medical School Assistant Prof; Leonid Chindelevitch – Simon Fraser University Assistant Prof; Po-Ru Loh – Harvard School of Public Health population genetics researcher; and Michael Schnall-Levin – 10X Genomics VP of Computational Biology & Applications. Postdocs include David Brudno – U. of Toronto tenured; Amy Keating – MIT tenured; Jian Peng – U. of Illinois UC Assistant Prof; Dana Ron – Tel Aviv University tenured; Jerome Waldishpuhl – McGill tenured; and Jinbo Xu TTI U. of Chicago tenured.

In addition to research and mentorship, I continually participate in community service, including currently playing an active role as Vice President of ISCB, Head of the RECOMB Steering Committee, and member of the NIH NIGMS Advisory Council. I have also served as both Proceedings and Conference Chairs for the two top conferences in the field—RECOMB and ISMB. In just the last year, I have given keynote addresses and distinguished lectures at RECOMB, ACM-BCB, RECOMB-RSG; RECOMB-BE (I contributed a MOOC), EPFL, UC San Diego, and Stanford Biomedical Data Science.

B. Positions and Honors

Positions and Employment

1990-1992 NSF Mathematical Sciences Postdoctoral Research Fellowship.
1992-1993 Radcliffe Bunting Institute, Science Scholar.
1992- Member of Computer Science & Artificial Intelligence Laboratory (CSAIL), MIT.
1992-1997 Assistant Professor of Applied Mathematics, MIT.
1997-1999 Associate Professor of Applied Mathematics, MIT.
1999-2002 Associate Professor of Applied Mathematics, tenured, MIT.
2002- Professor of Applied Mathematics, MIT.
2004-2012 Affiliated Faculty, Harvard-MIT Health Sciences and Technology (HST).
2004- Affiliated Faculty, Computational and Systems Biology (CSBi) at MIT.
2008- Beth Israel Deaconess Board of Overseers and Medical Advisory Committee.
2010- Joint Appointment, Dept. of Electrical Engineering and Computer Science, MIT.
2010- Associate Member, Broad Institute of MIT and Harvard.
2012- Affiliated Faculty, Harvard Medical School
2014- Faculty Member, Harvard-MIT Health Science and Technology.
2015- Member, Center for Microbiome Informatics and Therapeutics.
2016-2021 Simons Professor of Mathematics, MIT.

Other Experience and Professional Memberships

1995 Organizer for DIMACS Workshop: Sequence-based methods for protein folding.
1996-2003 BOD for Program in Mathematics and Molecular Biology (PMMB).
1998 NSF selection panel for the Protein Data Bank (PDB).
2001- Creator and organizer of MIT Math/CSAIL Bioinformatics Seminar.
2002- HST Graduate; Bioinformatics & Integrative Genomics; and Curriculum Committees.
2003-2006 ACM Nominating Committee.
2003-2014 NIH Scientific Review Group: Comparative modelling, BCMB & BDMA, ad-hoc member.
2004- Brandeis University Science Advisory Council.
2006-2012 NIH NCBI Board of Scientific Counselors, 3 time ad-hoc member.
2008-2014 Beth Israel Deaconess Board of Overseers and Medical Advisory Committee.
2009-2014 NIH NIGMS Advisory Council, 3 time ad-hoc member.
2009-2018 RECOMB Steering Committee, Chair (2015-2018) and member.
2010 RECOMB 2010 Program Chair; ISMB 2010 Area Chair.
2010 PloS Computational Biology Guest Editor.
2011-2017 Sloan Fellowship Selection Committee, Computational & Evolutionary Molecular Biology.
2012 ISMB 2012 Proceedings Chair & Steering Committee.
2013 ISMB 2013 Conference Chair & Steering Committee.
2013-2016 ISMB Area Chair; Data Theme Chair (2016).
2015-2018 FASEB Excellence in Science Award Committee.
2015- RECOMB Steering Committee, Chair.
2015-2016 NIH NIGMS Advisory Council.
2015-2018 ISCB Vice President, Member of Board of Directors, Awards Chair, Fellows Chair, Senior Member.
2016 Cold Spring Harbor Lab's Biological Data Sciences Program Committee (with 2 others).
2016 NIH BD2K Multi-Council Working Group, NIGMS representative.

Current Editorial Boards: JCB (Executive Editor), Genome Biology (2011-), Cell Systems (2015-), IEEE TCBB (2004-), SIAM J on Disc Math (2002-), Bioinformatics (2015-), and Annual Reviews for Biomedical Data Science (2016-).

Selected Awards and Honors

1990 Ph.D. thesis won MIT George M. Sprowles Prize for best research in computer science.
1995-1998 NSF Career Award.
1999 Biophysical Society's Dayhoff Award for research (1 award per year).
1999 Technology Review's Inaugural TR100 Award for 100 top young innovators for the 21st century.
2004 Elected as a Fellow of the Association for Computing Machinery.

2010 RECOMB Test of Time Award.
2012 NIH Margaret Pittman Lecture for Outstanding Scientific Achievement & Lectureship.
2012 Elected as a Fellow of the International Society for Computational Biology.
2013 Elected to the American Academy of Arts and Sciences.
2013 Brandeis University Alumni Achievement Award.
2015 École Polytechnique Fédérale de Lausanne (EPFL) Honorary Doctorate.
2016 Elected to the American Institute for Medical and Biological Engineering (AIMBE).

C. Contributions to Science (* for corresponding author or † for my student is 1st author)

1. Compressive algorithms that scale. The last two decades have seen an exponential increase in genomic and biomedical data, which will soon outstrip advances in computing power. Extracting new science from these massive datasets will require not only faster computers; it will require algorithms that scale sublinearly in the size of the datasets. With my students, I introduced 'compressive genomics', a novel class of algorithms that take advantage of redundancy in biological data to compress the data in such a way as to operate directly on the compressed data, enabling algorithms that scale sublinearly in the size of the data set. Very recently, we formalized and generalized this framework to develop compressive algorithms that scale with the entropy and fractal dimension of the dataset by taking advantage of the unique structure of biological data, thus enabling sublinear time and space algorithms. These algorithms can be used to address challenges in large-scale genomics, protein structure search and chemogenomics. There has been keen interest in this work by the computational biology research community (International Society of Computational Biology, letter of support): I have given multiple keynotes and a MOOC on this work; several associated Proceedings and Highlight papers have appeared in ISMB and RECOMB conferences, including an entire workshop focused on this area at ISMB 2016; and, importantly, the work is an invited contribution to *Nature Reviews Genetics* (2013), *J. of the ACM* (2016), and "Voices of Biotech" in *Nature Biotech*'s May 2016 20th Anniversary Issue.

a. P-R. Loh, M. Baym, and B. Berger *. "Compressive Genomics." *Nature Biotech* **30** (2012): 927-930. Most downloaded *Nat Biotech*, July 2012.
b. Y.W. Yu, D. Yorukoglu, J. Peng and B. Berger *. "Quality Score Compression Improves Downstream Genotyping Accuracy." *Nature Biotech* **33** (2015): 240-3.
c. Y. W. Yu, N. M. Daniels, D. C. Danko and B. Berger *. "Entropy-Scaling Search of Massive Biological Data." *Cell Systems* **1**, 2 (Fall, 2015):130–140. Cover image; focus article of Journal, Commentary, and Perspectives.
d. Deniz Yorukoglu†, Yun William Yu, Jian Peng, and Bonnie Berger*, **"**Compressive Mapping for Next-Generation Sequencing." *Nature Biotech* **4** (2016): 374-376.

2. Biological network analysis and function annotation. I pioneered the highly active field of global network alignment in the last ten years. I introduced global biological network alignment (over 1000 citations)—a critical step for transferring functional knowledge across different species— and set the standard for its use in functional orthology prediction, primarily through our Isorank suite of programs; these are based on a novel Eigenvalue formulation on the product graph of networks (*US Patent 8000262 B2,* 2011). Our Isorank algorithm and Isobase tools have been incorporated into many other webservers including Norbert Perrimon lab's DiOPT. I have further developed approaches to integrate RNAi data with protein interaction data in order to uncover signaling relations. Recently, we have newly characterized prevalence of microRNAs in coding regions, which was later experimentally verified. In the last five years, in experimental collaborations with Drs. Norbert Perrimon (HMS, HHMI) and Susan Lindquist (WI, HHMI), we have developed methods to integrate RNAi [Friedman et al. Sci Signaling, 2011], masspec and lumier data to shed light on the genetics of disease, including a new paper accepted to *Cell Systems* where I am co-corresponding author with Lindquist.

e. P. Uetz, Y. Dong †, C. Zeretzke, C. Atzler, A. Baiker, B. Berger †, S. Rajagopala, M. Roupelieva, D. Rose, E. Fossum and J. Haas *. "Herpesviral Protein Networks and their Interaction with the Human Proteome." *Science* **311**, 5758 (2006): 239-242. 340 citations
f. R. Singh, J. Xu, and B. Berger *. "Global Alignment of Multiple Protein Interaction Networks with Application to Functional Ortholog Detection." *Proc Nat Acad Sci USA* **105**, 35 (2008): 12763-68. Over 800 combined citations (with RECOMB, Bioinformatics 2009 and Isobase database).

g. M.Schnall-Levin †, O. Rissland, W. Johnston, N. Perrimon, D. Bartel * and B. Berger *. "Unusually Effective MicroRNA Targeting within Repeat-Rich Coding Regions of Mammalian mRNAs." *Genome Research* **21,** 9 (2011); including full cover and *Nature Reviews Genetics* highlights. 167 citations with *PNAS* (2010)

h. N. Sahni, S. Yi, M. Taipale, et al. "Widespread Specific Macromolecular Interaction Perturbations in Human Genetic Disorders." *Cell* **161,** 3 (2015): 647-660. Also *Cell* (2014). My group performed computational analysis of Massspec and Lumier HTP protein-protein interaction data.

3. Structural bioinformatics. My earlier work introduced pairwise probabilistic modeling to protein fold recognition and was implemented in our programs (e.g., Paircoil, Multicoil, Learncoil), which have thousands of citations and have been used to make important biological discoveries, including implicating coiled-coils in the membrane fusion mechanism for large classes of viruses and informing development of inhibitory drugs for HIV with Dr. Peter S. Kim, Stanford. I also solved a difficult theoretical problem central to the biophysics and protein folding communities (i.e., HP-lattice folding is NP-complete, 475 citations). Moreover, I showed that the self-assembly of viral shells—though seemingly a complex procedure—can be explained purely by local rules (400 citations); this work led to widespread application of similar approaches in biophysics and materials science engineering. In the last decade, we were the first in early 2006 to incorporate protein structure data to predict protein interactions in the Struct2Net webserver; we introduced TreePack for fast and accurate side chain packing (JACM, 2006), which was incorporated into the state-of-art SCWRL program. We have contributed the primary RNA structure analysis (with my postdoc S. Will) to modENCODE, an international consortium whose goal is to provide the biological research community with a comprehensive encyclopedia of genomic functional elements in the model organisms C. elegans and D. melanogaster. We introduced Matt, a structure alignment program, which newly allowed full backbone flexibility in the alignment phase and has been shown to outperform other aligners in independent tests.

i. M. Menke †, B. Berger * and L. Cowen *, "Matt: Flexibility Aids Protein Multiple Structure Alignment." *PLoS Computational Biology* **4**, 1 (2008). Also in ISMB 2008 Highlights track. 156 citations

j. R. Singh, D. Park, J. Xu, R. Hosur and B. Berger *. "Struct2Net: A Web Service to Predict Protein-Protein Interactions Using a Structure-based Approach." *Nucleic Acids Research* **38** suppl 2 (2010): w508-w515. 80 combined citations with PSB (2006) 11:403-414.

k. The ModEncode Consortium, "Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE." *Science* **330**, 6012 (2010): 1787-1797. 600 citations

l. S. Will, M. Yu, and B. Berger *. "Structure-based Whole Genome Realignment Reveals Many Novel Non-coding RNAs." *Genome Research* **23,** 5 (2013). Also RECOMB 2012.

4. Population genomics. I founded and developed conservation-based methods for comparative genomics, together with my PhD students (Serafim Batzoglou, Lior Pachter, and Manolis Kellis.) With Dr. Eric Lander, we performed the first whole-genome alignments for human and mouse as well as comparisons to detect exonic regions; we also performed the first comparisons of yeast genomes to identify genes and regulatory regions (nearly 2000 combined citations). My students and I have more recently spearheaded exciting work in population genomics. In response to a challenge from Nick Patterson, David Reich, and later Alkes Price, we capitalized on our observations about the structure of genome-wide association studies in GWAS and population data, as well as statistical and algorithmic advances, to improve the power and speed of such studies. As an example, we analyzed genome-wide data from 56 populations and showed that all sampled Southeast Asian Austronesian groups harbor ancestry that is more closely related to aboriginal Taiwanese than to any present-day mainland population, thereby resolving a controversial question. The methods we developed have also newly enabled the inference of population flow with admixture and allowed us to uncover population flow in Roma, Indian and African populations. Importantly, we have developed the first method for efficiently handling population stratification, which had been shown to be essential in GWAS studies. In addition, we have made significant progress on haplotype phasing, allowing haplotype reconstruction of a single sequenced individual using NGS data and applying these methods to Autism datasets (last two RECOMBs.) Most recently, we have turned to addressing privacy concerns with large cohort studies.

a. P. R. Loh †, M. Lipson †, N. Patterson, P. Moorjani, J. K. Pickrell, D. Reich *, and B. Berger *. "Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium" *Genetics* **193**, 4 (2013): 1233-1254; including full cover. 67 citations

b. I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, B. Berger, et al. "Ancient Human Genomes Suggest Three Ancestral Populations for Present-Day Europeans." *Nature* **513** (2014): 409-413. 135 citations.
c. E. Berger, D. Yorukoglu, J. Peng, and B. Berger *. "HapTree: A Novel Bayesian Framework for Single Individual Polyplotyping using NGS Data." *PLoS Computational Biology* **10**, 3 (2014): e1003502. Also RECOMB 2014 & 2015.
d. P-R. Loh †, G.J. Tucker †, B. Berger, N. Patterson, and A. Price. "Efficient Bayesian Mixed-model Analysis Increases Association Power in Large Cohorts." *Nature Genetics* **47** (2015): 284–290. 28 citations

5. Data privacy. I first studied cryptography as a PhD student working on randomized algorithms with Dr. Silvio Micali, who received the 2012 Turing Award for cryptography. I have since devised novel methods to anonymize location information used in epidemiological studies, achieving higher levels of accuracy and privacy than standard, insecure zip code based techniques. Though my focus in recent years has been in biological data science, I have recently come full circle, addressing privacy issues arising in public health and genomics. I have newly demonstrated that genomic structure can be used to better protect genomic privacy.

e. S.C. Wieland †, C. Cassa, K. Mandl * and B. Berger *. "Revealing the Spatial Distribution of a Disease While Preserving Privacy." *Proc Nat Acad Sci USA* **105**, 46 (2008): 17608-17613. 40 citations.
f. S. Simmons † and B. Berger *. "One Size Doesn't Fit All: Measuring Individual Privacy in Aggregate Genomic Data." 2015 *IEEE Security and Privacy Workshops [SPW]* (2015): 41-49. ISBN: 978-1-4799-9933-0S.
g. Simmons † and B. Berger *. "Realizing Privacy Preserving Genome-wide Association Studies." *Bioinformatics* **32**, 9 (2016): 1293-1300.
h. S. Simmons †, C. Sahinalp and B. Berger *. "Enabling Privacy-Preserving GWAS in Heterogeneous Human Populations." *Cell Systems* **3**, 1 (2016): 54–61. Also appeared in RECOMB 2016 (Springer).

Complete List of Published Work in My Bibliography:
http://www.ncbi.nlm.nih.gov/sites/myncbi/bonnie.berger.1/bibliograpahy/41140693/public/?sort=date&direction=ascending

D. Research Support

Ongoing Research Support

NIH 1-R01-GM108348 Berger (PI) September 5, 2013 – May 31, 2016
Compressive Genomics for Large Scale Omics Datasets: Algorithms, Applications, & Tools
$217,000/yr Annual Direct Costs to Berger
The major goal of this project is to develop methods for "compressive genomics," which allow efficient analysis of compressed sequencing and omics data on thousands of individuals and terabyte-sized datasets; this will better inform clinicians through more-scalable downstream genotyping, mapping, and searching of data. Renewal application received a percentile of 3.0.

NIH 1-R01-GM081871 Berger (PI) September 15, 2012 – May 15, 2017 NCE
Structure-Based Prediction of the Interactome
$225,000/yr Annual Direct Costs to Berger
This research develops improved algorithms for threading protein complexes and investigates whether such data enhances systems-level data in genome-scale protein-protein and protein-RNA interaction prediction. Project in No Cost Extension.

MIT internal Berger (PI) Center for Microbiome Informatics and Therapeutics Pilot Grant
January 1, 2016 – December 31, 2016
Compressive Metagenomics
$50,000 for 1-year Direct Cost to Berger
The goal of this 1-year small pilot grant is to initiate a formal collaboration with the Microbiome Center to lay the foundations for a software engineering effort that will facilitate our collaborative effort; in particular, this will address only the storage issue for metagenomic read data, an urgent need, by producing modified, more compressible FASTQ files using our Quartz tool to compress quality scores and EMBL's CRAM to compress reads.