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 advances in modern biology revolve around automated data collection and the large resulting data sets. I am considered a pioneer in the area of bringing computer algorithms to the study of biological data, and a founder in this community that I have witnessed grow so profoundly over the last 20 years. I have made major contributions to many areas of computational biology and biomedicine, largely, though not exclusively through algorithmic insights, as demonstrated by ten thousand citations to my scientific papers and widely-used software. My research group works on diverse challenges, including Structural bioinformatics, High-throughput Technology Analysis and Design, Network Inference, Data Privacy, and Computational Genomics. 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 exploration. Over the past five years I have been particularly active analyzing large and complex biological data sets; for example, my lab has played integral roles in modENCODE (non-coding RNA analysis), MPEG (biological data compression standard), and the Broad Institute's sequence analysis efforts.

I have trained more than 100 students and postdoctoral fellows, many of whom now hold 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. More recent PhD students include: Michael Baym – Stanford Assistant Prof; Alan Bryan – UAB Birmingham 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; Noah Daniels – URI Assistant Prof; Amy Keating – MIT tenured; Jian Peng – U. of Illinois UC Assistant Prof; Dana Ron – TAU tenured; Jerome Waldishpuhl – McGill tenured; and Jinbo Xu TTI U. of Chicago tenured.

I continually and actively engage in community service, including recently as Vice President of ISCB, Head of the RECOMB Steering Committee, and member of the NIH NIGMS Advisory Council. I have served as both Proceedings and Conference Chairs for the two top conferences in my field—RECOMB and ISMB. I am also proud to have headed a workshop at ISMB 2016 on Gender Equality and been ISCB Fellows Chair (2015-2017), focusing on minority inclusion. In just the last year, I have given keynote addresses and distinguished lectures at RECOMB, ACM-BCB, RECOMB-RSG; RECOMB-Bioinformatics Education (contributed a MOOC), EPFL, UC San Diego, Stanford Biomedical Data Science and a Gordon Research Conference.

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-	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.
2004-2014	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.
2010	RECOMB 2010 Program Chair.
2010-2016	ISMB Proceedings Chair (2012), Conference Chair (2013) & Steering Committee (2012-13); Area Chair (2010, 2012-16).
2011-2017	Sloan Fellowship Selection Committee, Computational & Evolutionary Molecular Biology.
2015-2018	FASEB Excellence in Science Award Committee.
2015-2018	RECOMB Steering Committee Chair, and member since 2009.
2015-2016	NIH NIGMS Advisory Council.
2015-2017	ISCB Vice President, Member of Board of Directors, Awards Chair, and Fellows Chair.
2016	ISMB 2016 Gender Equality Workshop Leader.
2016	Cold Spring Harbor Lab's Biological Data Sciences Program Organizer (with 2 others).

Selected Awards and Honors

1990	Ph.D. thesis won MIT George M. Sprowls 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. Computational genomics. The last two decades have seen exponential increases in genomic and genetic data that will soon outstrip advances in computing power. Extracting new science from these massive datasets will require not only faster computers, but algorithms that scale sublinearly to the size of the data. I therefore introduced 'compressive genomics', a novel class of algorithms able to take advantage of data redundancy to compress biological data. This compression permits operations and computations directly on the compressed data, thereby enabling algorithms that provably scale sublinearly with the size of the data. These algorithms can be used to address challenges in large-scale genomics, metagenomics and chemogenomics. There has been keen interest in this work by the computational biology research community, including an entire workshop focused on this area at ISMB 2016. Importantly, the work was an invited contribution to *Nature Reviews Genetics* (2013), "Voices of Biotech" in *Nature Biotech*'s May 2016 20th Anniversary Issue, and *Communications of the ACM* (Cover, 2016). In earlier work, I founded and developed conservation-based methods for comparative genomics, together with my PhD students (S. Batzoglou, L. Pachter, and M. Kellis.) Using these methods in collaboration 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 (P.R. Loh, M. Lipson, and G.J. Tucker) and I recently spearheaded exciting work in population genomics, which has resulted in high-profile papers in *Nature*, *Nature Genetics*, and *Genetics*.

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 (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. Network inference and functional annotation. I pioneered the highly active and rapidly evolving field of global network alignment. I introduced global biological network alignment (over 1000 citations) — a critical step to the transfer of functional knowledge across species — and set the standard for its use in functional orthology prediction, primarily through our Isorank suite of programs based on a novel Eigenvalue formulation of the product graph of networks (*US Patent 8000262 B2,* 2011). Our Isorank algorithm and Isobase tools have been incorporated into numerous external webservers including Norbert Perrimon lab's DiOPT. I have further developed, with the Perrimon lab, approaches to integrate RNAi data with protein interaction data to uncover signaling relations. Our characterization of prevalence of microRNAs suggested additional function of transcribed coding regions, a finding later experimentally verified. In preliminary studies, we have most recently designed a machine learning framework for incorporating heterogeneous sources of information from which networks with different connectivity patterns can be constructed, and successfully applied this framework to functional annotation problems (*RECOMB*/*Cell Systems* in press).

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 1000 combined citations (with RECOMB, Bioinformatics 2009, 2013, 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. Nearly 200 citations with *PNAS* (2010)

h. S. Wang, H. Cho †, C. Zhai, B. Berger † and J. Peng *. "Exploiting Ontology Graph for Predicting Sparsely Annotated Gene Function." ISMB/ECCB, *Bioinformatics* **31**, 12 (2015):i357-i364.

3. Structural bioinformatics. My earlier work introduced pairwise probabilistic modeling to protein fold recognition as implemented in our programs (e.g., Paircoil, Multicoil, Learncoil). These programs have thousands of citations and have resulted in important biological discoveries. I also solved a difficult theoretical problem central to the biophysics and protein folding communities (i.e., HP-lattice folding is NP-complete, 500 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 applications to biophysics and materials science engineering. In 2006, I was the first to incorporate protein structure data to predict protein interactions in the Struct2Net webserver. I, with postdoc Jinbo Xu, introduced TreePack for fast and accurate side chain packing (JACM, 2006), later incorporated into the state-of-art SCWRL program. I introduced Matt, a structure alignment program, newly allowing full backbone flexibility in the alignment phase that has been shown in independent tests to outperform other aligners in independent tests (*PLoS Computational Biology*, 2008, 160 citations). I 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 model organisms. I recently developed algorithms for structure-based prediction on both protein-protein and RNA-protein interactions whose predictions, tested experimentally, uncovered cancer-related interfaces and novel non-coding RNAs in fly, human and mouse. These results are described in a manuscript, "Structure-based Prediction of RNA–protein Interactions on a Genome-wide Scale," submitted to *Nature Biotechnology*.

i. 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. 100 combined citations with PSB (2006) 11:403-414.

j. The ModEncode Consortium, "Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE." *Science* **330**, 6012 (2010): 1787-1797. 600 citations. My group performed the structure-based non-coding RNA analysis.

k. R. Hosur, J. Peng**,** A. Vinayagam**,** U. Stelzl**,** J. Xu**,** N. Perrimon**,** J. Bienkowska* and B. Berger *. "A Computational Framework for Boosting Confidence in High-throughput Protein-Protein Interaction Datasets", *Genome Biology* **13** (2012): R76, PMID: 22937800. We integrated novel structure-based predictions with HTP data, uncovering cancer interactome.

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. High-throughput technology analysis and design. Long-running experimental collaborations with Drs. Norbert Perrimon (HMS, HHMI), David Bartel (WI, HHMI) and Susan Lindquist (WI, HHMI) have served to shed light on the genetics of disease though the development of methods to analyze RNAi [Friedman et al. Sci Signaling, 2011], Mass Spec, Lumier, and CLiP-seq data. My group is responsible for the computational analyses in these joint collaborations. Moreover, we have designed experiments for measuring RNA-protein binding (including Orenstein & Berger, WABI, 2015), and developed the first algorithm to infer RNA structure binding preference from experimental data.

m. M. Taipale, G. Tucker †, J. Peng †, I. Krykbaeva, Z. Y. Lin, B. Larson, H. Choi, B. Berger †, A. C. Gingras * and S. Lindquist *. "A Quantitative Chaperone Interaction Network Reveals the Architecture of Cellular Protein Homeostasis Pathways." *Cell* **158**, 2 (2014): 434-448.

n. N. Sahni, S. Yi, M. Taipale, et al. "Widespread Specific Macromolecular Interaction Perturbations in Human Genetic Disorders." *Cell* **161,** 3 (2015): 647-660.

o. Y. Orenstein, Y. Wang and B. Berger *. "RCK: Accurate and Efficient Inference of Sequence and Structure-based Protein-RNA Binding Models from RNAcompete Data." *Bioinformatics/ISMB* **32,** 12 (2016): i351-i359.

p. V. Khurana, J Peng †, CY Chung, Auluck PK, Tardiff DF, Fanning S, Bartels T, Koeva M, Eichhorn SW, H. Benyamini, Y. Lou, A. Nutter-Upham, V. Baru, Y. Freyzon, N. Tuncbag, M. Costanzo, B.J. San Luis, D.C. Schöndorf, M.I. Barrasa, S. Ehsani, N. Sanjana, Q. Zhong, T. Gasser, D.P. Bartel, M. Vidal, M. Deleidi, C.

Boone, E. Fraenkel *, B. Berger *, and S. Lindquist *. "Genome-scale Networks Link Diverse Molecular Pathways and Neurodegenerative Disease Genes to Alpha-synuclein." *Cell Systems*, in press.

5. Data privacy. I began my study of cryptography as a PhD student working on randomized algorithms with Dr. Silvio Micali, recipient of the 2012 Turing Award for cryptography. I have since devised methods to anonymize location information used in epidemiological studies. While my focus in recent years has been biological data science, I have come full circle in addressing privacy issues arising in sharing this data. I introduced differential privacy—slightly perturbing the data ensures privacy and with minimal loss in accuracy— that takes into account population stratification in GWAS analysis and developed homomorphic encryption methods for secure genome analysis. Our differential privacy work was recently highlighted in dozens of news stories, including *Nature* news.

q. 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.
r. 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.
s. Simmons † and B. Berger *. "Realizing Privacy Preserving Genome-wide Association Studies." *Bioinformatics* **32**, 9 (2016): 1293-1300.
t. S. Simmons †, C. Sahinalp and B. Berger *. "Enabling Privacy-Preserving GWAS in Heterogeneous Human Populations." *Cell Systems* **3**, 1 (2016): 54–61. Also RECOMB 2016; *Nature news* (Aug. 15, 2016).

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) Sept. 5, 2013 – May 31, 2020
Compressive Genomics for Large Scale Omics Datasets: Algorithms, Applications, & Tools
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) April 1, 2008 – May 15, 2017 NCE
Structure-Based Prediction of the Interactome
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.

NIH NHGRI 5T32HG004947-05 Berger (interim co-Director)
Training in Computational Genetics
The mission of our training program is to produce a new breed of interdisciplinary scientist who can create fundamentally new computational and mathematical approaches that enable significant forward progress on biological and health related problems using computational approaches to genetics.

MIT internal Berger (PI) Center for Microbiome Informatics and Therapeutics Pilot Grant
January 1, 2016 – December 31, 2016
Compressive Metagenomics
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.