Indicators of Successful Career Transitions from Physical Sciences and Engineering to Biomedical Research

Michael A. Beer*

1Department of Biomedical Engineering, and McKusick-Nathans Department of Genetic Medicine, Johns Hopkins University School of Medicine, 733 N. Broadway, Baltimore, MD 21205, USA

1. INTRODUCTION

Quantitative and computational biomedical research continues to evolve rapidly, and there remains a significant need for rigorous modeling in the tradition of the physical sciences. The fields of computational genomics and interdisciplinary biology are addressing fundamental scientific questions in cellular and developmental biology, evolution, and genetics. The robust and continuing development of these fields requires additional talented researchers trained in physics, mathematics, statistics, and computer science. Previous articles have addressed many factors contributing to successful biomedical research in general [1]. However, this essay focuses specifically on encouraging career transitions into biomedical research and will attempt to point out examples demonstrating areas of greatest need and likelihood of success. Although there is a long tradition of physicists and chemists making seminal contributions to the life sciences (Schrödinger, Delbrück, Crick, etc.), earlier reviews on alternative career paths for physicists tend not to emphasize the very real possibility of transitioning to biomedical research [2]. In this essay, I will describe specific elements of training and personal interest that anticipate a successful transition from physics or engineering into computational biology and genomics, and other areas of quantitative biology. The aim is to reduce some of the uncertainty associated with entry into biomedical fields and to encourage talented researchers to consider such a move.

These comments are largely informed by my own personal experience. I transitioned from a Ph.D. in theoretical plasma physics to computational biology and genomics just before the completion of the human genome sequence [3, 4]. While deciding on a research focus for graduate school, I was interested in both biophysics and plasma physics and fusion science, and I was drawn to the fact that both fields had a significant potential positive societal impact. In the end, I chose computational plasma physics, as, at the time, biophysics was much less quantitative. Theoretical physics was the field that interested me most and choosing it served me well. While day-to-day research often involves sitting in front of a computer for long periods of time, solving difficult problems, working on the type of problem that you intrinsically find most interesting can turn this potentially tedious activity into an exciting intellectual endeavor. Although the specific problems you study may change over time, it is most important to develop and hone a core set of analytical and problem-solving tools that you enjoy applying. Although I eventually left plasma physics, I do not regret my choice of training in plasma physics at all since my research interests continue to be very mathematical, and I learned how to work as a computational scientist from some of the best in the field. If I had entered biophysics earlier, my research would likely have settled on less computational problems. My doctoral and postdoctoral training in computational plasma physics in the area of controlled fusion energy research was excellent preparation in almost all respects for my eventual transition to biomedical research. Based on this personal experience and observations of my colleagues, I feel that the types of quantitative training that can most readily be successfully applied to research in quantitative biology share common characteristics that can be distilled into four major themes.

1.1. Strong Interplay Between Modeling and Experiments

Although biology has historically been descriptive, experiments are now largely driven by technological advances (e.g., DNA sequencing-based assays [5] and CRISPR [6]), and the field is entering an era where predictive and theoretical modeling will play an ever-increasing role. Training in fields of physics and engineering where experimental results and theoretical models are both evolving rapidly and informing each other will be needed. As an example, in the mid-and late 1990s, several major fusion experiments were performing the first experiments with a deuterium-tritium fuel mixture and generating record levels of fusion power [7, 8]. These experiments were also measuring transport and turbulent fluctuations with unprecedented accuracy, and theories and sophisticated nonlinear computer simulations of these fluctuations were being quantitatively compared. Initially, the theories contradicted experiments in important ways, but as they increased in complexity and evolved, they eventually began to agree with many important aspects of the observations for the first time [9-11]. The comparisons and agreement between simulations and experiments [9, 10, 12] dramatically increased confidence in projections of future fusion experiment performance. Although these new models ultimately did not yield optimistic projections for fusion’s economic competitiveness,
there was a feeling that this hypothesis-driven modeling could now more accurately constrain future fusion reactor designs, and I learned first-hand the excitement and real-world impact that computational and theoretical modeling could have. The interplay between stellar evolution modelling and experimental astrophysics, cosmology, and even particle physics (e.g., neutrino oscillations [13]) is another excellent example. Coming from complex research areas where measurements and nonlinear computational models and simulations are being rigorously tested is an invaluable experience for future success in biomedical research. To be successful working in an environment where technology and theory are evolving rapidly, one must develop the ability to quickly and reliably identify which questions are worth asking, which datasets are worth modeling, and which datasets should be most heavily weighted when they do not all agree. This most useful skill is immediately transferrable to genomics and biomedical research.

In genomics, technologies are always evolving rapidly, and understanding how to internally check the consistency of a dataset is extremely useful. When I first started working in genomics, my first project involved analyzing microarray gene expression data. After performing PCA on the data matrix, I realized that the two “t=0” time points in two different time courses were actually just identical copies of one experiment. In addition to this type of statistical analysis, it is also useful to have a set of genes that can be used as a positive control to assess the quality of the data. On similar microarray datasets, I could test the overall quality of each dataset by comparing the correlation of a set of genes I knew should be tightly co-regulated across the conditions, like the ribosomal proteins or other relevant protein complexes. Working in any data-rich field leads to the development of similar sets of analysis tricks and tools, which will be valuable in biomedical research.

Another benefit of working in an active field is exposure to a wider range of the psychological dimensions of research. The proponents of each model and experimental system have a vested interest in their success, and these factors often influence how results are reported. Knowing how to detect these factors helps one efficiently navigate the exponentially growing literature and build a more accurate reflection of reality. The behavior of the underlying complex biological system clearly is not affected by such bias, however slight it may be. Carl Sagan called this “having a good baloney detector,” and it is also a completely transferrable skill across disciplines.

2. THE ART OF BUILDING AND ANALYZING APPROXIMATE PHYSICAL AND BIOPHYSICAL MODELS

Genomics is becoming increasingly quantitative and is looking more like physics every day. For example, we are beginning to model cell-type, or cell-state transitions in development and disease (such as cancer) with dynamical models [14-16], which admit bifurcations and oscillations quite similar to those in other highly nonlinear physical systems, for example, the suppression of turbulence by the generation of sheared flows or pressure gradients [17-19]. Predictive quantitative models of how these cellular systems transition between normal and disease states will necessarily be built upon higher-level abstractions of the biochemical interactions between protein complexes and gene regulatory states. While the underlying elements are often combinations of DNA or protein sequence motifs [20, 21], their interactions are best described by biophysical and/or polymer models [22, 23]. A thorough understanding of the statistical mechanics behind such biophysical modeling gives investigators a significant advantage.

3. UNDERSTANDING THE MATHEMATICS BEHIND COMPUTATION, MACHINE LEARNING, AND STATISTICAL MODELING

Statistical and machine learning will continue to play an important role in the development of such quantitative cellular and organismal models. Historically, artificial intelligence has suffered from exaggeration of its impact, and it seems increasingly likely that the biomedical field is currently experiencing an inflated expectation bubble of this type regarding “deep learning.” Deep learning uses multiple network layers to detect and build hierarchical models of complex features built out of simpler ones. This is an appealing approach, but other algorithms typically have quite similar overall accuracy, and any differences should be understood to develop improved models. Rigorous and systematic experimental and computational evaluation of these methods is essential to be able to use machine learning to make robust and accurate predictions. Systematic testing of computational models is best done in blind assessments of multiple models submitted prior to scoring on datasets explicitly reserved for model evaluation, as has recently been performed for several regulatory genomics predictions [24-26]. Inferring the direction of causal relationships is generally more challenging than inferring statistical dependencies between categorical data [27], and dynamical models whose structure is designed to be consistent with real physical constraints will likely be needed to infer causality in complex biological systems and to make more accurate predictions. Training in the physical sciences and engineering will likely continue to be among the best preparations to contribute to these important questions.

Applying machine learning and statistical model inference correctly requires a thorough understanding of the mathematics behind the algorithms. For statistical learning, graduate-level linear algebra [28] and multivariable calculus are among the most important courses. Even some well-cited genomics papers [29] were eventually found to be of little or no predictive value because of a lack of understanding of potential errors in statistical learning and model evaluation [30, 31]. Especially important is an understanding of overfitting, which is a potential problem in the exponentially large parameter space of deep learning algorithms. While many techniques to manage overfitting have been developed in the machine learning community (e.g., weight decay, weight sharing, dropout, and data augmentation), the application of these methods to genomic training datasets of biologically limited size and correlation structure has not yet been fully explored.
4. WORKING WITH LARGE CONSORTIA AND PUBLIC DATASETS

Much progress has been made by focusing a broad set of researchers with complementary experimental technologies on a standardized set of biological substrates in consortium efforts [5, 32, 33]. Learning how to access, model, combine, and leverage these large datasets is itself an important skill. Working in large groups also requires a distinct set of communication, presentation, and organizational abilities, and experience gained in one consortium largely transfers to another. Genomics has moved strongly toward open access to these large datasets and compendia, so researchers with experience in similar fields will be needed and welcome.

CONCLUSION

In summary, the data-rich fields of genomics and computational interdisciplinary biology remain in great need of talented researchers from the quantitative fields of physics, mathematics, statistics, and computer science. One daunting aspect of biological research is the relatively large amount of jargon and detail of biological systems, but since there are ~20,000 genes in the human genome [3, 4], there is a bound to this complexity at a level comparable to learning a language. If willing to take on this challenge, one will be pleased to find that most other common characteristics of successful training in quantitative fields transfer surprisingly directly to biomedical research.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Gemayel, R. Scientific career paths - part II. FERS J., 2017, 284(17), 2700-2704. http://dx.doi.org/10.1111/febs.14178 PMID: 28887868
[2] McNeil, L.; Heron, P. Preparing physics students for 21st-century careers. Phys. Today, 2017, 70(11), 38-43. http://dx.doi.org/10.1063/PT.3.3763
[3] Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwen, P.; McKernan, K.; Meldrim, J.; Mesirov, J.P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rossetti, L.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, Y.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Durbin, R.; French, L.; Graham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J.C.; Mungall, A.; Plumb, R.; Ross, M.; Showmanne, R.; Sims, S.; Waterston, R.H.; Wilson, R.K.; Hillier, L.W.; McPherson, J.D.; Marra, M.A.; Mardis, E.R.; Fulton, L.A.; Chinwalla, A.T.; Pepin, K.H.; Poplin, R.; Rest, L.; Scherer, S.'.; N.; Schuler, G.; Stal, A.; Lii, S.; Miner, T.L.; Delehaunty, A.; Kramer, J.B.; Cook, L.L.; Fulton, R.S.; Johnson, D.L.; Minx, P.J.; Clifton, S.W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wigen, S.; Sleazak, T.; Doggett, N.; Cheng, J.F.; Olsen, L.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R.A.; Muzny, D.M.; Scherer, S.'.; Bouck, J.B.; Sodergren, E.J.; Worley, K.C.; Rives, C.; Riddles, J.H.; Metzker, M.L.; Naylor, S.L.; Kucherlapati, R.S.; Nelson, D.L.; Weinstock, G.M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Smith, D.R.; Doucette-Stamm, L.; Rubenfeld, M.; Weinstock, K.; Lee, H.M.; Dubois, J.; Rosenthal, A.; Platz, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R.W.; Federspiel, N.A.; Abola, A.P.; Proctor, M.J.; Myers, R.M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D.R.; Olson, M.V.; Kaul, R.; Raymond, C.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G.A.; Athanasiou, M.; Schultz, R.; Roe, B.A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W.R.; de la Bastide, M.; Dedhia, N.; Blake, H.; Hornischer, K.; Nordsieck, G.; Agarwala, R.; Aravind, L.; Bailey, J.A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D.G.; Burge, C.B.; Cerutti, L.; Chen, H.C.; Church, D.; Clamp, M.; Copley, R.R.; Cumber, R.; Doerks, T.; Eddy, S.R.; Eichler, E.E.; Furey, T.S.; Galagan, J.; Gilbert, J.G.; Harmon, C.; Hayashizaki, Y.; Hauser, D.; Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L.S.; Jones, T.A.; Kasif, S.; Kasprz, A.; Kennedy, S.; Kent, W.; E.; King, D.; Kohler, L.; Lander, E.S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Ther interface, D.; Bieda, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramoulswaran, I.; Charlies, R.; Chaturvedi, K.; Deng, Z.; Di Francesco, V.; Dunn, F.; Eilbeck, K.; Evangelista, C.; Gabrielian, A.E.; Han, W.; Ge, W.; Gong, W.; Lander, E.S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Ther interface, D.; Bieda, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramoulswaran, I.; Charlies, R.; Chaturvedi, K.; Deng, Z.; Di Francesco, V.; Dunn, F.; Eilbeck, K.; Evangelista, C.; Gabrielian, A.E.; Han, W.; Ge, W.; Gong, W.; Lander, E.S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Ther interface, D.; Bieda, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramoulswaran, I.; Charlies, R.; Chaturvedi, K.; Deng, Z.; Di Francesco, V.; Dunn, F.; Eilbeck, K.; Evangelista, C.; Gabrielian, A.E.; Han, W.; Ge, W.; Gong,
Shigaki, D.; Adato, O.; Adhikari, A.N.; Dong, S.; Hawkins-Hooker, A.; Inoue, F.; Juven-Gershon, T.; Kenlay, H.; Martin, B.; Patra, A.; Penzar, D.D.; Schubach, M.; Xiong, C.; Yan, Z.; Boyle, A.P.; Kreimer, A.; Kulakovskiy, I.V.; Reid, J.; Unger, R.; Yosef, N.; Shendure, J.; Ahituv, N.; Kircher, M.; Beer, M.A. Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. *Hum. Mutat.*, **2019**, *40*(9), 1280-1291. http://dx.doi.org/10.1002/humu.23797 PMID: 31106481

Beer, M.A. Predicting enhancer activity and variant impact using gkm-SVM. *Hum. Mutat.*, **2017**, *38*(9), 1251-1258. http://dx.doi.org/10.1002/humu.23185 PMID: 28120510

Pearl, J. *Causality: Models, Reasoning, and Inference*, 2nd ed; Cambridge University Press, 2013.

Friedberg, S.; Insel, A.; Spence, L. *Linear Algebra*, 5th ed; Pearson, 2018.

Whalen, S.; Truty, R.M.; Pollard, K.S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.*, **2016**, *48*(5), 488-496. http://dx.doi.org/10.1038/ng.3539 PMID: 27064255

Xi, W.; Beer, M.A. Local epigenomic state cannot discriminate interacting and non-interacting enhancer-promoter pairs with high accuracy. *PLOS Comput. Biol.*, **2018**, *14*(12), e1006625. http://dx.doi.org/10.1371/journal.pcbi.1006625 PMID: 30562350

Cao, F.; Fullwood, M.J. Inflated performance measures in enhancer-promoter interaction-prediction methods. *Nat. Genet.*, **2019**, *51*(8), 1196-1198. http://dx.doi.org/10.1038/s41588-019-0434-7 PMID: 31332378

Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **2013**, *45*(10), 1113-1120. http://dx.doi.org/10.1038/ng.2764 PMID: 24071849

Hawrylycz, M.J.; Lein, E.S.; Guillozet-Bongaarts, A.L.; Shen, E.H.; Ng, L.; Miller, J.A.; van de Lagemaat, L.N.; Smith, K.A.; Ebbert, A.; Riley, Z.L.; Abajian, C.; Beckmann, C.F.; Bernard, A.; Bertagnolli, D.; Boe, A.F.; Cartagena, P.M.; Chakravarty, M.M.; Chapin, M.; Chong, J.; Dalley, B.A.; David Daly, B.; Dang, C.; Datta, S.; Dee, N.; Dolbeare, T.A.; Faber, V.; Feng, D.; Fowler, D.R.; Goldy, J.; Gregor, B.W.; Haradon, Z.; Haynor, D.R.; Hohmann, J.G.; Horvath, S.; Howard, R.E.; Jeromin, A.; Jochim, J.M.; Kinnunen, M.; Lau, C.; Lazarz, E.T.; Lee, C.; Lemon, T.A.; Li, L.; Li, Y.; Morris, J.A.; Overly, C.C.; Parker, P.D.; Parry, S.E.; Reding, M.; Royall, J.J.; Schulkin, J.; Sequeira, P.A.; Slaughterbeck, C.R.; Smith, S.C.; Sodt, A.J.; Sunkin, S.M.; Svanson, B.E.; Vawter, M.P.; Williams, D.; Wohnoutka, P.; Zielke, H.R.; Geschwind, D.H.; Hof, P.R.; Smith, S.M.; Koch, C.; Grant, S.G.N.; Jones, A.R. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, **2012**, *489*(7416), 391-399. http://dx.doi.org/10.1038/nature11405 PMID: 22996553