GRAND CHALLENGE
published: 15 September 2009
doi: 10.3389/neuro.01.016.2009

A fully implantable 96-channel neural data acquisition system. *J. Neural Eng.* 6, 026002.

Shadmehr, R., and Krakauer, J. W. (2008). A computational neuroanatomy for motor control. *Exp. Brain Res.* 185, 359–381.

Shpigelman, L., Lalazar, H., and Vaadia, E. (2009). Kernel-ARMA for Hand Tracking and Brain-Machine Interfacing During 3D Motor Control Advances in Neural Information Processing Systems (NIPS) 21. MIT Press, Cambridge.

Strehl, U., Kotchoubey, B., Trevorrow, T., and Birbaumer, N. (2005). Predictors of seizure reduction after self-regulation of slow cortical potentials as a treatment of drug-resistant epilepsy. *Epilepsy Behav.* 6, 156–166.

Strehl, U., Leins, U., Goth, G., Klinger, C., Hinterberger, T., and Birbaumer, N. (2006).Self-regulation of slow cortical potentials – a new treatment for children with attention-deficit/hyperactivity disorder. *Pediatrics* 118, 1530–1540.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nat. Neurosci.* 7, 907–915.

Weiskopf, N., Sitaram, R., Josephs, O., Veit, R., Scharnowski, F., Goebel, R., Birbaumer, N., Deichmann, R., and Mathiak, K. (2007). Real-time functional magnetic resonance imaging: methods and applications. *Magn. Reson. Imaging* 25, 989–1003.

Wolpert, D. M., and Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nat. Neurosci.* 3(Suppl), 1212–1217.

Wu, W., Gao, Y., Bienenstock, E., Donoghue, J. P., and Black, M. J. (2006). Bayesian population decoding of motor cortical activity using a Kalman filter. *Neural Comput.* 18, 80–118.

Wu, W., and Hatsopoulos, N. G. (2008). Real-time decoding of nonstationary neural activity in motor cortex. *IEEE Trans. Neural Syst. Rehabil. Eng.* 16, 213–222.

Zach, N., Inbar, D., Grinvald, Y., Bergman, H., and Vaadia, E. (2008). Emergence of novel representations in primary motor cortex and premotor neurons during associative learning. *J. Neurosci.* 28, 9545–9556.

Received: 19 May 2009; published: 15 September 2009

Citation: Front. Neurosci. (2009) ***3****, 2: 151-154. doi: 10.3389/neuro.01.015.2009*

Herding cats: the sociology of data integration

Robert W. Williams*[1]

[1] Department of Anatomy and Neurobiology, Center for Integrative and Translational Genomics, University of Tennessee Health Science Center, Memphis, TN, USA

** Correspondence: rwilliam@nb.utmem.edu*

Upon this gifted age, in its dark hour,
Rains from the sky a meteoric shower
Of facts…they lie unquestioned, uncombined.
Wisdom enough to leech of our ill
Is daily spun; but there exists no loom
To weave it into fabric;

—Edna St. Vincent Millay (from Huntsman, What Quarry? 1939)

THE CHALLENGE OF PREDICTION

The age of personalized medicine and genomics is upon us and we are facing a grand challenge – or a brick wall. Once we have finally gained a near complete compendium of fundamental mechanisms, connections, and developmental sequences – an encyclopedia of biology, bodies, brains, and behavior – can we achieve the data density and integration needed to develop holistic and robust models that generate useful predictions? Will we be able to distinguish between personalized genomics and a horoscope? What new types of resources, data sets, and synthetic frameworks are needed to make correct prognoses and recommend actions? What is my personal risk for Alzheimer's disease, and what should I do about it today?

The complexity of biological systems implies that a parts list of mechanisms and processes, however complete, will not be up to the task of making good predictions. We need a way to test drive our models using a system that has the same level of complexity as human populations. I will describe an effective approach that relies on genetic reference panels (GRPs) that can be used to make and test predictions from base pair to behavior. I will describe how scientists can retain their independence while explicitly contributing to a fabric of tightly woven quantitative data.

THE COLLECTIVE COST OF SCIENTIFIC INDEPENDENCE

Scientists are trained to think independently and critically. It is inevitable that we like to do things our own way, generating and using data from experiments we designed ourselves. This approach is not a self-indulgent luxury – it is an essential attribute of innovative science, enshrined in the ways we evaluate and fund new and ongoing research. Independence contributes to the stirring cacophony of competing ideas that moves us toward a deeper understanding of biological processes.

Yet, independence has a cost. The scope of studies from single groups is limited by their technical and analytic proficiency and by modest budgets. The collective result is a fragmented, half-hidden literature and a fragmentary and rapidly evaporating collection of raw data, generated using different species and strains raised under different conditions, treated using varied paradigms, and measured using different equipment. Of course the pieces do not fit together! They were never intended to fit into any unified design. It is no surprise that integration of data sets and of key results is difficult; sometimes impossible. Ronald Fisher pointed out that "a competent overhauling of the process of collection, or of experimental design, may often increase the yield (precision of results) ten or twelve fold, for the same cost in time and labour" (Rao, 1992). Fisher meant this in the context of a single lab, but his observation applies with equal or greater force across the research efforts of entire communities. The sad truth is that the whole is less than the sum of its parts. The poet Edna St. Vincent Millay summarizes our plight; we are both blessed and cursed with "a meteoric shower/Of facts… they lie unquestioned, uncombined." Our collective consolation is that crucial discoveries and methods thrive and catalyze new generations of more advanced research. From this point of view, progress is inexorable and – who can deny it – rapid. The optimist's view is that everything lost in the process was mere scientific chaff.

The hidden cost of this fragmented style of biological research is that current models of complex biological processes are *ad hoc* and underpowered, and they will remain so until we develop new paradigms and looms that allow us to weave together massive data sets across all levels of biological organization – from base pair to social behavior. Our current models are mannequins we dress up with sparse data from a few experiments and send down a publication catwalk, praying they don't toddle and trip. What we need are powerful, diverse, and well dressed models that can navigate a landscape, not a catwalk. As a community we have not begun to grapple with this issue, but one certainty is that we will require extraordinarily well structured data sets for humans and for experimental models, to make accurate predictions.

In the case of our massive neuroscience community, there have rarely been grand multi-lab meta-experimental designs that recognize this requirement. There have been far too few efforts to systematically generate large volumes of data to support community efforts to generate and test hypotheses. Some feel that such efforts violate our pioneering spirit and invite the derision of critics who see big science as big boondoggle. The almost visceral – and in retrospect, clearly wrong-headed – response to the Human Genome Project as an egregious waste of money is the most obvious example of this cultural aversion to top-down science. One of the single largest projects in neuroscience, the Allen Brain Atlas, a free and complete compendium of data on gene expression in the mouse central nervous system, came in for more than its fair share of criticism, even though it was a gift to the research community that did not expropriate funds intended for individual investigators. Both, the Human Genome Project and the Allen Brain Atlas, are large projects, but they are not multiscalar – they focus on the analysis of DNA and RNA expression, respectively. Neither program gets us to the point of predicting relations between genes and behavior. We have fabulous tools now, but we apparently have neither the appetite nor the vision to apply them on an appropriate scale and in a systematic way across multiple levels of organization. This needs to change.

The consequence of failing to develop more robust quantitative models is about to become very painful. To give this assertion some teeth, consider how well we can predict outcomes of simple single gene deletions and mutations in any organism. The answer is extremely poorly, even when using fully inbred and isogenic lines reared in tightly controlled environments (e.g., Crabbe et al., 1999; van Swinderen and Greenspan, 2005). We need to understand why this is the case, what we need to do to specify the problems, and how to devise solutions. In an era in which many of us will soon be sequenced, it is essential to develop models of causal relations between genes, environments, and neuronal phenotypes. Neuroscientists must answer questions of this type: where are the promised cures and treatments for neurodegenerative diseases, autism, and a host of ills for which genes and mechanisms are now well known? Long term support for neuroscience and genomics depends on proving the personal and social utility of the results.

To reframe the challenge, can neuroscientists retain the real benefits of independence while working together within an integrated framework that gives us an increased yield of multiscale data sets that we can use to generate and test models? The answer is yes; the more systematic exploitation of GRPs retains the best elements of both approaches.

GENETIC REFERENCE PANELS AS PLATFORMS FOR PREDICTIVE BIOLOGY

In order to make strong predictions that can be refuted, refined, and verified, we need a test bed that has a level of genetic, molecular, and cellular complexity that matches that of human populations, but over which we have precise genetic and experimental control. The solution to this problem, the GRP, consists of a large set of isogenic strains of animals that can be used by a community of researchers to study an almost unlimited range of phenotypes (Chesler et al., 2003). Each member of a GRP is an inexhaustible and stable "clone" of animals, members of which can be studied at different stages of life, in different environments, using any number of techniques. Harvesting data across a large GRP can give us the right material to develop sophisticated models that account for genetic and environmental complexity.

Using a GRP, data from different labs can be readily compared and combined without the need to explicitly collaborate. Work separated by decades can be compared at the level of an entire GRP. The only modest concession that scientists need make is to use a GRP in the first place. The enticing compensation is access to an open and massive collection of highly useful data on all members of a GRP (e.g., Mozhui et al., 2008; www.genenetwork.org).

Each GRP needs to be large enough so that correlations and causes that link genotypes, phenotypes, and environmental perturbations can be quantified, modeled, and used to generate predictions. The most widely used GRP consists of a set of 80 BXD mouse strains (Peirce et al., 2004). Strains in this panel all trace back to matings between two of the most venerable inbred strains of mice – C57BL/6J and DBA/2J. Both parent strains have been sequenced. They differ at ~4 million sites scattered across the genome (Frazer et al., 2007; Roberts et al., 2007; Williams et al., unpublished sequence data), close to the number of single nucleotide polymorphisms (SNPs) segregating in human populations (www.hapmap.org). Members of this BXD family are genetically stable – all can be rederived from cryopreserved stock and all are available from the Jackson Laboratory. Much larger GRPs are on the horizon, with a set of up to 1,000 strains now in development (Chesler et al., 2008; Threadgill et al., 2002).

What makes a GRP truly a *reference* set is that each member of the panel can be replicated as an inbred and isogenic strain. We can accumulate massive multiscalar databases of phenotypes using precise instruments, and we can study many individuals from each of the GRP's strains at different stages of development, in different environments, and following different treatments.

What makes a GRP a *genetic* panel is that it is an extended family we can use to track down gene variants that influence phenotypes of almost any kind, from retinal ganglion cell number to ocular dominance plasticity (Heimel et al., 2008; Williams, 2000). Once all members of a GRP have been genotyped (the 80 BXD strains have been typed at 580,000 SNPs), the panel becomes a valuable platform for genetic prediction. And once the parents of a GRP have been sequenced, we can also use reverse genetic methods (going from gene variant to phenotype variant) similar to those used to study gene function in knockouts and mutants. However, unlike a study of knockouts, we do not study a gene variant on a single genetic background, but rather across an entire panel of strains – approximately half of which will have one allele or the other (Carneiro et al., 2009). The comparison amounts to a high power *t*-test between two genotype groups (in the BXD, between the parental *B* and *D* alleles). Many thousands of genetic differences can be studied in this way using a single GRP. In other words, GRPs scale well to study gene function in a more realistic and complex genetic context.

Extracting causal relations between DNA differences, mRNA expression, cellular physiology, and behavioral differences is a demanding genetic, biological, statistical, and computational problem. But our collective ability to develop and test hypotheses will continue to improve as we accumulate deeper and more diverse phenotype data across GRPs, at many levels of CNS structure and function, and from different labs and environments.

Lab-to-lab variation, once thought of as a major problem, can actually be turned to our advantage when studying a common GRP. For example, we can expose and measure the critical effects that experimental perturbations have on neuronal development and physiological responses (Heimel et al., 2008). We can do this because the entire panel is a stable platform upon which we build and test models. Each GRP can be its own control. We can think of this approach as a merging of systems neuroscience and systems neurogenetics – in which the term *system* is defined both classically, as the complex of neural networks that generates behavior, and as the multiscalar system that extends from variation in sequence to variation in behavior.

THE PRIMACY OF DATA

Should we be optimistic about the near-term prospects for generating efficient and effective models that make helpful predictions? Without data of the right type and scale, we will not get far. I can say with confidence that we will have the raw computational power to handle data and models. With encouragement, we can generate a social compact among scientists to make it possible to generate the massive multiscalar data sets we need to build and test models.

It is relatively easy to build a raft to float across a wide river; once on the other side, a small band can make important discoveries, often with simple tools. It is far more complex to build a permanent bridge to transport large volumes of goods and people across the river to allow a new economy to flourish. Building a bridge requires the output of industries, the work of many specialists, and a collective sense of delayed gratification. Building infrastructure at this level also requires leadership and superb management. It is time to begin building massive data bridges that will enable predictive biology to thrive. Perhaps the real challenge will be convincing both the community and its leaders that it is doable today.

ACKNOWLEDGMENTS

Thanks to Glenn D. Rosen and Kathryn Graehl for comments and editorial improvements. Thanks also to the NIH for support of many data sets for GRPs now being incorporated into GeneNetwork, a pilot project in systems genetics.

REFERENCES

Carneiro, A. M., Airey, D. C., Thompson, B., Zhu, C. B., Lu, L., Chesler, E. J., Erikson, K. M., and Blakely, R. D. (2009). Functional coding variation in recombinant inbred mouse lines reveals multiple serotonin transporter-associated phenotypes. *Proc. Natl. Acad. Sci. USA* 106, 2047–2052.

Chesler, E. J., Miller, D. R., Branstetter, L., Galloway, L., Jackson, B., Philip, V. M., Voy, B., Culiat, C. T., Threadgill, D. W., Williams, R. W., Churchill, G. A., Johnson, D. K., and Manly, K. F. (2008). The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. *Mamm. Genome* 19, 382–389.

Chesler, E. J., Wang, J., Lu, L., Qu, Y., Manly, K. F., and Williams, R. W. (2003). Genetic correlates of gene expression in recombinant inbred strains: a relational model to explore for neurobehavioral phenotypes. *Neuroinformatics* 1, 343–357.

Crabbe, J. C., Wahlsten, D., and Dudek, B. C. (1999). Genetics of mouse behavior: interactions with laboratory environment. *Science* 284, 1670–1672.

Frazer, K. A., Eskin, E., Kang, H. M., Bogue, M. A., Hinds, D. A., Beilharz, E. J., Gupta, R. V., Montgomery, J., Morenzoni, M. M., Nilsen, G. B., Pethiyagoda, C. L., Stuve, L. L., Johnson, F. M., Daly, M. J., Wade, C. M., and Cox, D. R. (2007). A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* 448, 1050–1053.

Heimel, J. A., Hermans, J. M., and Sommeijer, J. P., Neuro-Bsik Mouse Phenomics Consortium, and Levelt, C. N. (2008). Genetic control of experience-dependent plasticity in the visual cortex. *Genes Brain Behav.* 7, 915–923.

Mozhui, R. T., Ciobanu, D. C., Schikorski, T., Wang, X. S., Lu, L., and Williams, R. W. (2008). Dissection of a QTL hotspot on mouse distal chromosome 1 that modulates neurobehavioral phenotypes and gene expression. *PLoS Genet.* 4, e1000260.

Peirce, J. L., Lu, L., Gu, J., Silver, L. M., and Williams, R. W. (2004). A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC Genet.* 5, 7.

Rao, C. R. (1992). R. A. Fisher: the founder of modern statistics. *Stat. Sci.* 7, 34–48.

Roberts, A., Pardo-Manuel de Villena F., Wang, W., McMillan, L., and Threadgill, D. W. (2007). The polymorphism architecture of mouse genetic resources elucidated using genome-wide resequencing data: implications for QTL discovery and systems genetics. *Mamm. Genome* 18, 473–481.

Threadgill, D. W., Hunter, K., and Williams, R. W. (2002). Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mamm. Genome* 13, 175–178.

van Swinderen, B., and Greenspan, R. J. (2005). Flexibility in a gene network affecting a simple behavior in *Drosophila melanogaster*. *Genetics* 169, 2151–2163.

Williams, R. W. (2000). Mapping genes that modulate mouse brain development: a quantitative genetic approach. In Mouse Brain Development, A. F. Goffinet, P. Rakic, eds (New York, Springer), pp. 21–49.

Received: 01 June 2009; published: 15 September 2009

Citation: Front. Neurosci. (2009) ***3****, 2: 154-156. doi: 10.3389/neuro.01.016.2009*