Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Unpacking the Phenotype (UP)
Deciphering Genome to Phenome Relationships: Interdisciplinary Research at the Interface of the Biological and Mathematical Sciences

Report of a NSF-funded Workshop held in Arlington Virginia
on October 25-27, 2015

Summary

Biological systems are exceptionally complex, involving a multitude of interactions among a large number of components at different spatial and temporal scales. What these biological processes are, and how they operate to produce particular traits and properties has been a subject of intense interest among biologists. The past few decades have witnessed a revolution in our ability to sequence genomes, clone genes, identify gene products, manipulate gene expression and genetically engineer a variety of cells and model organisms. These powerful technologies have been used in almost every field of Biology, from Molecular Genetics to Cell and Developmental Biology and from Physiology to Behavior and Ecology.

The revolution in 'omics studies continues to create vast amounts of data on gene expression, and protein and metabolite profiles, in many species of animals and plants, and under a variety of experimental conditions. Although powerful techniques have been developed for deducing association networks among genes and a variety of traits, there has been little progress in understanding the causal relationship between genes and phenotypes. Understanding the causal processes that link genes and traits, and genomes and phenomes, is one of the Grand Challenges facing us today.

The processes that generate phenotypes are complex, nonlinear, multivariate and multi-level. And although these processes involve gene products, their kinetics and outcomes are determined by the system in which they operate. These processes lead to a number of virtually universal emergent properties of phenotypes: phenotypic robustness (insensitivity of phenotype to genetic and environmental variation); phenotypic plasticity (a single genotype can lead to different phenotypes in different environments); variable penetrance (a mutation alters a phenotype in some individuals but not in others) and variable expressivity (a mutation leads to different phenotypes in different individuals). These properties degrade a clean and simple relationships between genotype and phenotype.

Because of their extreme complexity and diversity, the interactions by which genes affect the development, properties and functions of complex traits cannot be fully understood without quantitative analyses and mathematical models. This

understanding will require the development of foundational mathematical, statistical, and computational approaches and the development of new mathematical tools. In particular, it will require a close collaboration of biological and mathematical scientists and leverage their complementary expertise to address these important challenges.

Introduction

Investigating and understanding emergent properties of biological systems are key to addressing one of the important grand challenges in biology, *Deciphering the Genomes-to-Phenomes (G2P) Relationship.* Examples of emergent properties include such diverse phenomena as protein folding, cell division, a stable body temperature, and Monarch butterfly migration. Each of these emergent properties displays stability of form within one species that is resistant to both genetic and environmental variation. Yet the developmental processes that produce this stability are also capable of giving rise to the vast diversity of forms across species: organisms exist as stable entities, as well as dynamic ones capable of responding to genetic and environmental change.

These biological properties are emergent in the sense that they are not encoded in the genome but arise from a multivariate non-linear interactions among many components at different spatial scales (molecules, cells, tissues, organs, organ systems, whole organisms) and temporal scales (from molecular interactions that occur in fractions of a second, to developmental and physiological time scales measured in hours to months).

Understanding what these processes are, how they are affected by genes and environment, and how their interactions give rise to emergent properties of biological systems is an exceptionally challenging enterprise that is difficult if not impossible to do by experimentation alone. Experiments are always based on conceptual models of what a system is and how it operates; this is the basis of hypothesis-driven research. Such conceptual models are always small, local and incomplete, and can be imprecisely formulated or biased by preconceived notions of how the system should operate [1]. Simplification is essential because the structure and kinetics of real systems are too complex to be fully understood by even the most experienced and astute experimenter. A rigorous understanding of gene-to-phene relationships will require the development of foundational mathematical, statistical, and computational approaches that integrate closely with experiments. The collaboration between the biological and the mathematical sciences communities will leverage their complementary expertise to address this grand challenge.

<u>Genome to phenome (G2P) relationships are complex in two very different ways.</u> First is the problem of <u>big heterogeneous data</u>. The various high-throughput 'omics approaches are producing increasing amounts of heterogeneous data on genome structure, gene expression patterns, protein patterns, metabolite patterns and many higher level phenotypes, particularly those associated with disease. For instance, it is estimated that 10-40% of the genes in the genome are expressed in any one tissue

amounting to roughly 2,000 to 8,000 genes expressed, in different combinations in different tissues. What do all these genes "do" and are they all equally relevant to the phenotype under consideration? Proteomic and metabolomic studies can identify the prevalence of thousands of proteins and metabolites [2]. A single RNAseq expression profiling run, can generate more than 10 million "reads" and needs to be replicated many times to obtain statistically valid data. The number of possible associations that can potentially exist within and among these various high-throughput data sets is staggeringly large. Much of modern statistical bioinformatics is geared toward discovering significant and interesting association patterns within and among those data sets.

Second, the causal mechanisms by which gene products and environmental variables interact to generate phenotypes form dynamically complex interaction networks that operate at multiple scales of organization. Because the interactions are multivariate and non-linear it is difficult if not impossible to deduce the operation of even a simple network without mathematical modeling. A rich and ever growing array of analytic tools has been developed to tease out the statistical associations among genes and traits. There is no equivalent initiative or body of knowledge that deals with the elucidation and analysis of the causal relationships among genes and traits. Dealing with the elucidation and analysis of the causal relationships among genes and traits requires carefully designed perturbation experiments and a close collaboration between mathematicians and statisticians and biologists to overcome the challenges non-identifiability poses.

The Big Challenges in understanding G2P relationships

The greatest challenge in coming to a complete understanding of the casual relationship between genes and traits is that most gene-trait associations do not obey Mendel's laws of segregation in the sense that there is no unique relationship between the genotype and phenotype. Instead the effects of genes on traits are characterized by variable expressivity (the same mutation can produce a range of phenotypes) and incomplete penetrance (a given mutation is associated with a phenotype in some individuals but not in others). These phenomena arise because genes do not cause traits directly. Genes do not code for phenotypes but affect traits via complex networks of interactions that involve many other genes, proteins, metabolites and various environmental variables, as well as the structural and functional context in which a particular gene expression takes place. Of course the expression of genes itself is part of the phenotype of a given cell in a given context.

Two widely observed features of the phenotype further degrade a clean and unambiguous association between genotype and phenotype: phenotypic robustness and phenotypic plasticity. Robustness refers to mechanisms that stabilize the phenotype against genetic and environmental variation. Phenotypic plasticity is the opposite: it is the sensitivity of phenotypes to environmental variables. Many robustness mechanisms are evolutionary adaptations including the multitude of homeostatic mechanisms in biochemistry, development and physiology that

stabilize form and function against environmental and genetic variation [3-8]. It is not clear yet whether the same mechanism can provide both environmental robustness as well as genetic robustness, or whether stability against each factor that affects the phenotype has evolved separately.

Phenotypic plasticity can be neutral or maladaptive as well as adaptive. Environmental variables such as temperature affect the rates of different molecular and cellular processes to different degrees and this can have deleterious consequences for the traits that depend on those processes, creating one of the conditions required for the evolution of mechanisms that stabilize traits. But if a phenotypic variant that is induced by an environmental variable is more fit under those novel environmental conditions, mechanisms may evolve that stabilize the alternative phenotype in that environment [9]. This is the foundation of adaptive phenotypic plasticity in which alternative phenotypes develop under different environmental conditions. Examples of such adaptive phenotypic plasticity are the seasonal forms of insects (some being so different phenotypically that they had originally been described as different species), the shade and sun leaf shapes of trees, the winged and wingless morphs of aphids, the soldier and worker castes in ants and termites, and winter plumage in birds. Behavior is the ultimate plastic phenotype [9].

In summary, the mechanisms that generate phenotypes degrade a close correspondence between genotype and phenotype. One genotype can correspond to many phenotypes, and one phenotype can correspond to many different genotypes. The phenotype that develops depends as much on environment as it does on genetics.

<u>The problems of non-linearity.</u>

Because gene products operate in complex multi-level networks, the effect that a gene has on a phenotype is not a property of the gene itself but depends on the structure and kinetics of the system in which it is embedded. The "activity" of a gene product is determined by its structure, which is affected both by the genetic sequence and its level of expression, which is determined in turn by transcriptional regulators controlled by other genes. Activity of a gene product can also be affected by a variety of allosteric activators or inhibitors which are, in turn, controlled by other genes. The activity of a gene product, in turn, contributes to the structural and kinetic properties of the network in which it is embedded, and a multivariate model is necessary to account for these nonlinearities. The kinetics of networks and their component parts are almost always non-linear and multidimensional. As a consequence, the effects of genes on phenotypes are almost always non-linear. This non-linear multidimensionality accounts for many of the indirect and context-dependent effects of genes on phenotypes.

For example, consider a common sigmoidal relationship between the activity of a gene product and a phenotypic variable: at low activity there is little or no effect on

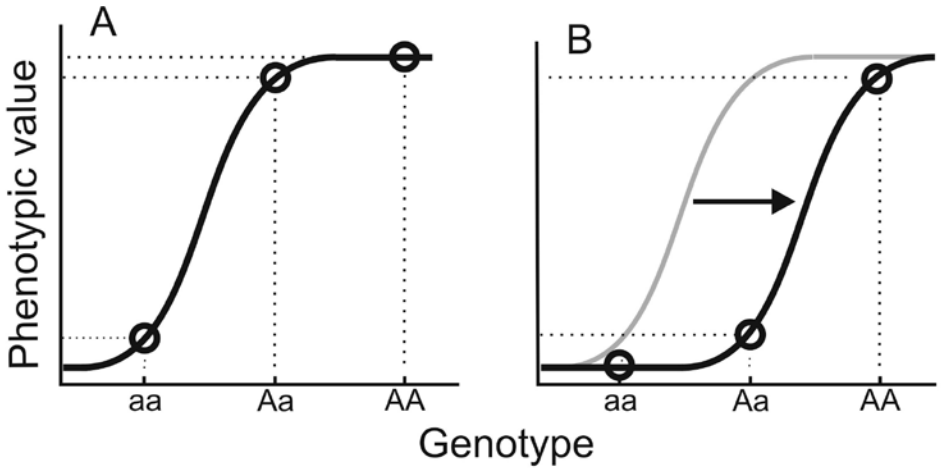

Figure 1. **A**. Relation between genotype and phenotype. The non-linear relationship makes the *Aa* heterozygote phenotype more like that of the *AA* homozygote, so *A* is dominant over a. **B**. In a different genetic or environmental background the relationship is shifted (arrow). In such an individual the heterozygote phenotype is more like that of the *aa* homozygote, so now the *a* allele is dominant.

the phenotype; at intermediate activities the effect is proportional, and at high activities the effect saturates (Figure 1).

The exact shape of the sigmoidal or other more complex relationship is not a property of the gene but of the entire network: it is a systems property that depends on the genetic and environmental background. In Figure 1A we show how a sigmoidal relationship controls the effect of a bi-allelic genotype on a phenotype. Changes elsewhere in the network can alter the shape or position of the sigmoidal relationship and will thus alter the relationship between the genotype at the gene and the phenotype (Figure 1B). The figure illustrates how a system-dependent shift in the sigmoidal relationship can cause the phenotypic effect of a gene to change from dominant to recessive [10]. In other words, the system determines whether a particular allele will have a dominant or recessive effect on the phenotype [11, 12]. Different populations, different sub-samples of a population, and different individuals can have different genetic and environmental backgrounds that can control the effect of a gene in a similar manner.

Non-linearities, such as those illustrated above, can explain the context-dependent effects of genes on phenotypes [13]. In Figure 1A the implicit assumption is that there is no variation in the genetic or environmental background, so the relationship can be graphed as a single line. In effect, it illustrates the conditions that could obtain in a single individual at a single time. In nature no two individuals have the identical genetic and environmental background, so a population representation of the same relationship could look like that illustrated in Figure 2A, with a broad range of possible associations between genotypic and phenotypic values. The correlation between genetic variation and phenotypic variation in the case shown in this figure will depend on the actual distribution of genotypic values in the sample

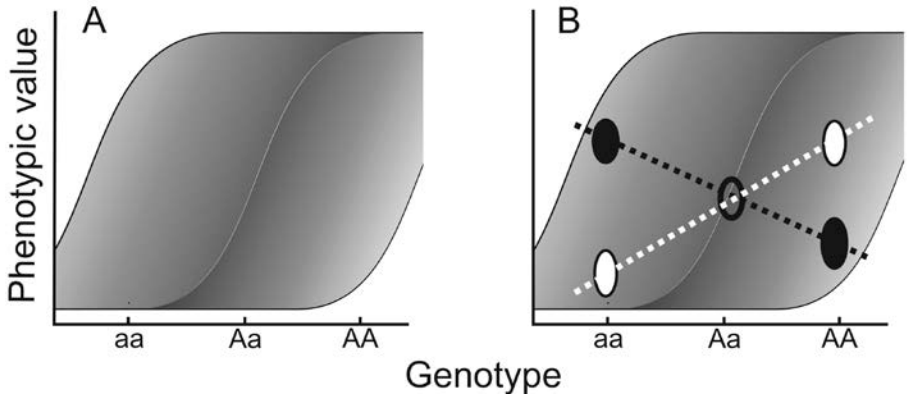

Figure 2: Population versions of the graphs shown in Figure 1. **A**. The relationship between genotype and phenotype in a population with genetic and environmental variation is potentially different for each individual. **B**. Allelic effects on the phenotype depend on the genetic and environmental background. In a population illustrated by the white ellipses there is a positive correlation of the A allele with the phenotypic value, whereas in a different population illustrated by the black ellipses the correlation is negative.

being studied, and could be positive, or negative, or not significantly different from zero (Figure 2B).

The relationships shown in Figure 2 offer a possible explanation of incomplete penetrance and variable expressivity. For incomplete penetrance a population could have an allelic distribution illustrated in Figure 2A, so that some individuals with the "dominant" AA genotype have a high phenotypic value corresponding to a disease state, for instance, and others a low value. For variable expressivity a population could have an allelic distribution illustrated in Figure 2B, where the same genotype (AA) has a range of phenotypic values. Again, exactly what the phenotypic value for a genotype like AA is, will depend on the genetic and environmental background.

It is also possible to visualize how robustness and phenotypic plasticity could arise in such a system. Robustness requires that there are no phenotypic effects of mutations (or of environmental variation). In Figure 2 this could be visualized as a system in which all genotypes produce the same phenotypic value. Dominance is a simple kind of robustness to genetic variation at a single locus. Phenotypic plasticity, where one genotype is associated with many phenotypic values can arise if variation in the relationship between genotype and phenotypic value is due to variation in the environmental background.

These simple examples are just one possible way in which variable expressivity, incomplete penetrance, robustness and plasticity could arise due to a particular kind of non-linearity in the phenotypic effects of one gene in a variable genetic and environmental background. There are certainly many other ways in which these effects could arise, but there is at present no general way to discover, investigate or

classify this diversity. If we now consider that most phenotypes are affected by many genes and environmental factors that act in large complex multi-level networks, it becomes clear that understanding the causal relationship between genomes and phenomes presents an exceptionally difficult conceptual and mathematical challenge. To address this challenge will require close cooperation between biologists, statisticians and mathematicians.

Research Goals and Targets

A. Pathways to Phenotypes. The overarching goal of a G2P program is to elucidate the causal connections between genotypes and phenotypes by predicting how a particular mutation in a gene causes a specific change in a trait, and how multivariate genetic and environmental factors interact to produce a phenotype. Such challenges can today only be addressed for simple molecular and biochemical phenotypes. In a biochemical system, for instance, enzymes are gene products and their kinetic properties arise from the gene sequence that determines enzymatic efficacy and the network that can include allosteric activators or competitive inhibitors that affect the rate of the reaction. The resulting phenotype could be a reaction rate (e.g. the rate of release or reuptake of serotonin at a nerve terminal), or a metabolite concentrations (e.g. the intracellular or plasma levels of folic acid). For higher-level phenotypes, like development or behavior, the causal pathway by which a mutation alters a trait can be very complicated, involving many gene products, structural features of cells and organs, transport mechanisms between tissues and interactions among cells within tissues among many others.

Intermediate phenotypes occur at all points along the casual chain. Every structure and function is a phenotype, from the subcellular machinery to a physiological regulatory mechanism to a fully formed organism. In order to understand higher level phenotypes it is necessary to also understand the intermediate phenotypes. The importance of this chain of phenotypes is becoming recognized as we move away from a statistical view of gene-trait associations to develop an understanding of the complex multi-level causes of phenotypes, a view that is becoming known as deep phenotyping in the biomedical literature [14].

Phenotypes develop and function at many different spatial and temporal scales. Here we list three concrete examples of how phenotypes are defined by multiscale processes. (1) Notch in *Drosophila* is a point mutation that causes the notch to be cut out of the wing. The notch gene product is involved with inter-cellular communication: the final phenotype is due to spatially patterned cell death in a large population of cells. (2) Homeotic mutations are point mutations that change the fate of one tissue into another (antennapedia causes the antenna to develop as a leg; ophtalmoptera causes the eye to develop as a wing); the causal networks are very complicated and little-studied. (3) Polydactyly is a dominant trait in which affected individuals have additional fingers and toes. Not all individuals who carry the mutant allele exhibit the trait (incomplete penetrance), and different individuals exhibit different degrees of digit multiplication and digit fusion (variable

expressivity). The mutation is in the GLI3 gene which codes for a transcriptional regulator; the developmental biology of digit formation involves cell migration, patterned cell death and a complex cascade of tissue differentiation [15]. Multiscale multivariate simulations could show how rapid molecular events at the gene level are connected to cellular events in an organ (the hand) composed of more than a dozen interacting cell types that develops over a period of days to weeks.

These are just three examples out of thousands of cases where point mutations lead to profound alterations in higher-level phenotypes. Many are known from studies of model organisms and genetic diseases in humans. In most cases, however, we know little about the molecular effects of the gene product and even less about the processes that produce the phenotype. All we typically have are statistical associations showing that an allele is correlated, to some degree, with a phenotype.

B. Can 'omics data be used to deduce causal pathways and networks? High-throughput 'omics studies are producing vast amounts of data on gene expression and protein and metabolite levels under a great variety of natural and experimental conditions. Much of this information is used to detect associations between genetic and phenotypic variables and to then deduce association networks of genes, proteins and metabolites[16], and how they are associated with traits of interest.

This vast amount of data is a powerful resource that can be leveraged to enhance our understanding the relationships between genes and traits. But are they sufficient to uncover or deduce causal relationships? Phenotypes are dynamical systems, but 'omics data are typically static snapshots, and, accordingly, current methods produce static networks. To understand causation it is necessary to get away from a "networks view" that has no kinetics. For that we need dynamic, not static data. In particular we need accurate time-series and longitudinal studies of gene expression, and the accompanying proteomics and metabolomics, and image-generated visual data on the accompanying phenotypes. In addition to new data, we also need mathematical models and effective abstract representations that better capture complex biological interactions.

Many attempts to deduce function from network reconstruction assume that connections between nodes have certain functional properties. Gene network reconstruction techniques pioneered by Reinitz [17] assume a sigmoidal relationship between the input and output of each node and machine learning algorithms can be used to discover functional networks [18]. These networks typically deal with only a single level of the hierarchy (e.g. gene regulatory networks, protein interaction networks) and current methods are not suited for multi-level systems. Parameter estimation for the properties of networks make several simplifying assumptions that may not be realistic: networks are at steady state (which is seldom the case); parameters are optimized (unlikely because real networks are the products evolution, and they are likely to just work "well-enough" rather than optimally); the kinetics largely of a single kind (linear, sigmoidal or Boolean logic), whereas in real life they are diverse. Thus although they produce

"results," current network reconstruction methods cannot be expected to find, nor represent, the actual networks structure and kinetics.

Some problems with current methods of elucidating the networks that connect genes to phenotypes are conceptual; others come from inadequacy of the data. Much of the required data are noisy and incomplete. Noise comes from fact that some processes that lead to phenotype are stochastic because of the low concentrations of gene products and small numbers of interactors. Noise and inadequacy also come from the fact that systems are imperfectly known and some of the observed variation is due to factors that cannot be controlled experimentally. Noise also comes from the fact that no two individuals are genetically identical nor have the same environmental exposure and history. This can cause uninterpretable noise when experimental samples are pooled from different individuals. Finally patterns of gene expression differ from cell type to cell type and change with time and conditions. This causes noise in data that are taken from tissues with heterogeneous cell populations or tissues that are not perfectly synchronized.

In short, to understand G2P relationships we need much more detailed data on the kinetics of the processes that make or constitute a phenotype. Instead of data mining, we need carefully designed hypothesis driven science, consisting of detailed longitudinal experiments, aided by carefully designed high throughout data. We also need believable network modeling techniques for 'omic-wide scale data, which go beyond correlation and incorporate detailed biological information to make predictions of the biological mechanistic process. New formal statistical methods of inference for parameters (such as rates and initial conditions) in partially observed multiscale multivariate stochastic dynamical systems are needed.

C. Quantification of Phenotypes. Phenotypes are diverse and occur at all levels of the organizational hierarchy. In order to develop models that connect genotypes to phenotypes we need accurate and unambiguous ways of quantifying phenotypes. Molecular phenotypes are perhaps the easiest because we can quantify amounts of chemicals and reaction rates. Cell interactions and the morphology and geometry of cells, tissues, organs, appendages and bodies are far more difficult to describe or quantify. Methods need to be developed to quantify and compare such phenotypes in biologically informative ways. For instance, although a morphological shape can be described quantitatively by an elliptical Fourier transform, the coefficients have no biological meaning and are not helpful in understanding how that morphology arose, or how it is biologically related to others. Image analysis is the route being chosen by many groups in cellular phenotyping. The analytics associated to such phenotypes involve a combination of statistics and differential geometry. Mathematical methods can provide crisp representations of similarity between phenotypes through metrics chosen in collaboration with biologists that codify carefully what the relevant landscapes looks like. Landscapes of genotype-environment interactions can thus be mapped into the relevant space of phenotypes. Previous examples of effective mappings have enabled computationally efficient representations of phylogenetic trees [19, 20], genome re-arrangements

[21], phenotypic landscapes [8] and fitness landscapes [22].

D. Principled bottom-up models. While one route to G2P insights is top-down -- through the acquisition of new high-bandwidth data from 'omics and high-throughput experiments -- another route is through bottom-up mechanistic models. And, while one class of mechanisms entails biochemical pathways and circuits, another class of mechanisms entails other variables – how environmental stressors can control the whole background state of the organism; or how the health of the whole proteome (the folding and aggregation and chaperoning state of large fractions of a cell's proteome) affects phenotype. As noted above, it is often not a single gene that gives rise to a phene. But more than that, sometimes the best predictor of P (phene) may not be G at all. For example, as cells age, they undergo changes in phenotype due to oxidative damage. How should we make models to understand the change in phenotype that results from oxidative aging? Maybe such models would not begin with particular genes at all; maybe such models would start from a description of the oxidative and folding status of a whole proteome at a time. In general, a broader perspective than G2P would be GE2P, where E represents environmental variables. And even beyond that, predicting phenotypes will also require understanding the dynamics of evolutionary change and the factors that contribute to evolutionary fitness landscapes. This may help us understand the "whys" behind the "whats".

There are several ways to support this enterprise. We encourage synthetic biologists to introduce "knobs" to control the relationships between G and P. We encourage a culture of model-driven experiments. We need models of G2P that capture environment and evolution. We need multi-scale models from fine-grained, such as at the biochemical pathway level, to coarse-grained, where phenotype can be more readily expressed. We need models of ensembles of evolutionary trajectories. We need to understand the possible "phase transitions" in evolutionary trajectories, such as in speciation. We need to understand the speed of adaptation of populations. We need ways to understand the "believability" of a model: When can we trust a model when its predictions go beyond the known data? The key value of a bottom-up model is in what predictions it makes that go beyond the data, that are novel and important, and that define an untested hypothesis or extrapolation. Toward this end, what are the roles of identifiability and "sloppy model" concepts? And, because principled mechanistic biology is a long-term dream, we need more support of *deep innovation,* of the most basic possible new science. Current examples are INSPIRE/ EAGER or NIH Pioneer Awards.

The Need and Benefit for Mathematics

The preceding sections make it clear that the practical and conceptual problems associated with uncovering and understanding the relationships between genes and traits are so vast, so diverse and complex that they cannot be addressed through laboratory experimentation alone. An effective collaboration between biological

and mathematical scientists will be essential for progress.

We have shown that biologists, and the solution of biological problems, will benefit from mathematics. But if mathematicians (and their funding agencies) are going to invest time and resources it is reasonable to ask what is in it for the mathematician? How will Mathematics benefit? There is excellent reason to believe that mathematics will benefit extensively, simply by considering the track record of how much has already been gained. We quote here from Reed (2015) (omitting references therein): "*In the twentieth century there were three main influences of biology on mathematics. The theory of evolution and genetics stimulated the fields of statistics, probability, and stochastic processes. The Hodgkin-Huxley equations and Turing's paper on morphogenesis inspired research in reaction-diffusion equations, pattern formation, and traveling waves. Sequencing and reconstruction of the human genome created new questions in probability, statistics and combinatorics. All three of these major influences continue today. In this century, the development of new core mathematics stimulated or inspired by biology has been increasing rapidly as more core mathematicians have gotten acquainted with and involved in biological problems. Biology has created fundamentally new questions in statistics and stimulated the field of algebraic statistics. The issue of how to compare teeth in paleontology led to new questions in conformal geometry. The transport of materials in axons led to new phenomena and theorems in partial differential equations. The theory of biochemical reactions stimulated new theorems in dynamical systems, and in queueing theory. The problem of how to compare different proposed phylogenetic trees led to the development of geometric central limit theorems on nonsmooth spaces. Since biological dynamics is very complicated and often parameters are known only approximately or not at all, one needs new coarse-grained methods for the classification of dynamical systems. The issue of how to detect the shape features of proteins stimulated new methods for the shape analysis of surfaces. The effort to understand central pattern generators in the nervous system led to new work exploiting groups of diffeomorphisms to characterize symmetries in the solutions of dynamical systems. The problem of providing low dimensional approximations for very large data sets has led to new questions in harmonic analysis.*"

These are but a few examples of how mathematics has been challenged and enhanced by addressing problems in biology. Although no one can predict the future, it seems reasonable to suppose that the very difficult problems that arise from the need to understand causal G2P relationships will pose new and diverse challenges for mathematics. Many new mathematical methods have been developed over the last 20 years which are not taught other than at the graduate mathematics level, although they would be very useful in addressing some of the G2P problems: (1) Differential geometry for the analysis of phenotypes through imaging thus allowing the study of images through deformations and invariance properties. (2) "Shape analysis" for the study of cells. (3) Algebraic geometry for identifiability (for instance for Gaussian Bayesian networks). (4) Sparsity and optimization techniques for over-parametrized systems. (5) Bayes factors and Bayesian information criteria for model generalizability and inference. (6) Data integration methods through

generalized singular value decomposition and the design of specific, biologically meaningful metrics for heterogeneous data. (7) Graph Limit theory for large graphs.

Many biological problems push the limits of current research and require new Mathematics to be developed. These will take longer to develop and the benefits of doing so will only be recognized later, but underdeveloped areas seem to include: a) Topology of Dynamical systems. b) Geometric measure theory for incorporating non-uniform distributions into persistent homology research. c) Asymptotic theory for exponential random graph models (linked to computations of partition functions in statistical physics). d) Identifiability in systems of stochastic differential equations.

A major challenge for mathematics comes from the realization that our ability to understand G2P relationships suffers from a lack of concrete approaches to modeling multiscale processes in biology (different time-scales, different spatial scales, different processes): no one is really doing this well, nor do we even have good ideas about how to approach this. It is likely that we will not be able to fully understand G2P unless this can be done, and we see a major role for mathematicians in developing techniques and concepts for multiscale modeling whose usefulness would far exceed the problems arising from G2P.

Resource and Infrastructure Needs

A. Bringing Biological and Mathematical Sciences Together: Challenges and Opportunities in the Study of G2P Relations

Human Resource Infrastructure. Significant potential exists for collaboration and synergism among the biological and mathematical sciences in investigating and resolving the difficult problems associated with the relationships between genes and traits. Indeed, such collaboration is essential to move our understanding of causal G2P relationships forward.

There are, however, major challenges in developing the right questions to ask, the methodologies to address the diversity of problems, and developing the right kinds of collaborations among biologists and mathematical scientists.

Collaborations among biologists and statisticians are already widespread, and despite excellent examples of collaboration among biologists and mathematicians these are substantially more scarce. Part of the reason for this is that mathematicians are typically enrolled late in the process, after the experiments have been designed and largely completed. They are then asked whether it would be possible to "model" the system under investigation. This makes mathematicians look like "helpers" rather than primary investigators, which is unattractive considering the potentially large time investment involved.

Biologists are asking mathematicians to "take a bet" on (what for the biologist is) an interesting problem; but it is difficult for a mathematician to get a biologist interested in doing the right kinds of experiments. To be effective, and to give mathematicians an investment in the problem at hand, they should be involved from the very beginning in the design of experiments, the definition of questions and the choice of data to be gathered. Another reason for the scarcity of collaborations is that most biologists know little mathematics and have little appreciation of what math can bring to the table. Statisticians in a way have an easier role because they deal with practical issues and thus have a different relationship with biologists than mathematicians do who (rightly or wrongly) seem, to a biologist, to be more abstract and theoretical. Conversely, if a mathematician has a model, can she get a biologist to test it? This is unlikely, for the same reason that mathematicians would be reluctant to dive into a biological problem; the overhead and investment are just too large.

There are currently few reward systems for this kind of collaboration.
A long-term program to enhance and facilitate the kinds of Bio-Math collaborations that are essential for moving G2P problems forward would be to develop foundations by promoting teaching of a broad array of mathematics in the context of introductory biology classes (and not just see mathematics as a co-requisite for a biology major, nor as just calculus and linear algebra, but geometrical and topological intuition as communicated in useful `pictures'). It should be a two-way street, involving cross-training of young "life scientists" and "mathematical scientists" so they can talk effectively with each other, and understand what can and cannot be done experimentally and theoretically.

There are already many applied mathematicians who work on biological problems, although, apart from work on gene regulatory networks, few if any work on problems relevant to G2P. These investigators are often motivated by discovering the general principles by which a biological system operates, which is exactly what is needed. It is among this group that, in the short term, effective collaborations on G2P problems could also be developed, broadening the type of mathematics used beyond the simple framework of PDEs to geometrical, topological and algebraic approaches.

B. The need for an interactive community. The vast field of G2P encompasses molecular biology, genetics, physiology, biochemistry, biophysics, development, morphogenesis, evolution and many other areas of biology. Just as individual biologists don't know all of Biology, mathematicians may not be aware of the diversity of techniques and algorithms that might suitable to address the diverse problems that arise in understanding how genes affect higher level traits. A user-friendly environment, perhaps on-line, that enables communication, problem sharing and brain-storming among investigators working on G2P problems is therefore desirable.

C. An infrastructure for encoding and archiving phenotypic data. Methods need to be developed to describe or encode phenotypes in a way that is relevant to mathematical methods that study the causes and consequences of those phenotypes. A resource perhaps analogous to the Protein Data Bank (PDB: http://www.wwpdb.org/) or to BioConductor (BioC: http:// https://www.bioconductor.org) could be useful, but needs to be developed in consultation with the diversity of users.

D. A software environment for dynamical multiscale systems. We need a flexible, expandable open source software environment for analyzing and simulating G2P mechanisms. The need for such software is evident by the fact that several research groups have attempted to develop software specifically designed to simulate complex biological mechanisms, especially at the cellular and subcellular level (e.g. *Virtual Cell* http://www.nrcam.uchc.edu/, *BioNetGen* http://bionetgen.org/index.php/, *CellBlender* http://www.mmbios.org/index.php/cellblender-1-0-1 and *E-Cell* http://www.e-cell.org/), but none of these have been widely adopted, nor do they provide for the broad diversity of multi-scale biological processes and interactions that lead to complex phenotypes [23].

Some multiscale problems lend themselves to agent-based modeling [24], but current techniques are far too slow for the kinds of problems that need to be addressed. Methods need to be developed to speed up agent-based models.

What is needed is a programming environment that can handle diverse multiscale problems, support rapid prototyping of models (perhaps something like Julia [MIT http://julialang.org/], or an open-source version of Simulink [Mathworks], or extra modules added to the already exemplary "R" environment, for instance an R4G2P Systems Biology project similar to the successful 'omics system BioConductor [25]. It would contain algorithms and libraries, be extensible open-source, and accessible from different programming languages. It should be a platform in which anyone can develop kinetic models that operate at many scales of the organizational hierarchy, from molecules to cells to organisms. It should also contain programs that can translate conceptual and mental models into appropriate mathematics.

Conclusions

Phenotypes are extremely diverse, they include both structure and function, and occur at all levels of organization. Phenotypes are generated by complex, multivariate, multilevel, nonlinear processes in which gene products, environmental factors and context all play significant roles. This makes the relationship between genes and traits, and between genetic variation and phenotypic variation, exceptionally difficult to understand and unravel.

Specific goals and challenges we identified are the following. Understanding the causal pathways by which genes connect to traits will require mathematical modeling and a close collaboration between biological and mathematical scientists.

Mechanisms and infrastructure need to be developed to identify fruitful areas of collaboration and to encourage and reward such cooperation. Phenotypes are dynamical systems but there is a sparsity of kinetic data on the development and function of most phenotypes. Much of the data available today has insufficient spatial and temporal resolution to develop adequate mathematical models. In particular we need accurate time-series and longitudinal studies of gene expression, with the accompanying proteomics and metabolomics, and visual data on the accompanying phenotypes. An additional pressing need are mechanisms and tools to describe and quantify phenotypes, at all levels of the organizational hierarchy, which are relevant to mathematical modeling. Although high-throughput 'omics data are a voluminous and ever-growing resource, it is not clear that they are sufficient to uncover or deduce the causal pathways that link genes and traits. There is a need for developing techniques for multi-scale modeling and improved agent-based modeling. Research on the causal processes that lead to phenotypes would be much enhanced by the availability of programming environment, perhaps analogous to R, that is open-source, expandable, accessible to users with different skill levels, and capable of multiscale programming.

Workshop Organizers

Fred Nijhout, Biology, Duke University
Fred Adler, Mathematics, University of Utah
Ken Dill, Laufer Center for Physical and Quantitative Biology, Stony Brook University
Susan Holmes, Statistics, Stanford University

Workshop Participants

Fred Adler, Mathematics, University of Utah
Kenneth Lange, Biomathematics, UCLA
Elizabeth Purdom, Statistics, University of California, Berkeley
Fred Nijhout, Biology, Duke University
Gheorghe Craciun, Biomolecular Chemistry, University of Wisconsin
Grzegorz Rempala, Matematics**,** Ohio State University
Jané Kondev, Physics, Brandeis University
Jeff Gore, Physics, MIT
Jeremy Gunawardena, Systems Biology, Harvard Medical School,
John Tyson, Biological Sciences, Virginia Tech.
Ken Dill, Laufer Center for Physical and Quantitative Biology, Stony Brook University
Kiisa Carla Nishikawa, Bioengineering, Northern Arizona University
Konstantin Mischaikow, Mathematics, Rutgers, The State University of New Jersey
Melina Hale, Organismal Biology and Anatomy, University of Chicago

Neda Bagheri, Engineering, Northwestern University
Sally Mood, Anatomy and Regenerative Biology, George Washington University
Stanislav Shvartsman, Chemical and Biological Engineering, Princeton University
Susan Holmes, Statistics, Stanford University
Wolfgang Huber, Multi-omics and Statistical Computing, EMBL Heidelberg, Germany
Yuliy Baryshnikov, Mathematics, University of Illinois

References

1. Gunawardena J: **Models in biology: 'accurate descriptions of our pathetic thinking'**. *BMC Biol* 2014, **12**(1):29.
2. Savage N: **Proteomics: High-protein research**. *Nature* 2015, **527**(7576):S6-S7.
3. Wagner A: **Robustness and Evolvability in Living Systems**. Princeton: Princeton University Press; 2005.
4. Whitacre JM: **Biological robustness: paradigms, mechanisms, and systems principles**. *Frontiers in Genetics* 2012, **3**:67.
5. Lehner B: **Genes confer similar robustness to environmental, stochastic, and genetic perturbations in yeast**. *PLoS ONE* 2010, **5**(2):e9035.
6. Nijhout HF: **The nature of robustness in development**. *Bioessays* 2002, **24**(6):553-563.
7. Stewart AJ, Parsons TL, Plotkin JB: **Environmental robustness and the adaptability of populations**. *Evolution* 2012, **66**(5):1598-1612.
8. Nijhout HF, Reed MC: **Homeostasis and dynamic stability of the phenotype link robustness and plasticity**. *Integr Comp Biol* 2014, **54**(2):264-275.
9. West-Eberhard MJ: **Developmental Plasticity and Evolution**. New York: Oxford University Press; 2003.
10. Klingenberg CP: **Dominance, nonlinear developmental mapping and developmental stability**. In: *The biology of genetic dominance.* Edited by Veitia RA. Austin, TX: Landes Bioscience; 2004: 37-51.
11. Gilchrist MA, Nijhout HF: **Nonlinear developmental processes as sources of dominance**. *Genetics* 2001, **159**(1):423-432.
12. Kacser H, Burns JA: **The molecular basis of dominance**. *Genetics* 1981, **97**(3-4):639-666.
13. Kim Y, Iagovitina A, Ishihara K, Fitzgerald KM, Deplancke B, Papatsenko D, Shvartsman SY: **Context-dependent transcriptional interpretation of mitogen activated protein kinase signaling in the Drosophila embryo**. *Chaos* 2013, **23**(2):025105.
14. Delude CM: **Deep phenotyping: The details of disease**. *Nature* 2015, **527**(7576):S14-S15.

15. Tickle C: **Making digit patterns in the vertebrate limb**. *Nat Rev Mol Cell Biol* 2006, **7**(1):45-53.
16. Craciun G, Kim J, Pantea C, Rempala GA: **Statistical model for biochemical network inference**. *Communications in statistics: Simulation and computation* 2013, **42**(1):121-137.
17. Reinitz J, Sharp DH: **Mechanism of eve stripe formation.** *Mech Dev* 1995, **49**:133-158.
18. Fogelberg C, Palade V: **Machine learning and genetic regulatory networks: a review and a roadmap**. In: *Foundations of Computational, Intelligence Volume 1.* Edited by Hassanien A-E, Abraham A, Vasilakos A, Pedrycz W, vol. 201: Springer Berlin Heidelberg; 2009: 3-34.
19. Diakonis PW, Holmes SP: **Matchings and phylogenetic trees**. *Proceedings of the National Academy of Sciences* 1998, **95**:14600-14602.
20. Billera L, J., Holmes SP, Vogtmann K: **Geometry of the space of phylogenetic trees**. *Advances in Applied Mathematics*, **27**:733-767.
21. Pevzner P, Tesler G: **Genome rearrangements in mammalian evolution: lessons from human and mouse genomes**. *Genome Res* 2003, **13**:37-45.
22. Evans SN, Steinsaltz D: **Estimating some features of NK fitness landscapes**. *Annals of Applied Probability* 2002, **12**:1299-1321.
23. Chase K, Carrier DR, Adler FR, Jarvik T, Ostrander EA, Lorentzen TD, Lark KG: **Genetic basis for systems of skeletal quantitative traits: Principal component analysis of the canid skeleton**. *Proc Natl Acad Sci U S A* 2002, **99**(15):9930-9935.
24. Heard D, Dent G, Schifeling T, Banks D: **Agent-based models and microsimulation**. *Annual Review of Statistics and Its Application* 2015, **2**(1):259-272.
25. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T *et al*: **Orchestrating high-throughput genomic analysis with Bioconductor**. *Nat Methods* 2015, **12**:115-121.