The grand challenge of predicting phenotype from genotype and environment

Amanda Wilson Carter[1], Kostya Kornev[2], Laura Rusche[3]

1. University of Tennessee, 2. Clemson University, 3. The State University of New York, University at Buffalo

Summary

The phenotype, or traits, of an organism is the manifestation of its genome, environment, and their multidirectional interactions, and it dictates how the organism interacts with its world (e.g. behavior, physiology, fitness). As such, it is the key descriptor of a biological organism. Biologists have attempted for decades to accurately and meaningfully predict phenotype given genotypic and environmental inputs. Past efforts have focused on manipulating individual genes or environmental conditions in isolation and measuring the response of a single phenotypic trait. Although this approach has enabled significant strides in our understanding of gene by environment interactions, we are not yet able to accurately and meaningfully forecast complex phenotypes in nature. At first glance, such an achievement would require infinite genetic and environmental inputs, endless phenotypic outputs, and exhaustive modeling that are beyond the scope of current datasets and computational capabilities. We view this barrier as an issue of complexity. In this vision paper, we propose a strategic approach to tackling this age-old question using emerging technologies. Specifically, we encourage systematic studies that hone in on the genetic and environmental factors that disproportionately affect phenotype and the phenotypic traits that disproportionately affect fitness with the goal of **reducing overall complexity without significantly compromising our ability to predict phenotype in nature.** Given that these key determinant factors are likely to vary across the tree of life, we also propose making the prediction pipeline modular such that the modeling parameters are suitable for the taxonomic group. This effort will help reintegrate biology, as phenotype is an important explanatory concept across biology and uncovering how it emerges will require contributions of scientists from multiple subfields.

What would predicting phenotype entail?

Predicting phenotype is one of the grand challenges in biology. In essence, the goal is to predict the traits of an organism based on its sequenced genome and known environment. This might initially be possible for an individual of a well-characterized species. Eventually, with an advanced prediction model, it might also be possible to reconstruct an unknown organism, such as a microbe that has not been successfully cultured or an extinct animal that left behind DNA in fossil bones. It might also be possible to predict the phenotypes of many organisms in a community. For example, one could analyze a sample of pond water and describe the ecology of the pond based on the chemical composition of the water and the genomic sequences of resident microbes and sloughed cells from larger organisms. Phenotype prediction could also occur at smaller scales. For example, one might predict the metabolic and physiological properties of individual cells or tissues within a multicellular organism. Doing so would require an understanding of the developmental program of that organism.

The ability to predict phenotype would have theoretical and practical benefits

There are both theoretical and practical benefits to predicting phenotype. On the theoretical side, given that phenotype is the observable manifestation of underlying biological processes, a deeper understanding of the dynamics that shape phenotype would provide insights into fundamental biological processes, i.e. the Rules of Life. In addition, this predictive ability could reveal how phenotypes evolve. For example, to understand the path by which a particular suite of adaptive phenotypes evolved, one could computationally evaluate how a sequence of genotypic changes altered phenotypic potential. Similarly, understanding how phenotype arises might reveal novel trait spaces (i.e. the biology of what does not exist).

The ability to predict phenotype also has practical applications, including in medicine. An individual's genotype could indicate their risk of developing particular diseases and allow prophylactic strategies (changes in environment) to be implemented. Similarly, genotyping of tumors could enable selection of drugs and other therapies with greater specificity and effectiveness, and genotyping of infectious microorganisms could indicate which antibiotics would be most potent.

The ability to predict phenotype would also be useful in the face of climate change. In contrast to the examples above, in which organisms of varying genotype are compared in similar environments, in this case the goal would be to predict how environmental change shapes phenotype given a particular genotype. Uncovering the potential impacts of climate change for organisms, species, communities, and populations would help prioritize limited management and conservation resources.

Finally, the ability to predict phenotype could expand the possibilities and solutions within synthetic biology and biotechnology. Bioengineers could create recombinant organisms with the appropriate genotype and raise them in an appropriate environment to achieve a desired beneficial phenotype. Towards this goal, researchers have been developing genome-scale models (GEMs) of metabolite, redox, and energy fluxes in microorganisms and identifying the genetic contributions to complex traits in agriculturally important plants and animals. Ultimately, engineered recombinant organisms could help improve the global food supply, produce renewable energy, or synthesize medicines cost effectively.

The project of learning to predict phenotype would help reintegrate biology

Since its creation, biology has been divided into subfields that operate independently. Nevertheless, the genotype/environment/phenotype triangle remains an important explanatory concept for all biology, meaning that practitioners in many subfields have a stake in uncovering the principles that govern relationships within this triangle. Moreover, uncovering these principles will require orchestrated efforts of scientists from across biology, synthesizing their knowledge, methodologies and techniques. Working towards this common goal, a greater understanding of the language, principles and culture developed in other fields will emerge, stimulating further collaboration. For example, biophysics and biochemistry can provide physico-chemical mechanisms of self-assembly of biological molecules, molecular biology and genetics reveal how genomic information is interpreted to produce varying sets

of proteins in response to environmental signals, developmental biology explains how genomic information is interpreted as a morphology, and ecology describes how organisms interact in an environment to develop a specific phenotype. Through the theory, computational modeling, and artificial intelligence algorithms, one can bridge together the knowledge gained from these fields to identify the pathway of development of the genotype/phenotype/ environment triangle to predict its evolution in an unknown environment. Thus, synthesizing the knowledge generated by the fields of biophysics, biochemistry, molecular biology, genetics, development, organismal biology, and ecology, this study will reintegrate the biological subfields.

The quest to predict phenotype from genotype and environment will also have broader impacts on the biological community and infrastructure. For example, the excitement of solving this grand challenge may attract new people to the field, and the interdisciplinary nature of the problem will provide excellent training opportunities for these people to develop a holistic view of biology. In addition, this research is expected to boost future innovations as scientists strive to collect data that accurately describes genotype, environment, and phenotype and to develop computational tools to analyze and model these data.

Emerging technologies make this a good time to undertake the challenge of predicting phenotype

The quest to predict phenotype is as old as the field of genetics, but scientists have not yet been able to satisfactorily predict whole-organism level phenotype given seemingly infinite inputs from genes and environment. However, we are currently in the wake of a technological revolution that may allow us to tackle this problem strategically. In particular, there have been significant advances in the tools with which we can measure each player in the genotype/environment/phenotype triangle, as well as advances in computational modeling and artificial intelligence. In terms of *genotype*, genomes can now be sequenced and assembled in a short time at relatively low cost due to the development of high throughput sequencing technologies, including approaches that generate long reads. Moreover, it is increasingly possible to alter genomes using techniques such as CRISPR, which makes it possible to assess the relevance of particular genotypes to phenotype. In terms of *environment*, abiotic data can be collected by Lidar, GIS and other spatio-temporal mapping techniques, and biotic information can be evaluated using social networking models to characterize the contributions of conspecific interactions. Finally, in terms of *phenotype*, large-scale data collection is increasingly possible at multiple scales. At the molecular level, chromatin modifications, RNA abundance, and protein levels and modifications, and metabolites can be measured using high-throughput sequencing and mass spectrometry. At the cellular level, advanced immunological and metabolic assays are available. And at the organismal level it is possible to conduct complex morphometric analyses and behavioral assays. Nanotechnology offers unprecedented opportunity to intervene and trace minute changes of phenotype at the nanoscale level of macromolecules and cellular membranes. Nanotechnology made it possible to develop new physico-chemical methods allowing one to probe phenotype variation caused by minute changes in genotype or environment at different scales at the molecular, nano-micro and organismal level. Computational hardware and software abilities enable to boost experimental diagnostics and to bridge multiscale phenomena with different types of stochastic and deterministic models.

Overcoming the challenges to predicting phenotype

One issue that has daunted would-be predictors of phenotype is the sheer complexity of the problem. Complexity compounds across levels of biological organization, from the gene to the organism and beyond. A typical genome encodes thousands of functional macromolecular complexes, each of which will contribute to phenotype. The genes encoding these macromolecular complexes vary among individuals of a species and are only expressed at certain times and in certain amounts. In multicellular organisms, the same genome is expressed differently in each cell type, and therefore decoding the developmental program is also necessary to predict phenotype. The environment also presents complexity. It is composed of multiple abiotic components whose abundance and state varies, as well as biotic components, such as other organisms that may influence the organism cooperatively or competitively. In addition, phenotype itself feeds back on environment, as an organism can alter its surroundings. Therefore, it will be necessary to strategically reduce the overall complexity of models by identifying factors that have a disproportionate effect on phenotype. These factors will likely have some level of taxonomic specificity.

A few key strategies that might be used to reduce complexity are offered here. First, predictive models could focus on *key phenotypic traits*, depending on the application. In the context of medicine, models would represent disease likelihood for people and susceptibility to drugs for pathogens and cancer cells, whereas in the context of ecology, traits that have an outsize effect on fitness would be key. These traits could be identified by evaluating genetic signatures of selection to identify genes and hence pathways that have been under selective pressure. A second strategy to reduce complexity would be to identify the *principal environmental conditions* that underlie phenotypic variation. These could be deduced through systematic experiments, and likely include the availability of nutrients and habitable temperature. Third, a similar strategy could be used to identify those *genes that have outsize effect* on phenotype and model building could rely on these genes. By developing predictive models using these select inputs, we can take the first steps in predicting relevant phenotypes.

Another issue that could hinder *de novo* prediction of phenotype starting with the genome of an unknown species is that the understanding of gene function and gene-gene interactions would be lacking for that organism. To address this issue, we propose building a database of guiding genetic annotations. This would consist of a catalog of genes, their functions, and their interactions in strategically selected species across the tree of life. This “guide information” would allow the genetic potential of other species to be deduced based on informed assumptions that homologous proteins will behave similarly. To build such a database, it would be necessary to identify and characterize guide species at strategic points throughout the tree of life.

It will also be necessary to determine how allelic variation within a single gene affects its function. Alleles with premature stop codons or missing splice sites are likely to be non-functional. However, for simple amino acid substitutions, it will be necessary to understand when these substitutions are likely to interfere with the function of the protein. Clues may come from the extent to which an amino acid is conserved across homologous proteins and our developing understanding of how proteins fold.

An important consideration in developing predictive models is whether to use a deterministic or a stochastic model or both. A deterministic model would imply that, given a particular genotype and environmental state, there is one dominant phenotypic outcome. In contrast, a stochastic model would allow for several possible phenotypes, each with a certain probability of occurring. It is already clear that stochastic heterogeneity is normal in biology, due to noise in gene expression and other physiological processes. Therefore, a stochastic model is more representative of reality; however, to increase the level of predictability and decrease the uncertainty of input data, this stochastic model should incorporate the experimentally validated deterministic models based on the fundamental physico-chemical and biological laws.

Just as a given set of genotypic and environmental conditions could produce more than one phenotype, a single phenotype could be produced by more than one set of genotypic and environmental conditions. One manifestation of this phenomenon is convergent evolution, in which different species separately evolve the same trait. However, this phenomenon also occurs at other scales. For example, an animal might have the same behavioral response to several different situations, and a cell might differentiate similarly in response to several different stimuli. Given this situation, the challenge in developing predictive models is to recognize all the causal factors that contribute to a given phenotype. Therefore, it will be necessary to apply a rich arsenal of mathematical methods and computer learning algorithms so that the genotypic and environmental cues can be elucidated.

Another challenge is that phenotype is not static. Individuals may express transient phenotypes when acclimating to their current environment or exhibit shifts in phenotype across ontogeny. Further, environmental conditions such as temperature and nutrient availability can alter the course of development, or how an organism can respond to future environmental conditions (i.e. developmental plasticity). Similarly, the environmental experiences of an organism can potentiate the future expression of particular genes through epigenetic memory. To address this issue, it will be necessary to include three types of inputs in prediction models. In addition to genotype and environmental conditions, past experiences will need to be incorporated. As with the genetic and environmental factors, it will be important to identify those historical experiences that disproportionately affect phenotype and include those in the model.

Finally, it will be necessary to build an ethical and legal framework to guide the use of predictive capabilities. This is particularly true in the case of human traits. For example, a person could be denied health insurance because they have a high likelihood of developing a disease that is costly to treat, or they could be directed into particular occupations based on inferred capabilities. Moreover, recent reports indicate that DNA profiling is being used to track and regulate Chinese citizens.