Towards a theory of human intraspecific variation for ergonomics and human modeling

By Bradly Alicea

Department of Telecommunication, Information Studies, and Media

Cognitive Science Program

Michigan State University, East Lansing, MI. 48824 USA

Keywords: Biological Variation, Digital Modeling, Intelligent Systems,
Anthropometry, Evolutionary Modeling, Morphometrics

ABSTRACT

Human intraspecific variation is a complex problem, but may be better understood by using computational models in tandem with knowledge about the genetic bases of phenotypic traits. These results can be used in a multitude of settings. To move closer to this goal, biologically-realistic mappings between genotype and phenotype are constructed using genetic algorithm and neural network-like models. These models allow for gene-gene and gene-environment interactions to be characterized in the resulting phenotype. Two types of model are introduced: a simple, two-layer model, and a more complex model. The final section will focus on trends of growth and development in relation to relationship to modeling anthropometric traits and other morphological phenomena.

INTRODUCTION

Decoding human variation is a complex problem, which becomes even more difficult when attempting to characterize large numbers of traits. Sequences of the Human [1,2] and Chimpanzee [3] genomes along with knowledge about how genes are expressed and advanced population-level modeling methods provide some of the necessary tools and raw data to characterize multivariate biological variation as a set of models and/or design principles.

In this paper, a computational model will be introduced that characterizes the problem at hand. While this model has not been formally implemented, its representational aspects can be used to answer a number of related questions.

This big picture understanding of human variation has a multitude of applications: drug discovery, forensics, biometrics, and ergonomics represent only a few potential application domains [4]. One way to solve this problem may be to focus on intraspecific variation, which deals with observable variation across *Homo sapiens* in the context of evolutionary processes and mechanisms.

METHODS

The working hypothesis is that models inspired by evolutionary theory and the genetic bases of traits can be used to derive or suggest ways of simulating and simplifying multi-trait human variation for use in digital modeling. This may be especially relevant to ergonomic design, where principles already exist regarding varieties of anthropometric traits without reference to their genetic underpinnings. However, there are several components that need to be in place before the benefits, drawbacks, and practical applications of the approach become clear.

P-TO-G AND G-TO-P MAPPINGS

The first such component is to link observable traits, which in the application domain are generally phenotypic, to their presumed genetic bases. In this way, a clear set of P (phenotype)-to-G (genotype) and G (genotype)-to-P (phenotype) mappings can be created from purely anthropometric and more complex datasets alike.

ESTABLISHING THE MAPPINGS: MAJOR FEATURES

To establish P-to-G mappings in models of the human operator, static anthropometric variables must be distilled into a set of binary or otherwise discrete genotypic representations. In cases where gene expression data is available, using the established P- to-G mapping may also allow for G-to-P mappings to be established that are more than simply the inverse of P-to-G mappings.

Conceptually, the inverse G-to-P mappings resemble quantitative trait loci (QTL) analyses that have been performed on a number of animal models. QTL analyses on phenotypic traits in dogs [5] and Japanese black cattle [6] have determined that multiple genetic loci control the growth and development of pelvic and limb bones in the former example and various phenotypic traits in the latter. These results are relevant to digital modeling in that specific complex traits such as the upper body, head and neck, and lower body can be reproduced using computational models that are sensitive to complexity and evolutionary dynamics.

The P-to-G mappings are based on a standard method in genetic algorithm research [7, Chapter 2]. Genetic algorithms (GAs) utilize a representational approach to achieve this mapping. In their simplest form, phenotypic attributes are mapped to single loci, or short sequences of continuous/discrete values each arranged

along a one-dimensional lattice. In GA research, this structure is called a chromosome, and a system composed of either single or multiple chromosomes is called a genotype.

Such a system can be loosely referred to as a population; in this paper, there will be a subtle distinction between groups of agents and subdivision that is meaningful to the production of variation. Functionally speaking, adjacent loci are often said to be linked in the sense that these loci co-evolve and are conserved as a single unit. An essential feature of each locus is that they can undergo mutation and take on alternate forms. These alternate forms at a single locus are also known as alleles; each allele acts as a discrete state relative to an entire system of genotypes.

In a population of digital agents, each with their own genotype, each allele exists at relative frequencies. The relative frequencies of the alleles in a population determine variability expressed in the phenotype. Changes in these frequencies can be enforced or otherwise altered using a fitness function; typically in GA research, this function governs fitness in the entire population of genotypes. In nature, fitnesses vary between local populations and over time.

FEATURES IN MAPPINGS: P-TO-G EFFICIENCY

One problem with establishing P-to-G mappings is that of computational efficiency. This has been already been investigated in the context of determining the most efficient mappings between phenotype and genotype [8]. The key to establishing efficient and realistic representations of variation is to exploit this digital linkage mechanism and make it salient at the phenotypic level. The phenotype can be best characterized computationally as a hierarchical and modular system. This yields a functional

and structural phenotypic representation similar to what is expressed by *Hox*-type genes in the animal genome.

Using a model inspired by homeotic function to approximate human variation provides us with two potential mechanisms that bridge the gap between genotype and phenotype: a temporally and topologically co-linear genetic representation and a system that is robust to combinatoric expression [9, Chapter 21].

A co-linear genetic representation refers to closely related traits being linked together on the same chromosome. These genetic units get expressed or actively mapped in temporal order as well; this provides an open architecture that is sensitive to modeling developmental parameters. Co-linear representations preserve the functionality of genetic units at multiple scales as well as interrelationships between complex traits.

Preserving the potential for combinatoric expression in digital representations allows for two alternate ways to manipulate the production of variation: unconstrained expression leads to the production of emergent patterns from relatively simple initial P-to-G mappings, while constraining expression selectively leads to the production of specific traits. Specific mechanisms for modeling gene expression in a G-to-P mapping will be discussed in the section on growth patterns.

ADVANCED FEATURES OF MAPPINGS: LINKAGE DYNAMICS

Once P-to-G mappings are modeled using genetic algorithms, statistical linkage and linkage disequilibrium can be addressed in more detail. While linkage refers to loci that evolve in tandem, linkage disequilibrium involves groups of loci that exist at allelic frequencies different than the normal equilibrium for that population [10]. This has implications for how specific traits

are distributed in various populations, which may provide insights into individual differences.

Population-based subdivision may provide insight into two ways. First, it may tell us how variants are influenced by natural selection at the microevolutionary level. Second, it might also reveal under what conditions various classes of traits become associated with one another [10]. Any model of the phenotype derived from genotypic inputs must include multiple agents, each with their own distinct genotype but forming distinct populations subject to different selective and environmental pressures. These agents are allowed to mate and reproduce, so that populations occasionally admix, or exchange members.

This may also allow us to replicate the forces of evolution at the microevolutionary level, especially when selection, drift, and recombination are uneven across groups and produce disequilibrium between populations for specific trait variants in or within various hot spots among specific populations.

GxG and GxE INTERACTION MODELS

Gene-environment (GxE) and gene-gene (GxG) interaction models for multi-locus genotypes [11, Chapter 1] can assess the amount of epistasis in a given G-to-P mapping. Epistasis can be thought of as the interaction term that results from the activity of multiple genes during their expression. In theory, gene-gene interactions involve the interaction of genes as they relate to overall fitness. Gene-environment interactions involve genes and environmental as they relate to overall fitness. The relative degree of nonlinearity in this mapping determines the amount of epistasis in a given system.

Given this observation, the degree of

isomorphism between the genotypic and phenotypic representations can be either maximized or minimized as a function of combinatoric expression. Any mapping from phenotype and genotype and vice versa will involve complex interactions [12], and can be considered to be an inverse problem.

ONE POTENTIAL IMPLEMENTATION

This section will introduce two variants of a basic model that could be used to produce digital human biological variation. These models are at various stages of implementation; however, they represent a step closer to effectively modeling human biological variation using digital tools.

The implementation presented here will focus on the potential for epistatic relationships between genes in the G-to-P mapping. In purely theoretical cases, epistatic can be either additive or multiplicative. To make things even more complex, an array of interactions between four or more genes can exhibit multiple stable states so that some interactions have additive effects while others have multiplicative effects on the system [13].

SIMPLE EPISTATIC MODEL

The simple epistatic model involves connecting the G-to-P mappings to a model resembling a two-layer feed-forward neural network (Figure 1). The model presented here might be used to simulate both GxG and GxE interactions.

The input layer in the presented example is composed of all the genes on a single GA chromosome. An additional input unit can be used to introduce stochastic noise, and approximates simple environmental inputs. These inputs are equivalent to simple thermal, haptic, or energetic inputs. These inputs tend to deliver bursts of stress to the system that can act to disrupt interactions

between the input and output layers.

Each element in the output layer represents a phenotypic property of a complex trait. Epistatic interactions are characterized an array w_{ij} representing a connectivity matrix; these links from the input to output layer represent the weighted contributions of each gene to each feature in the phenotype. This connectivity can be modified by pruning the network based on a threshold value; this threshold represents the weighted contribution of each gene in the system to each phenotypic unit.

Figure 1: An example of the simple epistatic model (connections between layers arbitrary).

Epistatic interactions can also be determined by a set of rules that enforce a specific fitness function. For example, only certain genes can map to certain phenotypic elements much in the way *Drosophila* specimens has been shown to have abdominal segments that correspond to tightly linked genes.

A single gene or family of linked genes might map to a corresponding phenotypic segment in a way that resembles the expression of repeatable elements. Examples of such modular systems include the human vertebral column or a striping pattern on the lateral surface of a fish.

There are actually two ways to mimic the modulatory effects of environment in

the simple epistatic model. One way is by using the input element labeled S_n in Figure 1 as a stochastic source of non-genomic variation. Unfortunately, this only serves as a very general stand-in for environmental factors. The other is to vary the threshold value on the connection matrix in real-time so that these values represent the up- and down-regulation of genes relative to the production transcripts for a specific gene. In other words, some connections are silenced while others are reinforced while the phenotype is being constructed.

COMPLEX EPISTATIC MODEL

The simple epistatic model may often be too simple to produce any consistent or robust results. A more complex model was then developed to account for gene products and signaling dynamics. This model differed from the model introduced for the simple case in that several hidden layers were introduced (Figure 2).

Figure 2: An example of the complex epistatic model (connections between layers arbitrary).

The input and output layers are identical to that used in the simple model. In the complex epistatic model, the input and output layers are identical to the stripped-down model. The first hidden layer consists of two sets of subunits: gene products (T_n) and environmental inputs (E_n). The benefit

of using multiple units for environmental inputs is that particular ecological or task-specific contexts can be modeled. The second hidden layer also consists of two sets of subunits: receptors (R_n) and interference elements (I_n). The benefits for including receptors and interference elements into the G-to-P mapping is that genes can be silenced through indirect mechanisms, which may allow for fine tuning of the phenotypic elements in the output layer.

In the current formulation, the GxG and GxE interactions are not explicitly separated out. In fact, this may be a drawback in the context of some applications. For example, in some implementations the environmental inputs may not provide the same degree of control as seen in nature. Alternate implementations may include modifying the model so that the effects of one interaction type or another can be isolated. These models provide information regarding the effects of different amounts of nonlinearity in the mapping between genotype and phenotype. Manipulating these models may provide information about what elements of variation are due to developmental trajectories and/or environment and what elements are strictly controlled by genetic markers.

In both of these models, as in genetic algorithms, fitness values can be used to enforce which connections in the epistatic interaction array are maintained over time. In theoretical treatments of gene-gene and gene-environment interactions, multiplicative fitness are indicative of additive epistasis, while additive fitnesses are equal to multiplicative epistasis. Therefore, the fitness function cannot be enforced directly on elements that make up the chromosome. In both of these models, the array f_{ij} can be used to weight all corresponding elements in w_{ij} for every iteration of the model.

FURTHER ASPECTS OF THE EPISTATIC MODEL: HUMAN DIGITS AS A “TOY” PROBLEM

A concrete example of how this mapping might be implemented is to walk through a toy problem. This problem involves representing variation seen in the digits of a human hand, specifically creating mappings from a simple epistatic model to a specific morphology. The first step is to the major components of this system to genes on a chromosome and to units in the network.

Multiple encoding strategies can be taken to represent these structures using gene and environment units. However, two general strategies can be taken regarding the mapping. One is to use as few genes/elements as possible in the representation. This strategy would involve two genes; one for producing self-similar elements (i.e. the first through the fifth digit) and the other regulating the relative variance exhibited by individual elements (i.e. the second digit).

While this approach is good for self-similar elements such as human digits, it leaves little room for natural variation. Therefore, a competing approach is to use as many elements as possible, but to maintain some efficiency of encoding so that large-scale patterns such as ratios between digit sizes may be maintained.

Table 1 demonstrates a strategy that has features of both approaches. Five human digits on a single are considered. Both the proximal and distal aspect are taken as individual units, providing ten degrees of anatomical freedom with which to work. Three genetic elements (i,j,k) are used to model the digits. Element i specifies the distal segments, while elements j and k specify the proximal segments. Five environmental elements are used to simulate hormonal and other stimuli that affect the expression of the genetic

elements.

It is of note that multiple units can act to produce a phenotypic element. In fact, their interaction is ensured if multiple input and hidden elements in the model are linked to output elements.

CONSEQUENCES, PRACTICAL APPLICATIONS, AND FUTURE WORK

Implementing digital representations of phenotypic data and then using genetic models to probe this variation *in silico* allows us to treat human variation as an "inverse" problem.

Table 1: table showing G-toP mapping for digits on a human hand.

P-level feature	G-level representation
First Digit, proximal	Gene j x Gene k x Enviroment l
First Digit, distal	Gene i x Gene k x Environment i
Second Digit, proximal	Gene j x Gene k
Second Digit, distal	Gene i x Environment k
Third Digit, proximal	Gene j x Gene k
Third Digit, distal	Gene i x Environment j
Fourth Digit, proximal	Gene j x Gene k
Fourth Digit, distal	Gene i x Environment h
Fifth Digit, proximal	Gene j x Gene k
Fifth Digit, distal	Gene i x Environment i

There are several advantages and

disadvantages to this approach. Producing a representative sample of possible phenotypes as opposed to relying upon static measurements allows us to capture some of the more complex features of the human phenotype. As genomic data becomes available to inform ergonomic research on the human phenotype, this approach may be refined to predict the entire range of human variation, allowing us to characterize the genetic and morphological components of individual differences in populations that lie beyond the 95th percentile [14]. This approach may also provide great insight into non-sampled human variation which may have existed in the past or exists as latent possibilities not immediately apparent given static anatomical measurements.

There are also two possible disadvantages to this approach. It might be too complex for simple assessments of human shape and size that involve less than three of four measurements. The other problem is that this is an idealized method; in some cases, the results may not be biologically meaningful.

ADDITIONAL CONSIDERATIONS

In future studies, it may be possible to characterize growth patterns due to developmental processes. By varying only a few key developmental variables, extra variation can be created in the adult form. Different combinations of these settings might allow us to infer the effects of aging on specific morphological traits. To fully capture the effects of development, hundreds of variables may need to be used. For example, attempts have recently been made in the genetics literature to capture variation and complexity related to development using a hyperspace model [15].

Developmental mechanisms serve to determine how genomic variation maps to the phenotype *in vivo*. The activational

effects of hormones and environmental factors such as nutritional and energetic constraints often interact with the growth of morphology during development. In particular, the rate of growth and overall length of the developmental period determine the shape and size of a given trait in adulthood [16]. It may not be necessary to replicate the entire process of development; a third hidden layer might be used to approximate developmental differences and constraints.

Microevolutionary change and variation is achieved in development as anatomical features are being crystallized. Variation observed in specific traits is often expressed in development through two mechanisms: secular change and allometry. Secular change refers to changes in the growth rate which result in traits which become larger or smaller over the course of several human generations mainly due to environmental factors [17]. Secular change impacts both allometric relationships and trait sizes. Allometric growth can be defined as the proportionality of one trait to another, usually defined as a linear or logarithmic function. A successful method of determining allometry from G-to-P mappings might aid in scaling the outcomes of these simulations to digital models of the human body.

DIGITAL MODELS OF PHENOTYPIC TRAITS

This model should not be thought of as a replacement for empirical measurements. In most cases, real measurements may be needed to initially train the model so that it can be calibrated to reflect the mean and variance for a specific set of traits. The strength of this model lies in the ability to estimate non-expressed or unmeasured variation; Ultimately, implementations of these models may become even more complex, and include information about developmental trajectories in determining the geometry of specific morphological

features in different adult populations. An attempt to map these results to three-dimensional anatomical models is another next step in this approach.

Much as with the QTL studies on cattle and dogs, the ultimate goal of these models is to map specific genes or sets of genes to specific anatomical locations or segments. Using these models, a complex trait can be any significant subsection of the human anatomy; future simulations will compare performance for different body segments. One example would be to compare how the model produces variable but faithful representations of the torso and arms versus those of the pelvis and legs. A recent paper by [18] attempted to use the CAESAR dataset to parameterize and classify static shapes for constructing virtual clothing models. The G-to-P and P-to-G models presented here also allows for the virtual modeling of human populations. The GENESIS approach [19] allows for faces and bodies of avatars of a realistic-looking human population with certain statistical parameters to be modeled. Aspects of the GENESIS system might refine the operational application of the epistatic models presented here. Finally, application to the domain of forensic science might be also be useful, since craniofacial identification and modeling methods currently rely upon empirical measurements of landmarks and soft tissue depths in specific populations [20]. In this way, not only will the simulations inform human factors-type issues, but also allow for a better understanding of how adaptations affect observable and non-observable human variation.

REFERENCES

- [1] Venter, J.G., Adams, M.D. Myers, E.W., and Celera Genomics Sequencing Team (2001). The Sequence of the Human Genome. *Science*, 291(5507), 1304- 1351.
- [2] McPherson, J.D. and International Human Genome Mapping Consortium (2001). A physical map of the human genome. *Nature*, 409, 934-941.
- [3] Waterston, R.H., Lander, E.S., Wilson, R.K. and Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437, 69-87.
- [4] Collins, F.S., Green, E.D., Guttmacher, A.E., and Guyer, M.S. (2003). A vision for the future of genomics research. *Nature*, 422, 835-847.
- [5] Carrier, D.R., Chase, K., and Lark, K.G. (2005). Genetics of Canid skeletal variation: size and shape of the pelvis. *Genome Research*, 15, 1825-1830.
- [6] Malau-Aduli, A.E.O., Niibayashi, T., Kojima, T., Oshima, K., Mizoguchi, Y. and Komatsu, M. (2005). Mapping the QTL for body shape and conformation measurements on BTA1 in Japanese black cattle. *Animal Science Journal*, 76(1), 19.
- [7] Goldberg, D.E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. New York: Addison-Wesley.
- [8] Grefenstette, J. (1986). Optimization of control parameters for genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics*, 16(1), 122-128.
- [9] Gilbert, S.F. 1997. *Developmental Biology*. Sunderland, MA: Sinauer.
- [10] Pearson, J.C., Lemons, D., and McGinnis, W. (2005). Modulating *Hox* Gene Functions During Animal Body Patterning. *Nature Review Genetics*, 6(12), 881.
- [11] Mourant, A.E. (1980). *Linkage*

equilibrium and disequilibrium in human population studies. *Annals of Human Biology*, 7(2), 109-114.

[12] Hartl, D.L. and Clark, A.G. (2006). *Principles of Population Genetics*. Sunderland, MA: Sinauer.

[13] McKinney, B.A., Reif, D.M., Ritchie, M.D., Moore, J.H. (2006). Machine Learning for Detecting Gene-Gene Interactions: a review. *Applied Bioinformatics*, 5(2):77-88.

[14] Ewens, W.J. (2004). *Mathematical Population Genetics*. Berlin: Springer.

[15] Ahituv, N., Kavaslar, N., Schackwitz, W., Ustaszewska, A., Martin, J., Hebert, S., Doelle, H., Ersoy, B., Kryukov, G., Schmidt, S. Yosef, N., Ruppin, E., Sharan, R., Vaisse, C., Sunyaev, S., Dent, R., Cohen, J., McPherson, R., and Pennacchio, L.A. (2007). Medical Sequencing at the Extremes of Human Body Mass. *The American Journal of Human Genetics*, 80, 779-791.

[16] Wu, R. and Hou, W. (2006). A hyperspace model to decipher the genetic architecture of developmental processes: autometry meets ontogeny. *Genetics*, 172(1), 627-637.

[17] Jantz, L.M. and Jantz, R.L. (1999). Secular change in long bone length and proportion in the United States, 1800-1970. *American Journal of Physical Anthropology*, 110(1), 57-67.

[18] Magnenat-Thalmann, N., Seo, H., and Cordier, F. (2004). Automatic modeling of virtual humans and body clothing. *Journal of Computer Science and Technology*, 19(5), 575-584.

[19] Seo, H., Yahia-Cherif, L., Goto, T., and Magnenat-Thalmann, N. (2002). GENESIS: Generation of e-population

based on statistical information. *Proceedings of the Computer Animation*, IEEE Computer Society, 81.

[20] Domaracki, M. and Stephan, C.N. (2006). Facial soft tissue thickness in Australian adult cadavers. *Journal of Forensic Sciences*, 51(1), 5.