

INVITED PRESENTATION

Database Representation of Phenotype: Research Issues

Prakash M. Nadkarni

Yale University School of Medicine, New Haven, CT, USA

Correlation of genotype with phenotype is one of the major goals of clinical genetics. Devising a standard approach to database representation of phenotype facilitates electronic storage and exchange, and eventual analysis, of phenotypic information. To be useful, an electronic representation must be detailed enough to be understood and interpreted unambiguously by scientists who are not experts in a particular disease. While genotype is reasonably straightforward to represent computationally, as variations from a consensus sequence, the extreme variability of phenotypic parameters make devising this representation a challenge.

Phenotypic parameters can be characterized at molecular, organelle, cellular, organ-system or whole-organism level. Because they vary with the gene/s of interest, the total number of parameters would range in the hundreds of thousands. Further, phenotype does not necessarily imply a single causative gene: in multigenic diseases such as diabetes, we have numerous gene-gene and gene-environment interactions.

Phenotypic parameters are typically represented as columns of data in a flat file or database table. To be interpreted by anyone who has not originated that data requires a “data dictionary” – a document that describes in detail what each represents. *The challenge of standardizing the computational representation of phenotype then reduces to the problem of standardizing the descriptions in a dictionary.*

For study-specific experimental parameters: even the metadata descriptors depend on the nature of the parameter. Standard clinical laboratory test descriptors that have been devised for the LOINC (Logical Observations, Identifiers, Names and Codes) vocabulary include, in addition to the parameter name and brief description, the source of the biological sample used for measurement, the timing of the sample (e.g., random, fasting vs. post-prandial, cumulative 24-hr sample) units of measurement, the lower and upper limit of “normal” values, if known, a bibliographic reference to the test method, if standard, and so on. For non-clinical parameters, the situation gets even more complex, because the same parameter may be studied by a variety of laboratory techniques by different investigators, and each technique requires its own descriptors.

While requiring the supply of extensive detail for every single parameter in a phenotypic dataset can be a highly onerous, publication of papers in scientific journal generally mandates such details, and for electronic-publication of results that others may wish to replicate, the standards cannot be dramatically less stringent. Nonetheless, to encourage submission of phenotypic data, the designers of public phenotypic repositories must strive to reduce the amount of manual labor required of submitters. To do this, such tools must support reuse through controlled vocabularies, reuse of definitions across multiple submissions by the same investigator, as well as reuse within the same data set.

While the software-engineering task of creating a user-friendly data submission tool is relatively straightforward, a bigger challenge is arriving at a consensus among researchers as to what descriptors constitute a “minimum acceptable” set with respect to different types of parameters. Standardization efforts such as those among microarray researchers (such as MIAME – “minimum information about a microarray experiment”) must therefore be emulated. Existing standards such as LOINC, and clinical vocabularies such as SNOMED, must be utilized for clinical phenotype descriptors.