STRVILLAGE: A LARGE-SCALE VIRTUAL PEDIGREE GENERATOR

Anthony Lapadula, PhD, Anna Shcherbina, BS, Vamsee Pillalamarri, MS, Jeffrey Palmer, PhD
Massachusetts Institute of Technology Lincoln Laboratory, 244 Wood Street, Lexington, MA 02420

STRvillage is a software simulator that generates DNA profiles for large sample sets of virtual individuals to enable a variety of human DNA–based kinship experiments. Each individual is represented by an STR (Short Tandem Repeat) profile. By generating virtual sample sets, STRvillage obviates the need to collect or access sensitive personal genetic data in support of familial analyses and expert software evaluation.

To run the software, users provide a list of STR loci to include in the simulation, along with allele frequency distributions for one or more subpopulations of interest. Using agent-based modeling, STRvillage simulates the mating of virtual individuals across multiple generations to yield a village-sized pedigree. Each virtual individual is tracked through his or her reproductive life cycle, where they make choices that affect that person's number of mating events and offspring.

Users may customize a number of parameters to change the behavior of the individuals during this process, including: (1) marriage and divorce rates (which are adjustable across a varying number of marriages and divorces), (2) the probability of producing illegitimate children outside of marriage, (3) the degree of inbreeding, and (4) the prevalence of cross-generational mating and marriage.

The model has a number of other parameters, which focus on the population as a whole instead of individual agent decision-making. These include: (1) the number of founders of the population, (2) the maximum size of any generation, (3) the male/female ratio for offspring, (4) rates for identical siblings, and (5) immigration rates from multiple specified subpopulations. The model also includes simple tunable parameters for locus-specific mutation rates.

The resulting population is written as a Common Message Format (CMF) file for compatibility with existing software. The file is encoded so that the full pedigree is preserved and can be reconstructed.

Future work will involve the validation and comparison of existing kinship analysis algorithms, as well as the development of additional algorithms for use in a variety of specific circumstances. For example, we are investigating approaches to localize a suspect in a large pedigree with incomplete STR profile coverage. Such an algorithm would prove useful in situations where a subset of a population has volunteered DNA samples and documented familial relationships, and where the suspect is assumed to be a member of that pedigree. Future experiments will involve degrading the pedigree in various ways, for example, removing a subset of the individuals' profiles to simulate an incomplete canvass of the population, or misdirecting a fraction of the kinship linkages to simulate the real-world phenomenon where the social father is in fact not the biological father.

STRvillage will also be able to perform these analyses using a subset or superset of the STR loci in use today (including the 13 core CODIS loci), and will give empirical evidence with which to estimate the value of including additional loci.

This research is sponsored by DDR&E under United States Air Force under Contract FA8721-05-C-0002. The views expressed are those of the author and do not reflect the official policy or procedure of the United States Government.