Markus Riester

Genealogy Reconstruction
Methods and applications in cancer and wild populations

Summary

Genealogy reconstruction is widely used in biology when relationships among entities are studied. Phylogenies, or evolutionary trees, show the differences between species. They are of profound importance because they help to obtain better understandings of evolutionary processes. Pedigrees, or family trees, on the other hand visualize the relatedness between individuals in a population. The reconstruction of pedigrees and the inference of parentage in general is now a cornerstone in molecular ecology. Applications include the direct infer- ence of gene flow, estimation of the effective population size and parameters describing the population's mating behaviour such as rates of inbreeding.

In the first part of this thesis, we construct genealogies of various types of cancer. Histopathological classification of human tumors relies in part on the degree of differentiation of the tumor sample. To date, there is no objective systematic method to categorize tumor subtypes by maturation. We introduce a novel algorithm to rank tumor subtypes according to the dissimilarity of their gene expression from that of stem cells and fully differentiated tissue, and thereby construct a phylogenetic tree of cancer. We validate our methodology with expression data of leukemia and liposarcoma subtypes and then apply it to a broader group of sarcomas and of breast cancer subtypes. This ranking of tumor subtypes resulting from the application of our methodology allows the identification of genes correlated with differentiation and may help to identify novel therapeutic targets. Our algorithm represents the first phylogeny-based tool to analyze the differentiation status of human tumors.

In contrast to asexually reproducing cancer cell populations, pedigrees of sexually reproduc- ing populations cannot be represented by phylogenetic trees. Pedigrees are directed acyclic graphs (DAGs) and therefore resemble more phylogenetic networks where reticulate events are indicated by vertices with two incoming arcs. We present a software package for pedigree reconstruction in natural populations using co-dominant genomic markers such as microsatel- lites and single nucleotide polymorphism (SNPs) in the second part of the thesis. If available, the algorithm makes use of prior information such as known relationships (sub-pedigrees) or the age and sex of individuals. Statistical confidence is estimated by Markov chain Monte Carlo (MCMC) sampling. The accuracy of the algorithm is demonstrated for simulated data as well as an empirical data set with known pedigree. The parentage inference is robust even in the presence of genotyping errors. We further demonstrate the accuracy of the algorithm on simulated clonal populations. We show that the joint estimation of parameters of inter- est such as the rate of self-fertilization or clonality is possible with high accuracy even with marker panels of moderate power. Classical methods can only assign a very limited number of statistically significant parentages in this case and would therefore fail. The method is implemented in a fast and easy to use open source software that scales to large datasets with many thousand individuals.