Phylogenetic detection of protein sites associated to a phenotype, at the genome scale

Louis Duchemin

Abstract

Phenotype variations across species — morphological, physiological, or functional traits — are a manifestation of variations in their genomes at the molecular level. Long-term variations between species are a consequence of multiple interacting processes of different nature. Mutation is the source of diversification that occurs at the level of individual organisms, and is generally considered to be a stochastic process. Through generations, a novel genetic variant diffuses within a population under the combined effect of a non-adaptive process that is genetic drift, and selection that promotes or represses its transmission, depending on the reproduction advantage it provides. Adaptation of species to a perpetually changing environment emerges from the interaction of these processes, from which the massive diversity of life unfolds.

Extent species and their genomes share a common history that stems from the ancestral ascent they share, and separated into distinct species through the accumulation of divergences between populations of the ancestral species. By gathering genomic sequences that originate from the same ancestral sequence, and analyzing their divergence, it is possible to interpret traces left by their evolutionary history and infer parts of it. Among the variety of modifications that may alter genome sequences, I focus on substitutions, i.e. point modifications at one position within protein coding genes, that may result in changes in the structure and function of the protein they encode, with consequences in terms of adaptation. By analyzing the signal in these substitutions, in combination with the history of a phenotypic trait, one may attempt to detect correlations between the evolutionary history of a coding site, and that of the phenotype. The identification of such correlations might then be the signature that a genotype site is involved in the emergence or the stability of the trait under consideration, and more generally hint at its implication in the adaptation of a species to a particular environment.

Many models of substitutions in gene sequences that exploit this comparative approach already exist, and are widely used to develop our knowledge of molecular evolution. However, they are difficult to apply at the scale of whole genomes for systematic detection of sites associated to a phenotype, because of the large amount of data involved, and limited computing power. In this thesis, I search for a solution to allow this kind of analyses at large scale, that would involve shorter computation times, while preserving the quality of the resulting predictions.

After some unfruitful attempts at adapting linear models used in GWAS at the population scale to study genotype-phenotype associations, in order to make them applicable at the level of species, I identified an approach that seems to be a satisfactory solution. It is based on a model of amino acid sequence evolution — thus working directly at the level of protein sequences, after translation from DNA — that was previously published, but whose potential had not been recognized yet. I have shown, using simulations, that our implementation of this model enables fast and accurate detection of changes in the substitution dynamics that are associated to phenotype variations, just as well as several other more complex and computationally intensive models. Although it might not be a lot faster than other implementations based on phylogenetic models, that we could also evaluate, it appears to be the fasted among so-called "profile" methods, which provide estimates for the direction of selection at one site.

A part of this thesis is dedicated to exposing the details of this method, which we call Pelican, including its model, implementation and some of its limitations. An alternative strategy for fitting the model, using GPU computation to exploit the highly parallel nature of the problem, was also

explored to attempt improving the throughput of analysis further. I then describe an extension of the model based on continuous traits, which were initially limited to discrete categories; more efforts are yet required to evaluate the validity of this alternative model. I also investigate several ways to predict genes associated to a phenotype, using site-level predictions obtained at each position of their sequence. This is not an easy task, because of statistical issues inherent to the method, but I came up with an approach that allows to exploit site-level predictions with good statistical power, although its robustness may be lacking in some cases.

Finally, to further validate our approach using empirical data, I applied it to a genome-scale dataset of coding sequence alignments of mammals, to identify sites and genes associated to several discrete phenotypes. The predictions we obtained, when compared to the existing gene annotations and literature, suggest that this method is able to identify sites associated to the trait quite reliably. The result of this work is a software implementation for Pelican that, although it is in an early-stage, is proposed as a solution to detect inter-species genotype-phenotype associations at the genome scale.