Computational Pattern Recognition for the Identification of Transposases in Prokaryotic Genomes: Challenges and Advances.

Gonzalo Riadi\textsuperscript{1,2} and David Holmes\textsuperscript{1}

\textsuperscript{1}Center for Bioinformatics and Genome Biology, Fundación Ciencia & Vida and Facultad de Ciencias Biológicas, Andres Bello University, Santiago, Chile.

\textsuperscript{2}Bioinformatics Engineering School, Faculty of Engineering, Talca University, Talca, Chile.

Transposases (Tnps) are enzymes that are encoded by Insertion Sequences (ISs). Tnps are one of the commonest proteins found in nature and play an important role in gene and genome evolution. However, they are difficult to predict bioinformatically and given the increasing availability of prokaryotic genomes and metagenomes, it is incumbent to develop rapid, high quality automatic annotation of Tnps. In addition, a database of carefully annotated Tnps could potentially be used to reveal novel biology. This thesis: (i) describes methods for improved Tnp prediction and classification; (ii) generates a web service based on these improvements so that biologists can annotate new genomes and gene sequences; (iii) describes the construction of a searchable database of over 1150 genomes preannotated for Tnp content and (iv) describes novel biology revealed by high-dimensional analysis of over 210,000 Tnps.

A novel bioinformatics pipeline is described for Tnp discovery that is based upon the generation and use of Hidden Markov Model (HMM) profiles built using known Tnps deposited in the public database ISFinder. As part of this talk, ISFinder was first improved by removal of annotation errors. HMM sequence profiles were then used to predict about 210,000 Tnps in 1150 prokaryotic genomes including about 7,000 novel ones. High-dimensional analysis of all Tnps revealed: (i) a genome wide tendency for integration centered around the termination point of DNA replication in circular genomes; (ii) many genomes contained a plethora of different families of Tnps but, with a few exceptions, there were no positive or negative correlations between the presence or absence of Tnps within a particular genome; (iv) or with genome variables including %GC, genome size and number of Tnp families present in the genome; (v) Certain families of Tnps were more abundant in specific ecological niches (e.g. low pH) of the Prokaryotic domains; (vi) Certain families of Tnps are more abundant in specific taxonomic levels. (vii) The extended Tnp families show discrepancies that suggest there are around 14,000 protein sequences in NCBI’s Protein database for which their annotation can be improved.

This talk lays the conceptual framework for the robust identification of Tnps in new genomes and provides the most rigorous and extensive database of existing Tnps for other investigators to explore. This database is available in the form of a web server for the prediction of the most probable IS family for a query protein sequence.

With the data generated in this work, seventeen analyses were proposed in order to further develop this research topic. Selected analyses include a proposed method to bioinformatically predict types of transposition in IS
families; prediction of complete ISs and a method to assess the prediction of more complex mobile DNA elements like Integrative Conjugative Elements, Conjugative Transposons and genomic islands.

Acknowledgments: Fondecyt 1090451