Assignment of P-element Genomic Sequence Tags to Expressed Genes Sequences in *Drosophila*

Jason Goncalves [1, 2, 3] jgonca@oci.utoronto.ca, Jacob Mouka [2], Ozgur Huner [2], Jamie Cuticchia [3] and James R. Woodgett [1,4]

[1]Department of Medical Biophysics, University of Toronto, Canada
[2]Bioinformatics, Ontario Cancer Institute, Toronto, Canada
[3]Bioinformatics Supercomputer Center, Hospital for Sick Children, Toronto, Canada
[4]Experimental Therapeutics, Ontario Cancer Institute, Toronto, Canada

In the model organism *Drosophila melanogaster* genetic analysis provides one of the most powerful approaches to understanding gene function. However, genetic analysis relies on the availability of mutants with a single defined aberration. In *Drosophila* the P transposable element has been engineered for use in gene ablation and gene misexpression. However, insertion of the P element into the genome is relatively random, that is, not targeted at a gene. In order to map the P element insertion to an expressed gene, genomic sequence tags derived from the region flanking the P element insertion are being generated. There are currently over 3400 mutant lines for which a genomic sequence tag, derived from the region flanking the P element insertion, is available. Careful mapping of these sequence tags to genes provides a method for the high throughput identification of *Drosophila* mutants.

We were interested in optimizing the methods to map short genomic sequence tags to expressed genes. A common first step in this type of assignment is a BLAST search of *Drosophila* EST and expressed gene databases using the genomic sequence tag as a query. Thus we explored the question: given a BLAST result between a genomic sequence tag and an expressed sequence, what parameters or combination of parameters best predicts a true positive and excludes false positive assignments. In order to do this we generated a validated test set of P element mutants in which the disrupted gene had been characterized biologically. The genomic sequence tags from all P elements in the test set were compared to a *Drosophila* gene database using WU-BLAST. The BLAST results were filtered and those gene hits which passed the filter were assigned to the query P element. The computed assignments were compared against the validated test set, the number of true positives and the number of false positives were calculated from this comparison. The BLAST filter parameters included the probability score, the percent identities, the identities' length and a derived value of percent of maximum possible overlap. Each parameter was varied throughout its total range of values and the filter settings included all combinations of parameters, that is, the parameter space was exhaustively searched.

Results from the analysis indicate that the use of a single parameter, including the probability score, works poorly. For example, filtering with the probability set at a high stringency results in a large number of true positives being missed, while setting the stringency low results in a large number of false positive assignments. However, several combinations of different parameters (such as percent identity, percent of maximum overlap and identity length) could effectively identify almost all true positives with no false positive assignments. Thus it is possible to assign P-element genomic sequence tags to expressed genes or ESTs with a high degree of confidence using a simple filtered BLAST search. We are currently in the process of applying this method to all the P-elements with genomic sequence tags and archiving the results in a publicly accessible database called PEGAP or P Element Gene Assignment Project (http://pegap.oci.utoronto.ca). The test set and results of this analysis will also be available at this site.