

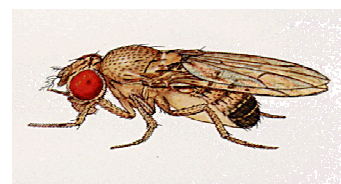
Structural and Functional Annotation of the *Drosophila Melanogaster* Genome using GeneAtlas™

A case study for the “kinase” annotation

Sunil Patel¹, Mikhail Velikanov²

¹Accelrys, 334 Cambridge Science Park, Cambridge, CB4 0WN, UK.

²Accelrys Inc, 9685 Scranton road, San Diego, CA 92121.



In recent years, several genomes have been completed. The major emphasis now is the identification and characterisation of protein function through knowledge of protein structure using various methodologies. In this study, we have used the automated GeneAtlas™ pipeline for structural and functional annotation of the *Drosophila Melanogaster* genome (a central model for the human genome) versus the GadFly genome annotation. This automated pipeline allows creation of database DS AtlasStore™. DS AtlasStore is a relational schema based on Oracle designed to store sequence data, family information, output from GeneAtlas which includes 3D structure prediction as well as functional annotation of the genomes. The GeneAtlas pipeline consists of identification of functional domains of protein sequences, homology searching using PSI-BLAST, fold recognition using SeqFold, high throughput homology modelling using MODELER (HTM), and function annotation using 3D motif searches. Models qualities are measured against Profiles-3D (verify) scores and Potential Mean Force scores (PMF).

Approximately, 14,322 ORFs were downloaded from FlyBase and analysed using the GeneAtlas. Using text search as “kinase”, all the sequences that contained a kinase annotation, were retrieved. All genes that contained an Interpro domain associated with kinases annotation were identified in the original GadFly genome. The results between the two genomes were compared. Additionally, analyses of randomly selected protein sequences were examined. These results were categorised with PSI-BLAST scores versus PMF scores.

Using the GeneAtlas pipeline, 99% of the genome was functionally annotated in DS AtlasStore. 59% of sequences have structural annotations, while 40% of sequences have sequence-based annotations. Using a text search for a “kinase” annotation, 1340 sequences were retrieved. 99.6% of these sequences were found to have homology with a known structure. 0.4% sequences were found to have homology with other sequences in the NRDB database. The results also revealed that a total of 333 sequences had unique HTM hits and 4 sequences had Seqfold hits. 385 sequences with an Interpro domain associated with “kinase” annotation were identified in the original GadFly genome.

Comparison of the annotation results demonstrated that DS AtlasStore identified a significant proportion (92.4%) of the kinases identified in the GadFly genome (with the exception of 14 protein sequences that were not in the original database and 20 sequences that were also annotated but not as kinases). Detailed analysis of randomly selected protein sequences demonstrated that DS AtlasStore has a very high level of annotations and it is possible to prioritise protein sequences that are annotated with high confidence. Comparison of protein sequence identity in the GadFly genome vs PMF scores showed that PMF scores can be used as a method to prioritise protein sequences that are annotated with high confidence in DS AtlasStore.

Using the GeneAtlas pipeline and the *Drosophila Melanogaster* genome as an example, shows that integration of 3D structural information with standard 1D sequence information adds value, clearly indicating structure is frequently more closely related to function than sequence. The GeneAtlas pipeline, therefore, not only annotates a higher number of protein sequences than those currently available in the public domain but also prioritises protein sequences that are annotated with high confidence allowing acceleration of both the target and drug discovery processes.