GeneAtlas^{TM} - An Automatic High-Throughput Pipeline for Structure Prediction and Function Assignment for Genomic Sequences (Part II)

Lisa Yan **Zhan-Yang Zhu** **Azat Badretdinov**
lly@msi.com
David Kitson **Krzysztof Olszewski** **David Edwards**

Molecular Simulations, Inc., 9685 Scranton Road, San Diego, CA 92121, USA

Keywords: homology modeling, high throughput, fold recognition, genomic sequence, functional annotation

With the vast amount of protein sequences determined from the genomic sequencing project, there is an emerging need to use high throughput method to predict structures and assign functions of the protein sequences. GeneAtlas is an automated, high throughput pipeline for the prediction of protein structure and function using sequence similarity detection, homology modeling, and fold recognition methods. It uses PSI-BLAST and SeqFold to search for homologous structures from PDB database and MODELER to build 3D models for the sequences based on the template structure. The quality of the 3D-model is validated using Profile-3D/verify score. The accepted model gives the correct fold and indicates the possible function of the genome sequence from the known template. Furthermore, protein 3D structure contains much richer information for its function than sequence along. Functional annotations based on the 3D-model using a suite of methods give further details of the protein function which are crucial for target discovery, protein engineering, and inhibitor design.

Using a "virtual" genome, a subset of PDB structures from SCOP database which consists of protein structures of less than 40% sequence identity, as a benchmark, we demonstrate that GeneAtlas detects additional functional relationships by building 3D-models for genomic sequences in comparison with the widely used sequence searching method, PSI-BLAST. The method was applied to 22 publicly available genomes, including *C. elegans*, *Drosophila Melanogaster*, *Saccharomyces cerevisiae*, *human*, *Arabidopsis thaliana*, etc. The modeling results of a small genome, *Mycoplasma genitalium*, and the comparison with PSI-BLAST and Hidden Markov Model (HMMer/pfam) on function assignment will be discussed.