GeneAtlas$^{\mathrm{TM}}$: A High-Throughput Pipeline for Automated Model Building and Functional Annotation of the Genome (Part I)

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

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

Keywords: genome analysis, protein function, sequence annotation, bioinformatics, molecular modeling

To add value to the information derived from genome projects, we have developed an automated high-throughput pipeline - GeneAtlas$^{\mathrm{TM}}$ for model building and functional annotation of protein sequences expressed by the genome. In GeneAtlas, we emphasize automation, quality control and efficiency (to be able to run on a large number of protein/DNA sequences). We use PSI-BLAST to search for homologous structures. Several PSI-BLAST search schemes were analyzed to obtain not only greater accuracy but also efficiency. We also use MODELER to build 3D models. To reduce errors commonly found in sequence/template alignments, we compare sequence/template alignments obtained using sequence profiles with those obtained using pair-wise sequence alignments by comparison of their 3D models. We have also developed procedures to validate the 3D models produced by MODELER through Profiles-3D scores, sequence/template similarity and others. The 3D models and related information are stored in an Oracle database for query. Examples of 3D technology for functional annotations include binding site analysis, 3D residue clusters (such as exposed hydrophobic surfaces) and neighboring sequences and structures.