# Structural Prediction and Functional Annotation of Proteomic Sequences using GeneAtlas™

**Dana Haley-Vicente, Velin Spassov, Tina Yeh, Ken Butenhof, Christoph Schneider, Azat Badretdinov, Lisa Yan**

Accelrys Inc., 9685 Scranton Road San Diego, CA 92121

We have used GeneAtlas™ to provide functional annotation of proteomic sequence data including structural prediction. 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. Using template searching, GeneAtlas searches for relationships between query sequences and known protein structures, motifs, and folds.

Subsequent inferences and assignment of the target protein's function is based on its homology to the experimentally derived template protein and the models generated as part of the pipeline.

Using CASP5 targets as query sequences, we demonstrate that GeneAtlas detects additional relationships, via its high-throughput modeling component, in comparison with the sequence searching method PSI-BLAST only. Furthermore, functionally related proteins with sequence identity below the twilight zone can be recognized correctly.

In addition, some targets were selected to test two new methods that we have developed, ChiRotor and Looper, for side-chain and loop prediction. ChiRotor is a fast algorithm that predicts the conformation of all or part of amino-acid side chains with an average RMSD of about 1Å for the core residues. The loop-modeling program, Looper, produces a number of energy minimized loop backbone conformations ranked according to force-field energy terms. Both algorithms are a combination of a discrete search in dihedral angle space and CHARMm energy minimization.

## GeneAtlas: High Throughput Functional Annotation Pipeline

GeneAtlas™ is an automated protein annotation pipeline for analyzing protein sequences and identifying their biochemical function. The GeneAtlas pipeline automates and integrates several steps into one seamless operation, collapsing the genomic information explosion and converting it into information and knowledge. In Figure 1 below, the protein sequences are run through a series of methods.[1]

• **Domain Analysis**: For sequence domain analysis we use the Hidden Markov Model (HMMer) algorithm to identify to perform a comparison to PFAM.

• **Similarity Search**: Before the search for similar sequences, a number of filters are applied including the masking of low sequence complexity regions. The sequence similarity searching component is comprised of a modified version of PSI-BLAST including a forward and reverse search method. Optimization of this component has been performed in a variety of ways to minimize the rate of false positives.[1]

• **High throughput Modeling**: There are several steps in this method, which are based on the work of Dr. Andrej Šali and his lab at Rockefeller University.[2]

• **SeqFold** Is a fold recognition method from Accelrys, originally developed in the laboratory of Dr. David Eisenberg.[4] As with the similarity search method, optimization of this component has been performed in a variety of ways to minimize the rate of false positives.[1]

•**Annotations** In addition to including the active site annotation from the PDB SITE record, Accelrys has developed algorithms for the location of potential binding sites on the basis of a structural template method, which identify three-dimensional features known to confer function

• **3D-** (e.g. serine protease catalytic triad, metal binding site, ATP binding site). The extracted structural patterns form a library of 3D pharmacophore-style templates that can in turn be used to characterize new protein structures in a manner similar to how Prosite is used to characterize sequences.

• **DS AtlasStore™:** The results are initially output in flat file format and then loaded into DS AtlasStore™, designed to store protein sequences, 3-D structures, and related functional annotations that have been derived using the methods contained in the GeneAtlas pipeline.

A putative homology relationship between a query sequence and a template from PDB is confirmed on the basis of the quality of the resulting homology model rather than solely on the basis of the level of sequence identity between the sequence and template.

Accelrys' PSI-BLAST protocol is used to search between query sequences and known protein structures stored in the RCSB (PDB) database. Protein models are generated using MODELER with the PSI-BLAST alignment. The last step is validation of the models where currently, GeneAtlas employs both the patented technology of Profiles 3D/Verify.[3] in addition to an algorithm developed by Andrej Šali to test whether the protein is reasonably folded and is a valid model.

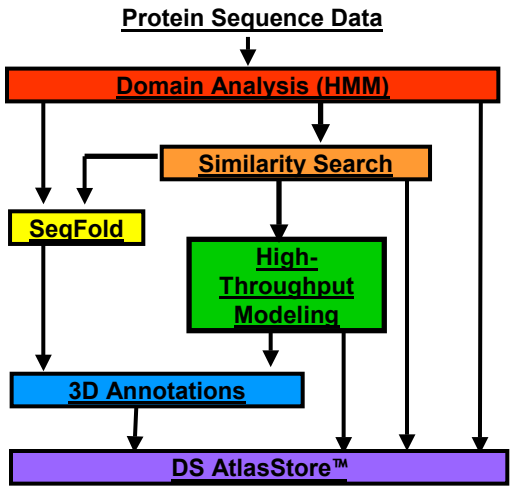



**Figure 1:** Schematic representation of the GeneAtlas™ pipeline, a high throughput pipeline for functional annotation of protein sequences. The resulting annotations are stored into Discovery Studio AtlasStore™.

## Structure Modeling of CASP5 Targets

We have processed selected CASP5 targets through the GeneAtlas™ high throughput functional annotation pipeline to identify potential PDB templates. The sequence alignments results from the GeneAtlas were adjusted using the help from several alignment tools including

- Align123, a method based on ClustalW and augmented with a secondary structure match term added to the alignment score
- Combinational Extension (CE) method (http://cl.sdsc.edu/ce.html)
- MALIGN3D, which aligns a set of structures (structure block and sequence block)
- ALIGN2D, which aligns two sets of sequences (structure and sequence block).

Multiple models were built using MODELER, and the side-chains and loop regions were further refined using its CHIROTOR and LOOPER, respectively. The models were checked for proper stereochemistry and evaluated by comparing the restraint violations reported by MODELER, and by the Profiles-3D Verify[3] method that measures the compatibility of each residue in the model with its environment.

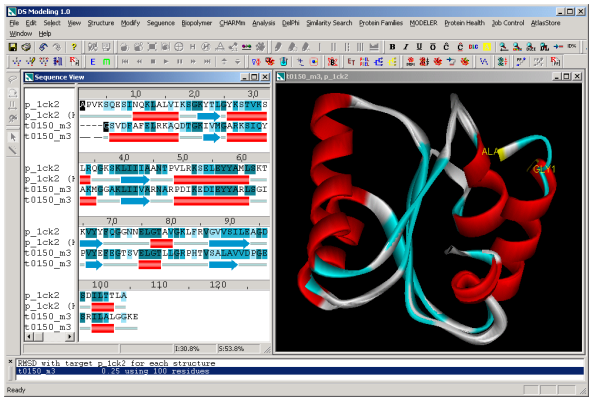


**Figure 2:** CASP5 target #150 3D model and template 1ck2 (60S ribosomal protein L30) structures and alignment displayed Discovery Studio Modeling (Accelrys new software for Modeling and Simulations on Windows®)

### Accelrys' Side-chain and Loop Modeling Algorithms

**Figure 3: CHIROTOR** is an algorithm based on CHARMm that optimizes the position of the side-chains.

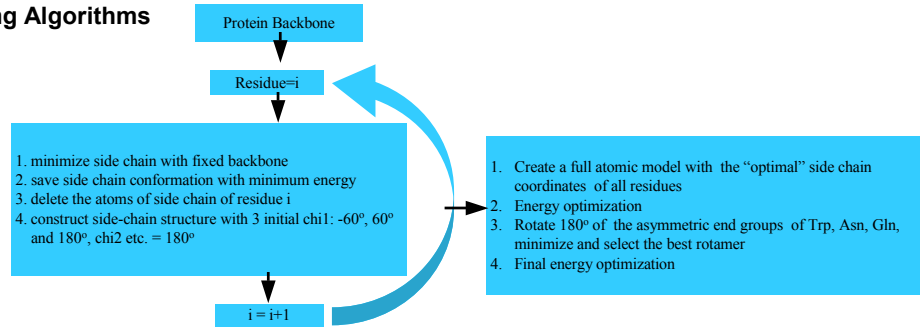
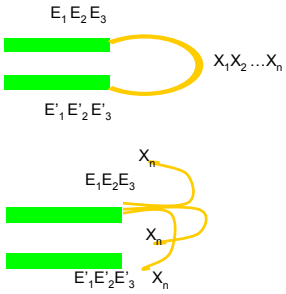



**Figure 4: LOOPER** is an algorithm based on CHARMm that optimizes the position of the loops.

1. Generate Nc conformers of peptide $E_1E_2E_3X_1X_2\ldots X_nE'_1E'_2E'_3$ by combining maximum four basic Ramachadran states of each $X_i$. The loop residues are without side chains.
2. 20 steps minimization of the loop $X_1\ldots X_N$ region
3. Calculate distance R between C atom of $X_N$ and N atom of $E'_1$.
4. If R < Rmax, minimize the peptide with stem residues fixed. Calculate CHARMm energy E. If $E\text{-}E_0 < E_{tr}$, accept the structure
5. Minimize the accepted conformers in the environment of the rest of protein
6. Select the low-energy conformers.
7. Generate the side chains of the residues in loop region using CHIROTOR algorithm.
8. Minimize and rank the structures.




### Example of GeneAtlas™ output for CASP5 target #TO142

casp5c000002 **casp5c000002 T0142 Salivary nitrophorin, C. lectularius**

| #Hit | Region | Template | Function | PSI-BLAST e-value | Method | AS | BP |
|---|---|---|---|---|---|---|---|
| 1 | 2-282 | 1i9z SCOP | HYDROLASE 21-MAR-01 1I9Z CRYSTAL STRUCTURE OF INOSITOL POLYPHOSPHATE 5-PHOSPHATASE DOMAIN (IPP5C) OF SPSYNAPTOJANIN IN COMPLEX WITH INOSITOL (1,4)- … | 1e-83 | HTM and PB90 | | 3/117 |
| 2 | 12-279 | 1ako Dbix 1hd7 | DNA repair endonuclease Hap1 AND/OR DNA-repair enzyme exonuclease III | 6.7e-18 | HTMM and PB90 | +/+ | 2/157 |
| 3 | 150-280 | 1hd7 SCOP | DNA REPAIR 09-NOV-00 1HD7 A SECOND DIVALENT METAL ION IN THE ACTIVE SITE OF A NEW CRYSTAL FORM OF HUMAN APURINIC/APYRIDINIMIC ENDONUCLEASE, APE1, AND ITS IMPLICATIONS FOR THE CATALYTIC MECHANISM … | 6.7e-15 | PB90 only | 0/0/2 | 2/58 |
| 4 | 2-280 | 1ako Dbix 1hd7 | DNA repair endonuclease Hap1 AND/OR DNA-repair enzyme exonuclease III | 2e-13 | HTMM and PB90 | +/+ | 1/95 |

**Legend:**
**AS** - this column indicates the presence of an active site annotation for given pdb hit. The +/+ and -/+ values refer to multiple template hits (MTM). +/+ implies there are active site residues in at least one of the templates AND some of these residues have been mapped to target sequence via alignment. -/+ implies there are active site residues in at least one of the templates AND none of these residues have been mapped to target sequence via alignment. The numerical indices pertain to the single template hits (STM). The 3rd number is the total number of amino acid residues that are annotated as an active site in the template. The 2nd number is those residues from the active site that are found in the alignment frame. The 1st number is the number of active site residues that are identical between the target and template.
**BP** - this column reports binding pocket annotation for given pdb95 hit The 1st number is the total number of binding pockets identified in the model structure The 2nd number reports the volume of the largest binding pocket in cubic angstroms
**PB90** - sequence profile-based searching protocol utilizing optimized PSI-BLAST
**HTM** - High Throughput Modeling protocol with subsequent verification with Profiles-3D and PMF Verify
**HTMM** - High Throughput Multiple-template Modeling protocol with subsequent verification with Profiles-3D and PMF Verify
**SeqFold™** - based searching protocol on the PSI-BLAST-defined domains
**HMMER** - Hidden Markov model profile searching protocol on PFAM-A domain database
BLAST

**Domain prediction results**

| #Domain | Region | Template function or SWISS-PROT accession number | Method | E-value | Bit score |
|---|---|---|---|---|---|
| 1 | 1-282 | Endonuclease/Exonuclease/phosphatase family | HMMER | 1e-16 | 69.0 |

1. D. H. Kitson, A. Badretdinov, Z-Y Zhu, M. Velikanov, D. J. Edwards, K. Olszewski, S. Szalma, L. Yan (2002) Functional annotation of proteomic sequences based on consensus of sequence and structural analysis. Briefings in Bioinformatics, 3, 32-44.
2. Sánchez, R. & Šali, A. (1998). Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proc. Natl Acad. Sci. USA, 95, 13597-13602.
3. Sánchez, R. & Šali, A. (1998). Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proc. Natl Acad. Sci. USA, 95, 13597-13602.
4. Fischer D. and Eisenberg D. (1996). Protein fold recognition using sequence-derived predictions,. *Protein Sci*., 5, 947.