

The ANNOTATOR software suite

Georg Schneider, Michael Wildpaner, Miklos Kozlovsky, Werner Kubina, Florian Leitner, Maria Novatchkova, Alexander Schleiffer, Sun Tian and Frank Eisenhaber
Research Institute of Molecular Pathology (IMP), Dr. Bohr-Gasse 7, A-1030 Vienna, Austria

To a great extent, collaboration of bioinformatics researchers with experimental life-science groups takes the form of functional prediction for protein targets that result from screens searching for process-relevant genes (expression profiling, mutational screens, interaction assays, etc.). There are more than 30 important sequence-analytic tools for proteins and, even for a single target, they can easily generate ~100 MB of ASCII-text with results. The manual analysis of dozens of targets can easily become impossible, especially if the information has to be collected from distributed websites.

The ANNOTATOR software suite is the ultimate solution for this problem. The ANNOTATOR accepts query sequence sets from the user, analyses them with all major academically available sequence analytic tools including also comparisons with motif, domain and protein sequence libraries, parses the outputs and presents the positive results in a user-friendly way both at the protein sequence and query set levels. Thus, the ANNOTATOR greatly enhances the productivity of protein sequence analysis in an applied setting.

Functional assignment to a protein requires a multi-step process in which individual segments are analyzed using a variety of algorithms. This is a tedious and inefficient process because individual tools often require cryptic parameters and options and present their results in incompatible output-formats which are then unavailable for further processing. The ANNOTATOR automatically recognizes the need for format transformations and knows the rules for calling tools in different contexts.

We have taken a proven segmentation approach routinely used by human experts and integrated all used algorithms into a multi-stage pipeline that undertakes the following task:

1. Identification of non-globular regions: This first step is achieved by taking into account regions with a specific compositional bias as well as differing levels of complexity. Subsequent processing identifies post-translational modifications as well as targeting signals for sub-cellular localization or secretion. Special emphasis is being put on the characterization of membrane-embedded regions. In a last step secondary-structure prediction is undertaken.
2. Determination of known domains and motifs: A wide range of algorithms and corresponding motif and domain libraries are searched to identify already known domains.
3. Sequence-Similarity-Searching: The remaining segments are analyzed using sequence similarity search techniques that include multi-step iterative family searches.

The results of the individual analysis are stored in an object-relational database in a common format and are then visualized for each individual protein in a coherent manner. The ANNOTATOR system independently decides whether additional sequence-analytic procedures are reasonably applied to the given target.

A unique feature of the system is the possibility to treat sets of proteins as analyzable entities. In this way it is possible to view the distribution of particular segments within a collection of proteins. This is accomplished by presenting a special histogram view which also allows for the selection of subsets of sequences.

Another useful capability is the preservation of taxonomic information throughout the pipeline. In combination with the above mentioned histogram feature this allows for powerful multi-dimensional selection of proteins.

Due to heavy usage the system is being run as a web-service with a cluster of distributed computing nodes as a backend processing facility. Nevertheless all necessary components can be installed on a laptop running a Linux/Unix operating system.

The system is available for academic use. Questions regarding licensing and availability should be directed to Dr. Frank Eisenhaber (Frank.Eisenhaber@imp.univie.ac.at).