45. The ANNOTATOR sequence analysis suite

Georg Schneider, Michael Wildpaner, Florian Leitner, Birgit Eisenhaber, Miklos Kozlovsky, Werner Kubina, Sebastian Maurer-Stroh, Maria Novatchkova, Alexander Schleiffer, Sun Tian and Frank Eisenhaber

Research Institute of Molecular Pathology (IMP), Dr. Bohr-Gasse 7, A-1030 Vienna, Austria
Presenter: Florian Leitner

Analyzing experimentally uncharacterized protein sequences is a regular research task. In-depth study of sequences requires dozens of programs, analyzing their output for many proteins is a huge effort. The ANNOTATOR is an intelligent environment for sequence annotation generation, evaluation and candidate selection including integration with MS/MS interpretations (MASCOT).

Large-scale prediction and annotation of protein function is a major, if not the most important creative contribution of computational biology to life science research. The need for analyzing experimentally discovered, yet uncharacterized protein sequences regularly occurs in the everyday research practice and the respective support is often requested in collaborations with experimental groups.
In-depth study of a protein sequence requires the invocation of about three dozens of programs and the comparison with many sequence, motif and domain databases. Thus, even generation and parsing of this output composes a huge effort, especially in cases of sets consisting of hundreds of proteins or even full proteomes. Likewise, selection of candidates from proteomes by sequence annotation features is impossible without generating and parsing the complete functional annotation of all proteins. We designed the ANNOTATOR suite as intelligent support environment for sequence annotation generation, evaluation and candidate selection, where routine sequence-analytic operations are automated. As a result, the productivity of protein sequence analysis work in a collaborative setting is greatly improved. The ANNOTATOR suite is also prepared to upload results from tandem mass-spectrometra interpretations (e.g., MASCOT) and to present the protein hits in context with their functional annotation.

The ANNOTATOR software suite implements all requirements for protein analysis into one single, multipurpose bioinformatics research workbench. Query sequences or complete sequence sets can be obtained conveniently by querying major databases such as UniProt, NCBI NR, GenBank and PDB – all of which are completely integrated into the ANNOTATOR environment. Alternatively, sequences can be imported as user-defined FastA-formated sets. More than 30 academically available sequence-analytic tools as well comparisons with many motif and domain libraries can straightforwardly be initiated on a sequence (set). In this way, the user avoids conflicts with a wide variety of input formats, cryptic parameter settings and complex outputs not prepared for postprocessing. The ANNOTATOR contains rules for applying tools in different context (i.e. parameter settings) and has knowledge of transforming various in- and output formats. There are also integrated procedures available that analyze conclusions from coincidences in findings from individual algorithms. Outputs from sequence-analytic tools (including those from interpreting tandem mass spectra) are automatically parsed and positive findings are presented in an user-friendly manner: results at both protein
sequence and query set level are made accessible through an integrated view. This output visualization enables effortless recognition of hot spots and noteworthy results to the investigator. The result set of each individual analysis is stored in an object-relational database using an internal, unified data format

and can be visualized through an intelligible display. Taxonomic affiliations and gene ontology assignments are preserved throughout the process. ANNOTATOR-based investigation of protein sequences is initiated by applying a staged pipeline to the set, adapted from a proven segmentation approach frequently applied by experts on protein function annotation. Following stepwise approach is executed by the system.

1. Identification of non-globular regions (by considering compositional bias and complexity, recognition of post-translational modifications and targeting signals, membrane-embedded regions and coiled coil segments),

2. Determination of known globular domains (a range of algorithms and corresponding motif and domain libraries are searched), 3. Similarity searches in sequence databases.

For any given target, the ANNOTATOR system subsequently decides whether continued sequence analytic procedures can be reasonably applied. Viewing the distribution of annotation features within a collection of proteins is another feature provided by the ANNOTATOR. These sets are presented in a special histogram view allowing for selection of sequence-subsets. In combination with the taxonomic information, this enables multi-dimensional selection of new query protein sets from the current results. The system is available for academic use at http://www.annotator.org. Questions regarding licensing and availability should be directed to Dr. Frank Eisenhaber (Frank.Eisenhaber@imp.univie.ac.at).