Peter Ernst
Dr. sc. hum.

Complex analysis pipelines in bioinformatics

Geboren am 26. August 1971 in Mannheim
Diplom der Fachrichtung Physik am 20. November 1997

Promotionsfach: DKFZ (Deutsches Krebsforschungszentrum)
Doktorvater: Prof. Dr. rer. nat. Sándor Suhai

The development of efficient DNA sequencing methods has made an enormous amount of sequence data available in recent years. For the analysis of this overwhelming amount of data the use of bioinformatics methods in life science research has become indispensable. This fact has been the motivation for the development of the W2H web interface using web technologies to enable access to high-performance computing resources and regularly updated databases from every PC, and for the development of the W3H task framework for the implementation of customised, complex analysis pipelines for high-throughput data analysis.

The W2H web interface enables platform-independently full access to command-line driven bioinformatics applications, such as those contained in the HUSAR, GCG or EMBOSS packages. By using a meta-data approach to describe the parameters of applications and other properties like input/output data types, an interface to hundreds of programs is provided, which is used at many academic and commercial sites world-wide to provide WWW access to their applications. W2H generates all application specific pages such as input masks from this meta-data, that is distributed along with the applications in the program packages mentioned above.

In order to provide fast response to user actions and considering low network bandwidth, several computational tasks have been transferred to the client in the most portable way by using JavaScript embedded in HTML. Several techniques have been developed and applied in order to handle the stateless nature of the HTTP protocol and the need to keep working sessions on the server. In order to extend and replace the features of the traditional command-line interface, mechanisms have been developed to handle batch queues, regular repeated jobs, graphics conversion and output post-processing including the integration of Java applets. By default, the data is stored persistently and secure in UNIX user accounts, which makes it also possible to work in parallel via the web and the command-line. W2H contains a loop feature that allows an analysis of a high number of data-sets with just a few mouse clicks. As a full-featured graphical user interface, results, files, projects and upload managers, sequence and alignment editors as well as different modes for experienced and novice users have been implemented.

The experiences and features of the W2H web interface have been used to create the W3H task framework in order to address the demand for customised, high-throughput data analysis pipelines. The *W3H task framework* allows the combination of different computational methods, algorithms and data sources from an existing software environment for the analysis and validation of results from biological experiments. This framework assists bioinformaticians in the design of analysis tasks invented in collaborations with lab scientists. It allows a high degree of flexibility in enabling the addition of new methods and databases as they become available. Using an object-oriented design, analysis tasks can be assembled from building blocks, that are based on external applications, by describing their work and data

flow. Using this configuration data, a task is automatically web-enabled by means of W2H's meta-data capabilities. In order to allow all kinds of post-processing of the computed results, tasks in the W3H task framework store their results in XML. This XML file can be used for further analysis (i.e. direct integration in user's databases, additional pipeline analysis) or can be refurbished for web users by means of XSLT. Based on these developments, a library of application output parsers and analysis building blocks was created.

To proof the power of the W3H task framework, some automatic analysis pipelines (tasks) are presented and described. These tasks were developed to solve specific biological problems like the annotation of ESTs (ESTAnnotator), the inference of phylogenies (PATH), the mapping and annotation of cDNAs (cDNA2Genome) and the analysis of the domain architecture of proteins (DomainSweep).