Hans-Peter Pohle
Dr. sc. hum.

Design and implementation of a flexible system for analysis, annotation, and classification of biomolecular sequences

Geboren am 10.05.1972 in Springe
Reifeprüfung am 29.05.1991 in Bad Nenndorf
Vordiplom am 16.03.1994 an der Universität Oldenburg
Diplom am 05.11.1996 an der Universität Oldenburg

Promotionsfach: Medizinische Biometrie und Informatik
Doktorvater: Prof. Dr. rer. nat. T. Wetter

Two pieces of software have been developed comprising a system to support important tasks during the work with biomolecular sequences. An environment was built which provides a user interface to control various analysis tools. Additional tools can be integrated easily into it by specifying their properties in a configuration file. This considerably improves the flexibility of the system. An annotation editor is integrated into the analysis environment, which provides a way to visualise the analysis results and attach them to the sequences. In addition, the environment supports the automatic execution of several tools one after another. Results from one tool can be passed on to the next. The second piece of software deals with classifying proteins according to their function. It can learn any wanted classification scheme by providing a set of training data to it. This makes the system very flexible.

Both pieces of software have been implemented in a scripting language, because development times are much shorter when using these languages. The analysis environment was implemented in Tcl/Tk, the classification system in Perl. The analysis environment, which is a rather big program, was designed object-oriented to keep it maintainable. The [incr Tcl] extension was used, which adds object-oriented programming capabilities to the Tcl/Tk language. It was found to have some shortcomings, namely the lack of copy constructors and exceptions. The ways to work around these problems are shown. To ensure a fast execution, it was sometimes necessary to exploit special properties of the Tcl/Tk language. The techniques used comprise idle tasks, the capabilities of the canvas widget, and C extensions.

For the classification system tests have been carried out to assess its prediction accuracy. The sensitivity, specificity, and correctness of the prediction have been determined using leave-one-out tests on four different data sets. In addition, two tests have been performed where the training and test set were taken from different organisms. When comparing the test results, it shows that there are big differences between the data sets. A detailed analysis of the classification errors revealed that the majority of them was caused by misleading homologies, a factor that mainly lies outside of the scope of the system.

The analysis environment provides a large quantity of functionality that is important for the analysis and annotation of biomolecular sequences. Its flexibility is a major advantage, because it allows to adapt the system to changing requirements without writing code. The integrated annotation editor enables to get a better insight in the biological function of the sequences. The possibility to define analysis strategies, which is not present in most other systems, facilitates the annotation of larger amounts of sequences. On the other hand, the system requires a skilled user to achieve high-quality results. The choice of the Tcl/Tk language has been proven to be right, because the development time was much shorter than it would have been with C++, for example.

The main features of the protein classification system are its flexibility and the possibility to handle hierarchical classification schemes. It doesn't make the manual assessment of protein function superfluous, but it drastically reduces the work for the expert.

All of the software has proven its usefulness in practice. It is in routine use at LION.