

Evaluation of the performance of IE systems in Molecular Biology

By Christian Blaschke and Alfonso Valencia.

The practice of evaluating the performance of a system in terms of precision and recall on limited test sets addressing a particular problem is standard in computer linguistics and computer science in general. In most cases annotated pieces of text are provided as input and programs trained with this data compete in the solution of specific problems based on additional pieces of text. For the evaluation the results are compared with solutions provided by expert human annotators.

Even if it is possible to imagine the application of this approach in any field, including Molecular Biology and Biomedicine, there are several reasons why we think that alternative approaches may be preferable here.

In our view, the current discussion about method evaluation can be put in direct comparison with the experience in the evaluation of protein structures (CASP and CAFASP experiments for the Critical Assessment of Structure Prediction methods) organized by J. Moult over the last 8 years. In this case, the success for the field has been based on very clear ideas, including the focus in well defined application areas (levels of structure prediction), a clear message to the potential users of the technology (experimental biologist), and the creation of standards that are required for the publication of new methods. The current homogenous view of the field has also brought very interesting technical developments in the evaluation of the prediction methods, including the development of automatic evaluation servers.

We have discussed some of issues related with evaluation of I.E. Systems in Molecular Biology in previous publications (Blaschke and Valencia, 2002, Blaschke et al, 2002). By raising them here, we would like to stimulate the discussion in the community.

First, there are many different possible tasks in Molecular Biology, given that each subfield has specialized languages and particularities (the diversity and specific features of the fields is characteristic of molecular biology). Given the early stage of most applications, it would be difficult, and probably detrimental, to concentrate all of them in a specific task.

Second, it is difficult (if not impossible) to obtain accurately annotated, and large enough, corpora, for the many different possible tasks.

Third, it would be difficult to attract the attention of the biologist (the potential user), if the tasks analyzed are not of direct biological interest, and/or the analysis is not done according to the biologist's perception of what is important.

Fourth, researchers are now able to build systems that rely on only very coarse-grained annotation (annotation at the article level, e.g., facts entered in a database with a pointer to the article). These data mining systems do not require annotation at the phrase or sentence level (as commonly done in linguistically-based approaches). This means that evaluation must also be done at this coarse-grained level.

Fifth, biology is extraordinarily influenced by the techniques of massive production. It is difficult to justify the evaluation of methods for automatic information extraction if the results are not comparable (and extrapolated) to the evaluation of large data sets.

In the past we have proposed an approach to the problem of evaluating the results of information extraction systems by applying our system to detect protein-protein interactions to experimental results contained in a specialized database (Blaschke and Valencia, 2001). Each entry in the database contains two protein names, their type of interaction and a pointer to the corresponding entries in a sequence database (the molecular composition characteristics of each protein). We

believe that an evaluation based on this type of task can be of interest for biologist since it addresses simultaneously the problem of validation of experimental data and analysis of massive interaction networks. It is possible to find other similar external annotated repositories in many other areas of Molecular Biology, and each one of them can provide a valid evaluation set for the corresponding methods.

References

Blaschke, C. and Valencia, A. (2001). "Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study". *Comp Funct Genom* 2, 196–206.

Blaschke, C., and Valencia, A. (2002) The frame-based module of the Suiseki information extraction system, *IEEE Intelligent Systems* 17: 14–20.

Blaschke C, Hirschman L and Valencia A (2002) Information extraction in Molecular Biology, *Briefings in Bioinformatics* 3: 154–165.