Integration of EnsEMBL with BioAgent - (session: Other)

Ezio Bartocci*, Steffen Möller+, Luca Toldo§, E.Merelli

Dipartimento di Matematica e Informatica, Università di Camerino, +University of Rostock, Proteome Center, Rostock, §Merck KGaA, Pharma Preclinical R&D, Scientific Information Services, Darmstadt

The human, murine and other eukaryotic genomes are presented by the Open Source project EnsEMBL[1,7]. Besides facilitating access through the web or by directly querying the relational database, EnsEMBL may present itself by the BioDAS [2] interface. Moreover, external data sources may be integrated with EnsEMBL seamlessly via the Internet while obeying to BioDAS. The interface facilitates the storage and retrieval of arbitrary properties of the genome, referred to as features that represent the annotation of the genome.
While interacting with EnsEMBL via its web interface, the series of manual interactions with the remote system by the researcher determines the information/experiment flow and it is by his or her mental capacity that the results of multiple experiments are integrated. The challenges of the post-genome area, i.e. with a huge amount of transcriptomics and proteomics data, demand an increased automation of the data analysis process. However, with semantics of publications and their interpretation not being transferable to machines and limited CPU power prohibiting the genome-wide precalculation of algorithms, we search for the directed autonomous execution of in-silico experiments and their communication to and between researchers.
The current bioinformatics approach is to facilitate customised workflows [3,4], which need to be adapted to the respective application, i.e. these are customer and data dependent. Automation starts by employing wrappers and „glue‰ code that parse the inputs/outputs, exploiting so called non-mobile agent technology. Not being mobile has a consequence of restraining bioinformatics algorithms from the use of long lasting computations on remote servers. The WWW information gathering architecture has been designed and implementd for interactive browsing of pre-generated data. Therefore, the choose of a connectionless architecture. The WWW paradigm was extended to access computational engines by the „Common Gateway Interface (CGI1.2) ‰[8] However, the CGI mechanism can only be exploited if the remote jobs can be performed rapidly or else rely on other mechanisms (e.g. SMTP) or on client polling of a predefined URL to return the result. If a bioinformatics algorithm requires a long lasting job to be executed remotely, then it will grow very much in complexity and is difficultly amenable of workflow mode.The data in bioinformatics may be too huge to be transferred (e.g. raw images of microarrays, of proteomics, or other) or unavailable (e.g. full text of scientific journals) hence, inmobilisation in code also means the limitation on accessing data sources. Furthermore, relevant information may have been carried out within a collaborative research environment, which remains unretrievable by the static web environment.
While EnsEMBL offers a BioDAS interface of the available information sources, it remains difficult to fully exploit them, due to their huge data structure and their dynamic nature. We here describe a first approach to perform the prior mentioned directed automated search by delegating it to specialized software, and to eventually further improve it by employing mobile agents. A mobile agent is a computational unit capable of migrating to different places from any location. An agent can behave in an opportunistic and reactive way. Agents do not require the user's presence and can be assigned a task to be exploited over distributed resources [5]. In the case that EnsEMBL be integrated with agent-based application, one could graphically compare, in the same Contig-View panel, annotations and features extracted from EnsEMBL database with those from generic BioDAS source and with those coming from the agent's task. From this integration we obtain two main advantages

1. a visual tool to check the result of the agent's task. EnsEMBL puts all BioDAS sources in a graphical format in its contig view, highlighting annotations and features found by the agent in a physical chromosomal position;
2. a useful way for a bioscientist to compare new results with other BioDAS sources and EnsEMBL annotations.

Furthermore, this integration offers providing a to cache computationally costly results.
Current implementations of DAS sources have certain drawbacks:

- No dynamics. Current BioDAS sources are stateless. This is in a way a direct consequence on the demand for instant replies. Primary data is displayed as available in a pre-computed manner, not its interpretation or context-sensitive information, e.g. by investigating prior results from the cache.
- Limited specification of features in the query. The DAS interface facilitates the query for features within a specific chromosomal range or for specific identifiers, and also allows the retrieval of all features of a

source. However, EnsEMBL currently does not offer an interface to present results from BioDAS sources on multiple chromosomes. One could think of an application like a BLAST sequence similarity search and the results of which being presented as a DAS source or of other indirect specifications for a selection of features which is beyond current implementations.

 Inmobility. Moreover, EnsEMBL does not allow the use of remote services within the databases access.

In this work we propose the integration of EnsEMBL with the mobile agent system BioAgent [6]. This was performed by wrapping the BioDAS interface around a single agent for visibility within EnsEMBL. Conversely, the agents may contact any DAS server for information, which includes EnsEMBL itself. The combination of EnsEMBL with the agent system BioAgent gives all advantages of dynamic data generation and data integration to EnsEMBL, in particular of computationally costly algorithms that may not be feasible to precompute. We also addresse an application for inhouse knowledge management, as partial results of analyses and their contexts may be communicated between researchers.

References

1. Hubbard, T., D. Barker, et al. (2002). "The Ensembl genome database project." Nucleic Acids Res. 30(1): 38-41.
2. Stein, L.D., Eddy,S., Dowell, R. (1999-2002) „Distributed Annotation System‰ http://www.biodas.org/documents/spec.html.
3. Möller, S., Schroeder, M., Apweiler, R., (2001). "Conflict-resolution for the automated annotation of transmembrane proteins." Comput. Chem.; 26(1):41-46
4. Toldo, L., Rippmann, F. (2001) „Method for Determining Nucleic And/Or Amino Acid Sequences‰(Pat WO0120024).
5. Hall, D., Miller, J., Arnold, J., Kochut, K., Sheth, A., and Weise, M. (1999). "Using Workflow to Build an Information Management System for a Geographically Distributed Genome Sequencing Initiative‰, Genomics of Plants and Fungi, R.A. Prade and H.J. Bohner, Editors .
6. Merelli, E., Culmone, R. and Mariani, L. (2002) „BioAgent: A Mobile Agent System for Bioscientists‰, NETTAB02-Agents in Bioinformatics, Bologna.
7. Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Hubbard T, KasprzykA, Keefe D, Lehvaslaiho H, Iyer V, Melsopp C, Mongin E, Pettett R, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Birney E. ≥Ensembl 2002: accommodating comparative genomics‰ Nucleic Acids Res. 2003 Jan 1;31(1):38-42.
8. Coar K, „The Common Gateway Interface‰ http://cgi-spec.golux.com/.