The Distributed Annotation System for Integration of Biological Data

Andreas Prlić[1], Ewan Birney[2], Tony Cox[1], Thomas A. Down[1], Rob Finn[1], Stefan Gräf[2], David Jackson[1], Andreas Kähäri[2], Eugene Kulesha[1], Roger Pettett[1], James Smith[1], Jim Stalker[1], and Tim J.P. Hubbard[1]

[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK
ap3@sanger.ac.uk,
http://das.sanger.ac.uk/registry/
[2]EMBL - European Bioinformatics Institute, Hinxton, UK

Abstract. The Distributed Annotation System (DAS) is a protocol for sharing of biological data which allows for dynamical data integration. It has become widely used in both the genome and protein bioinformatics communities. Here we provide an overview of the available DAS infrastructure and present our latest developments, including a registration server that facilitates service discovery by DAS clients while automatically monitoring service availability. Currently there are 108 registered DAS servers, provided by 24 institutions in 10 countries.

1 Introduction

Annotation of biological data, such as genome and protein sequences, is one of the central tasks in biological research. This is done by different means, for example manually, computationally and experimentally. There are a number of centralized resources available that are working on the integration of these data. They are facing the problems of how to manage the vast amount of data that is available, the need for frequent updates and releases, and how to exchange data with other institutions and users.

The Distributed Annotation System (DAS) is a protocol that addresses these issues and facilitates the sharing of biological data [1]. It is based on the idea that annotation data is not aggregated into large centralized databases, but instead is spread over multiple sites, generally maintained by the original data creators. DAS is frequently used for

1. integration of personal data into bioinformatics resources,
2. integration of the annotations from external sources into local applications,
3. access to most recent data versions without the need for local installations,

DAS is a web service protocol built upon well established open technologies (HTTP and XML), with some similarities to SOAP-based services. Where SOAP services use XML requests and responses for the transport of information, DAS

U. Leser, F. Naumann, and B. Eckman (Eds.): DILS 2006, LNBI 4075, pp. 195–203, 2006.

provides a data model, a query model, and a transport. The returned XML documents contain objects like *sequence* or *feature*. All data are provided by DAS servers and it is up to a DAS client to retrieve the annotations from multiple servers and to integrate these into a visualization that is presented to the user (see Fig. 1). For a detailed description of the DAS protocol see `http://www.biodas.org/documents/spec.html`.

The DAS protocol was originally designed to serve annotation for genomes. Resources like the *Ensembl* genome browser utilize this protocol to visualize new or personal data in the context of other annotations [2]. Different web pages, *"views"*, provide access to annotation data for e.g. chromosomes, transcripts, genes, or proteins. Each of these views acts as a DAS client. A management interface allows users to configure a list of DAS servers from which annotation should be retrieved. Once a new server has been added in the configuration, Ensembl establishes the contact to the server, fetches the data, and displays it together with other annotations. In this setup the Ensembl web server acts as a data-proxy and the users can access all data via their web browsers.

Over the last few years DAS has also been used to share annotations of proteins. We recently presented *SPICE*, a browser of protein structures, sequences, and their annotations, which is built on DAS [3]. SPICE is a Java application that installs and runs locally using the Java Web-Start technology. It can be launched by simply following a link on a web page. SPICE provides an integrated view of protein sequence and structure and can project annotations from one coordinate system onto another. This, for example, allows it to display protein sequence annotations with respect to their position on the protein structure. SPICE is integrated with Ensembl (see Fig. 2).

Dasty is another protein DAS client [4]. It is a Java application with a Macromedia Flash front-end, and all DAS communication is done via a dedicated server. Other DAS clients that can be easily integrated into web pages are ProView (`http://www.sanger.ac.uk/proview/`) or the CBS DAS Viewer [5].

DAS has been widely adopted in the bioinformatics community, because it is simple to use and simple to set up. Both DAS servers and client software are available with implementations in multiple languages: In Perl there is support for setting up a DAS server using ProServer (`http://www.sanger.ac.uk/proserver/`) or LDAS (`http://www.biodas.org/servers/LDAS.html`), while users who prefer Java can use Dazzle (`http://www.derkholm.net/thomas/dazzle/`). Client libraries are also available in Perl, e.g. the Bio::DasLite library (`http://search.cpan.org/~rpettett/Bio-DasLite/`), and in Java (`http://www.biojava.org/`, `http://www.spice-3d.org/dasobert/`), making integration of DAS support into new and existing bioinformatics tools easy.

Several collaborations are providing support for DAS. The BioSapiens Network of Excellence (`http://www.biosapiens.info/`) is providing a large number of DAS sources, which are listed at the BioSapiens Information Resource (`http://www.biosapiens.info/page.php?page=biosapiensdir`). BioSapiens also provides a Portal that can query UniProt and provides access to several DAS clients (`http://www.biosapiens.info/page.php?page=das_portal`). Another