Using the Distributed Annotation System

http://www.ebi.ac.uk
Introduction

This half day course is designed for those with a biological background that are relatively new to the use of the Distributed Annotation System for utilising further biological information. The practical sessions will start with a short lecture and be punctuated by others to run over approximately half a day. Please ask questions. There will inevitably be a mixture of abilities in this course. If you find that we are going too fast or not making ourselves clear, feel free to interrupt!

Our aim is to present a hands-on approach to the use of the distributed annotation system, focusing on using the system to compare protein annotation from different sources and as a starting point for investigations.

Practical instructions are presented as indented text. Follow these instructions to complete the practical. Surrounding text is designed to be explanatory; both to clarify why you are following specific instructions and what your answer should look like.

The three main DAS clients are:

- Dasty  [http://www.ebi.ac.uk/dasty/](http://www.ebi.ac.uk/dasty/)

Suggested Further Reading

**DAS:**

The DAS website:  [http://www.biodas.org/](http://www.biodas.org/)


**Context:**


# Table of Contents

[Introduction](#) ........................................................................................................................ 2  
  Suggested Further Reading ........................................................................................................ 2  

[The Distributed Annotation System](#) ......................................................................................... 4  
  Anatomy of a DAS feature request .............................................................................................. 4  

[The e-protein DAS servers](#) .................................................................................................... 7  
  3D-Genomics ................................................................................................................................. 7  
  Catalytic Site Atlas (CSA) .............................................................................................................. 7  
  Genomic Threading Database (GTD) ............................................................................................ 7  
  PDBsum ........................................................................................................................................ 7  

[A simple query using Dasty](#) .................................................................................................... 8  

[Ensembl and Spice – protein structure](#) ...................................................................................... 10  

[Conflicting annotation](#) .......................................................................................................... 14  

[Versioning](#) ............................................................................................................................... 14  

[Adding new DAS sources](#) ........................................................................................................ 15  

[Types and annotation quality](#) ................................................................................................ 17
The Distributed Annotation System

The Distributed Annotation System is a simple protocol to query servers about what annotation they possess about a particular biology entity, proteins in the case of the e-protein project but DAS was original designed for genomic data. DAS sits between the client and the database, so annotation can be retrieved and displayed without having to know anything specific about the database in question --- no custom scripts are required to extract and read the data.

While a DAS server can be queried for many things, like what databases it provides, the remainder of this section concentrates on the “feature” request that actually returns annotation about a given protein. All this is usually handled by the DAS client, invisible to the user, and it is not necessary to understand it in order to use DAS.

Open a web browser and go to the following URL

http://www.ebi.ac.uk/das-srv/das/proteindas/csaexended/features?segment=P11766

You should see an XML document with sections looking like those in the section “Anatomy of a DAS feature request” below. Some browsers will hide the structure of the XML, in which case select view->source (view->page source in Firefox/Mozilla).

The URL is a DAS feature request, asking the server www.ebi.ac.uk/das-srv/das/proteindas if the database csaexended has any features about the protein with the Uniprot accession number P11766.

Anatomy of a DAS feature request

```xml
<?xml version="1.0" standalone="yes"?>
<!DOCTYPE DASGFF SYSTEM "http://www.biodas.org/dtd/dasgff.dtd">
<DASGFF>
  <GFF version="1.01" href="http://web59-node1.ebi.ac.uk:9001/das/csalit/features">
    <SEGMENT id="P11766" version="6c676e99aff470dd72b585fab49ae963" start="1" stop="">
      The “segment” that this response is about is the protein with Uniprot accession P11766. The version is a check-sum on the protein’s amino acid sequence, so the client can tell whether the response is about an older or new version of the protein than the request. Use of the version will be described in more detail later. The start and stop co-ordinates specify the region of the protein that this response annotates, in this case: from the first residue to the end.
    </SEGMENT>
  </GFF>
</DASGFF>
```

A response from a DAS server may contain many features; this one describes a literature reference. The id of the feature is given, along with a human-readable label.

- **<TYPE>** Information describing the type of annotation. A human readable label is given between the tags, along with a more controlled “id”, and optionally “category”, qualifiers which will form part of a controlled vocabulary (see Types and annotation quality).

- **<METHOD>** The method used to produce the annotation. In this case the feature is derived from the primary literature but, in general, should describe how the feature was found. For example: domains might be found using models from SCOP or Pfam, or hand annotated.

- **<START> & <END>** The region which the annotation covers, defined by its start and end residues. Non-positional features, like this literature reference, have a start and end of zero.

- **<NOTE>** A textual description of the feature to display in the DAS client for human consumption. In this case, bibliographical information about the literature reference.

- **<LINK>** A web address (URL) and label to be displayed. A link to the Pubmed page for the reference, giving the abstract and further bibliographic information. "pubmed:12484756" is given as a label to displayed in the client.
A feature record describing a putative catalytic residue at site 46 of the sequence. The notes give details of the annotation, where they are from and the residues involved. Here there has been a mutation from site on 1GUF (chains A/B) from which the annotation was derived: serine to the chemically similar threonine. Links are provided to: the relevant page of the catalytic site atlas, to view the structure from which the annotations in the SPICE DAS client, and to two references from which the original annotation was derived.

- **<SCORE>** A measure of how accurate the annotation is. Although the actually measure is specific to each DAS server, it can be used by the client to rank the features returned by each server. In the case of annotations from the CSA, this is the percentage ID between the amino acid sequence of the query and that from which the original annotation is from.

- **<GROUP>** Used to group features together. Here, this site is grouped with all putative catalytic residues in the same CSA record “REP00702” which is comprised of all annotated residues on the same PDB structure.
The e-protein DAS servers

The e-protein project provides several DAS servers, focusing on annotations about the structure and function for proteins. Each of the three institutes involved in the project provides its own servers, demonstrating the distributed nature of the system.

Imperial College and University College London provide annotation about a protein’s fold and features of its tertiary structure, like domains. The European Bioinformatics Institute provides predictions about functional regions derived from homology to proteins reported in primary sources.

3D-Genomics
Imperial College. [http://www.sbg.bio.ic.ac.uk/3dgenomics/](http://www.sbg.bio.ic.ac.uk/3dgenomics/)

3D-GENOMICS is a database containing structural annotations for over 260 sequenced genomes. Representatives from the major branches of Life are included: Prokaryota, Eukaryota and Archaea. 3D-Genomics annotates both SCOP and Pfam domains, protein secondary structure, and sequence features such as repeats and regions of low complexity.

Catalytic Site Atlas (CSA)
European Bioinformatics Inst. [http://www.ebi.ac.uk/thornton-srv/databases/CSA/](http://www.ebi.ac.uk/thornton-srv/databases/CSA/)

The Catalytic Site Atlas is a hand-annotated database of enzyme active sites and sites involved catalytic function, derived from primary literature. The site-wise annotations are extended by homology to cover all sequences in Uniprot, with confidence that each putative site is indeed homologous.

Genomic Threading Database (GTD)
University College London. [http://bioinf.cs.ucl.ac.uk/GTD/](http://bioinf.cs.ucl.ac.uk/GTD/)

The Genomic Threading Database (GTD) contains structural annotations of proteomes, translated from the genomes of key organisms. Annotations are made using a modified version of the GenTHREADER software. GenTHREADER is a fast and powerful protein fold recognition method, which can be applied to either whole, translated genomic sequences (proteomes).

PDBsum
European Bioinformatics Inst. [http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/](http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/)

Crystallised protein structures contain much information about function: binding sites of ligands, including factors and cofactors, and those regions of the protein involved in protein-DNA and protein-protein interactions. As with the CSA, the experiment derived annotations are extended by homology to all sequences in Uniprot.
A simple query using Dasty
Starting at the e-protein web site, this section introduces the Dasty DAS client. The Dasty client is produced by the EBI written in Macromedia Flash™.

Go to the e-protein DAS portal at [http://www.e-protein.org/](http://www.e-protein.org/) and click on the link marked “Protein DAS client”. This should take you to a form which allows the e-protein servers to be searched.

Enter P53041 into the box marked “ID/Accession” and press submit.

Information about the protein PPP5_HUMAN should be displayed; it is a serine/threonine-protein phosphatase. If the accession number does not correspond to protein in Uniprot, “No hits found” will given. The form allows the Uniprot description of the protein to be searched, restricting the scope of the search to particular species.

Next to the entry for PPP5_HUMAN are two symbols: a “D” to open the Dasty DAS client, and a “U” to go to the Uniprot page for this protein.

Click on the “D” to view annotation about this protein in Dasty.

You should now be faced with screen like the one in Figure 1. The main Dasy interface is split between windows, the navigator and the visualiser. The navigator window displays an overview of the protein and the region selected in the navigator is magnified in the visualiser. As the mouse move over annotation in the visualiser window, the relevant fragment of sequence is highlighted in the protein sequence window.
For P53041, Uniprot annotates one active site (ACT_SITE) His304. If you move the mouse over the site in the features window, the text box now contains a description telling us that this site is a protein donor. Move the box in the navigator window down until it covers the results from the “CSA – EXTENDED” server.

The visualiser window contains several annotations of potential catalytic sites from three different structures contained in the Catalytic Site Atlas. The site involved in proton donation that is annotated by Uniprot (His304) is well supported, occurring in all three. In addition the structures 1AUI and 1USH contain other sites which might plausibly be involved in catalytic activity; moving over these sites shows that the annotated residues are conserved and there is a high confidence that the sites are correctly aligned.

Non-positional features, like the Uniprot description of the protein or literature references for catalytic sites can be viewed by selecting the “Non Positional” or “Both” radio buttons on the left of Dasty.
Ensembl and Spice – protein structure

The Ensembl and Spice clients are integrated into other services, the Ensembl client is integrated into the Ensembl genome browser and the Spice client maps annotation on to a protein’s structure (if known).

Go to [http://www.ensembl.org/](http://www.ensembl.org/) and search for P53041. This Uniprot accession corresponds to Ensembl gene ENSG00000011485, click on the link to this gene.

You should now see an Ensembl gene view for ENSG00000011485. The transcripts section shows that this gene only has one annotated transcript, ENST00000012443. There are, of course, many different ways to get to this page (and others like it) using the Ensembl browser and the web site contains many tutorials.

Click on “peptide info” in the transcripts section of the page.

This is the Ensembl “protview” for viewing protein features. This page has a DAS client integrated into it: the “protein features” section of the page shows positional features on a graph, the bottom of the graph show the scale in residues. Clicking on any of the features gives information about it, including links to web sites that contain further information about the annotation.

Further down the page, there is a section marked “Protein DAS report”. This contains all the non-positional annotations about the protein and tools to manage the available DAS sources, which will be described further in “Adding new DAS sources”. The DAS features of the protview client are shown in Figure 2.
Figure 2 The DAS client features of the Ensembl protview page, showing annotation for P53041.
The "CSA – extended" server provides an annotation that consists of several sites close together (ID: REP00682). Clicking on this annotation reveals that it was derived by homology from annotation on the protein structure with PDB code 1USH; none of the residues mutated from the original source.

There are four links (to the CSA entry, two pubmed references, and one marked View:1USH). Click on the link marked "view:1USH”.

The "view" link provided by the CSA server starts the Spice protein structure viewer and DAS client. This is a Java webstart application and, if this is the first time that it has been run on your machine, you will be asked whether or not to trust it; click run. A screenshot of the client is shown in Figure 3.

**Figure 3:** the Spice DAS client showing the structure of 1USH and DAS annotation. Highlighted in space-fill are the putative MN binding sites from the PDBsum_ligands server (originating from the structure 1HPU), also highlighted on the right in the sequence based DAS client.
Click on a feature in the DAS window of Spice, a domain for example.

The residues corresponding to the feature are selected on the structure; the Spice client automatically maps annotation from the Uniprot sequence to residues in the PDB structure.

Find the CSA – literature DAS server and click on the “Catalytic Re” label. This should highlight the entire line.

Go to the Display menu and select wireframe. The catalytic residues are now picked out.

At the bottom left of the screen, there is a text box with “Enter RASMOL like command”. Type in “wireframe 0.5”. The catalytic residues should now be thicker and more obvious.

Play with protein structure. The commands to manipulate the structure are Rasmol –like (left-click+mouse to rotate, shift+left-click+mouse up or down to zoom in or out). Explore the relationship between the catalytic residues and the various putative ligand contacts from the PDBsum_ligands server.

It is not clear which metal is bound by the enzyme in its natural state, although ZN and MN are strong candidates. There appears to be a binding site for ATP, or an analog like A12 or ADN, above the metal binding site.
Conflicting annotation

Not all annotation from by DAS servers is perfect. In fact, one of the strengths of the DAS is that it allows annotation from many sources to be easily compared to see if they support each other.

Find the protein with accession Q9UIL8 (this is ENSP0000343708 in Ensembl v37, Feb 2006).

Several DAS servers return the prosite pattern PS00146, a beta-lactamase active site. However the Uniprot server notes that this pattern is a known false-positive, which can be confirmed by looking at the Prosite website linked from the Uniprot annotation: this pattern is only active in prokaryotes. On reflection, there is no support from other DAS sources for this protein having catalytic activity and we should have been sceptical even if the annotation was not labelled as a false-positive.

Versioning

The e-protein project servers use Uniprot accession numbers to identify proteins. While the purpose of each entry is to represent one protein, the representative amino acid sequence may change between Uniprot versions -- perhaps because new information reveals the original entry to contain a mutation. More severely, all splice variants of a protein are contained within a single entry and a switch of splice variant, the addition or deletion of an exon, will considerably change the sequence and co-ordinates annotations to sites of the old one may be off by a considerable margin.

Each DAS server returns a version, describing which protein sequence it annotates. The version is string of characters which changes randomly if a single amino acid of the sequence is changed, so if two versions strings agree there is a high probability that the annotations are for the same sequence; if they are different, then the definitely not from the same sequence but it is not possible to tell how different (technical detail: the version is cryptographic hash of the amino acid sequence using the MD5 algorithm\(^1\)).

The Dasty client reveals which servers have given annotation for a sequence which conflicts with the one it is displaying. Most of the annotation from the dissenting servers is probably valid since it is supported by annotation from other servers which do annotate the correct sequence.

\(^1\) Invented by R. Rivest (1991). Although serious flaws have been discovered in the algorithm that mean it is no-longer recommended for cryptographic use, its use within the e-protein project does not require such assurances and so is safe.
Adding new DAS sources

Depending on the client used, a user may not be restricted just to the DAS servers installed by default and sources of annotation may be added or removed as required: perhaps the some sources suggest that the query may be a transmembrane protein and you wish to view to add a server that annotates transmembrane helices, or the protein belongs to a very promiscuous family and the number of potential interacting ligands is too great to be useful.

Go to the peptide view of a protein in Ensembl, see “Ensembl and Spice – protein structure” for an example.

Scroll down the page until you reach a box marked “DAS Sources”, which contains a list of source that the client currently knows about. Tick the boxes next to one or two of the source and press the update button.

Only those sources checked in this list are queried by the Ensembl client and so have results displayed in the "Protein Feature" graph. On clicking update, the screen should refresh and data from the newly checked sources should now appear in the “Protein Features” and “Protein DAS report” section of the web page.

Go back to the “DAS Sources” section of the web page and click on “Manage Sources”. A new window will open containing a list of DAS sources.

Click “Add Data Source” on the left-hand side of the page, under the heading “Manage Sources”.

By default, the Ensembl client queries the DAS registry of sources and the page now opened contains a long list of every registered data source, along with a short description and a link to further information. Alternatively, sources can be specified by hand by entering the correct URL and so non-public or experimental sources can be added to the view.
Check boxes next to a couple of sources, scroll back to the top of the page and click the red “Next” button. This gives a page asking which of the Ensembl views the sources should be displayed in — protview is fine since this is what we are currently using. If you are adding a source by hand, the co-ordinate system (query by: Ensembl gene ID, Uniprot accession number, etc) must also be specified here. Click next.

The “Display configuration” window should now have opened. The options here affect how the client displays the annotation it receives, whether to use style sheets (server hints about the annotation should be displayed) and what to do with the scores for each feature.

Click on “Finish” to take you back to the “DAS Sources” page. The new sources you selected should now have appeared at the bottom of the sources list, along with two buttons: edit and remove.

Click on “Close Window” in the top right-hand corner of the page

The protview of your protein should now have reloaded with the new sources already selected.

Open [http://das.sanger.ac.uk/registry/](http://das.sanger.ac.uk/registry/) in a web browser. Click on “list sources” in the right-hand menu.

The purpose of the DAS registry is to keep track of which DAS servers are around, whether they reliably up, and what capabilities they support. The full list of servers is similar to that which we have just looked at in Ensembl but may be searched using the facilities at the top of the page. The little letter icon next to each source allows you to email the source to a friend; the email contains a URLs which open up either the Ensembl or Spice clients and attaches the new source.

Some servers may have “It has not been possible to validate this server for > 2 days” where the URL should be. The DAS registry regularly checks that the servers contained within it are working and servers which fail this check for a prolonged period of time are flagged as such.
Types and annotation quality

While the <TYPE> and <METHOD> tags for each annotation provide some information about what the annotation is and how it was derived, the content of these tags is defined by the server from which the annotation came and may not be consistent with other sources. For example: a domain may be annotated simply as a domain, or by whichever method was used to find it (Smart, Pfam, SCOP, etc). All these sources are annotating the same biological feature, using a variety of different methods (e.g. protein structure, sequence motifs) to do so, but this not immediately obvious and makes any form of analysis of the data difficult. Work is underway to cluster the currently extant <TYPE> tags into meaningful groups and produce a controlled vocabulary to describe them.

A further problem is that there is no associated confidence with the data; experimentally verified results have the same weight as uncurated computational predictions and one can only tell one from the other by looking at the description of each annotation. The BioSapiens\(^2\) (http://www.biosapiens.info/) project is encouraging servers to include an <EVIDENCE> tag, based on the evidence codes used by the Gene Ontology Consortium (http://www.geneontology.org/), to describe the confidence in the annotation. The valid codes for the <EVIDENCE> tags can be grouped into four broad categories: manually curated, experimental verified, computationally predicted, or no assignment possible.

The stricter control of the <TYPE> tag and the new <EVIDENCE> tag will enable future DAS clients to filter results for confidence and display annotations in biologically meaningful clusters rather than grouping by server, as is currently the case, helping the user to compare different annotations about the same biological feature.

---

\(^2\) BioSapiens is a virtual institute for genome annotation funded by the European Union. “The objective of the BioSapiens network of Excellence is to provide a large scale, concentrated effort by laboratories distributed around Europe to annotate genome data, using both informatics tools and input from experimentalists.”