TEPIDAS: A DAS Server for Integrating T-Cell Epitope Annotations

M. García-Boronat, C.M. Díez-Rivero, and Pedro Reche

Abbreviations

CMV	Cumulative phenotypic frequency
DAS	Distributed annotation system
HLA I	Human leukocyte antigen class I
PSSM	Position-specific scoring matrix

Introduction

Recent years have witnessed the birth of Immunoinformatics, an emerging subdiscipline of Bioinformatics. With the burgeoning explosion of immunological data, computational analysis has become an essential element of immunology research, facilitating the understanding of the immune function by modeling the interactions among immunological components (Petrovsky and Brusic 2006). Another major role in Immunoinformatics is the efficient management, storage, and annotation of such data. Following those principles, a large number of immunoinformatics resources, including immune-related databases and sophisticated analysis software, are available through the World Wide Web (Davies and Flower 2007). Collectively, these resources contribute to the advances made in immunological research. Yet, there is still a major step to be taken toward the integration of all these resources, as ideally, multiple research groups should be able to exchange and compare their data, in a quick and efficient fashion.

In this chapter, we show an example of how an epitope database can be integrated to other database resources using the Distributed Annotation System (DAS) (Dowell et al. 2001). For that we describe the TEPIDAS server, a DAS Annotation Server of HLA I-restricted CD8 T-cell epitopes specific of human pathogenic organisms.

P. Reche (✉)
Facultad de Medicina, Departamento de Immunología (Microbiología I), Universidad Complutense de Madrid, Pabellón 5º, planta 4ª, 28040, Madrid, Spain
e-mail: parecheg@med.ucm.es

The Distributed Annotation System

Introduction

The distributed annotation system defines a communication protocol used to exchange biological annotations from a number of heterogeneous distributed databases. The key idea behind the DAS concept is that annotations should not be provided by single centralized databases but instead be spread over multiple sites. This distribution of data encourages a divide-and-conquer approach to annotation, where experts provide and maintain their own annotations.

The Protocol

Currently, there are two versions of the DAS protocol. The original DAS protocol (DAS/1) was designed to serve annotation of genomic sequences. That protocol was later extended (DAS/2) to be applicable to alignments and 3D structure information (Prlic et al. 2005). It is very likely that further extensions of the protocol will appear in a near future, such as the new extension for electron microscopy data recently published by Macias et al. (2007).

The DAS protocol is a simple http-based client–server system. DAS clients make requests in the form of a URL to the servers and receive simple XML responses (Crook and Howell 2007). The architecture of the system will be next described in the following subsection.

The Architecture of the System

The basic system is composed of a reference server, one or more annotation servers, and an annotation viewer. The reference server is responsible for serving genome maps, sequences and information related to the sequencing process. Annotation servers are responsible for returning the annotations on a defined region (given a start and stop position coordinates) of the genome. The annotation viewer can either be a simple web browser, which will visualize the raw XML data provided by the server, or a graphical client such as the Center for Biological Sequence Analysis (CBS) DAS viewer (Olason 2005) accessible at http://www.cbs.dtu.dk/cgi-gin/das. This viewer translates the XML annotations to aligned graphical tracks making it easier to visualize the features along the length of the protein. Additional information about the annotations is shown in a pop-up window when the mouse points to an annotation track.

Although the servers are conceptually divided between reference and annotation servers, there is in fact no key difference between them. A single server can provide both reference sequence information and annotation information. The only functional

difference is that the reference sequence server is required to serve the coordinate map and the raw DNA, while annotation servers have no such requirement. Our TEPIDAS server falls into the category of annotation servers.

The DAS Registry

The DAS Registry is a public server (http://www.dasregistry.org) dedicated to the registration, validation, and listing of worldwide DAS servers. One can browse the list of available DAS sources at the Registry, as well as register his own DAS server for public use. The Registry automatically validates the DAS server when it is being registered, ensuring that it returns well-formed XML responses. In addition, it periodically tests DAS sources and notifies their administrators if they are unavailable.

When you register your DAS server, you have to specify the Coordinate System of your source in order to describe the kind of data that are being made available. This information is important for the DAS clients to deal with data correctly, as they often can accept data served in multiple coordinate systems. The Coordinate System is described by the following four fields: "Authority," "(assembly) Version," "Type," and "Organism." The assembly version is important for genome assemblies, but not really applicable for other datasets like UniProt sequences; therefore, this field is optional. The "authority" is the name of an authority/institution that defines the accession codes of a coordinate system or that provides a gene build. In the latter case this field also contains the "version" number of the assembly. The "type" or category of the coordinate system refers to the physical dimension of the annotated data. Some examples include: Chromosome, Clone, Protein Sequence, and Protein Structure. The last field is the "organism" the data refer to. Not every DAS source is organism specific, and therefore this field is optional.

During the registration process, you also have to specify the capabilities of your DAS source, that is the types of queries that your server will be able to serve a response to. Some basic queries that can be used by a client to interrogate a DAS server are: "dna," "features," and "types." The "dna" query can be used to fetch a segment of DNA from a reference server. "features" is the query used to retrieve the actual annotations, and the "types" query returns a summary of the available annotation types. These three are just some examples of DAS queries. Readers can access the full list and specification of query types at the DAS web page (http://www.biodas.org).

The TEPIDAS server has been registered at the DAS registry since February 2008 and has the unique id DS_545. The coordinate system defined for TEPIDAS is Uniprot (Wu et al. 2006), as the "authority," and Protein Sequence, as the "type." As for TEPIDAS capabilities, our server implements the "types" and "features" queries. Note that our server is just an annotation server, and therefore it does not provide the "dna" query, served only by reference servers. A comprehensive description of the TEPIDAS server follows next.

TEPIDAS

TEPIDAS is a DAS annotation server that provides annotations for CD8 T-cell epitopes consisting of the distinct HLA I molecules to which that epitope binds, following the UniProt coordinates system. TEPIDAS is implemented using ProServer (Finn et al. 2007), a lightweight Perl-based DAS server that does not depend on a separate HTTP server. The annotations are precalculated and the results stored in a relational database, allowing for fast retrieval and update of data. When a client makes a query to the TEPIDAS server, ProServer simply retrieves the relevant information from the relational database and composes the XML response.

Annotations Served by TEPIDAS

TEPIDAS annotates CD8 T-cell epitopes according to the HLA I molecules that restrict them. Epitopes were obtained from the EPIMHC (Reche et al. 2005) and IMMUNEEPITOPE (Peters et al. 2005) databases, and were selected to be experimentally defined in humans infected with the pathogen or immunized with the relevant source antigen. HLA I-restriction annotations can be classified as experimental, when determined experimentally, or predicted. Predictions of the epitopes binding HLA I molecules were obtained using a set of 72 position-specific scoring matrices (PSSMs), also known as weight matrices of profiles, which are obtained from aligned peptides known to bind to the relevant HLA I molecules. This predictive method is described in full detail at (Reche et al. 2002, 2004). In addition to the experimental and predicted data, the cumulative phenotypic frequency (CMV) of the T-cell epitope HLA I restriction is also provided for five ethnic groups (Black, Caucasian, Hispanic, North American natives, and Asian). CMV was computed using the gene and haplotype frequencies of the relevant HLA I alleles (Reche et al. 2006). The potential population protection coverage of a T cell epitope-based vaccine is determined by the percentage of the population that could elicit a T cell response to the epitopes, which in turn is given by the CMV of HLA I molecules restricting these epitopes.

TEPIDAS Query Capabilities

As we mentioned before, TEPIDAS capabilities include the “types” and “features” queries. An explanation and an example for each query follow next.

The “types” query returns a list of all the distinct HLA I molecules that are used to annotate the epitopes. A total of 125 different HLA I restriction elements are included in TEPIDAS. To make this query to the server, you simply have to access the following URL through your web browser:

http://imed.med.ucm.es:9000/das/tepidas/types

and the XML response you will get is shown as follows.

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE DASTYPES SYSTEM "http://www.biodas.org/dtd/dastypes.dtd">
<DASTYPES>
  <GFF version="1.0" href="http://imed.med.ucm.es:9000/das/tepidas/types">
  <SEGMENT version="1.0">
    <TYPE id="HLA-A*02" method="Experimental" category="default"></TYPE>
    <TYPE id="HLA-A*0201" method="Experimental" category="default"></TYPE>
                                .
                                .
                                .
    <TYPE id="HLA-B*02706" method="Predicted" category="default"></TYPE>
    <TYPE id="HLA-B*02709" method="Predicted" category="default"></TYPE>
    <TYPE id="HLA-B*027" method="Predicted" category="default"></TYPE>
  </SEGMENT>
</GFF>
</DASTYPES>
```

Only a part of the XML response file is shown due to length constraints. Each type has an "id" that corresponds to the name of the HLA I molecule. There is also a "method" attribute that distinguishes between experimental and predicted annotations. In addition, a third attribute named "category" can be used to group different types, although we have not used that attribute, and therefore *default* is the "category" shown in the response.

The other type of query supported by TEPIDAS is the "features" query, which returns the actual annotations made on a reference UniProt sequence. An annotation feature includes the following information: the start and end position of the feature annotated, the method used to annotate it (experimental or predicted), the type of the annotation (the HLA I molecule to which it binds), a link to the UniProt page of the reference protein sequence, and a note field with additional complementary information. The information on the note varies depending on the feature's method. Common fields in the note of both methods are: the epitope source species name and taxonomy identifier, the name of the source protein, the cumulative phenotypic frequency (CMV) of the T-cell epitope HLA-I restriction for five ethnic groups (Black, Caucasian, Hispanic, North American natives, and Asian), and the immunogen type. Specific fields for the features with an experimental "method" are: T-cell epitope activity assays, the experimental HLA I restriction element, its binding level (low, moderate, high, or unknown), and the predicted HLA I restriction elements. As for the features with a predicted "method" the note also includes the predicted HLA I restriction element, as well as an extended prediction with additional HLA I restriction elements for that epitope.

The "features" query has several arguments that can be optionally used to restrict the results. For example, the following URL string:

http://imed.med.ucm.es:9000/das/tepidas/features?segment=P26664

will return all the features annotated on the UniProt protein sequence identified with the accession number P26664 (which will also be the features id).

If we want to restrict our query to the annotations on a particular region of the protein sequence, we could use:

http://imed.med.ucm.es:9000/das/tepidas/features?segment=Q9WMX2:885,893

which returns all the features for the protein sequence with accession number Q9WMX2 that lie within the region defined by the start and end positions 885 and 893. The XML response to this query is shown as follows.

```
<?xml version="1.0" standalone="yes"?>
<!DOCTYPE DASGFF SYSTEM "http://www.biodas.org/dtd/dasgff.dtd">
<DASGFF>
  <GFF version="1.01" href="http://imed.med.ucm.es:9000/das/tepidas/features">
    <SEGMENT id="Q9WMX2" version="1.0" start="885" stop="893">
    <FEATURE id="Q9WMX2" label="Q9WMX2">
      <TYPE id="HLA-A*2402" reference="no" subparts="no" superparts="no">
       HLA-A*2402</TYPE>
      <METHOD id="Experimental">Experimental</METHOD>
      <START>885</START>
      <END>893</END>
      <ORIENTATION>0</ORIENTATION>
      <NOTE>
       Epitope Source Species: Hepatitis C virus; TaxID: 11103
       Epitope Source Protein: Genome polyprotein
       T cell Epitope Activity positive on: 51 Chromium Release,
       Cytokine bioassay
       MHCI Restriction Element: HLA-A*2402 (Experimental)
       MHCI Binding level: unknown
       Predicted MHCI Restriction: HLA-A*24,  HLA-A*2402
       Cummulative Phenotypic Frequency of MHCI(%):
       5.5(Black),12.8(Caucasian),22.9(Hispanic),
       40.3(North American Natives),34.3(Asian)
       Immunogen: Infection</NOTE>
     <LINK href="http://www.ebi.uniprot.org/unipro t-
srv/uniProtView.do?proteinAc=Q9WMX2">http://www.ebi.uniprot.org/uniprot-
srv/uniProtView.do?proteinAc=Q9WMX2</LINK>
    </FEATURE>
    <FEATURE id="Q9WMX2" label="Q9WMX2">
      <TYPE id="HLA-A*24" reference="no" subparts="no" superparts="no">
       HLA-A*24</TYPE>
      <METHOD id="Predicted">Predicted</METHOD>
      <START>885</START>
      <END>893</END>
      <ORIENTATION>0</ORIENTATION>
      <NOTE>
       Epitope Source Species: Hepatitis C virus; TaxID: 11103
       Epitope Source Protein: Genome polyprotein
       T cell Epitope Activity: predicted
       MHCI Restriction Element: HLA-A*24 (Predicted)
       MHCI Binding level: unknown
       Extended predicted MHCI Restriction: HLA-A*24, HLA-A*2402
       Cummulative Phenotypic Frequency of MHCI(%):
       5.5(Black),12.8(Caucasian),22.9(Hispanic),
       40.3(North American Natives),34.3(Asian)
       Immunogen: Infection</NOTE>
     <LINK href="http://www.ebi.uniprot.org/uniprot-
srv/uniProtView.do?proteinAc=Q9WMX2">http://www.ebi.uniprot.org/uniprot-
srv/uniProtView.do?proteinAc=Q9WMX2</LINK>
    </FEATURE>
    </SEGMENT>
  </GFF>
   </DASGFF>
```

Example: Access TEPIDAS from the SPICE Graphical Client

In the previous section we have described how to access TEPIDAS annotations using formatted queries from a web browser, and we have also shown examples of the XML responses to the queries. We will now describe a different way of accessing TEPIDAS from a graphical client such as SPICE (Prlic et al. 2005). We hope that this example will illustrate the integration capability of DAS.

SPICE is a Java program that can be used to visualize annotations of protein sequences and protein structures. It is available at: http://www.efamily.org.uk/software/dasclients/spice. SPICE accepts either a PDB (Berman 2008) or a UniProt code, and integrates information from four different types of DAS servers: (1) a protein sequence server that provides the sequence (typically UniProt), (2) an alignment server that provides the alignment between the protein sequence and its structure, (3) a structure server that serves the 3D coordinates displayed, and (4) several feature servers that provide precalculated annotations, as for example TEPIDAS among others.

The SPICE viewer window consists of (1) a left structure panel, which provides a 3D visualization of the molecule using the open source Jmol library (http:/www.jmol.org), and (2) a right 2D feature panel that displays the annotations provided by the distributed servers. This is illustrated in Fig. 1 using the protein sequence with UniProt code P35961 as an example. As we can appreciate in Fig. 1, SPICE has automatically mapped that protein sequence to PBD "19GN" using its default alignment server. Figure 1 clearly shows how different annotations from several DAS servers can be integrated and collectively visualized through a graphical client such as SPICE. Users can choose which DAS annotations servers to use, as well as add new local DAS sources that are still under development or have not been registered with the DAS registry.

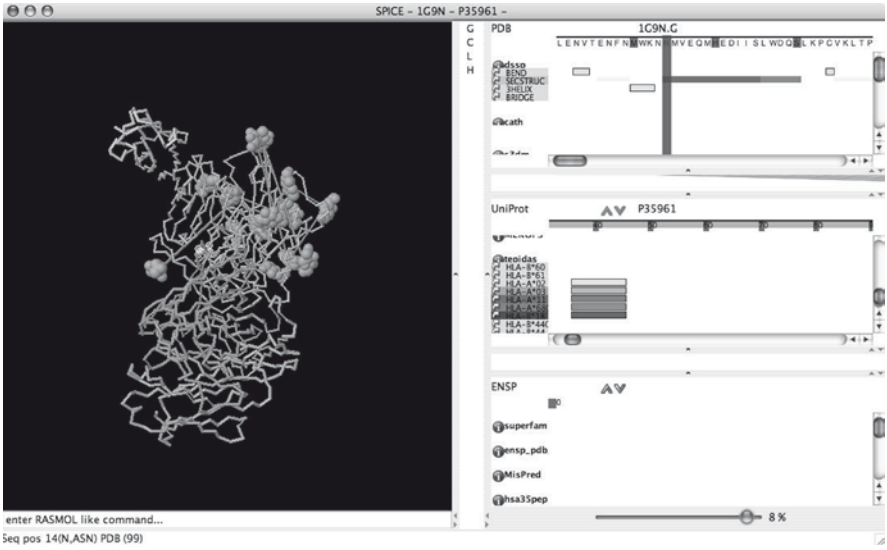

Fig. 1 SPICE viewer window. *Left panel* provides a 3D visualization of the molecule. *Right panel* displays the annotations provided by the distributed serves. This figure was generated using the UniProt code P35961 as the reference sequence. SPICE's alignment server automatically maps the protein sequence to a 3D structure (1G9N in this example). Feature annotations from TEPIDAS are displayed in the *right center panel* as *rectangular* tracks colored as the HLA I molecules on their *left* under the tepidas source descriptor

SPICE retrieves the protein sequence pertaining to the selected UniProt code and displays it as a ruler with relative position numbers, although there is a zoom feature that allows it to be expanded up to amino acid level as shown in Fig. 2 TEPIDAS annotation features are listed below the sequence in that figure. On the left of the panel, below the "tepidas" descriptor, appears the type of HLA I molecule of the corresponding feature shown as a colored rectangle on the right. When the user clicks on a feature, a pop-up window appears, containing all the information of the feature, including the explanatory note. In addition, the PDB coordinates of the selected feature will be highlighted at the left panel, enabling the location of the epitope at the 3D structure. Figure 3 shows an example of a pop-up window with feature information.

Fig. 2 SPICE zooming capability. Protein sequence visualized at amino acid level

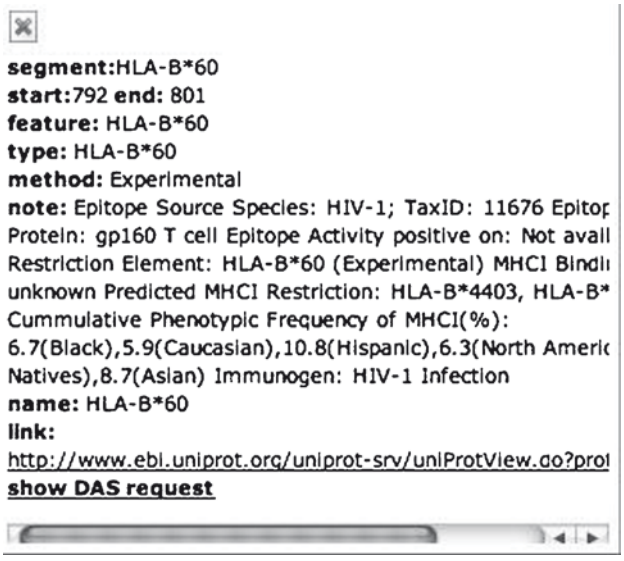

Fig. 3 Pop-up window containing all the information for feature HLA-B*60 annotated for protein sequence referenced by UniProt code P35961

Conclusion

DAS is an important, simple, and yet a powerful system for exchanging and viewing biological data that are already being used in real-world bioinformatics applications. The TEPIDAS annotation server described in this chapter is a clear example of how epitope data can be integrated and shared by the research community using the DAS architecture. The complexity of immune interactions and the data-intensive nature of immune research make Immunoinformatics a suitable area that could greatly benefit from the advantages of using such a powerful integration and annotation system, allowing to gain a more insightful understanding of the complexities of the immune system.

Acknowledgments We would like to thank Alfonso Valencia, Osvaldo Graña, and Jaime Fernandez Vera from the Spanish National Cancer Research Center (CNIO) for their helpful advice on DAS and ProServer. Work and authors were supported by grant SAF2006-07879 from the "Ministerio de Educación y Ciencia" of Spain, granted to PR.

References

Berman HM (2008) The Protein Data Bank: a historical perspective. Acta Crystallogr A 64(Pt 1): 88–95

Crook SM, Howell FW (2007) XML for data representation and model specification in neuroscience. Methods Mol Biol 401:53–66

Davies MN, Flower DR (2007) Harnessing bioinformatics to discover new vaccines. Drug Discov Today 12(9–10):389–395

Dowell RD, Jokerst RM et al (2001) The distributed annotation system. BMC Bioinformatics 2:7

Finn RD, Stalker JW, Jackson DK et al (2007) ProServer: a simple, extensible Perl DAS server. Bioinformatics 23(12):1568–1570

Macias JR, Jimenez-Lozano N, Carazo JM (2007) Integrating electron microscopy information into existing Distributed Annotation Systems. J Struct Biol 158(2):205–213

Olason PI (2005) Integrating protein annotation resources through the Distributed Annotation System. Nucleic Acids Res 33(Web Server issue):W468–W470

Peters B, Sidney J et al (2005) The immune epitope database and analysis resource: from vision to blueprint. PLoS Biol 3(3):e91

Petrovsky N, Brusic V (2006) Bioinformatics for study of autoimmunity. Autoimmunity 39(8):635–643

Prlic A, Down TA, Hubbard TJ (2005) Adding some SPICE to DAS. Bioinformatics 21(Suppl 2): ii40–ii41

Reche PA, Glutting JP, Reinherz EL (2002) Prediction of MHC class I binding peptides using profile motifs. Hum Immunol 63(9):701–709

Reche PA, Glutting JP et al (2004) Enhancement to the RANPEP resource for the prediction of peptide binding to MHC molecules using profiles. Immunogenetics 56(6):405–419

Reche PA, Zhang H et al (2005) EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. Bioinformatics 21(9):2140–2141

Reche PA, Keskin DB, Hussey RE et al (2006) Elicitation from virus-naïve individuals of cytotoxic T lymphocytes directed against conserved HIV-1 epitopes. Med Immunol 5:1

Wu CH, Apweiler R, Bairoch A et al (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34(Database issue):D187–D191