Distributed Annotation System (DAS)

Osvaldo Graña Castro
ograna@cnio.es

Structural Computational Biology Group
CNIO

VII Jornadas de Bioinformática 2006
Zaragoza

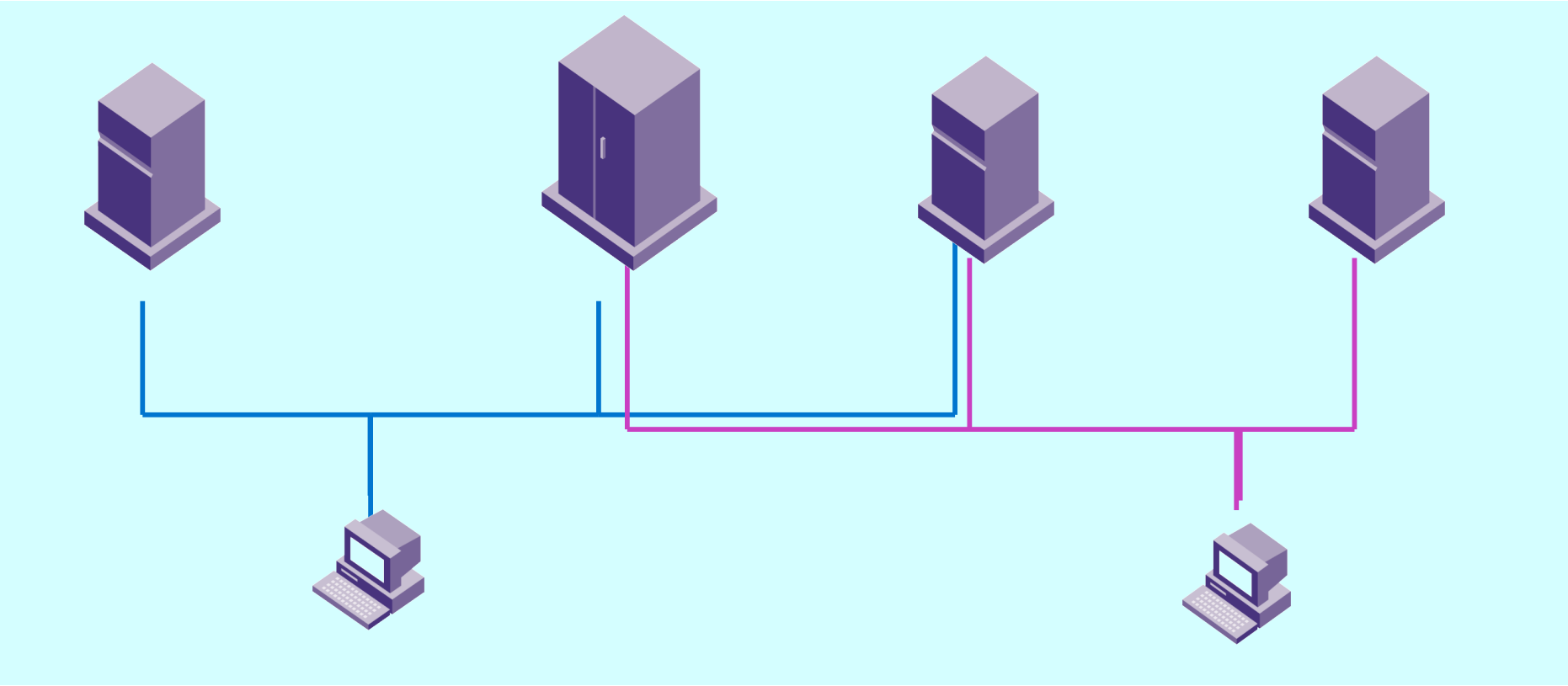
Distributed Annotation System (DAS)

The distributed annotation system (DAS) is a client-server system in which a single client integrates information from multiple servers. It allows a single machine to gather up genome annotation information from multiple distant web sites, collate the information, and display it to the user in a single view.

The original version 1 specification, written by Lincoln Stein and Robin Dowell, is the basis for a number of clients and servers. DAS/1 servers are currently running at WormBase, FlyBase, Ensembl, TIGR, UCSC, KEGG, INB, etc.

A single DAS server is designated as either an *annotation server* or as the *reference server*. The reference server provides essential structural information about the genome: the physical map which relates one entry point to another (where an "entry point" is an arbitrary segment of the sequence, such as a sequenced BAC, a contig, or a whole assembled chromosome), the DNA sequence for each entry point, and some standard authorship information. Using either a freestanding application or a web site, that acts like a DAS client, researchers can interrogate one or more annotation servers to retrieve features in a region of interest. The servers return the results using a standard data format, allowing the sequence browser to integrate the annotations and display them in graphical or tabular form.

Distributed Annotation System (DAS)

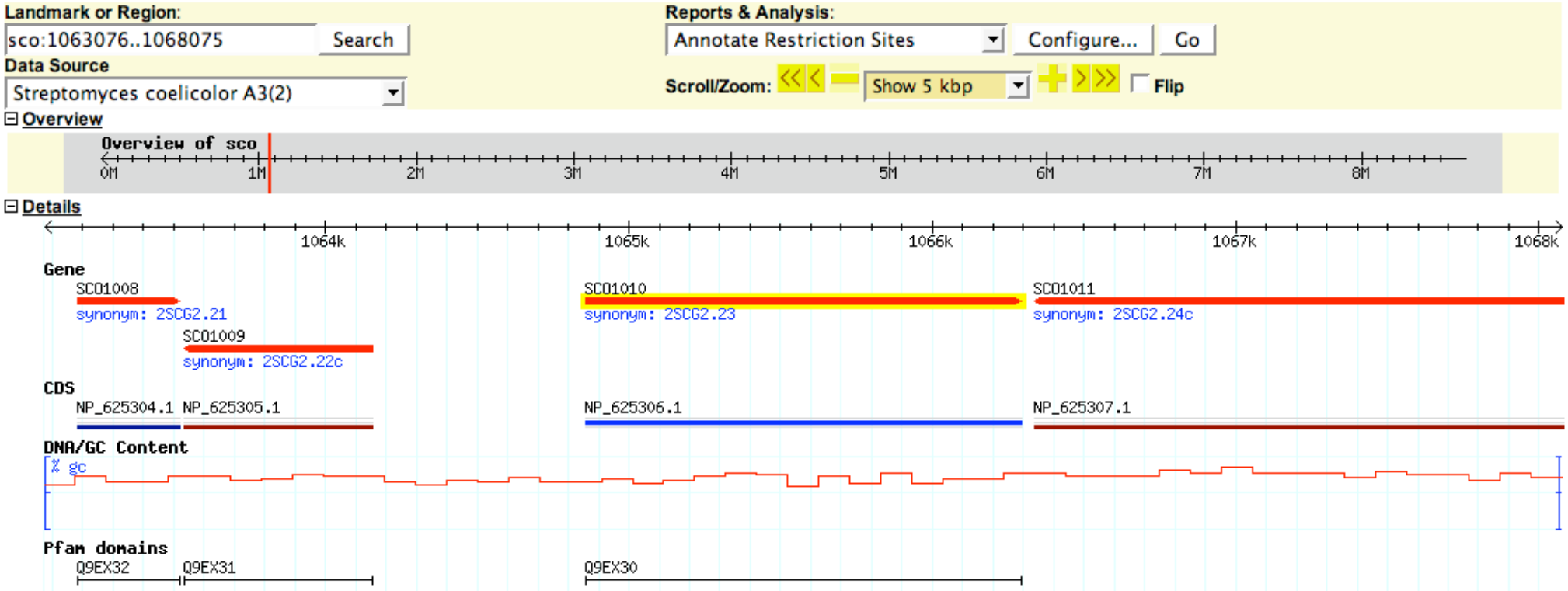


What is an annotation ?

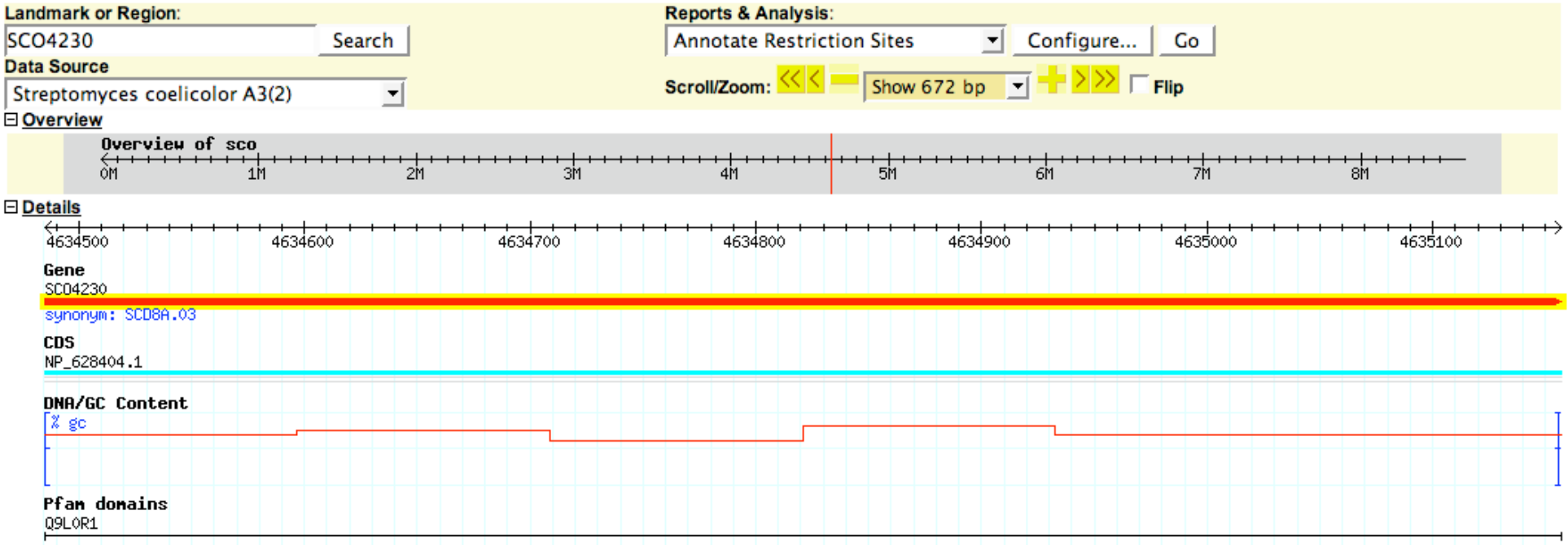
Any feature that is defined within two genomic coordinates:

- –Genes
- –Exons
- –CDS
- –Restriction Sites
- –Ribosome binding sites
- –Repeats
- –PolyA signals
- –…….
- –Scientific references (papers)

What is an annotation ?



What is an annotation ?



DAS implementations

There are currently two DAS specifications:

- –DAS v1 (Lincoln Stein & Robin Dowell, 2000)
- –DAS v2 (it officialy started in July 2004, 2 year NIH grant - Stein lab, Affymetrix, EBI/Sanger and Dalke scientific).

DAS v1

•The distributed annotation system (DAS) consists of a reference sequence server and one or more annotation servers.

•The reference sequence is the one on wich to base annotations. The reference sequence consists of a set of *entry points* into the sequence, and the lengths of each entry point.

•The entry points describe the top level items on the reference sequence map. It is possible for each entry point to have a substructure, basically a series of subsequences (components) and their start and end points. This structure is recursive.

•Each annotation is unambigously located by providing its position as the start and stop positions relative to a *reference sequence.* The reference sequence can be one of the entry points, or any of the subsequences within the entry point.

•Annotations have *types*, *methods* and *categories*. The annotation **type** is selected from a list of types that have biological significance, and correspond to EMBL/GenBank feature table tags (or our own feature tags as well). Examples of annotation types include “exon”, “intron”, “CDS”. The annotation **method** is intended to describe how the annotated feature was discoverd (it may include a reference to a software program). The annotation **category** is a broad of functional category that can be used to filter, group and sort annotations. “Experimental”, “Structural”, “Translation” are all valid categories.

DAS v1

Although the servers are theoretically divided between reference servers and annotation servers, there is in fact no key difference between them. A single server can provide both the reference sequence information and annotation information. The main functional difference is that the reference sequence server is required to serve the sequence map and the raw DNA, while annotation servers have no such requirement.

Client/Server interactions

All DAS requests take the form of a URL. Each URL has a site-specific prefix, followed by a standarized path and query string. This standarized path begins with the string /das. This is followed by URL components containing the data source name and a command.

```
http://www.wormbase.org/db/das/elegans/features?segment=CHROMOSOME_I:1000,2000
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^ ^^^ ^^^^^^^ ^^^^^^^^ ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  site-specific prefix    das  data   command  arguments
                           src
```

In this case, the site-specific prefix is **http://www.wormbase.org/db**. The request begins with the standarized path **/das**, and the data source, in this case **/elegans**. This is followed by the command **/features**, which requests a list of features, and a query string providing named arguments to **/features** command.

DAS v1

The queries

The following queries are recognized by reference and/or annotation servers. Each of these queries begins with some site-specific prefix. All of them return as output an XML document.

dsn

Returns the list of data sources that are available from a particular (reference/annotation) server.

PREFIX/das/dsn

http://www.wormbase.org/db/das/dsn

DAS v1

The queries

dsn response

Format:

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE DASDSN SYSTEM "http://www.biodas.org/dtd/dasdsn.dtd">
<DASDSN>
 <DSN>
  <SOURCE id="id1" version="version">source name 1</SOURCE>
  <MAPMASTER>URL</MAPMASTER>
  <DESCRIPTION>descriptive text 1</DESCRIPTION>
 </DSN>
 <DSN>
  <SOURCE id="id2" version="version">source name 2</SOURCE>
  <MAPMASTER>URL</MAPMASTER>
  <DESCRIPTION href="url">descriptive text 2</DESCRIPTION>
 </DSN>
 ...
</DASDSN>
```

<!DOCTYPE> (required; one only)
The doctype indicates which formal DTD specification to use. For the dsn query, the doctype DTD is "http://www.biodas.org/dtd/dasdsn.dtd".

<DASDSN> (required; one only)
The appropriate doctype and root tag is DASDSN.

<DSN> (required; one or more)
There are one or more <DSN> tags, one for each data source. Each <DSN> contains one <SOURCE> tag, one <MAPMASTER> tag, and optionally one <DESCRIPTION> tag.

<SOURCE> (required; one per DSN tag)
This tag indicates the symbolic name for a data source. The symbolic name to use for further requests can be found in the **id** (required) attribute. A source **version** attribute is optional, but strongly recommended. The tag body contains a human-readable label which may or may not be different from the ID.

<MAPMASTER> (required; one per DSN tag)
This tag contains the URL (site.specific.prefix/das/data_src) that is being annotated by this data source. For an annotation server, this is the reference server which is being annotated. For a reference server, this would echo its own URL.

<DESCRIPTION> (optional)
This tag contains additional descriptive information about the data source. If an **href** (optional) attribute is present, the attribute contains a link to further human-readable information about the data source, such as its home page.

DAS v1

The queries

entry_points
Returns the list of sequence entry points available and their sizes in base pairs.
Scope: reference servers.

PREFIX/das/DSN/entry_points

http://www.wormbase.org/db/das/elegans/entry_points

DAS v1

The queries

entry_points response

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE DASEP SYSTEM "http://www.biodas.org/dtd/dasep.dtd">
<DASEP>
 <ENTRY_POINTS href="url" version="X.XX">
  <SEGMENT id="id1" start="start1" stop="stop1" type="type" orientation="+">descriptive text</SEGMENT>
  <SEGMENT id="id2" start="start2" stop="stop2" type="type" orientation="+">descriptive text</SEGMENT>
  <SEGMENT id="id3" start="start3" stop="stop3" type="type" orientation="+">descriptive text</SEGMENT>
  ...
 </ENTRY_POINTS>
</DASEP>
```

<!DOCTYPE> (required; one only)
The doctype indicates which formal DTD specification to use. For the entry_points query, the doctype DTD is "http://www.biodas.org/dtd/dasep.dtd".

<DASEP> (required, one only)
The appropriate doctype and root tag is DASEP.

<ENTRY_POINTS> (required, only one)
There is a single <ENTRY_POINTS> tag. It has a version number (required) in the form "N.NN". Whenever the DNA of the entry point changes, the version number should change as well.

The **href** (required) attribute echoes the URL query that was used to fetch the current document.

 (optional; zero or more)
Each segment contains the attributes **id**, **start**, **stop** and **orientation**. The id is a unique identifier, which can be used as the reference ID in further requests to DAS. The start and stop indicate the start and stop positions of the segment. Orientation is one of "+" or "-" and indicates the strandedness of the segment (use "+" if the segment is not intrinsically ordered).

If the optional **subparts** attribute is present and has the value "yes", it indicates that the segment has subparts.

If the optional **type** attribute is present, it can be used to describe the type of the segment (for future compatibility with Sequence Ontology-based feature typing).

For compatibility with older versions of the specification, the tag can use a **size** attribute rather than **start** and **stop**, and can omit the **orientation** attribute:

```
<SEGMENT id="id" size="123456">
```

DAS v1

The queries

dna

Returns the DNA corresponding to the indicated segment.

Scope: reference servers

Arguments: segment (required; one or more).

PREFIX/das/DSN/dna?segment=RANGE[;segment=RANGE…]

http://www.wormbase.org/db/das/elegans/dna?segment=I:1,30

DAS v1

The queries

dna response

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE DASDNA SYSTEM "http://www.biodas.org/dtd/dasdna.dtd">
<DASDNA>
 <SEQUENCE id="id" start="start" stop="stop" version="X.XX">
  <DNA length="NNNN">
    atttcttggcgtaaataagagtctcaatgagactctcagaagaaaattgataaatattat
    taatgatataataataatcttgttgatccgttctatctccagacgattttcctagtctcc
    agtcgattttgcgctgaaaatgggatatttaatggaattgtttttgtttttattaataaa
    taggaataaatttacgaaaatcacaaaattttcaataaaaaacaccaaaaaaaagagaaa
    aaatgagaaaaatcgacgaaaatcggtataaaatcaaataaaaatagaaggaaaatattc
    agctcgtaaacccacacgtgcggcacggtttcgtgggcggggcgtctctgccgggaaaat
    tttgcgtttaaaaactcacatataggcatccaatggattttcggattttaaaaattaata
    taaaatcagggaaatttttttaaatttttcacatcgatattcggtatcaggggcaaaat
    tagagtcagaaacatatatttccccacaaactctactccccctttaaacaaagcaaagag
    cgatactcattgcctgtagcctctatattatgccttatgggaatgcatttgattgtttcc
    gcatattgtttacaaccatttatacaacatgtgacgtagacgcactgggcggttgtaaaa
    cctgacagaaagaattggtcccgtcatctactttctgatttttggaaaatatgtacaat
    gtcgtccagtattctattccttctcggcgatttggccaagttattcaaacacgtataaat
    aaaaatcaataaagctaggaaaatattttcagccatcacaaagtttcgtcagccttgtta
    tgtcaaccacttttatacaaattatataaccagaaatactattaaataagtatttgtat
    gaaacaatgaacactattataacattttcagaaaatgtagtatttaagcgaaggtagtgc
    acatcaaggccgtcaaacggaaaaatttttgcaagaatca
  </DNA>
 </SEQUENCE>
</DASDNA>
```

<!DOCTYPE> (required; one only)
The doctype indicates which formal DTD specification to use. For the dna query, the doctype DTD is "http://www.biodas.org/dtd/dasdna.dtd".

<DASDNA> (required; one only)
The appropriate doctype and root tag is DASDNA.

<SEQUENCE> (required; one or more)
There is a single <SEQUENCES> tag per requested segment. It has the attributes **id**, which indicates the reference ID for this sequence, **start** and **stop**, which indicate the position of this segment within the reference sequence, and **version**, which provides the sequence map version number. All four attributes are required.

<DNA> (required; one per SEQUENCE)
This tag surrounds the DNA data. It has the attribute **length** (required), which indicates the length of the DNA. The DNA is found in the body of the tag and is required. DNA will be lower-case and adhere to the IUPAC code conventions.

DAS v1

The queries

sequence

Returns the sequence (nucleotide or protein) corresponding to the indicated segment. Available from version 1.5.
Scope: reference servers.
Arguments: segment (required; one or more).

PREFIX/das/DSN/sequence?segment=RANGE[;segment=RANGE..]

http://servlet.sanger.ac.uk:8080/das/ensembl_Gallus_gallus_core_28_1d/dna?segment=1:1,1000
http://servlet.sanger.ac.uk:8080/das/ensembl_Gallus_gallus_core_28_1d/sequence?segment=1:1,1000

DAS v1

The queries

types
Returns the annotation available for a segment of sequence.
Scope: annotation and reference servers.
Arguments: segment (optional; sequence range).
type(optional; one or more type IDs to be used for filtering annotations on the type field).

PREFIX/das/DSN/types [?segment=RANGE][;type=TYPE]

http://www.wormbase.org/db/das/elegans/types

DAS v1

The queries

features

Returns the annotations across one or more segments of sequence.

Scope: reference and annotation servers.

PREFIX/das/*DSN*/features?segment=*REF:start,stop*[;segment=*REF:start,stop*...]
[;type=*TYPE*]
[;type=*TYPE*]
[;category=*CATEGORY*]
[;category=*CATEGORY*]
[;categorize=*yes*|*no*]
[;feature_id=ID]
[;group_id=ID]

Description: This query returns the annotations across one or more segments of sequence.

Arguments:

segment (zero or more)
If specified, the segment argument restricts the list of annotations to those that overlap the indicated range. Each segment argument uses the format format *reference:start,stop*, where *reference* is the ID of the reference sequence used to establish the coordinate system, and *start* and *stop* are the endpoints of the region to query, inclusive. Multiple segments may be specified.

type (zero or more)
Zero or more type IDs to be used for filtering annotations on the type field. If multiple type names are provided, the resulting list of features will be the logical OR of the list.

For compatibility with versions 0.997 and earlier of this protocol, servers are allowed to treat the type ID as a regular expression, but this feature is **deprecated** and should not be relied on.

category (zero or more)
Zero or more category IDs to be used for filtering annotations by category. If multiple categories are provided, they are treated as the logical OR.

For compatibility with versions 0.997 and earlier of this protocol, servers are allowed to treat the type ID as a regular expression, but this feature is **deprecated** and should not be relied on.

categorize (optional)
Either "yes" or "no" (default). If "yes", then each annotation must include its functional category.

feature_id (zero or more; new in 1.5)
Instead of, or in addition to, **segment** arguments, you may provide one or more **feature_id** arguments, whose values are the identifiers of particular features. If the server supports this operation, it will translate the feature ID into the segment(s) that strictly enclose them and return the result in the *features* response. It is possible for the server to return multiple segments if the requested feature is present in multiple locations.

group_id (zero or more; new in 1.5)
The **group_id** argument, is similar to **feature_id**, but retrieves segments that contain the indicated feature group.

DAS v1

The queries

features response

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE DASGFF SYSTEM "http://www.biodas.org/dtd/dasgff.dtd">
<DASGFF>
 <GFF version="1.0" href="url">
 <SEGMENT id="id" start="start" stop="stop" type="type" version="X.XX" label="label">
   <FEATURE id="id" label="label">
     <TYPE id="id" category="category" reference="yes|no">type label</TYPE>
     <METHOD id="id"> method label </METHOD>
     <START> start </START>
     <END> end </END>
     <SCORE> [X.XX|-] </SCORE>
     <ORIENTATION> [0|-|+] </ORIENTATION>
     <PHASE> [0|1|2|-]</PHASE>
     <NOTE> note text </NOTE>
     <LINK href="url"> link text </LINK>
     <TARGET id="id" start="x" stop="y">target name</TARGET>
     <GROUP id="id" label="label" type="type">
         <NOTE> note text </NOTE>
         <LINK href="url"> link text </LINK>
         <TARGET id="id" start="x" stop="y">target name</TARGET>
     </GROUP>
   </FEATURE>
   ...
 </SEGMENT>
 </GFF>
</DASGFF>
```

<!DOCTYPE> (required; one only)
The doctype indicates which formal DTD specification to use. For the features query, the doctype DTD is "http://www.biodas.org/dtd/dasgff.dtd".

<DASGFF> (required; one only)
The appropriate doctype and root tag is DASGFF.

<GFF> (required; one only)
There is a single <GFF> tag. Its **version** (required) attribute indicates the current version of the XML form of the General Feature Format. The current version is (arbitrarily) 1.0 The **href** (required) attribute echoes the URL query that was used to fetch the current document.

DAS v1

The queries

features response

 (required; one or more)
The tag provides information on the reference segment coordinate system. The **id**, **start** and **stop** attributes indicate the position of the segment. The **version** attribute indicates the current version of the sequence map. The id, start, stop, and version attributes are required. The optional **label** attribute provides a human readable label for display purposes. The optional type attribute describes the segment type, for future compatibility with Sequence Ontology-based feature typing.

<FEATURE> (optional; zero or more per SEGMENT)
There are zero or more <FEATURE> tags per , each providing information on one annotation. The **id** attribute (required) is a unique identifier for the feature. It can be used as a reference point for further navigation. The **label** attribute (optional) is a suggested label to display for the feature. If not present, the **id** attribute can be used instead.

<TYPE> (required; one per FEATURE)
Each feature has just one <TYPE> field, which indicates the type of the annotation. The attributes are **id** (required), which is a unique id for the annotation type and can be used to retrieve further information from the annotation server (see Linking to a Feature), and the **category** (optional, recommended) attribute, which provides functional grouping to related types.

The reference server's annotations can consist of additional overlapping landmarks (parents, children, and neighbors), which should be marked "yes" in the third attribute **reference** (optional, defaults to "no") to indicate that the feature is a structural landmark within the map (this feature can be annotated). The tag contents (optional) is a human readable label for display purposes.

If a **reference** annotation has either or both of the optional attributes, **subparts="yes"** and **superparts="yes"**, then in addition to being useable as a reference sequence, the feature contains subparts and/or superparts that themselves can act as reference features. This can be used to reconstruct reference server's assembly. See also Fetching Assembly Information.

<METHOD> (required; one per FEATURE)
Each feature has one <METHOD> field, which identifies the method used to identify the feature. The **id** (optional) tag can be used to retrieve further information from the annotation server. The tag contents (optional) is a human readable label.

<START>, <END> (required; one apiece per FEATURE)
These tags indicate the start and end of the feature in the coordinate system of the reference sequence given in the tag. The relationship between the feature start and stop positions and the segment start and stop is that the two spans are guaranteed to overlap.

<SCORE> (required; one per FEATURE)
This is a floating point number indicating the "score" of the method used to find the current feature. The number can only be understood in the context of information retrieved from the server by linking to the method. If this field is inapplicable, the contents of the tag can be replaced with a **-** symbol.

<ORIENTATION> (required; one per FEATURE)
This tag indicates the orientation of the feature relative to the direction of transcription. It may be **0** for features that are unrelated to transcription, **+**, for features that are on the sense strand, and **-**, for features on the antisense strand.

<PHASE> (required; one per FEATURE)
This tag indicates the position of the feature relative to open reading frame, if any. It may be one of the integers **0**, **1** or **2**, corresponding to each of the three reading frames, or **-** if the feature is unrelated to a reading frame.

<NOTE> (optional; zero or more per FEATURE)
A human-readable note in plain text format.

<LINK> (optional; zero or more per FEATURE)
A link to a web page somewhere that provides more information about this feature. The **href** (required) attribute provides the URL target for the link. The link text is an optional human readable label for display purposes.

DAS v1

The queries

features response

<TARGET> (optional; zero or more per FEATURE)
The target sequence in a sequence similarity match. The **id** attribute provides the reference ID for the target sequence, and the **start** and **stop** attributes indicate the segment that matched across the target sequence. All three attributes are required. More information on the target can be retrieved by linking back to the annotation server. See Linking to a Feature.

<GROUP> (optional; zero or more per FEATURE)
The <GROUP> section is slightly odd, as it is derived from an overloaded field in the GFF flat file format. It provides a unique "group" ID that indicates when certain features are related to each other. The canonical example is the CDS, exons and introns of a transcribed gene, which logically belong together.

The group **id** attribute (required) provides an identifier that should be used by the client to group features together visually. Unlike other IDs in this protocol, the group ID cannot be used as a database handle to retrieve further information about the group. Such information can, however, be provided within <GROUP> section, which may contain up to three optional tags.

The **label** attribute (optional) provides a human-readable string that can be used in graphical representations to label the glyph.

The **type** attribute (optional) provides a type ID for the group as a whole, for example "transcript". This ID can be used as a key into the stylesheet to select the glyph and graphical characteristics for the group as a whole.

<NOTE> (optional; zero or more per GROUP)
A human-readable note in plain text format.

<LINK> (optional; zero or more per GROUP)
A link to a web page somewhere that provides more information about this group. The **href** (required) attribute provides the URL target for the link. The link text is an optional human readable label for display purposes.

<TARGET> (optional; zero or more per GROUP)
The target sequence in a sequence similarity match. The **id** attribute provides the reference ID for the target sequence, and the **start** and **stop** attributes indicate the segment that matched across the target sequence. All three attributes are required. NOTE: although this tag is present in the GROUP section, it applies to the FEATURE, and it is preferred to place it directly in the <FEATURE> section. Earlier versions of this specification placed the TARGET tag in the GROUP section, and clients must recognize and accomodate this.

http://servlet.sanger.ac.uk:8080/das/ensembl2034/features?segment=1:1,50
http://servlet.sanger.ac.uk:8080/das/ensembl2034/features?segment=1:1,50;categorize=yes
http://servlet.sanger.ac.uk:8080/das/ensembl2034/features?segment=1:1,50;feature_id=AP006221.1.1.36731

DAS v1

The queries

Some more examples of possible queries:

http://servlet.sanger.ac.uk:8080/das/dsn
http://das.ensembl.org/das/ensembl_Homo_sapiens_core_32_35e/
http://das.ensembl.org/das/ensembl_Homo_sapiens_core_32_35e/dna?segment=2_NT_079503
http://das.ensembl.org/das/ensembl_Homo_sapiens_core_32_35e/sequence?segment=2_NT_079503
http://das.ensembl.org/das/ensembl_Homo_sapiens_core_32_35e/sequence?segment=2_NT_079503:35,98
http://das.ensembl.org/das/ensembl_Homo_sapiens_core_32_35e/features?segment=2_NT_079503:100,2500;type=exon
http://das.ensembl.org/das/ensembl_Homo_sapiens_core_32_35e/features?segment=2_NT_079503;type=transcript
http://das.ensembl.org/das/ensembl_Homo_sapiens_core_32_35e/features?group_id=ENST00000339816
http://das.ensembl.org/das/ensembl_Homo_sapiens_core_32_35e/features?group_id=ENST00000339816;type=exon
http://www.wormbase.org/db/das/
http://www.wormbase.org/db/das/dsn
http://www.wormbase.org/db/das/elegans
http://www.wormbase.org/db/das/elegans/entry_points
http://www.wormbase.org/db/das/elegans/types
http://www.wormbase.org/db/das/elegans/stylesheet
http://www.wormbase.org/db/das/elegans/dna?segment=I (reference servers)
http://www.wormbase.org/db/das/elegans/dna?segment=I:1,30
http://www.wormbase.org/db/das/elegans/dna?segment=I:1,1000;segment=II:5000,5200
http://www.wormbase.org/db/das/elegans/features?segment=I:1,1000;segment=II:5000,5200; type=gene:gene

DAS v1

The queries

stylesheet
Returns the servers recommendations on formatting annotations retrieved from it. They are not mandatory.
Scope: annotation servers.
Arguments: none

PREFIX/das/DSN/stylesheet

http://www.wormbase.org/db/das/elegans/stylesheet

DAS v1

The response (DAS status codes):

200	OK, data follows
400	Bad command (command not recognized)
401	Bad data source (data source unknown)
402	Bad command arguments (arguments invalid)
403	Bad reference object (reference sequence unknown)
404	Bad stylesheet (requested stylesheet unknown)
405	Coordinate error (sequence coordinate is out of bounds/invalid)
500	Server error, not otherwise specified
501	Unimplemented feature

DAS v1

Example of the difference between anottation and reference servers:

We can find examples of exclusively annotation servers at:

http://servlet.sanger.ac.uk:8080/das/dsn

See the first case with <source ID="ens1834trans"…..

We can ask its mapmaster server for the entry points and types (it is both annotation and reference server). We can ask the annotation server about types but not about entry points because it is only annotation server.

GenomicDAS vs Protein/ProteomicDAS

ProteinDAS uses the DAS protocol to exchange protein annotation. In this case, SwissProt aminoacid sequences are used as the common reference. Typical annotations are:

- –Protein domains (Pfam, SCOP, SMART)
- –Signal peptide
- –Transmembrane regions
- – Isoforms
- – Residue contacts
- –Low complexity regions
- – 2ary structure
- – 3D structure
- – Scientific references

http://www.ebi.ac.uk/das-srv/uniprot/das/
http://www.ebi.ac.uk/das-srv/uniprot/das/aristotle/sequence?segment=Q24488

3D-DAS

- A DAS protocol to annotate features over protein 3D structures.
- A DAS protocol to annotate alignned regions within different protein 3D structures.

The following new or modified services are being used by the SPICE - DAS client:

New services	
Capability Name	**Description**
alignment/1.1	The server supports the basic *alignment* request.
structure/1.0	The server supports the basic *structure* request.
entry_points/1.1	(update of old capability)The server supports the basic *entry_points* request.

3D-DAS

Alignment attributes description (example):

http://das.sanger.ac.uk/das/msdpdbsp/alignment?query=1a4a

```
- <dasalignment xsd:schemaLocation="http://www.efamily.org.uk/xml/das/2004/06/17/dasalignment.xsd http://www.efamily.org.uk/xml/das/2004/06/17/dasalignment.xsd">
  - <alignment alignType="PDB_SP">
    - <alignObject dbAccessionId="1a4a.A" intObjectId="1a4a.A" objectVersion="toBeRetreived" type="STRUCTURE" dbSource="PDB" dbVersion="200404" dbCoordSys="PDBresnum,Protein
      Structure">
        <alignObjectDetail dbSource="PDB" property="resolution">1.89</alignObjectDetail>
        <alignObjectDetail dbSource="PDB" property="experiment_type">XRAY</alignObjectDetail>
        <alignObjectDetail dbSource="PDB" property="header">ELECTRON TRANSPORT</alignObjectDetail>
      - <alignObjectDetail dbSource="PDB" property="title">
          AZURIN MUTANT WITH MET 121 REPLACED BY HIS, PH 6.5 CRYSTAL FORM, DATA COLLECTED AT 16 DEGREES CELSIUS
        </alignObjectDetail>
        <alignObjectDetail dbSource="PDB" property="molecule description">AZURIN</alignObjectDetail>
      </alignObject>
      <alignObject dbAccessionId="P00280" intObjectId="P00280" objectVersion="toBeRetreived" type="PROTEIN" dbSource="UniProt" dbVersion="200404" dbCoordSys="UniProt,Protein
      Sequence"/>
    - <block blockOrder="1">
        <segment intObjectId="1a4a.A" start="1" end="129"/>
        <segment intObjectId="P00280" start="21" end="149"/>
      </block>
    </alignment>
  - <alignment alignType="PDB_SP">
    - <alignObject dbAccessionId="1a4a.B" intObjectId="1a4a.B" objectVersion="toBeRetreived" type="STRUCTURE" dbSource="PDB" dbVersion="200404" dbCoordSys="PDBresnum,Protein
      Structure">
        <alignObjectDetail dbSource="PDB" property="resolution">1.89</alignObjectDetail>
        <alignObjectDetail dbSource="PDB" property="experiment_type">XRAY</alignObjectDetail>
        <alignObjectDetail dbSource="PDB" property="header">ELECTRON TRANSPORT</alignObjectDetail>
      - <alignObjectDetail dbSource="PDB" property="title">
          AZURIN MUTANT WITH MET 121 REPLACED BY HIS, PH 6.5 CRYSTAL FORM, DATA COLLECTED AT 16 DEGREES CELSIUS
        </alignObjectDetail>
        <alignObjectDetail dbSource="PDB" property="molecule description">AZURIN</alignObjectDetail>
      </alignObject>
      <alignObject dbAccessionId="P00280" intObjectId="P00280" objectVersion="toBeRetreived" type="PROTEIN" dbSource="UniProt" dbVersion="200404" dbCoordSys="UniProt,Protein
      Sequence"/>
    - <block blockOrder="1">
        <segment intObjectId="1a4a.B" start="1" end="129"/>
        <segment intObjectId="P00280" start="21" end="149"/>
      </block>
    </alignment>
</dasalignment>
```

3D-DAS

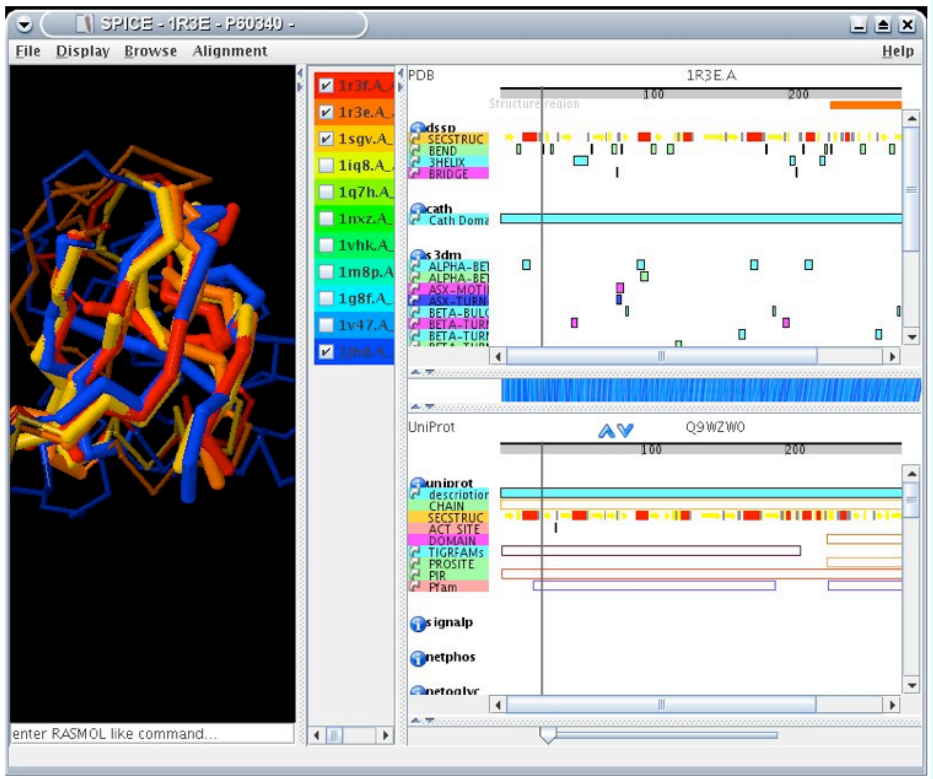
Structures attributes description (example):

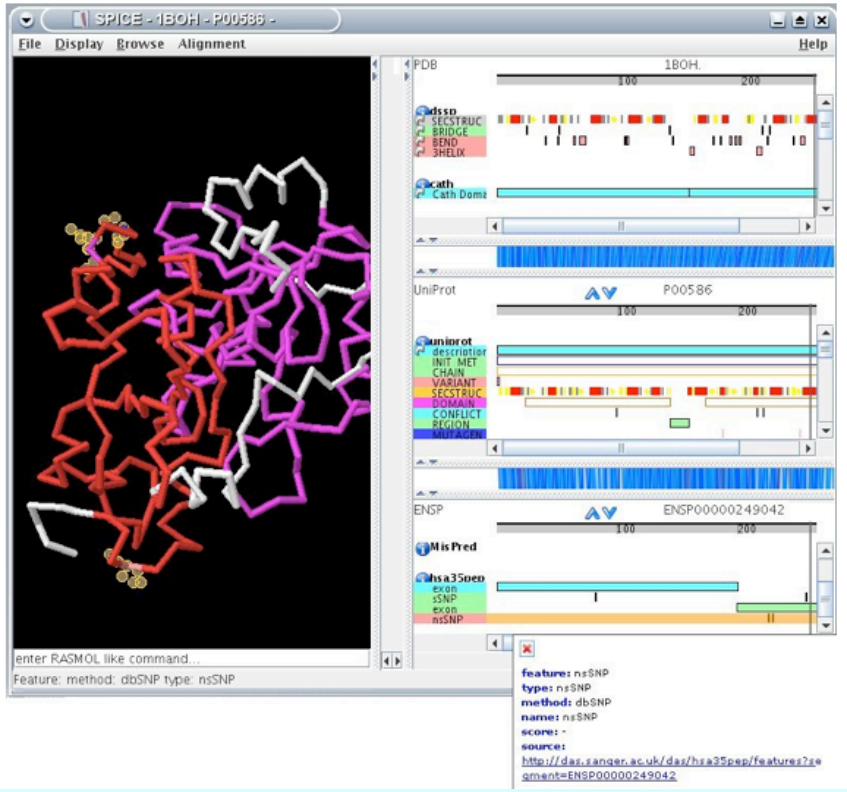
http://das.sanger.ac.uk/das/structure/structure?query=1a4a

```
- <dasstructure xsi:schemaLocation="http://www.efamily.org.uk/xml/das/2004/06/17/dasstructure.xsd http://www.efamily.org.uk/xml/das/2004/06/17/dasstructure.xsd">
    <object dbAccessionId="1A4A" intObjectId="1A4A" objectVersion="29-APR-98" type="protein structure" dbSource="PDB" dbVersion="20060707" dbCoordSys="PDBresnum"/>
  - <chain id="A" SwissprotId="null">
    - <group name="ALA" type="amino" groupID="1">
        <atom atomID="1" atomName=" N " x="-19.031" y="16.695" z="3.708"/>
        <atom atomID="2" atomName=" CA " x="-20.282" y="16.902" z="4.404"/>
        <atom atomID="3" atomName=" C " x="-20.575" y="18.394" z="4.215"/>
        <atom atomID="4" atomName=" O " x="-20.436" y="19.194" z="5.133"/>
        <atom atomID="5" atomName=" CB " x="-20.077" y="16.548" z="5.883"/>
        <atom atomID="6" atomName="1H " x="-18.381" y="17.406" z="4.081"/>
        <atom atomID="7" atomName="2H " x="-18.579" y="15.781" z="3.874"/>
        <atom atomID="8" atomName="3H " x="-19.018" y="16.844" z="2.68"/>
      </group>
    - <group name="GLN" type="amino" groupID="2">
        <atom atomID="9" atomName=" N " x="-21.002" y="18.711" z="2.981"/>
        <atom atomID="10" atomName=" CA " x="-21.135" y="20.03" z="2.331"/>
        <atom atomID="11" atomName=" C " x="-19.672" y="20.413" z="2.082"/>
        <atom atomID="12" atomName=" O " x="-18.933" y="19.513" z="1.657"/>
        <atom atomID="13" atomName=" CB " x="-21.931" y="21.119" z="3.194"/>
        <atom atomID="14" atomName=" CG " x="-21.64" y="21.845" z="4.542"/>
        <atom atomID="15" atomName=" CD " x="-20.828" y="23.143" z="4.53"/>
        <atom atomID="16" atomName=" OE1" x="-19.601" y="23.1" z="4.599"/>
        <atom atomID="17" atomName=" NE2" x="-21.387" y="24.34" z="4.397"/>
        <atom atomID="18" atomName=" H " x="-21.2" y="17.964" z="2.382"/>
        <atom atomID="19" atomName="1HE2" x="-20.813" y="25.128" z="4.518"/>
        <atom atomID="20" atomName="2HE2" x="-22.353" y="24.423" z="4.24"/>
      </group>
    - <group name="CYS" type="amino" groupID="3">
        <atom atomID="21" atomName=" N " x="-19.124" y="21.577" z="2.35"/>
        <atom atomID="22" atomName=" CA " x="-17.743" y="21.878" z="2.07"/>
        <atom atomID="23" atomName=" C " x="-16.886" y="21.629" z="3.306"/>
        <atom atomID="24" atomName=" O " x="-16.088" y="22.445" z="3.741"/>
        <atom atomID="25" atomName=" CB " x="-17.755" y="23.321" z="1.592"/>
        <atom atomID="26" atomName=" SG " x="-18.823" y="23.494" z="0.125"/>
        <atom atomID="27" atomName=" H " x="-19.617" y="22.287" z="2.792"/>
      </group>
    - <group name="GLU" type="amino" groupID="4">
        <atom atomID="28" atomName=" N " x="-16.988" y="20.423" z="3.838"/>
        <atom atomID="29" atomName=" CA " x="-16.306" y="19.998" z="5.053"/>
        <atom atomID="30" atomName=" C " x="-16.058" y="18.513" z="5.03"/>
        <atom atomID="31" atomName=" O " x="-16.812" y="17.79" z="4.372"/>
```

SPICE

Spice is a client to visualize 3D-DAS annotations.





DAS v2

•DAS v2 (it officialy started in July 2004, 2 year NIH grant - Stein lab, Affymetrix, EBI/Sanger and Dalke scientific).

•It describes features located on the genomic sequence. Future versions will add support for sharing annotations of protein sequences, expression data, 3D structures and ontologies. The genomic DAS interface is deliberately designed so there will be a large core shared with the protein sequence DAS.

•A DAS 2.0 annotation server provides feature information about one or more genome sources. Each source may have one or more versions. Different versions are usually based on different assemblies.

- –Genomic sequence and annotation
- –Stylesheet
- –Writeback
- –Region locking
- –Assay Retrievals
- –Ontology Retrievals

DAS public sofware/servers

DAS SERVERS

ProServer (perl)
LDAS (perl)
Dazzle (java)
GBrowse (perl)
Madas (perl)

DAS CLIENTS

Apollo (java)
Spice (java)
GBrowse (perl)
Synbrowse (perl)
Madas (perl)

MaDas: a specific DAS client/server

1. A requirement of the group was to have a DAS annotation server to annotate DNA and protein features.
2. At the same time we wanted the system to let the user to manually upload its own annotations.
3. Finally a web client would be suitable to display the annotations.

(http://madas.bioinfo.cnio.es/MaDas/cgi-bin/MaDas)

- MaDas was developed by Victor de la Torre (current version 0.5).
- It is a perl based DAS client/server.
- It serves DNA and protein annotations.
- As reference sequence server it uses Ensembl DAS and Uniprot DAS, the available genomes are those contained in Ensembl.
- It allows users to manually introduce their own annotations (through a web form, a GFF file or a XML file).
- Any DAS client can read the annotations served by MaDas.
- MaDas is also provided with a web client that let users extract and visualize annotations. Furthermore it has a tool to add sequence annotations to the server.
- A user can browse the system as a 'guest', or can register and lead or share annotations that belong to a particular project.

Available DAS servers

Examples of some DAS servers that are currently available:

WormBase (C. elegans)
Washington University @ St. Louis (C. elegans)
FlyBase (Drosophila)
Ensembl (human)
UCSC (human)
TIGR (C. elegans, human)
Cambridge University (C. elegans, human, Drosophila)
Lawrence Berkeley Laboratories (human, mouse)
NCBI (in development)
INB (*Corynebacterium diphtheriae*, *Mycobacterium tuberculosis* CDC1551, *Streptomyces avermitilis* MA-4680, *Streptomyces coelicolor* A3(2), Human_Herpes_simplex_4).

A repository of DAS servers:
http://www.dasregistry.org/listServices.jsp

DAS

An interesting web site about DAS that you should keep in mind:

http://www.biodas.org/

GFF files

The way that annotations are uploaded to annotation servers is through GFF files. **GFF** ('Gene-Finding Format' or 'General Feature Format').

GFF is a format for describing genes and other features associated with DNA, RNA and Protein sequences. It was firstly developed at the Sanger center (**their current version is 2.0** - http://www.sanger.ac.uk/Software/formats/GFF/).

Fields are: <seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes] [comments]

<seqname>

The name of the sequence. Normally the seqname is the identifier for a sequence in a public database, such as an EMBL/Genbank/DDBJ accession number.

<source>

The source of this feature. This field will normally be used to indicate the program making the prediction, or if it comes from public database annotation, or is experimentally verified, etc.

GFF files

<feature>

The feature type name. It is possible to use standard features or to define new ones if needed.

<start>, <end>

Integers. <start> must be less than or equal to <end>. Sequence numbering starts at 1, so these numbers should be between 1 and the length of the relevant sequence, inclusive.

<score>

A floating point value. When there is no score you should use '.'

<strand>

One of '+', '-' or '.' '.' should be used when strand is not relevant.

<frame>

One of '0', '1', '2' or '.', '0' indicates that the specified region is in frame, i.e. that its first base corresponds to the first base of a codon. '1' indicates that there is one extra base, i.e. that the second base of the region corresponds to the first base of a codon, and '2' means that the third base of the region is the first base of a codon. If the strand is '-', then the first base of the region is value of <end>, because the corresponding coding region will run from <end> to <start> on the reverse strand. As with <strand>, if the frame is not relevant then set <frame> to '.'

GFF files

[attribute]

From version 2 onwards, the attribute field must have an tag value structure flattened onto one line by semicolon separators. Tags must be standard identifiers ([A-Za-z][A-Za-z0-9_]*). Free text values must be quoted with double quotes. Note: all non-printing characters in such free text value strings (e.g. newlines, tabs, control characters, etc) must be explicitly represented by their C (UNIX) style backslash-escaped representation (e.g. newlines as '\n', tabs as '\t'). As in ACEDB, multiple values can follow a specific tag. The aim is to establish consistent use of particular tags, corresponding to an underlying implied ACEDB model if you want to think that way (but acedb is not required).

SEQ1	EMBL	gene	103	400	.	+	0
SEQ1	EMBL	gene	1060	1700	.	+	0
SEQ1	EMBL	gene	3400	3875	.	+	.

VERY IMPORTANT: All of the above described fields should be separated by TAB characters ('\t'). All values of the mandatory fields should not include whitespace (i.e. the strings for <seqname>, <source> and <feature> fields).

GFF files

Comments inside a GFF

It is possible to add comments at the top of a GFF file:

```
##source-version NCBI C++ formatter 0.2
##date 2005-06-20
##Type DNA BA000030.2
```

GFF files extensions

GFF authors propose to use file names ending with ".gff".

GFF files

GFF v3.0
(http://song.sourceforge.net/gff3.shtml)

- Created by Lincoln Stein.
- Compatible with previous versions.

1) adds a mechanism for representing more than one level of hierarchical grouping of features and subfeatures.
2) separates the ideas of group membership and feature name/id.
3) constrains the feature type field to be taken from a controlled vocabulary.
4) allows a single feature, such as an exon, to belong to more than one group at a time.
5) provides an explicit convention for pairwise alignments.
6) provides an explicit convention for features that occupy disjunct regions.

GFF 2.0 <seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes] [comments]
GFF 3.0 <seqid> <source> <type> <start> <end> <score> <strand> <phase> [attributes]

GFF files

GFF v3.0

[Atributes]

A list of feature attributes in the format tag=value. Multiple tag=value pairs are separated by semicolons.

These tags have predefined meanings:

- **ID** Indicates the name of the feature. IDs must be unique within the scope of the GFF file.
- **Name** Display name for the feature. This is the name to be displayed to the user. Unlike IDs, there is no requirement that the Name be unique within the file.
- **Alias** A secondary name for the feature. It is suggested that this tag be used whenever a secondary identifier for the feature is needed, such as locus names and accession numbers. Unlike ID, there is no requirement that Alias be unique within the file.
- **Parent** Indicates the parent of the feature. A parent ID can be used to group exons into transcripts, transcripts into genes, an so forth. A feature may have multiple parents. Parent can *only* be used to indicate a partof relationship.

GFF files

GFF v3.0

[Atributes]

- **Target** Indicates the target of a nucleotide-to-nucleotide or protein-to-nucleotide alignment. The format of the value is "target_id start end [strand]", where strand is optional and may be "+" or "-". If the target_id contains spaces, they must be escaped as hex escape %20.
- **Gap** The alignment of the feature to the target if the two are not collinear (e.g. contain gaps).
- **Derives_from** Used to disambiguate the relationship between one feature and another when the relationship is a temporal one rather than a purely structural "part of" one.
- **Note** A free text note.
- **Dbxref** A database cross reference.

GFF files

GFF v3.0

[Atributes]

- **Ontology_term** A cross reference to an ontology term.
- Multiple attributes of the same type are indicated by separating the values with the comma "," character, as in:

  Parent=AF2312,AB2812,abc-3

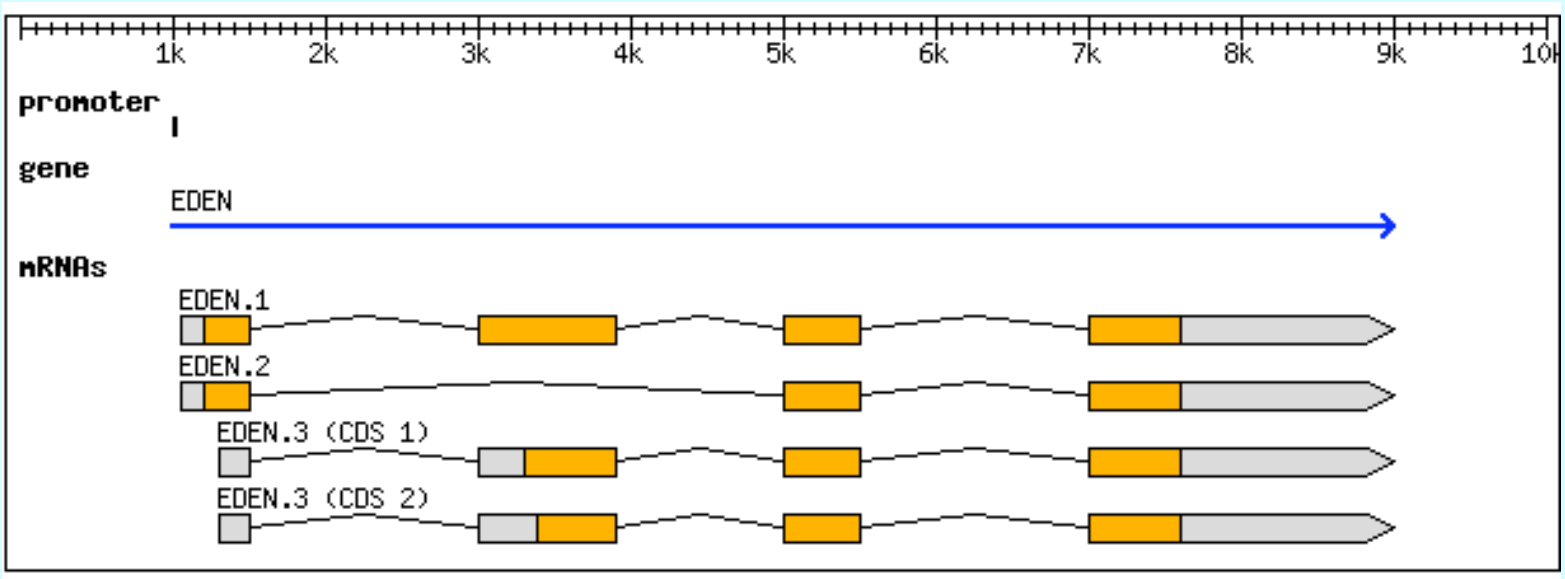
  Note that attribute names are case sensitive. "Parent" is not the same as "parent".

GFF files

GFF v3.0

Example: a gene named EDEN extending from position 1000 to position 9000. It encodes three alternatively-spliced transcripts named EDEN.1, EDEN.2 and EDEN.3, the last of which has two alternative translational start sites leading to the generation of two protein coding sequences.

There is also an identified transcriptional factor binding site located 50 bp upstream from the transcriptional start site of EDEN.1 and EDEN2.

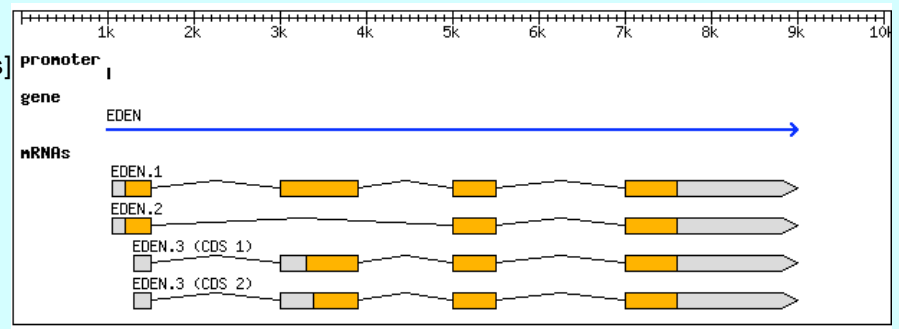


GFF files

GFF v3.0

<seqid> <source> <type> <start> <end> <score> <strand> <phase> [attributes]

Example:




```
##gff-version  3
##sequence-region  ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . 5'-UTR          1050  1200  .  +  0  ID=utr00001;Parent=mRNA00001;Name=5'UTR
ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             3000  3902  .  +  0  ID=cds00002;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             5000  5500  .  +  0  ID=cds00003;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             7000  7600  .  +  0  ID=cds00004;Parent=mRNA00001;Name=edenprotein.1
ctg123 . 3'-UTR          7611  9000  .  +  0  ID=utr00002;Parent=mRNA00001;Name=3'UTR
ctg123 . CDS             1201  1500  .  +  0  ID=cds00005;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             5000  5500  .  +  0  ID=cds00006;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             7000  7600  .  +  0  ID=cds00007;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             3301  3902  .  +  0  ID=cds0008;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             5000  5500  .  +  2  ID=cds0009;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             7000  7600  .  +  2  ID=cds00010;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             3391  3902  .  +  0  ID=cds00011;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS             5000  5500  .  +  2  ID=cds00012;Parent=mRNA00003;Name=edenprotein.4
Ctg123 . CDS             7000  7600  .  +  2  ID=cds00013;Parent=mRNA00003;Name=edenprotein.4
```

GFF files

GFF v3.0

Note that several features, including the gene, its mRNAs and the CDSs, all have Name attributes. This attributes assigns those features a public name, but it is not mandatory. The ID attributes are only mandatory for those features that have children (the gene and mRNAs), or for those that span multiple lines. The IDs do not have meaning outside the file in which they reside.

GFF files

GFF v3.0

Hence, a slightly simplified version of the previous GFF would look like this:

```
##gff-version  3
##sequence-region  ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000  1012  .  +  .  Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001
ctg123 . exon            1300  1500  .  +  .  Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .  Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .  Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             3000  3902  .  +  0  ID=cds00002;Parent=mRNA00001
ctg123 . CDS             5000  5500  .  +  0  ID=cds00003;Parent=mRNA00001
ctg123 . CDS             7000  7600  .  +  0  ID=cds00004;Parent=mRNA00001
ctg123 . CDS             1201  1500  .  +  0  ID=cds00005;Parent=mRNA00002
ctg123 . CDS             5000  5500  .  +  0  ID=cds00006;Parent=mRNA00002
ctg123 . CDS             7000  7600  .  +  0  ID=cds00007;Parent=mRNA00002
```

GFF files

GFF v3.0

Annotations are sometimes available in public databases (like GenBank) in a format that is not a GFF.

If we want to add these annotations to our own set of annotations it is possible to translate these annotations to a GFF format, meaning that we don't have to build our own script for this task.

Example of a GenBank file:

ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Streptomyces_coelicolor/AL645882.gbk

GFF files

GFF v3.0 ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Streptomyces_coelicolor/AL645882.gbk

```
FEATURES             Location/Qualifiers
     source          1..8667507
                     /organism="Streptomyces coelicolor A3(2)"
                     /mol_type="genomic DNA"
                     /strain="A3(2)"
                     /db_xref="taxon:100226"
     misc_feature    1..21653
                     /note="TIR-L. Left hand chromosome end terminal inveretd
                     repeat."
     RBS             435..440
     gene            446..1123
                     /gene="SCO0001"
                     /note="synonym: SCEND.02c"
     CDS             446..1123
                     /gene="SCO0001"
                     /note="SCEND.02c, unknown, doubtful CDS, len: 225aa"
                     /codon_start=1
                     /transl_table=11
                     /product="hypothetical protein"
                     /protein_id="CAD30875.1"
                     /db_xref="GI:20520987"
                     /translation="MTGHHESTGPGTALSSDSTCRVTQYQTAGVNARLRLFALLERRA
                     CPRARRTTWWPGRSARWWSWTAWRRLLGVCCVRGRLGRRRDGGERGPGGHRGPGLATA
```

GFF files

GFF v3.0

There is a bioperl script for this purpose: **bp_genbank2gff.pl**

```
Usage: /usr/local/bin/bp_genbank2gff.pl [options] [<gff file 1> <gff file 2>] ...
Load a Bio::DB::GFF database from GFF files.

 Options:
   --create               Force creation and initialization of database
   --dsn       <dsn>      Data source (default dbi:mysql:test)
   --user      <user>     Username for mysql authentication
   --pass      <password> Password for mysql authentication
   --proxy     <proxy>    Proxy server to use for remote access
   --stdout               direct output to STDOUT
   --adaptor   <adaptor>  adaptor to use (e.g. dbi::mysql, dbi::pg, dbi::oracle)
   --viral                the genome you are loading is viral (changes tag choices)
   --source    <source>   source field for features ['genbank']
    EITHER --file         Arguments that follow are Genbank/EMBL file names
    OR --gb_folder        What follows is a folder full of gb files to process
    OR --accession        Arguments that follow are genbank accession numbers (not gi!)
    OR --acc_file         accession numbers (not gi!) in a file (one per line, no punc.)
    OR --acc_pipe         accession numbers (not gi!) from a STDIN pipe (one per line)
```

GFF files

GFF v3.0

By executing the following command:

bp_genbank2gff.pl AL645882.gbk -stdout > AL645882.gff

We get a GFF file like the one below:

```
##gff-version 3
AL645882	Genbank	region	1	8667507	.	.	.
	ID=AL645882;Note=Streptomyces%20coelicolor%20A3%282%29%20complete%20genome.
AL645882	Genbank	region	1	8667507	.	+	.
	ID=Streptomyces%20coelicolor%20A3%282%29;mol_type=genomic%20DNA;db_xref=taxon%3A100226;strain=A3%282%29
AL645882	Genbank	RBS	435	440	.	+	.	ID=Misc
AL645882	Genbank	gene	446	1123	.	+	.	ID=SCO0001;note=synonym%3A%20SCEND.02c
AL645882	Genbank	CDS	446	1123	.	+	.
	Parent=SCO0001.t00;db_xref=GI%3A20520987;codon_start=1;protein_id=CAD30875.1;translation=MTGHHESTGPGTALSSD
	STCRVTQYQTAGVNARLRLFALLERRACPRA
RRTTWWPGRSARWWSWTAWRRLLGVCCVRGRLGRRRDGGERGPGGHRGPGLATARRRSGGATELAVHCADVRQRERADLVRLEGFVRES
	VLPRAHPHTTARRRVLEVLGEAGSLCTARTVNSDEDYILCTLGVGHYDPDDQPPFKDGKPGWQRAGASIWNGSGAACIPHAAIEGP
	RK;product=hypothetical%20protein;
transl_table=11;note=SCEND.02c%2C%20unknown%2C%20doubtful%20CDS%2C%20len%3A%20225aa
AL645882	Genbank	RBS	1238	1243	.	+	.	ID=Misc
……
```

GFF files

GBrowse http://www.gmod.org/?q=node/71

GBrowse is a genomic DAS client developed by Lincoln Stein's group. It lets the user to browse a genomic sequence (genome, chromosome, contig, etc.) and see the annotations that are available for this sequence in a suitable way.

GBrowse is perl-based.

We have applied GBrowse to a project we are involved in.

Streptomyces coelicolor

The INB is actively participating in the project 'Análisis Bioinformático del Genoma de *Streptomyces*' **(GEN2003-20245-C09-09)**, funded by the Spanish Ministry of Education and Science. This three year project involves eight molecular biology groups and a bioinformatics group (PDG, Prof. Alfonso Valencia's lab). The aim of the project is to study the specific and global regulatory pathways of *Streptomyces coelicolor* (SCO) related to the production of antibiotics. SCO was sequenced by the Sanger Institute in collaboration with Prof. David Hopwood of the John Innes Center **(Bentley et al., Nature 2002)**. It is the model representative of a group of soil-dwelling organism that are notable for their production of chemical useful compounds including anti-tumor agents, immunosuppressants and over two-thirds of all natural antibiotics currently available. The chromosome is 8,667,507bp and is predicted to contain 7825 genes. The bionformatics work started with the functional annotation of the complete SCO genome. For this purpose we used an in-house functional annotation algorithm, **FUNcut (Abascal & Valencia, Bioinformatics 2002, Abascal & Valencia, Proteins 2003)**. This program does (recursive or simple) **BLAST** searches to collect similar sequences that then aligns 'all versus all' to obtain a measure of the distance between all possible pairs. This way, a representation of the sequence space is obtained, and a clustering algorithm tries to detect differentiated groups of sequences in that space that ideally will belong to a common subfamily, in which function is expected to be conserved. The 'subfamily' of the target sequence is retrieved and its sequence annotations inspected to distill: 1) a representative function description; 2) enzymatic activities (EC numbers); and 3) keywords.

These annotations are available under a DAS server **DAS (Dowell et al., BMC Bioinformatics 2001)** and ready to visualize with **GBrowse (Stein et al., Genome Res.2002).**

GFF files

GBrowse

- We wanted to show annotations for the *Streptomyces coelicolor* genome. Initially we took all the annotations available in GenBank for this bacteria. Furthermore we generated additional functional annotations with FUNcut, an algorithm for functional annotation developed by Federico Abascal.

- All the information about annotations (GenBank + FUNcut) was finally converted to GFF.
- In this case we use Gbrowse as both reference server and annotation server.

GFF files

GBrowse

GBrowse requirements:

GBrowse runs on top of several software packages. These must be installed and configured before you can run GBrowse. Most preconfigured Linux systems will have some of these packages installed already.

A) MySQL
B) Apache Web Server
C) Perl 5.005
D) Standard Perl modules (CGI, DBI, etc)
E) Bioperl version 1.5 or higher

Other modules are optional.

GFF files

GBrowse

We have to set up a MySQL database with the data. As we wanted the server to act as both reference and annotation servers, it is required that the database contains the annotations we have and the genomic sequence of this bacteria in FASTA format.

```
> AL645882
CCCGCGGAGCGGGTACCACATCGCTGCGCGATGTGCGAGCGAACACCCGGGCTGCGCCCGGGTGTTGCGC
TCCCGCTCCGCGGGAGCGCTGGCGGGACGCTGCGCGTCCCGCTCACCAAGCCCGCTTCGCGGGCTTGGTG
ACGCTCCGTCCGCTGCGCTTCCGGAGTTGCGGGGCTTCGCCCCGCTAACCCTGGGCCTCGCTTCGCTCCG
CCTTGGGCCTGCGGCGGGTCCGCTGCGCTCCCCCGCCTCAAGGGCCCTTCCGGCTGCGCCTCCAGGACCC
AACCGCTTGCGCGGGCCTGGCTGGCTACGAGGATCGGGGGTCGCTCGTTCGTGTCGGGTTCTAGTGTAGT
GGCTGCCTCAGATAGATGCAGCATGTATCGTTGGCAGAAATATGGGACACCCGCCAGTCACTCGGGAATC
TCCCAAGTTTCGAGAGGATGGCCAGATGACCGGTCACCACGAATCTACCGGACCAGGTACCGCGCTGAGC
AGCGATTCGACGTGCCGGGTGACGCAGTATCAGACGGCGGGTGTGAACGCCCGGTTGAGACTGTTCGCGC
TCCTGGAGCGCCGGGCGTGCCCGCGAGCGAGGCGGACGACCTGGTGGCCGGGGCGCAGTGCGAGGTGGTG
GAGCTGGACGGCATGGCGCCGGCTTCTCGGGGTGTGCTGTGTTCGCGGACGGCTGGGACGAAGGCGTGAT
.....
```

GFF files

GBrowse

We first create a database in MySQL.

```
mysql> create database Streptomyces_coelicolor;
Query OK, 1 row affected (0.08 sec)

mysql> grant all privileges on Streptomyces_coelicolor.* to ograna@'localhost';
Query OK, 0 rows affected (0.49 sec)

mysql> grant select on Streptomyces_coelicolor.* to GBrowseUser@'ubio.cnio.es';
Query OK, 0 rows affected (0.00 sec)

mysql> flush privileges;
Query OK, 0 rows affected (0.14 sec)
```

GFF files

GBrowse

Once the database is created we have to upload all the data (annotations + fasta file). For doing that there are three bioperl scripts that we can use:

bp_load_gff.pl (very slow, not suitable for big GFFs)
bp_bulk_load_gff.pl
bp_fast_load_gff.pl

Before using them we have to define a path to a temporal directory:

export TMPDIR=/tmp (bash)
setenv TMPDIR /tmp (tcsh)

GFF files

GBrowse

We are now ready to upload the data:

```
ograna@ancares:~$ bp_bulk_load_gff.pl -d Streptomyces_coelicolor AL645882.fasta AL645882.gff
This operation will delete all existing data in database Streptomyces_coelicolor.  Continue? y
Preparing embedded sequence....
done....
Preparing DNA files....
done...
Loading feature data.  You may see duplicate key warnings here...
done...
19214 features successfully loaded
```

***** Database ready !!!

Other possibilities could be:

```
bp_bulk_load_gff.pl -d Streptomyces_coelicolor AL645882.fasta *gff (in case of several GFF files)
bp_bulk_load_gff.pl -d Streptomyces_coelicolor AL645882.gff -fasta AL645882.fasta
bp_bulk_load_gff.pl -d Streptomyces_coelicolor *.gff -fasta AL645882.fasta
```

GBrowse

The following step is to create a configuration file in the GBrowse configuration directory:

This file must be named as **'filename.conf'**

```
[GENERAL]
description   = Streptomyces coelicolor A3(2) Prueba Curso DAS
db_adaptor    = Bio::DB::GFF
db_args       = -adaptor dbi::mysql
                -dsn dbi:mysql:database=Streptomyces_coelicolor;host=ancares.cnio.es

user = GBrowseUser
pass =

# DAS reference server
das mapmaster = self
```

GBrowse

It is possible that GBrowse reads the data directly from the GFF file instead of creating a MySQL database.
We just have to specifiy in the configuration file that we are using a GFF file and the location where it is:

```
[GENERAL]
description   = FUNCut New Annotations
db_adaptor    = Bio::DB::GFF
db_args       = -adaptor memory
                -gff    '/pathToTheGFFfile/'
```

Using a GFF file is only possible when it contains a very few information. Otherwise queries would be completely inefficient (long time of response). If the GFF is really big the system is just unable to return a response.

My personal suggestion is to use a database.

GBrowse

All we have to do now is to define inside the configuration file the GFF features that we want to display with GBrowse:

```
################### TRACK CONFIGURATION ####################
# the remainder of the sections configure individual tracks
############################################################

[Gene]
feature      = gene
glyph        = generic
height       = 5
stranded     = 1
bgcolor      = red
fgcolor      = red
description  = 1
key          = Gene
das category = translation

[CDS]
feature      = CDS
glyph        = cds
height       = 5
stranded     = 1
bgcolor      = brown
fgcolor      = brown
height       = 10
description  = 0
key          = CDS
das category = translation

Etc……
```

GBrowse

We can even add links to external databases:

```
[scocyc]
glyph          = generic
stranded       = 0
height         = 5
fgcolor        = orange
bgcolor        = orange
key            = ScoCyc entry
link_target    = _blank
feature        = gene
link           = http://scocyc.jic.bbsrc.ac.uk:1555/SCO/NEW-IMAGE?type=NIL&object=$name
```

GBrowse

Disadvantages of using Gbrowse as a client for showing annotations:

- It is only appropriate for genomic data (DNA), but not for protein data (aminoacids).
- Users are allowed to upload their own annotations. There is an upolad function in GBrowse that works even if the GBrowse you are uploading to is located on a remote server. The uploaded files are stored in a private directory on the server away from the main data set. **Other users cannot see your data.**

Advantages of using GBrowse as a client for showing annotations:

Once GBrowse is set up as a DAS client it is straightforward to also have a DAS server.

GBrowse

Configuring Gbrowse to be a DAS server: The following flags must be specified within the configuration file.

das mapmaster

This option, which should appear somewhere in the [GENERAL] section, indicates that the database should be made available as a DAS source. The value of the option corresponds to the URL of the DAS reference server for this data source, or “SELF’’ if this database is its own reference server.

Examples:

das mapmaster = SELF

das mapmaster = http://www.wormbase.org/db/das/elegans

GBrowse

Configuring GBrowse to be a DAS server: The following flags must be specified within the configuration file.

das category

This option must appear in each of the track configuration stanzas that you wish to export as DAS-accessible data. Note that it is not sufficient to define a das category in the [TRACK DEFAULTS] section. The value of this option should be one of:

repeat
transcription
translation
variation
experimental
structural
miscellaneous

GFF files

It is possible to represent sequence alignments inside GFF files. The requirements are to have the fasta sequences of the species we want to align, and an alignment program.

Once we have performed the alignment we have to produce the appropriate GFF file with this information:

```
sco    conserved70    similarity    3539792 3539858 70    +    .    Target "Sequence:mtu" 95006 95072
sco    conserved60    similarity    3539862 3539909 60    +    .    Target "Sequence:mtu" 95073 95120
sco    conserved60    similarity    3539916 3539991 63    +    .    Target "Sequence:mtu" 95121 95196
sco    conserved80    similarity    3539993 3539997 80    +    .    Target "Sequence:mtu" 95197 95201
sco    conserved90    similarity    3539998 3540023 92    +    .    Target "Sequence:mtu" 95203 95228
sco    conserved80    similarity    3540024 3540029 83    +    .    Target "Sequence:mtu" 95230 95235
sco    conserved60    similarity    3540031 3540128 62    +    .    Target "Sequence:mtu" 95236 95333
sco    conserved50    similarity    3540129 3540176 58    +    .    Target "Sequence:mtu" 95337 95384
sco    conserved60    similarity    3540178 3540192 60    +    .    Target "Sequence:mtu" 95385 95399
sco    conserved60    similarity    3540193 3540262 64    +    .    Target "Sequence:mtu" 95407 95476
sco    conserved50    similarity    3540263 3540280 56    +    .    Target "Sequence:mtu" 95483 95500
sco    conserved80    similarity    3540281 3540287 86    +    .    Target "Sequence:mtu" 95504 95510
sco    conserved100   similarity    3540297 3540309 100   +    .    Target "Sequence:mtu" 95511 95523
```

Some web sites to keep in mind

GFF 2.0	http://www.sanger.ac.uk/Software/formats/GFF/
GFF 3.0	http://song.sourceforge.net/gff3.shtml
GBrowse web site and tutorials	http://www.gmod.org/?q=node/71

THANKS for your attention !!!