

Functional specification of ARCADE

Introduction

The existence of large amounts of genomic information from many species on different databases means that sophisticated tools are required to retrieve the information important to a user and to display it in a coherent manner. Furthermore, since information about one species can be used to aid research on closely related species there is great need for tools to be able to perform complex comparative queries. For example, the conserved gene order found in some cereals may be exploited to locate homologous genes in related species. A Real-time Comparative Analysis Display Environment (ARCADE) is a software tool for carrying out such comparative queries within a distributed environment.

Primary function of ARCADE

The primary function of ARCADE is to perform comparative queries across data from many different sources and to display the results of these queries in a suitable coherent format.

Comparative queries

Users will be able to perform queries at several levels from the most basic to highly complex queries. Simple comparative queries can be grouped into two distinct types.

1. Common properties

In this type of query objects are compared based on common properties. A simple query of this type would be to find all objects that have the same name. Alternatively, the user may wish to find all genes that are associated with a particular function such as flower development.

2. Membership of a group

Queries of this type rely on the existence of predefined groups of comparable objects. For example, the user may wish to find all genes that are homologous to a specified gene in which they are interested. In our current schema such queries would be performed using a homology group, which defines homologous loci. Users will also be able to define their own groups and use these groups in comparative queries.

More complex queries will make use of two or more of these queries combined by Boolean operators. For example, the user may wish to find all the genes that are homologous to a specified gene, which are also associated with flower development. This query is simply the combination of the above two example queries. Using just the basic Boolean operators AND, OR and NOT a user will be able to form any query as a logical combination of a number of these basic search queries.

Data sources and problems of compatibility

ARCADE will not be reliant on any specific database system or data schema, but should be able to access any data source, including user-inputted files, via an appropriate interface (discussed in the

ARCADE technical specification). The user should be given the option to specify which data sources should be queried.

A major problem in performing comparative queries across different data sources is that there can be incompatibility between different sources, which can exist within both the structure (data types and their properties) and the content of the data. ARCADE will overcome the problems associated with different data structures by adopting a standard schema and by providing interfaces from each of the different data sources as mentioned above, in order to map them onto this schema.

ARCADE will also be able cope with incompatibility at the data content level. This incompatibility can occur when different data sources use different names for the same biological object or even when the same name is used for different objects. Naming problems of this type can easily be overcome because any object can be uniquely identified using the name of an object, its type and the name of its source provided there are no naming conflicts within a specific data source.

Incompatibility can also occur when databases contain different keywords to denote a common property. This is not a problem when all the possible values that a property can take are specified with the data schema. However, this is not true, for example in the case of a text tag, which can take any alphanumerical value. In this case the data may contain several text keywords to denote the same property. This degeneracy may also be present within the same as well as different data sources, especially, for example when data are obtained from different laboratories that employ different nomenclatures. ARCADE therefore must have a mechanism to cope with this degeneracy.

User interface to ARCADE

ARCADE should have a suitable user interface(s), which will allow users to perform both simple and complex queries in a graphical and/or textual environment. This interface(s) will be platform independent and portable. A user will have the option to enter a text query directly or to construct a query via a 'Query Builder' tool, and to save queries for later use. The results of a query will be in the form of a result set (i.e. a list of data items that fit the criteria defined by the query), which can be displayed in appropriate format (see below), used in another query or saved for later use. ARCADE will be able to display data retrieved in whatever appropriate format required by the user. What constitutes an appropriate format is determined by the type(s) of the data to be displayed. For example, an appropriate display for a group of homologous loci that map on to different linkage groups would be a multi-map display

Configurability and defaults

ARCADE will be fully configurable by the user and any configurations can be saved for use in the next session. The user will have ability to accept the default parameters, define their own defaults or even set any or all parameters before each query. For example, users will have full control over which data sources should be queried. However, where not specifically defined ARCADE will make assumptions about which data sources are used.