A SYSTEM, METHOD AND INTERFACE FOR BUILDING BIOLOGICAL DATABASES USING TEMPLATES

Field of the Invention

This invention relates to a system, method and interface for building biological databases using templates and particularly, but not exclusively, to such systems, methods and interfaces for immunological databases or databases for MHC molecules.

Background to the invention

A considerable amount of biological data is available from public and other databases. Biological databases are characterised by various degrees of heterogeneity in that they:

- encode different views of the biological domain;
- utilise different data formats;
- utilise various database management systems;
- utilise different data manipulation languages;
- encode data of various levels of complexity;
- are constantly evolving, and are geographically scattered;

Genbank, SWISS-PROT, and other general-purpose databases, are the main source of biological information, but they have data entries from various species.

There is an increasing need for specialist databases that store more detailed subject-oriented data, compared to general-purpose databases. There is also an increasing need to enable database users (for example researchers) to have the ability to create their own databases. A common purpose for such a database is to combine one's own data with publicly available data for creating special-purpose and subject-oriented databases suitable for data mining and knowledge discovery.

Bioinformatics is a field dealing with biomolecular, related structural, related functional, related clinical, and related biochemical data. A standard practice for reusing databases is creation of a subset of an existing bioinformatics database (for example, creating a subset of Genbank that contains only swine sequences). Databases that encode different views to biological data exist, but they are heterogeneous and standards for data integration into an unified database are lacking (Markowitz and Ritter, 1995; Brusic and Zeleznikow, 1999). Attempts to build a unified bioinformatics database for storing biological data have failed to date.

Finally, there is a need to create specialist databases for the same subject across different species.

Consideration of the prior art

Bioinformatics databases:

a) Maslyn et al., US Patent 5,5953,727 relates to a relational bioinformatics database suitable for cataloguing and searching sequences according to association with one or more projects. The present invention provides a system for building and integrating of databases based on combining template modules using multiple dimensions, or views, to data and templates for database tools.

b) Eilbeck K. et al., ISM8199,87-105 is an object oriented database created to provide scientists with a resource for examining protein-protein interactions and inferring possible interactions from the data stored. The present invention is a system for building databases using templates. Some of the dimensions (views) can use, but are not restricted to, an object-oriented data design.

c) Nowacki et al., Nucleic Acids Res 1998 Janl; 26(1): 2205 describes a database of nucleotide variation using a relational data model. The present invention is a system for building of databases using templates.

Use of templates

d) Cruz I.F. and Lucas W.T., 1998. Automatic generation of user-defined virtual documents using query and layout templates. Theory and Practice of Object Systems 4(4), 245-260. An authoring, querying, and visualisation framework for multimedia information retrieved from distributed repositories. Users compose virtual documents by specifying visually templates that contain both layout information and query specification. The present invention is used for building bioinformatics databases.

e) Thalhammer-Reyero, July 27, 1999 US patent 5930154. Computer-based system and methods for information storage, modelling and simulation of complex systems organised in discrete compartments in time and space. US patent 5930154. An integrated computer-based system, methods, and graphical interfaces, providing an environment for development of visual models of complex systems organised in discrete time and space compartments, used for graphic information storage and retrieval, visual modelling and dynamic simulations of said complex systems. The present invention is used for building bioinformatics databases.

f) Barsalou T., 1989. An object-based architecture for biomedical expert database systems. Computer Methods and Programs in Biomedicine 30(2-3):157-168. This discloses an object-oriented system for database structuring and manipulation for expert systems. The present invention can use, but is not restricted to using an object-oriented design. The present invention is used for building and use of bioinformatics databases.

Reusable databases and systems:

g) Kojima T., Nakata H., Kawagishi M., Uehara T., 1998. A framework for constructing databases for supervisory control systems. Electrical Engineering

in Japan 123(1), 32-42. The proposed framework utilises a generation-based approach and object-oriented framework libraries. The present invention is used for building bioinformatics databases.

h) Nguyen J.H., Shahar Y., Tu S.W., Das A.K. and Musen M.A, 1999. Integration of temporal reasoning and temporal-data maintenance into a reusable database mediator to answer abstract, time-oriented queries: The Tzolkin system. Journal of Intelligent Information Systems 13(1-2), 121-145. Tzolkin system facilitates the expression of clinical queries to reduce the manual data processing that users must undertake to decipher the answers to their queries. This approach is general, facilitates software reuse, and thus decreases the cost of building new software systems that require this functionality. Tzolkin facilitates software reuse for generating clinical queries, while the present invention facilitates software reuse for building bioinformatics-related databases.

i) Gennari J.H., Cheng H.N., Altman R.B. and Musen M.A., 1998. Reuse, CORBA, and knowledge-based systems. International Journal of Human-Computer Studies 49(4), 523-546. They developed CORBA-based architecture for a library of platform-independent, sharable problem-solving methods and knowledge bases. The aim of this library is to allow developers to reuse these components across different tasks and domains. The present invention system can, but is not limited to, use CORBA for extraction of data. The present invention approach does not necessarily utilize CORBA standards.

Integration of heterogeneous data:

j) Davidson, S.B., Overton, C., Tannen, V., Wong, L., 1997. BioKleisli: a digital library for biomedical researchers. International Journal of Digital Libraries 1(1), 36-53. Kleisli system enables complex queries across multiple databases

and data integration. Kleisli does not specify the system for the user-end data integration using templates. The present invention can, but is not limited to, use Kleisli for accessing heterogeneous data sources and data extraction.

k) Macauley J., Wang H. and Goodman N., 1998. A model system for studying the integration of molecular biology databases. Bioinformatics 14(7):575-582. They tried to build a gene data warehouse by automatic extraction of entries from public databases and discovered numerous errors (up to 20% of entries were determined erroneous by a single criterion). The present invention allows use of templates for expert annotation and is not limited to automatic data acquisition.

l) Chen I.M., Kosky A.S., Markowitz V.M., Szeto E. and Topaloglou T., 1998. Advanced query mechanisms for biological databases. ISMB, 6, 43-51. This describes a system for integrating tools for exploring multiple heterogeneous databases using Object-Protocol-Model. The present invention allows integration of tools based on templates defined for each tool.

Data warehousing

m) Wu O.P., Seow K.T., Wong L., Chung S.Y. and Subbiah S. 1998. From sequence to structure to literature: the protocol approach to bioinformation. Pacific Symposium of Biocomputing, 747-758. They have described a system for data integration and building data warehouses by extracting information from heterogeneous sources. The present invention describes a system for building a bioinformatic database or a data warehouse by using a set of templates and integration with a set of tools for use of this database.

n) Eckman B.A., Aaronson J.S., Borkowski J.A., Bailey W.J., Elliston K.O., Williamson A.R., Blevins R.A., 1998. Bioinformatics. 1998;14(1):2-13

describe a database for storage and use of the expressed sequence tag (EST) data. The present invention is a system for building databases using templates.

o) Sorace J.M. and Canfield K. 1998 Collaborative bioinformatics: data warehouses for targeted experimental results. Journal of Interferon and Cytokine Research 18(9), 799-802. They describe a data warehouse that stores heterogeneous data on measurements of in vitro cellular functions using a single data model. The present invention is a general model for building bioinformatics databases using templates.

Dimensional data model

p) Bunardzic A., 1995. Dimensional modelling: beyond data processing constraints. Medinfo, 8 Pt 1, 520. This describes the dimensional model focusing on the knowledge of the relevant facts, which are reflecting the business operations and are the real basis for the decision support and business analysis. The present invention focuses on bioinformatics domain.

Knowledge discovery from databases

q) Kolchanov N.A., Ponomarenko M.P., Frolov A.S., Ananko E.A., Kolpakov F.A., Ignatieva E.V., Podkolodnaya O.A., Goryachkovskaya T.N., Stepanenko I.L., Merkulova T.I., Babenko V.V., Ponomarenko Y.V., Kochetov A.V., Podkolodny N.L., Vorobiev D.V., Lavryushev S.V., Grigorovich D.A., Kondrakhin Y.V., Milanesi L., Wingender E., Solovyev V. and Overton G.C. 1999. Integrated databases and computer systems for studying eukaryotic gene expression. Bioinformatics 1999 Jul;15(7):669-686. They describe an integrated database for integration of informational and software resources on the regulation of gene expression, navigation through them and discovery of related knowledge. The present invention is the general system for building

and using bioinformatics databases based on template use, suitable for knowledge discovery.

Brusic V. and Zeleznikow J., 1999. Knowledge Discovery and Data Mining in Biological Databases. Knowledge Engineering Review 14(3).

Markowitz V.M. and Ritter O., 1995. Characterising heterogeneous molecular biology database systems. Journal of Computational Biology 2(4), 547-556.

r) Further references

Altschul S.F. and Gish W. (1996). Methods Enzymol. **266**: 460-480. Bairoch, A., Apweiler, R., 1999. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res. 27, 49-54.

Ballard C., Herreman D., Schau D., Bell R., Kim, E. and Valencic A. 1998. Data Modeling Techniques for Data Warehousing. IBM Corporation, International Technical Support Organization, San Jose, California.

Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., Ouellette, B.F., Rapp, B.A., Wheeler, D.L., 1999. GenBank. Nucleic Acids Res. 27, 12-17.

Brusic V. and Zeleznikow J., 1999. Knowledge Discovery and Data Mining in Biological Databases. Knowledge Engineering Review 14(3).

Bunardzic A., 1995. Dimensional Modeling: beyond data processing constraints. Medinfo 8 Pt 1, 520.

Davidson, S.B., Overton, C., Tannen, V., Wong, L., 1997. BioKleisli: a digital library for biomedical researchers. International Journal of Digital Libraries 1(1), 36-53.

Fayyad U., Piatetsky-Shapiro G. and Smyth P., 1996. From data mining to knowledge discovery. AI Magazine 17(3), 37-54.

Fisher L., 1996. Along the infobahn. Strategy&Business, Third Quarter, 1996. Booz – Allen & Hamilton, Inc. <http://www.strategy-business.com/technology/96308>.

Objects of the invention

It is therefore the principal object of the present invention to provide a system which has a framework for relatively fast and relatively efficient:

a) building of specialist bioinformatics databases for different species using the same, or a similar, database structure; and/or

b) building specialist bioinformatics databases for different molecular forms using the same or a similar, database structure; and/or

c) building families of related bioinformatics databases allowing an arbitrary level of complexity; and/or

d) selective combination of private data with public data into new bioinformatics databases; and/or

e) building of data warehouses for data mining and knowledge discovery in life sciences; and/or

f) enabling of the building of data warehouses that integrate data with database search and analysis tools, without a requirement for a significant bioinformatics background.

Summary of the Invention

With the above and other objects in mind, the present invention provides a general system, method, and interfaces for building and integrating databases based on combining template modules using multiple dimensions, or views, to data and templates for database tools. The method may be applicable to domains characterised by complex data where multiple different views to data need to be combined for extraction of information. An example application is from bioinformatics, where multiple databases using this system and method may be built. The information extraction and data management is based on using templates, each of which is designed for a specific purpose.

This invention may also provide a system and method for a relatively efficient creation of bioinformatics databases that concentrate on a particular subject within the bioinformatics field, and may reuse templates related to various views of the subject-related data. The breadth and the depth of the coverage of the resulting database may depend on user specifications.

Description of the drawings

In order that the invention may be readily understood and put into practical effect, there shall now described by way of non-limitative example only a preferred embodiment of the present invention, the description being with reference to the accompanying illustrative drawings in which:

Figure 1 is an illustration of the general arrangement of the present invention;

Figure 2 is a representation of an example of the main page of an interface;

Figure 3 is a representation of a second example of the main page of an interface;

Figure 4 is a representation of the views and links of an exemplary database;

Figure 5 is a flow chart for the building of a database;

Figure 6 is an illustration of the design structure for the definition and implementation of a template;

Figure 7 is a flow-chart for the building of a template;

Figure 8 is a representation of a graphical user interface for template selection;

Figure 9 is an illustration of a graphical user interface for template building; and

Figure 10 is an illustration of a graphical user interface for the input of parameters for the main template.

Description of preferred embodiment.

The present invention system uses templates for building a database by loading data into a structure defined by the templates.

The general operation of the present invention is given in Figure1. Data sources can be external (e.g. "Genbank", or other databases) or local. Data can be acquired by a variety of means: manually or by automatic data extraction - for example using the techniques disclosed in (Davidson et al., referred to above). Templates given in Figure 1 include:

- a main template that defines data dimensions and relevant tools for the target database;
- templates for various data dimensions;
- templates for specific tools to be used for searching and analysis of the target database; and
- templates for search and analysis reports.

Templates for additional functions are to be defined and created as required.

Database building using present invention has two essential elements: database design, and the loading steps. Database design can be further divided into more specific steps:

a) decision on the database content;

b) selection of the main template and define the main interface page outlay;

c) selection of the view, tool, and report templates;

d) design new templates if required; and

e) template linking

Data loading steps are:

f) data acquisition; and

g) data loading into files.

These steps will be explained using an example that involves databases exemplary swine leukocyte antigen (SLA) (Figure 2), and which was used as the template for a database of functional immunology (Figure 3).

The contents of the two databases differ, but several templates may be used in them both. The template for the main page (Figure 2 and Fig 3) for each was reused. The SLAD modules 'Retrieve Allele Info', Retrieve Epitope Info', and 'Search References' and the FIMM modules 'Diseases', 'Antigens', 'Search FIMM', and 'References' all use the same family of templates that enable keyword searching. The variations of these templates include optional selection of data dimensions, selection of the number of output entries, or other possible search limiting criteria.

The BLAST search (Altschul and Gish, 1996) templates were used in modules 'BLAST MHC Databases' (SLA) as well as in 'Blast Antigens' and 'Blast HLA' (FIMM). Other modules used in SLAD and FIMM include modules providing physical maps of genes, sequence alignments for proteins and DNA, phylogenetic analysis, finding and analysis of peptide binding sites, display of 3-D structure of molecules, internet links, and motif searching. Templates for other analysis and search queries related to bioinformatics problems can be added and integrated into the template library.

A template preferably consists of an interface page, file formats for data storing, and a set of programs that allow data storing and data retrieval. The interface page may take a standard form such as, for example, the BLAST interface, widely used for the Internet BLAST services; or may be novel such as 'Phylogenetic analysis' of SLAD, used for inter-species sequence comparison. The format of files for data storing is flexible - it depends on the bioinformatics problem related to the question asked. For SLAD and FIMM, some of the files may be flat files, containing record fields, labels and delimiters, or bin-hexed files suitable for BLAST searches.

The list of possible templates is given in Table 1. The present invention is not limited to these templates and other templates for other purposes may be developed. The present invention allows users to build their own databases by selecting the appropriate templates, maintain the databases, and annotate new entries. It also allows users to combine sequence search and analysis tools within the database. It also allows database access and tools to be packaged in a single interface, and brings together the capacity for a user to build the databases and integrate sequence analysis tools. Integrated sequence analysis tools were previously available through packages like GCG (Genetics Computer Group, Wisconsin, USA) but these packages do not enable a user to build databases; they only enable user to create individual sequence entries as separate files and access them through lists of file names.

The present invention allows the building of bioinformatic data warehouses. A data warehouse is a database structured to facilitate analytical tasks, rather than operational purposes. The present invention provides the framework for building bioinformatics warehouses by combining and integrating various data views and analysis tools. Data warehouses are commonly used for performing Knowledge Discovery from Databases (KDD). KDD is defined as the non-trivial process of identifying valid, novel, potentially useful, and understandable patterns in data. Data warehousing has not previously been described in bioinformatics.

Table 1.

Table 1

Templates	Function
Main page	Defining and linking database content
Links	Internet links to other sites
Information page	Displaying database information
Report	Templates for various reports produced as results of database queries
Lists	Templates for various lists produced as results of database queries
Keyword searching I	Search the database content by keyword
Keyword searching II	Search the database content by keyword with predefined search limiting criteria
Sequence motif searching	Search sequences for presence of motifs and sequence patterns
BLAST searching	Search sequence databases using BLAST
Multiple sequence alignment I	Alignment of multiple sequences from the database
Multiple sequence alignment II	Alignment of multiple sequences from the database with a query sequence
Phylogenetic analysis	Analysis of inter-allelic or inter-species sequence similarity using phylogenetic analysis
Active site analysis I	Comparison of active sites from multiple sequences from the database and displaying of multiple active sites
Active site analysis II	Comparison of active sites from representative sequences from the database and from a query sequence and displaying of multiple active sites
Physical map	Display physical genetic map and allow extraction of further information by clicking the objects
Structure models	Displays 3-dimensional structure models of molecules
Mutation	Displays aligned sequence fragments of variant sequences
Varius database templates	Allows searching sequences from various sequence databases (e.g. non-redundant amino acids database, Swiss-Prot/TREMBL, MHC databases etc.)
Formatted file download	Downloads sequence files in various formats (e.g. Fasta, zip, tar)
Genomic sequence	Searching, analysis and display of genomic sequences
Automatic update	Automatic updating of databases
Expert update	Updating of database by human expert
Annotation I	New data annotation by automatic data extraction
Annotation II	New data annotation by inspection by human expert
Prediction I	Prediction of sequence structure, function, regions, domains etc. of database entries
Prediction II	Prediction of sequence structure, function, regions, domains etc. of database entries and query sequences

Sequence comparison	Templates for tools for sequence comparisons
Templates	**Function**
Multiple sequence comparison	Templates for tools for multiple sequence comparisons
Searching	Templates for database searching using various sequence comparison, pattern, motif, profile, and similar searches
Evolutionary analysis	Templates related to analysis of evolutionary relationships
Format interchange	Templates for conversion of various formats associated with biomolecular data
Functional sites	Templates for searching and analysis of various functional sites
Motifs and domains	Templates for searching and analysis of various motif sites and sequence domains
Display forms	Various graphical displays of results of search and analysis
Data mining	Templates for various tools for data mining
KDD	Templates for various tools for knowledge discovery from databases

FIMM and SLAD utilise dimensional modelling, which enables users to form multidimensional views of the relevant facts which are stored in a 'flat' (non-structured), easy-to-comprehend and easy-to-access database. Relational modelling appears too rigid to provide efficient extraction of data for analytical processing needs. Another alternative approach, using the object-oriented modelling can deal with complex data structure, but is difficult to build and has highly structured data. At the core of the dimensional modelling are fact tables that contain the non-discrete, additive data.

The multidimensional views of the FIMM database and their links are shown in Table 2. Data from various views in FIMM are linked, providing the ability to produce series related reports. The links of the FIMM database are given in Figure 4.

Table 2

Dimensions				
Protein antigens	**Peptides**	**MHC**	**Diseases**	**Sources**
ID Date Name Aliases Features Sequence Links -Peptides -Diseases -Sources -SwissProt -Genbank	ID Date Sequence MHC T-cell epitope Naturally processed Peptide binding Links -Prot. antigens -MHC -Sources -MHCPEP -SYFPAETHY - HIVdB - SWISSPROT	ID Date Name Aliases Sequence Disease assoc. Binding pockets Binding sites CD8 Binding sites TcR Alignments Sources Links -Peptides -Diseases -IMGT database	ID Date Name Aliases Etiology Disease assoc. Links -Protein antigens -MHC -Sources -OMIM -GeneCards -Genatlas	PMID Title Authors Journal Links -PubMed

Database building using the present invention may be a multi-step process. It preferably consists of template selection, template storage, template building, refinement (if necessary), and integration of the templates into the database.

The process of database building is given in Figure 5. The general template design structure is given in Figure 6. The process of building individual templates is given in Figure 7. Examples of various graphical user interfaces are given in Figures 8, 9 and 10.

Each database may have at least three dimensions selected from the list consisting of, but not limited to, sequence structural data, sequence functional data, gene expression data, protein expression data, relevant pathology associations, evolutionary data, data on biologically active sites within biomolecular sequences, data on biochemically

active sites within biomolecular sequences, pharmacological data, and sequence patterns and motifs.

To now refer to Figure 5, the steps are to:

- select a set of templates from a master list of templates to be integrated into the database. Refer to Figure 8 for the GUI;
- store the set of selected templates for the database;
- complete the specifications for each template. Each template consists of a set of sub-specifications (refer to the Figure 6). Further details of this step are given in Figure 7. Refer to the Figure 9 for the GUI;
- store the specifications and other required information;
- integration is then confirmed and specifications corrected if necessary; then
- conduct integration and building of the database based on the information collected from the storing step described above

In Figure 6 there is shown the design structure for the definition and implementation a template. Each of the templates may contain the full set of sub-specifications, or a partial set (i.e. not all templates will have data input-output).

To refer now to Figure 7 the user interface for the template is usually a HTML page which collects the input parameters from the users. An example is given in Figure 10 for the main template.

The source and the format of the input data for the template are then specified.

Since the input data will be in heterogeneous formats, they may need to be reformatted before storing into the database. The data output format refers to the format of the stored data records.

The format of the record to be displayed by the template after processing is then specified. The default format is based on the data output.

The tools and the procedures to be used to process the data are then specified. A set of tools serves as a master copy for the tool specification as a result, the system then generates the integration logic of the template with other templates based on the specifications.

The specifications are finally confirmed.

Figure 8 shows a graphical user interface for the template selection. The graphical user interface may have a first polygonal area and a plurality of contained polygonal areas and/or textual links within the first polygonal area; the contained polygonal areas including at least one second polygonal area to enable available templates to be listed, and a third polygonal area to enable the selected templates to be listed. The second polygonal area can display titles of selected templates, and the third polygonal area can specify the data to be entered on the database.

A fourth polygonal area may be provided for specifying additional data to be entered on the database. The second polygonal area may be in a plurality of segments, with there being one segment for each template title. The first, second, third and fourth polygonal areas are preferably rectangular, as are each of the plurality of segments.

The template shown in Figure 10 shows a preferred form of a graphical user interface for the specification of the input parameters for the main template. It has a first polygonal area which contains a second polygonal area for displaying a list of selected templates, and a third polygonal area for specifying the data to be entered on the database.

A contained fourth polygonal area for specifying additional data to be entered on the database may also be provided. The second polygonal area is preferably in a plurality of segments, there being one segment for each template title.

It is preferred that the first, second, third and fourth polygonal areas are rectangular, as are the segments. However, other shapes may be used, if desired.

Figure 9 shows the GUI for template building . Like the GUIs of Figures 8 and 10, it has a first polygonal area, preferably rectangular, and two contained polygonal areas, which are also preferably rectangular. The first contained area is used to select the template, and the second contained area is used to list the sub-specifications of the template selected.

Whilst there has been described in the foregoing description preferred embodiments of the present invention, it will be understood by those skilled in the technology that many variations or modifications in the specific details may be made without departing from the present invention.

The Claims

1) A computer system for creation of at least one bioinformatics database, other than creating subsets of an existing database, wherein:

a) the bioinformatics database has records that comprise sequence records using a dimensional model identifying at least one view to data,

b) a user interface allowing the extraction of information and analysis of data in the bioinformatics database, and,

c) a library of re-usable templates for establishing structure for the bioinformatics database.

2) The computer system of claim 1, wherein a new structure for the bioinformatics database can be produced by combining templates.

3) The computer system of claim 1, wherein new entries are added to the at least one bioinformatics database with new entries being linked by using update templates.

4) A computer system as claimed in claim 1, wherein a new structure the bioinformatics database can be created by combining templates; an new entries are added to the at least one bioinformatics database with new entries being linked by using update templates.

5) The computer system of claim 1, wherein the system is used to produce bioinformatics data warehouses.

6) The computer system of claim 1, wherein the at least one bioinformatics database is used for the purposes selected from the list comprising one or more of knowledge discovery and data mining.

7) A computer system as claimed in of claim 1, wherein the bioinformatics database is selected from the lists comprising immunological databases and MHC-molecules-related databases.

8) A computer system as claimed in claim 1, wherein there are at least three different views to data.

9) A computer system as claimed in claim 8, wherein there is a first view to data which is a nucleotide or protein sequence with basic annotation.

10) A computer system as claimed in claim 9, wherein there are second and subsequent views to data each of which contains at least one view in relation to the sequence of the first view, the second and subsequent views being selected from the list comprising structural data, functional data, peptide data, references, disease association, gene expression, relevant pathology associations, evolutionary data, MHC data, active sites data, pharmacological data, and biological pathways.

11) A computer system as claimed in claim 1, wherein each bioinformatics database has at least three dimensions selected from the list consisting of sequence structural data, sequence functional data, gene expression data, protein expression data, relevant pathology associations, evolutionary data, data on biologically active sites within biomolecular sequences, data on biochemically active sites within biomolecular sequences, pharmacological data, sequence patterns and motifs, and biological pathways.

12) A method for creating multiple related bioinformatics databases, other than creating subsets of existing databases, including:

a) selecting a main template;

b) defining a main interface page outlay;

c) establishing a library of re-usable templates to enable a structure for the bioinformatics database to be established; and

d) linking the templates.

13) The method of claim 12, wherein a new structure for the bioinformatics database is produced by combining templates.

14) The method of claim 12, wherein new entries are added to the at least one bioinformatics database with new entries being linked by using updated templates.

15) The method of claim 12, wherein a new structure the bioinformatics database is created by combining templates; and new entries are added to the at least one bioinformatics database with new entries being linked by using update templates.

16) The method of claim 12, wherein the method is used to produce bioinformatics data warehouses.

17) The method of claim 12, wherein the at least one bioinformatics database is used for the purposes selected from the list comprising one or more of knowledge discovery and data mining.

18) The method of claim 12, wherein the bioinformatics database is selected from the list comprising immunological databases and MHC-molecules-related databases.

19) The method of claim 12, wherein there are at least three different views to data.

20) The method of claim 19, wherein there is a first view to data which is a nucleotide or protein sequence with basic annotation.

21) The method of claim 20, wherein there are second and subsequent views to data each of which contains at least one view in relation to the sequence of the first

view, the second and subsequent views being selected from the list comprising structural data, functional data, peptide data, references, disease association, gene expression, relevant pathology associations, evolutionary data, MHC data, active sites data, pharmacological data, and biological pathways.

22) The method of claim 12, wherein each bioinformatics database has at least three dimensions selected from the list consisting of sequence structural data, sequence functional data, gene expression data, protein expression data, relevant pathology associations, evolutionary data, data on biologically active sites within biomolecular sequences, data on biochemically active sites within biomolecular sequences, pharmacological data, sequence patterns and motifs, and biological pathways.

23) A graphical user interface for use in creating multiple related bioinformatics databases, the graphical user interface having a first polygonal area and a plurality of contained polygonal areas within the first polygonal area; the contained polygonal areas including at least one second polygonal area to enable available templates to be listed, and a third polygonal area to enable the selected templates to be listed.

24) A graphical user interface for use in creating multiple related bioinformatics databases, the interface having a first polygonal area and a plurality of contained polygonal areas within the first polygonal area; the contained polygonal areas including a second polygonal area for displaying fitter of selected templates, a third polygonal area for specifying the data to be entered on the database.

25) A graphical user interface as claimed in claim 25, wherein the contained polygonal areas include a fourth polygonal area for specifying additional data to be entered on the database.

26) A graphical user interface as claimed in claim 25, wherein the second polygonal area is in a plurality of segments, there being one segment for each template title.

27) A graphical user interface as claimed in claim 24, wherein the first, second and third polygonal areas are rectangular.

28) A graphical user interface as claimed in claim 26, wherein the fourth polygonal area is rectangular.

29) A graphical user interface as claimed in claim 27, wherein each of the plurality of segments is rectangular.