Object-oriented Knowledge Bases for the Analysis of Prokaryotic and Eukaryotic Genomes

G. Perrière[1], F. Dorkeld[1], F. Rechenmann[2] and C. Gautier[1]

[1]Laboratoire BGBP — Univ. Claude Bernard Lyon I
43, bd. du 11 Novembre 1918
69622 Villeurbanne Cedex (France)
perriere@biomol.univ-lyon1.fr

[2]IRIMAG — LIFIA
46, av. Felix Viallet
38031 Grenoble Cedex (France)
rechenmann@imag.imag.fr

Abstract

The amount of biological sequences introduced in the general collections, and the growing complexity of the biological knowledge require the construction of models to formalize this knowledge and particularly the relationships between several data types. Two examples of such situations are presented here, they result from the biological research lead in our team in the field of molecular evolution. ColiGene is a modelling of *E. coli* genetics devoted to the analysis of relationships between genomic sequences and gene expressivity. MultiMap implements a new formalization of genome maps allowing manipulation of "maps of maps" in two species. Application of ColiGene and MultiMap are not restricted to molecular evolution and, for instance, MultiMap offers new capabilities for infering data on a genome from knowledge on another species. This could be essential for many mapping projects (human, mouse but also other mammals like pig). Development and implementation of those models have been done using an object-oriented knowledge base management system (SHIRKA) interfaced with a dedicated genomic data base management system (ACNUC). Graphical interfaces have been designed to give an environment similar to the biological representations used by biologists.

Introduction

Since 1980 our group has developed computer tools to handle genomic sequences. We have been among the first to propose genomic data bases (Gautier et al., 1981), then to develop a dedicated data base management system (DBMS): ACNUC (Gouy et al., 1985). ACNUC implements a first modelling of genome organization allowing access to coding parts of complex genes. However ACNUC, as other systems using the relational model (Kanehisa et al., 1984; Kuhara et al., 1984), is not able to represent complex biological structures. For example, it is unhandy to formalize, under the relational model, the various regulation pathways involved in gene expression. Morever, these systems introduced a somewhat artificial boundary between data itself and results of data analysis. It appears clearer and clearer that "methodological knowledge" cannot be separated from "biological knowledge". As an example, the expressivity of *E. coli* genes can be inferred from complex computations applied to sequences. A query to retrieve highly expressed genes needs that the system "knows" the method to estimate expressivity. Object models are particularly adapted to manage and to integrate these two kinds of knowledge.

Material and methods

The object-oriented knowledge base manager SHIRKA (Rechenmann & Uvietta, 1991) has been used to develop ColiGene and MultiMap. It has already been validated in various biological fields: species identification (Gautier & Pavé, 1990) or biological growth models (Rousseau et al., 1986). SHIRKA handles two kind of objects, *classes* and *instances*, under one common formalism based upon a notion derivating from Minsky's frames (Minsky, 1975): the *schemes*. Classes allow the formal definition of the model and instances are realization of the classes, one instance is attached to one or many classes. Classes are organized in a hierarchical structure, which is a common characteristic of most object-oriented systems. SHIRKA integrates various inference mechanisms: pattern-matching, inheritance, procedural attachment. Procedural attachment provides a simple way to link methods to biological objects. It allows association of methods to *slots*. Value inference is then realized by automatic call to calculus procedures.

Queries and navigation in the knowledge bases are made through graphical interfaces called IVAN and IACE (Grivaud, 1992). Under Ivan, navigation is made by mouse pointing in a graphical representation of the hierarchical structure of the base (Fig. 1). A similar mechanism allows navigating along links between instances that result from the existence of slot values belonging to another class of the hierarchy. Moreover, this interface provides management tools such as knowledge bases loading, creation or edition of instances. IVAN provides also a panel devoted to queries. Queries apply to instances of a selected class and result from user's choice of restrictive domains for slot values. The list of instances matching the query condition can be saved in a file and used as input for new queries. We can notice that, under IVAN, instance visualization and queries use all SHIRKA inference mechanisms.

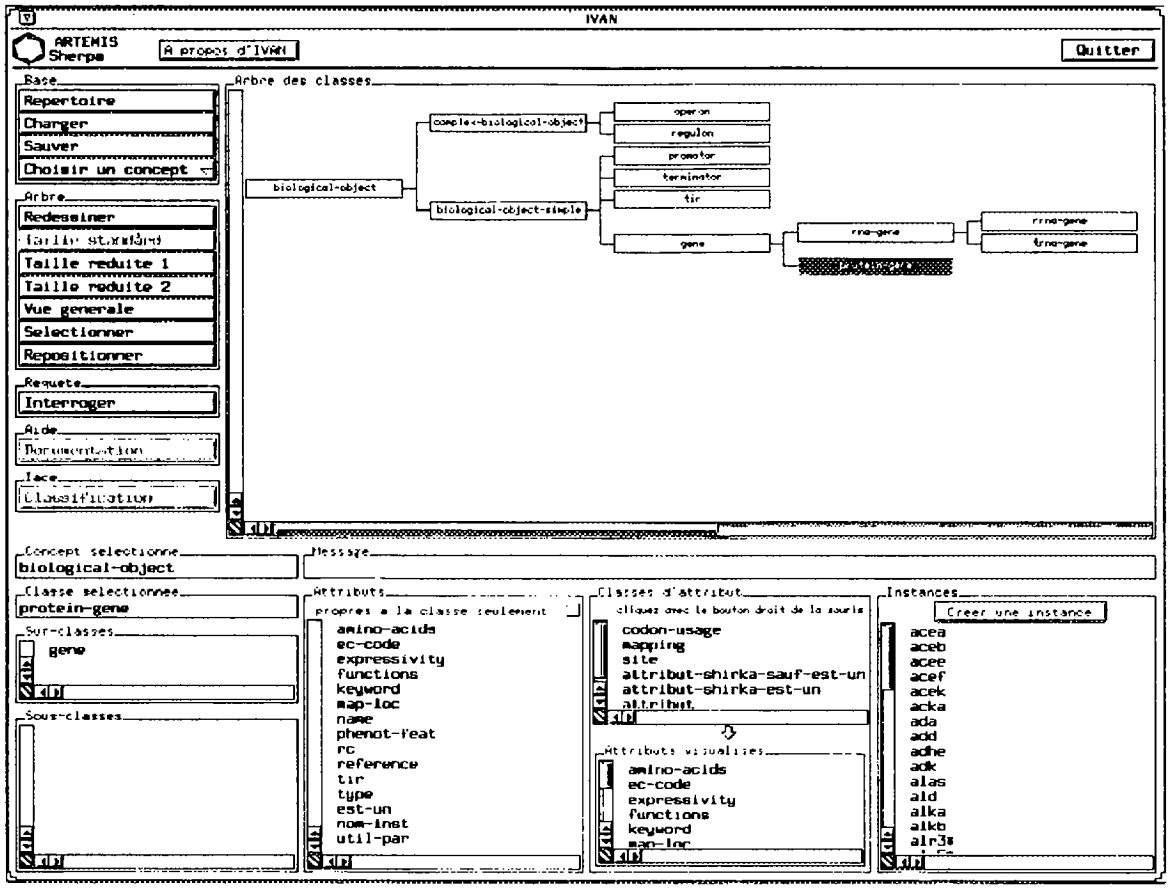

Figure 1 — Main window of the graphical interface IVAN developed for knowledge bases built upon SHIRKA. Panel "Arbre des classes" shows the hierarchical organization of biological objects integrated in the base, it is possible to select a given class at this level. Panel "Attributs" allows to visualize the attributes attached to the selected class. Panel "Instances" contains the list of all instances belonging to this class.

SHIRKA is written in Le_Lisp (ILOG, 1992a); graphical tools attached to SHIRKA (and to the knowledge bases) are developed using the interface generator Aida (ILOG, 1992b). Le_Lisp and Aida are available on a wide set of computers, from PC clones to UNIX workstations. The current versions of ColiGene and MultiMap run on SUN SPARCstations with at least 16 Mbytes in main memory.

Results

Both ColiGene and MultiMap follow the general organization presented in Fig. 2. The core of the structure is the SHIRKA system, associated with its dedicated interface IVAN. The management of the methods linked with the knowledge base is made through graphical interfaces developed under Aida, these methods are mainly written in C. Interfacing knowledge bases with ACNUC allows an efficient management of a large set of nucleotide sequences. ColiGene, in its first stage of development, used all *E. coli* sequences of the GenBank collection (Burks et al., 1991) and is presently connected to the EcoSeq collection (Rudd et al., 1991). MultiMap has access to all human and murid sequences of GenBank with regular updating.

ColiGene

E. coli is one of the best known living organism and about 50% of its genome has already been sequenced. It is the organism where the mechanisms that lead from DNA to proteins are best known, particularly with the work of Ikemura (1981) on tRNA cellular frequencies and the dynamic modelling of translation made by Gouy (1981). Relationships between the way genomic information is written and gene expressivity (the amount of proteins in the cell) have been already identified (Gouy & Gautier, 1982; Blake & Hinds, 1984; Médigue et al., 1991). However, more complex relationships remain to be analyzed, particularly in function of the new available knowledge on translation and transcription initiation processes. This requires a modelling of *E. coli* genetics that takes into account both local organization of genetic information and methods that have been proposed to estimate expressivity. This aim implies, for example, that

ColiGene contains nothing about structures like insertion sequences, recombination sites or replication origins. Moreover, the amount of data concerning the relationships between sequences and expressivity in *E. coli* is so huge, that we just took into account a few mechanisms intervening in the regulation of transcription initiation and translation initiation and elongation. Particularly, we focus on factors allowing to predict the efficiency of the regulation signals involved in these processes.

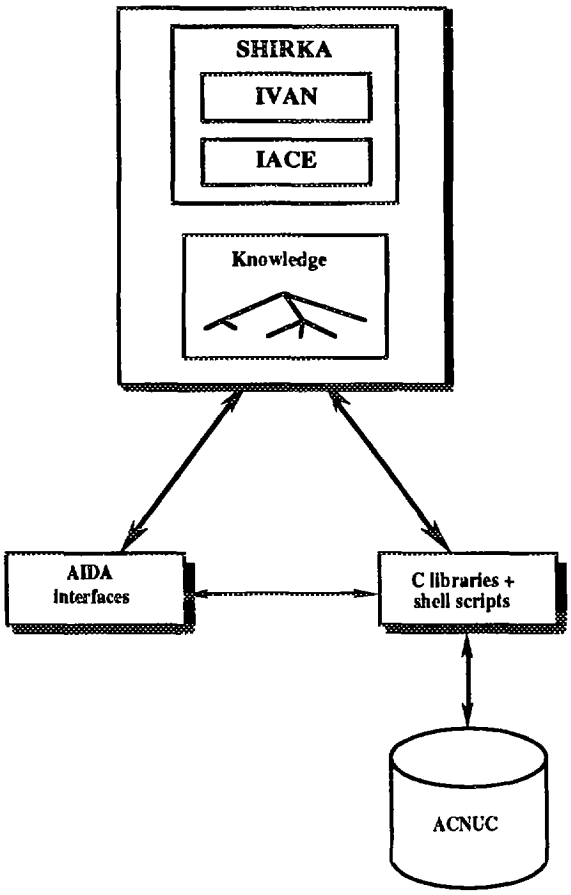

Figure 2 — General organization of ColiGene and MultiMap. The kernel consists in the object-oriented knowledge bases management system SHIRKA, associated to the navigation systems IVAN and IACE. The first one is for general purpose and the second one is dedicated to the use of the SHIRKA classification algorithm (Rechemann & Uvietta, 1991). Linked with the knowledge base, we can find the interfaces for methods management and for nucleotide sequence data base access (ACNUC).

Under ColiGene it is possible to access the structure of objects like transcription and translation initiation signals; genes coding for proteins, tRNA and rRNA and complex organizations like operons and regulons. Gene names, map locations and nucleotide sequences are taken from the EcoSeq/EcoMap/EcoGene collections version 6 (Kenn Rudd, pers. comm.), phenotypic traits associated to the genes, alternative names, and E.C. codes for the enzymes are taken from Bachmann (1990). Promoter data were collected mainly from the Collado-Vides, Magasanik & Gralla compilation (1991). Presently ColiGene allows access to the structure of 1,314 protein genes, 86 structural RNA genes, 178 promoters, 75 transcription terminators and 326 operons. It is also possible to access the 490 nucleotide sequences of the EcoSeq library.

In the class representing protein genes we have integrated generic information such as the systematic and alternate names of the gene (slots **name** and **alt-name**), E.C. code when the encoded protein is an enzyme (**EC-code**), map location in minutes on the *E. coli* chromosome (**map**), the number of amino-acids in the protein (**amino-acids**) and the mnemonic of the sequence associated to the gene (**reference**) in EcoSeq collection. We have considered that the translation initiation region should be integrated in the protein gene structure, so the slot **TIR** allows access to a scheme describing the initiation site attached to a particular gene. Finally we can find two slots corresponding to codon usage indexes: **MND** (Mean Number of tRNA Discriminations per elongation cycle) and **RC** (Right Choice in codon third position). The value of these slots is inferred by procedural attachment and, depending upon resulting indexes scores, the expressivity of the gene is deduced. The possible values for the slot **expressivity** are *high*, *medium* and *weak*, corresponding to three classes of protein genes, named Highly Expressed Genes (HE), Moderatly Expressed Genes (ME) and Weakly Expressed Genes (WE).

In the class representing operons, sequences are not directly associated to instances, so operons where genes and signals correspond to different entries in EcoSeq can be represented. The composition of an operon is summarized in its slot **structure**, so it is possible to query the base to retrieve operons that match a particular structure (e.g. number of genes in the operon, presence of translational coupling between two genes). Here is an example of an operon structure under our formalism: [p]G•GGpOU(t). *p* is for promoter, *G* for gene, *O* for ORF, *U* for URF and *t* for terminators. A dot indicates the occurence of translational coupling between two protein genes. A bracketed structure is potential (i.e. described as potential in the literature) and a square-bracketed structure is putative (i.e. detected using a prediction algorithm). Informations on regulatory factors that enhance or inhibit the transcription of an operon are also introduced.

Methodological knowledge, represented by a wide set of sequence analysis programs, has also been introduced in ColiGene. These programs are written in classical algorithmic languages (C or Fortran) and the communications between the knowledge base and the applications are made by temporary text files or by dynamic linking. Methods are invoked either by procedural attachment or directly by the user. In the first case, the starting of a method is automatic, and is

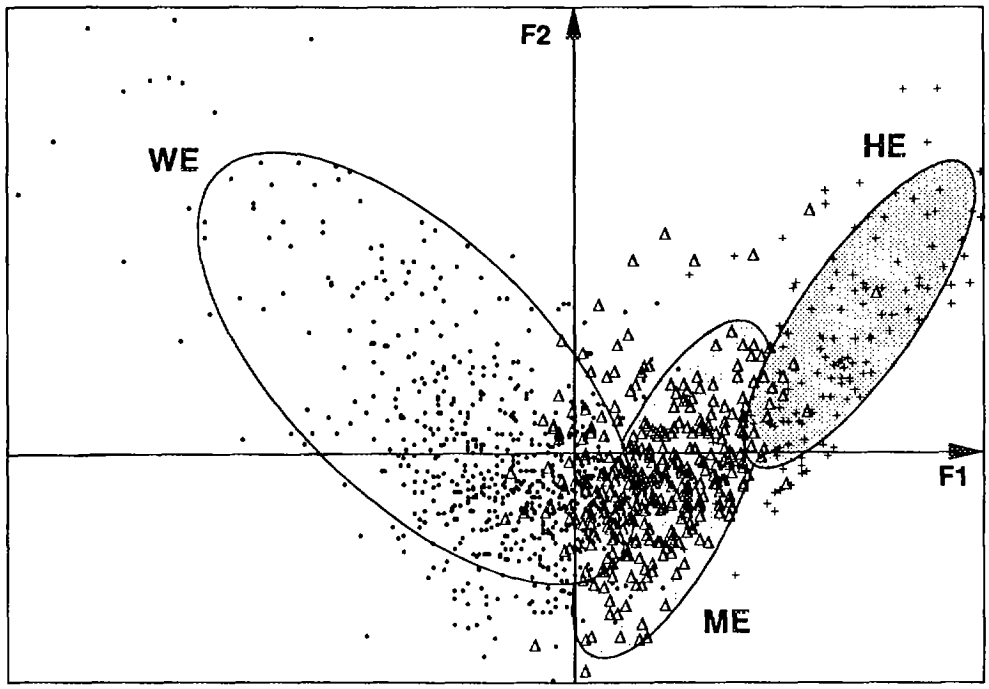

Figure 3 — Use of the CA method coupled with ColiGene to study the codon composition of *E. coli* protein genes. The two first factors of the CA are displayed on the graph, and the repartition of the genes along the first factor shows that their codon composition vary following their expressivity. The three ellipses are grouping genes belonging to the same expressivity level class. From right to left we can find highly expressed genes (HE), moderatly expressed genes (ME) and weakly expressed genes (WE).

realized each time a user wants to visualize an instance containing inferable slots. In the second case, all attached methods are accessible through a menu panel. We have integrated methods for coding sequence localization (proteins and tRNA), promoter and terminator search, etc. After an analysis is complete, it is possible to integrate revealed features in the knowledge base since all research and prediction procedures have an instance creation option available. After instance creation and integration, it is immediately possible to have ColiGene work with the new structures. General purpose tools have also been integrated. So, it is possible to access programs for statistical analyses and to graphical tools that are necessary to exploit the results furnished by these programs.

Biological results have been yet obtained with ColiGene in combination with these programs. For example, we have studied the codon composition of all the *E. coli* protein genes by the mean of the Correspondance Analysis (CA) method. On the graph with the two first factors of CA, we can see that it exists a repartition of the genes along the first factor that follows their expressivity (Fig. 3). This is a confirmation — on a wider set of genes — of the work of Gouy & Gautier (1982) which stated that the codon composition of the *E. coli* protein genes is highly correlated to the their expressivity.

MultiMap

Localisation of genes along the genome is one of the main challenge of modern genetics. Large projects have been initiated to build a genomic map for human but also for several other mammals (mouse, rat and pig as examples). Efficient manipulation of these data is essential to combine information of different types to build composite maps (Collins et al., 1992) more complete than individual maps. Application results mainly from the possibility of finding genetic markers of some critical genes (implied in genetic diseases or as selectionnable quantitive trait in agricultural research). Moreover, important features of gene functionning are linked to localization (DNA condensation, replication date, etc.) Our group is particularly implied in research on the isochore organization of mammalian genomes (Mouchiroud & Gautier, 1990; Mouchiroud et al., 1991) and its evolution. Mammalian genomes are compartimentalized in regions with different G+C% contents called isochores (Bernardi, 1989). This organization seems to have an implication on gene distribution since it appears that genes are preferentially localized in G+C rich isochores. To understand mechanisms responsible of this structure it necessary to compare genomes of several mammals. Such studies imply the manipulation of different kinds of information: i) biological objects such as chromosomes,

cytogenetic bands, DNA sequences and associated concepts usually handled in molecular biology (expression products, introns, exons); ii) sequence localization given by linkage and physical maps with their own units to express gene localization (unaccuracy must be taken into account too); iii) base composition of genes studied and treatments that can be applied to; iv) comparative information between different species, such as conserved chromosome segments (Sawyer, 1991). Part of this information is stored in databases. But these databases do not integrate methods to apply complex treatments on data they store.

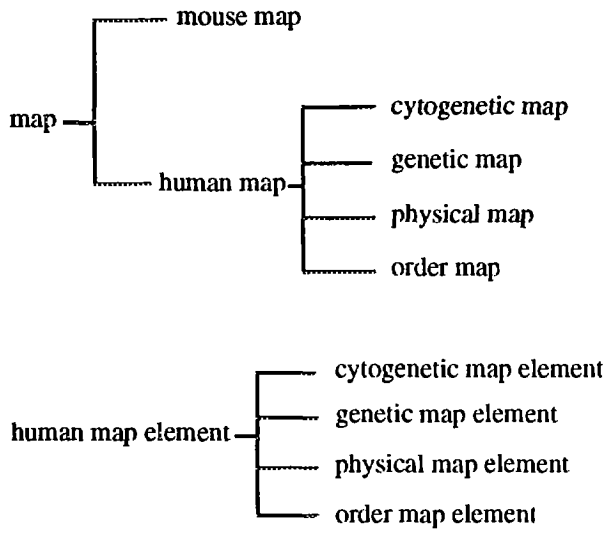

Figure 4 — Simplified hierarchy of classes representing mapping information in MultiMap. A possibility is given to include a map as an element of an other map.

We have then used the object approach of ColiGene to build a system able to handle heterogeneous information (i.e. composition, homology, localization and processing) to complete classical databases. MultiMap integrates data about the human and mouse genomes taken from several databases and personal files. Information on the mouse genome (linkage maps, cytogenetic maps and object nomenclature) is taken from EMG (EMG, 1990). In the same way, GDB (Pearson et al., 1992) and Genatlas (Frézal, 1991) provide data on human cytogenetics and gene definition. An advantage of MultiMap, in comparison with EMG and GDB, is that it allows direct handling of homologous loci and associated data. So base-composition studies implying mapping informations can be achieved. In the same way, the base composition of coding and non-coding sequences can be computed. Biological concepts are described using 27,000 objects.

We have represented chromosomes and links to available maps, genes and pseudogenes, anonymous DNA segments. Organization of gene clusters is described by 95 instances. Cytogenetic chromosome bands obtained with G staining method are described by 1,806 instances. For man, three stages of chromosome banding are represented according to the ISCN classification (Harnden & Klinger, 1985) and to T bands (Dutrillaux, 1973). For mouse, one stage of banding is integrated to the model. Information on homology between mouse and human is also represented and these loci can be handled with their GenBank sequences. Nucleotide sequences of 1,700 human loci and 330 mouse loci are directly accessed through links with Genbank.

Cartographic information is represented by two classes that describe maps and map elements (Fig. 4). In the map hierarchy each class represents a particular type of map (linkage, physical...). A map object includes slots giving the unit used to localize objects (such as CentiMorgan or Kilobase), the chromosome it is linked to, and a slot called **element** which allows the research of instance belonging to subclasses of the class **elements-of-maps**. The subclasses of **elements-of-maps** allow positionning of a large variety of genomic objects in a map. This representation of cartographic information by two classes allows the inclusion of a map as element of another one, which is important to order local maps along a chromosome, having unestimated gaps between them.

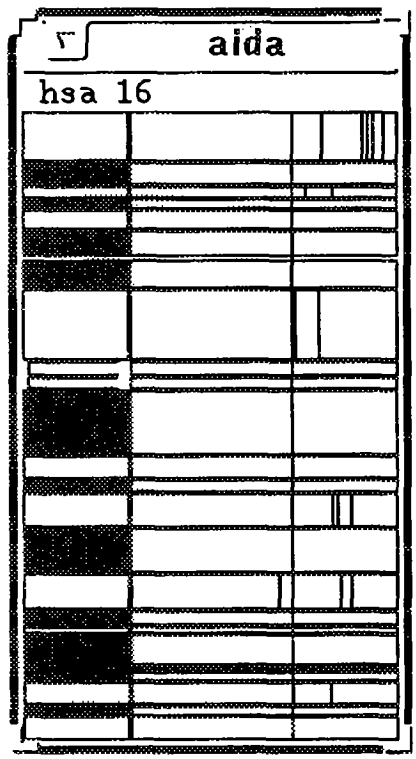

Figure 5 — Visualization of complex treatments integrating localization and base composition of genes. This figure shows a cytogenetic map of human chromosome 16. Vertical bars represent G+C contents in the third codon position of coding sequence for the genes

localized in each band. The main vertical bar stands for 60% G+C content (representation used by Ikemura & Aota (1991)).

Exploitation of the knowledge stored in MultiMap is provided by a set of integrated methods. Analyses of GenBank sequences can be achieved, with respect to their localization using access to ACNUC. A subset of GenBank has been choosen and integrated to the system to avoid problems in compositionnal studies due to redundance. Graphical editors allow the interrogation of maps represented, and the obtention of information on each elements included such as genes or cytogenetic bands. Different levels of cartography can be visualized simultaneously on the same screen. The use of map editors is possible to apply specific algorithms to each element of a map and to visualise the results on a graphical representation, without leaving the environment. For exemple the base Composition of human loci located on chromosome G bands can be directly obtained through GenBank access (Fig. 5). Also, graphical comparative mapping between man and mouse is currently under development.

Discussion

Strong relations exist between the model of a knowledge representation system and the aims assigned to the tools developed with it. Many systems using frames were previously developed in various fields of molecular biology and have proven useful. Two well known examples are MOLGEN (Friedland et al., 1982; Friedland & Kedes, 1985), which was built for helping experiment planning and representating genetic data, and GeneSys (Overton, Koile & Pastor, 1990), which was built for modelling eukaryotic genes expression regulation. Our developments have confirmed that object-oriented systems are powerful modelling languages that are able to represent complex biological systems that range from prokaryotes to eukaryotes. The possibility of using a slot which is an instance of another class is a natural way to describe complex ("embedded") biological organizations. This is clearly the case for the modelling of "maps of maps" structure in MultiMap. Moreover, description of a biological concept as an entity with slots eases its manipulation by non-specialists. In MultiMap cytogenetic bands have specific slots that allow the research of included objects by calling complex functions without any user intervention. Biological knowledge changes quickly, and a large part of knowledge in molecular biology is unsettled. In this context, the object model is powerful in the sense that it allows the construction of tools that are easy to be modified.

The query mechanism associated with SHIRKA is clearly less complex that the one provided by query languages associated with DBMS (like SQL). However, use of a graphical interface allows complex biological questions: under ColiGene, it is possible to make a query like "retrieve all protein genes that are highly expressed and which contain a weak translation initiation site" (Fig. 6). This kind of query is possible due to the integration of methodological knowledge in the bases. Integrating in the base the new knowledge that methods are able to produce allows round-trips between "formal" and "methodological" knowledge. An example of this situation is the following: it is possible to use the prediction methods linked to ColiGene to detect new protein genes in sequences, then it is possible to introduce these predicted genes in the base and to perform other analyses on them. These analyses leading to the creation of new objects in the base, such as the possible translation initiation sites associated to the gene.

A peculiarity of Le_Lisp language is that it allows to consider functions written in C or Fortran as Le_Lisp functions. So, some methods used in procedural attachment could be written in those languages instead of Le_Lisp. In the case of knowledge bases dedicated to molecular biology, the interest of such a possibility is evident: the use of any program from the numerous packages of sequence analysis is virtually possible.

This is an evidence that the genomic data bases will have, in the near future, to integrate more and more graphical interfaces. Firstly, they are necessary to simplify object manipulation: visualization of a map is the best way to have an immediate idea on relative position of genes. Then, they simplify the management of queries by scrolling menus. At last, they facilitate the use of methods included in the knowledge base by guiding the user with messages and by providing default options to avoid misuses.

Finally, some limitations to the SHIRKA model have been shown during the construction of ColiGene and MultiMap. This is why some developments are currently made in collaboration with the conceptors of the SHIRKA system in a way to improve its modelling capabilities. In summary, the main limitations we have encountered were: i) the difficulty to realize a whole knowledge base as a single hierarchy, due to the complexity of the biological objects represented. Biological structures are represented in ColiGene and MultiMap from only a functional perspective. Certain kinds of structures should however also be considered from other perspectives, such as an evolutionary perspective; ii) the impossibility to represent in an efficient way some kinds of knowledge such as dynamic knowledge, complex objects or textual informations; iii) the fact that all the knowledge and the tools for its representation and use have to be loaded in central memory, as it exists no manager for objects on disk. This limitation implies the use of computer systems with a least 16 Mbytes in main memory to have ColiGene and MultiMap work. However, despite these limitations, it is remarkable that SHIRKA was flexible enough to allow the representation of a wide variety of biological concepts that range from bacterial operons to chromosome and maps. It is also noteworthy that the

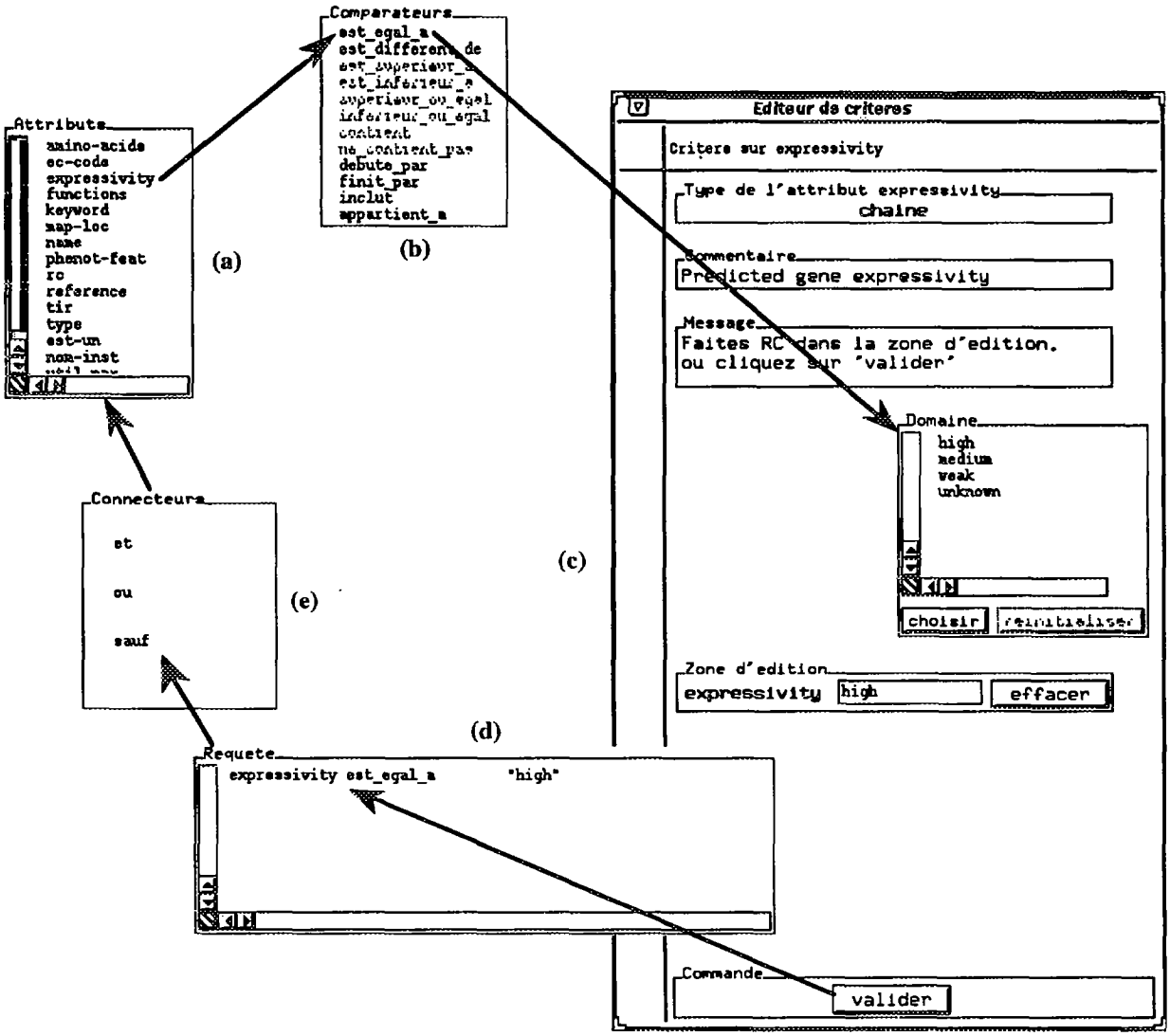

Figure 6 — Example of query under ColiGene allowing the retrieving of the protein genes highly expressed. In the first panel of the interface (**a**), the slots attached to the selected class are displayed. The user select the slot he wants to use to build his query (here **expressivity**). This action activate the comparator panel (**b**), in this panel only certain kinds of comparators are activated, depending on slot type. Selection of a comparator, in this case **est_egal_a** (which means "is equal to") activates a slot editor (**c**) for typing the values to retrieve. After validation, values are displayed in the query builder (**d**) and it is possible to activate the query or to extend it using a logical connector (**et** (= and), **ou** (= or) and **sauf** (= not)) (**e**). This query builder allows the access to complex objects, so it is possible to complete the preceding query by interrogating on the slots values for the TIR associated to the protein genes and so, to retrieve highly expressed genes with weak TIR.

robustness of this system has permitted the management of a huge number of objects, respectively 6,000 for ColiGene and 27,000 for MultiMap (classes and instances merged).

Conclusion

The use of an object-oriented modelling language like SHIRKA has allowed the building of two biological knowledge bases devoted to genome analysis. Publications of biological results obtained by using these bases has already validated the approach (Cortay et al., 1991; Mouchiroud et al., 1991). However, it must be noticed that such development needs large efforts particularly in the conception of the model and the development of the user interface. Presently this interface, allows complex data analysis in an environment relatively familiar to the biologist. The next step is to provide the user with the methodological knowledge of the domain in a simpler way. This will be achieved by using a task management system to model data analysis strategies. Such a system, fully compatible with SHIRKA already exists and is presently experimented in our laboratory in another project

(Chevenet, Jean-Marie & Willamowski, 1993). Linkage between biological knowledge modelling and methodological knowledge modelling is a very attractive continuation for the work presented here.

Personal accounts on a dedicated machine providing access to ColiGene (and soon MultiMap) could be set up upon request to the authors. SHIRKA code of ColiGene is also available at our anonymous FTP server: biomol.univ-lyon1.fr.

References

Bachmann, B.J. 1990. Linkage map of *Escherichia coli* K-12, Edition 8. *Microbiol. Rev.* 54: 130-197.

Bernardi, G. 1989. The isochore organisation of the human genome. *Annu Rev Genet.* 23: 637-661.

Blake, R.D. and Hinds, P.W. 1984. Analysis of the codon bias in *E. coli* sequences. *J. Biomol. Struct. Dynam.* 2: 593-606.

Burks, C., Cassidy M., Cinkosky M.J., Cumella K.E., Gilna P., Hayden J.E.-D., Keen G.M., Kelley T.A., Kely M., Kristofferson D. and Ryals J. 1991. GenBank. *Nucleic Acids Res.* 19: 2221-2225.

Chevenet F., Jean-Marie, F. and Willamowski, J. 1993. A development shell for cooperative problem-solving environments. Forthcoming.

Collado-Vides, J., Magasanik, B. & Gralla, J.D. 1991. Control site location and transcriptional regulation in *Escherichia coli*. *Microbiol. Rev.* 55: 371-394.

Collins, A., Keats, B.J., Dracopoli, N., Shields, D.C. and Morton, N.E. 1992. Integration of gene maps: chromosome 1. *Proc. Natl. Acad. Sci.* 89: 4598-4602.

Cortay, J.-C., Nègre, D., Galinier, A., Duclos, B., Perrière, G. and Cozzone, A.J. 1991. Regulation of the acetate operon in *Escherichia coli*: purification and functional characterization of the IclR repressor. *EMBO J.* 10: 675-679.

Dutrillaux, B. 1973. Nouveau système de marquage chromosomique: les bandes T. *Chromosoma* 41: 395-402 (in french).

EMG. 1990. Encyclopedia of the Mouse Genome version 1.0: reference manual. Bar Harbor, ME: the Jackson Laboratory.

Frézal, J. 1991. Genatlas: une banque de données sur la carte des gènes de l'homme. *Médecine/Sciences* 7: 595-601 (in french).

Friedland, P., Kedes, L., Brutlag, D., Iwasaki, Y. & Bach, R. 1982. GENESIS, a knowledge-based genetic engineering simulation system for representation of genetic data and experiment planning. *Nucleic Acids Res.* 10: 323-340.

Friedland, P. and Kedes, L. 1985. Discovering the secrets of DNA. *Comm. ACM* 28: 1164-1186.

Gautier, C., Gouy, M., Jacobzone, M. and Grantham, R. eds. 1981. Nucleic acid sequence handbook. London, UK: Praeger publishers.

Gautier, N. and Pavé, A. 1990. Object-centered representation for species systematics and identification in living systems in nature. *Comput. Applic. Biosci.* 6: 383-386.

Gouy, M. 1981. Etude des aspects dynamiques de la protéosynthèse chez *Escherichia coli*. Ph.D. thesis, Claude Bernard University, France (in french).

Gouy, M. and Gautier, C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* 10: 7055-7073.

Gouy, M., Gautier, C., Attimonelli, M., Lanave, C. and di Paola, G. 1985. ACNUC — a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. *Comput. Applic. Biosci.* 1: 167-172.

Grivaud, S. 1992. Navigation dans une base de connaissances à objets. Application aux bases de séquences génomiques. CNAM internal report, Grenoble, France (in french).

Ikemura, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* 146: 1-21.

Ikemura, T. and Aota, S. 1991. Global variation in C+G content along vertebrate genome DNA. Possible correlation with chromosome band structures. *Mol. Biol. Evol.* 2: 150-174.

ILOG. 1992a. Le_Lisp version 15.25: reference manual. Gentilly, France: ILOG S.A.

ILOG. 1992b. Aida version 1.65: reference manual. Gentilly, France: ILOG S.A.

Harnden, D. G. and Klinger, H.P. eds. 1985. Report of the Standing Committee on Human Cytogenetic Nomenclature. Manchester, UK and New York, New York: ISCN.

Kanehisa, M., Fickett, J.W. and Goad, W.B. 1984. A relational database system for the maintenance and verification of the Los Alamos sequence library. *Nucleic Acids Res.* 12: 149-158.

Kuhara, S., Matsuo, F., Futamura, S., Fujita, A., Shinohara, T., Takagi, T. and Sakaki, Y. 1984. GENAS: a database system for nucleic acid sequence analysis. *Nucleic Acids. Res.* 12: 89-99.

Médigue, C., Rouxel, T., Vigier, P., Hénaut, A. and Danchin, A. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.* 222: 851-856.

Minsky, M. 1975. A framework for representing knowledge. In The Psychology of Computer Vision, 211-277. Winston, P.H. ed., New York, New York: McGraw-Hill.

Mouchiroud, D. and Gautier, C. 1990. Codon usage changes and sequence dissimilarity between human and rat. *J. Mol. Evol.* 31: 81-91.

Mouchiroud, D., D'onofrio, G., Aissani, B., Macaya, G., Gautier, C. and Bernardi, G. 1991. The distribution of genes in the human genome. *Gene* 100: 181-187.

Overton, G.C., Koile, K. and Pastor, J.A. 1990. GeneSys: a knowledge management system for molecular biology. In Computers and DNA, 213-239. Bell, G.I. and Marr, T. eds., Reading, Massachusetts: Addison-Wesley.

Pearson, P.L., Matheson, N.W., Flescher, D.C. and Robbins, R.J. 1992. The GDB™ human genome data base anno 1992. *Nucleic Acids Res.* 20: 2201-2206.

Rechenmann, F. and Uvietta, P. 1991. SHIRKA: an object-centered knowledge based management system. In Artificial Intelligence in Numerical and Symbolic Simulation, 9-23. Pavé, A. and Vansteenkiste, G.C. eds., Lyon, France: Aléas.

Rousseau, B., Pavé, A., Rechenmann, F. and Landau, M. 1986. Edora project: Artificial Intelligence approach and workstation concept to aid dynamic modelling in biology and ecology. In Supplement of the Proceedings of the Summer Computer Conference, 14-20. Reno, Nevada: SCS.

Rudd, K.E., Miller, W., Werner, C., Ostell, J., Tolstoshev, C. and Satterfield, S.G. 1991. Mapping sequenced *E. coli* genes by computer: software, strategies and examples. *Nucleic Acids Res.* 19: 637-647.

Sawyer, J.R. 1991. Highly conserved Segments in Mammalian Chromosomes. *J. Hered.* 82: 128-133.