Prediction of primate splice junction gene sequences with a cooperative knowledge acquisition system

Engelbert MEPHU NGUIFO & Jean SALLANTIN
Laboratoire d'Informatique, de Robotique et de Micro-electronique de Montpellier
161 rue Ada, 34392 Montpellier Cedex 5 - France.
E.mail: mephu@lirmm.fr Fax: (33) 67 14 85 00 Tel: (33) 67 14 85 83/82

Abstract

We propose a cooperative conceptual modelling environment in which two agents interact : the machine and the human expert. The former is able to extract knowledge from data using a symbolic-numeric machine learning system, and the latter is able to control the learning process by accepting and validating the machine results, or by criticizing those results or the explanation that the system produces on them. The improvment of the conceptual modelling relies on the cooperation between the two agents.

Results obtained with our method on prediction of primate splice junctions sites in genetic sequences are far better than thoses reported in the literature with other symbolic machine learning systems, and are as better as thoses obtained with some artificial neural networks methods reported at present. But in opposite to neural networks which lack of argumentation, our system provides the user a plausible explanation of its prediction.

Introduction

Can one believe, without doubt, knowledge produced by a machine ? Certainly not, therefore our work is on the boundary between Machine Learning and Knowledge Acquisition. This paper especially deals with concept acquisition from similarity-based learning applied to the prediction of splice junction sites from genetic sequences. Our basic task is to build a concept as an abstraction from data, and to fit it to new data through an interaction with the expert, in a control step.

Our knowledge acquisition method LEGAL focusses only on one concept with imperfect domain theory, and the system builds a formulation and a set of objections defined in section 2. Then the expert is able to examine how the concept is efficient on data, and how it is possible to force an evolution of the concept formation or to suppress its validity. LEGAL is described in a data driven way and is based on the different levels defined by Russell (1956) to build an abstraction of data. We want to establish a cooperative control between the expert and the system. We show how effective our principle is in a cooperative concept modelling, where the algorithm structure kernel is the concept lattice or Galois lattice (Birkhoff, 1967), (Wille, 1992) - and the real-world problem is genetic sequences analysis.

In such domain, an active research area in the hierarchical approach to the protein folding problem is the prediction of primate splice junction sites from the DeoxyriboNucleic Acid (DNA) sequence. It is known that DNA sequence carries genetic information which dictates the production of proteins. Splice junction sites are points on a DNA sequence at which 'superflous' DNA is removed during the process of protein creation in higher organisms. Established approaches to this problem have involved handcrafted rules by experts, and statistical methods (Fichant, Quinqueton & Gautier, 1988). More recently, a variety of machine learning methods (Lapedes & al. 1990), (Noordewier, Towell & Shavlik, 1991), (Brunack, Engelbrecht & Knudsen, 1991) have been applied: both neural networks, and symbolic induction. The results of these methods are often more effective and accurate than their human-designed counterparts.

A comparison between some of both methods has been made by Towell, Noordewier and Shavlik. We have focused our interest on the same dataset in order to compare our results to theirs. It appears that our results are better than those of other symbolic methods, and nearly identical to those of neural networks methods. Moreover our main advantage over neural networks is our abitility to provide an understandable explanation to the expert, so that he is able to interact with the system to improve conceptual modelling.

The paper is organized as follows. Section 2 is an overview of our cooperative environment. We describe the materials used in our experiment in section 3. Section 4 shows our results and discusses its strengths.

Conceptual Modelling

We describe here the foundations of our method. Concept is expressed in terms of sets of examples. Knowledge acquisition consists in defining a language description, finding an abstraction, and validating it on data. The process of concept abstraction is progressive and requires different levels. Figure 1 shows an extended Russell definition of nine levels. The purpose of Russell was to extract the concept of number from ZF set theory. For this task, he defines the 8 first points, his point 9 was specific to the number theory, we point out his approach as a theoretical framework that outlines the concept formation in a general case. Our point 9 is due to Lakatos definition of control by proof and refutation.. Figure 2 illustrates the process of cooperative revision through a little concept example: "To be a good General", and figure 3 gives the formalism of our implementation.

#	Abstraction levels	Conceptual labels
1	determine a set of examples	Examples
2	describe the examples	Fact
3	select a relation of equivalence between the examples	Regularity
4	build the class of equivalence for the examples	
5	label each class of equivalence by a statement in the language	Hypothesis
6	determine a membership function for each class of equivalence	
7	formulate the concept as a relation which links together the labels of the classes	Concept formulation
8.a	determine a protocol which decides if an object can be or not a concept example	Empirical proof
8.b		Analogical proof
9.a	give an argumentation, to explain the success or failure in decision	General objections
9.b		Contextual objections

Figure 1 : Extended Russell's data abstraction

Reading downwards are the different levels of concept abstraction which is an extension of Russell definition. Column 3 indicates conceptual formalisms that can be used according to the levels. In level 8, one can choose (8.a) or (8.b). Level (9.a) is linked to level (8.a), while level (9.b) is linked to level (8.b).

The first level is related to the choice by the expert of the *examples* or counter-examples characterizing the concept. At the second level, we define the language used to describe objects. Objects are described by means of *facts*. Going through some examples denoting the concept "To be a good general", as Napoleon and so on, one's first task is to search for a **description language** of examples. Here we use some binary attributes[1] for example: small, fat, white ...

Due to the incompleteness of initial data, it is hard to find a single characterization or **abstraction** which corresponds to the concept formulation of Russell example (see figure 2). So one may agree to take a "good" one which would be possible to refine. To this end, levels 3 and 4 are grouped to define *regularity*, which is a feature that is retrieved among objects descriptions. Regularity is a relation between facts. Using the previous descriptors, a regularity can be a conjunction of attributes which often holds for a general, like to be small and fat and white - Napoleon was, but many others weren't. Levels 5 and 6 are related to *hypothesis*, i.e. a combination of a regularity and a subgroup of objects, such that those objects verify the regularity according to some constraints. *Concept formulation* in level 7 is related to an organization of hypothesis.

Concept formulation gives rise to a **cooperative revision** phase, where it is necessary to understand and control the decision, as shown in figure 2. For example, consider the assertion : "if someone is not white then he cannot be a good general". It is true until the expert refutes it as an explanation of the system decision. This allows to refine concept formulation by removing or modifying this knowledge. Levels 8 and 9 express decision and argumentation using concept formulation. A decision based only on learned regularities is an *empirical decision*, and *analogical decision* is based on the set of examples. It is well known that inductive learning methods are best suited to tasks where a considerable amount of data is available and knowledge about the domain is scarce. These methods do not work effectively when there is insufficient data since they rely on finding patterns among data. The use of analogical reasoning is a way of reducing this shortcoming (Mephu Nguifo & Sallantin, 1993).

These decisions initiate an interaction between the user and the system, so that the system becomes able to provide a *plausible explanation* to the user, who in turn can validate or refute it. These explanations allow knowledge acquisition since they builds some abstraction characterizing the learned concept, and also allow the expert to change some data and return to a previous level, according to the conjunctions or disjunctions of attributes that influence the decision.

The control is partially based on the notion of objection. In fact, explaining with objections is more understandable for an expert than using regularities. An objection is what is sufficient on an object to refute it as an example. In our method, *general objection* is linked to the concept formulation, while *contextual objection* is related to an example. In Russell's example, a general objection is : "a general who is not white can never be a good general". The system builds this assertion since it appears in many built regularities as the initial examples are white people - Napoleon, Wellington. The user can refute this argumentation by removing this fact in the language description, or by introducing in the data set some examples which do not verify it.

[1] The goal of Russell was to find how to express such concept with terms that describe human qualities of a general. For logical reason, he assumed that "to be a good general" could be express by a disjunction of conjunction of terms.

Phases	"To be a good general"	Ctual. labels		User Interaction
Examples	Napoleon, Wellington, ...	Examples		
Language	small, fat, clever, ..., white, age	Fact		
Abstraction	(small ∧ fat ∧ ...), (clever ∧ ...),	Regularity		
	(small ∧ white ..., {Napoleon, ...})	Hypothesis		
	?	Concept formulation		
Cooperative Revision	X is ~ because (small, fat, ...)	Empirical proof	⇒	return to level 6 e.g. change membership function
	X is ~ as Napoleon	Analogical proof	⇒	return to level 7 e.g: modify relation that links classes
	(age ≥ 70), or (¬white)..., ⇒ never ~ Soundiata Keita is not ~, due to (¬white)	General objections	⇒	return to level 2 e.g. : suppress fact (white)
	Liken to Napoleon, X is not ~, due to not (small and fat and ...)	Contextual objections	⇒	return to level 1 e.g.: modify objects

Figure 2 : Cooperative revision.

Abbreviation : (i) ~ : a good general (ii) ? : unknown (iii) Ctual. : conceptual

The cooperation between the expert and the system allow to return back to a previous levels if the expert critics the system justification. This is materialized par the symbol ⇒ due to a refutation of the expert.

For example, Soundiata Keita is not a good general because he verifies none of the regularities. In addition, he is objected because he is not white. The expert critics this objection because it is not a sufficient reason to be refuted as a good general. Giving this objection, he notices for example that all the initial examples are white people (acquired knowledge). So he can suppress the fact (white) in the description language, or add some good generals who are not white in the initial sets of examples, and in both cases he may rebuild concept formulation (improvment of conceptual modelling). He can also choice only to suppress this objection.

The purpose of explanation is to provide some abstraction of the learned concept through an interaction with the user. The expert can also use this process to compare two objects. The questions might be as follows:

i) why the object o_x is justified, refuted or ambigous?

ii) why o_x is similar or not to o_y?

Explanation principle

The system searches for a plausible argument to justify its decision, e.g an object is refuted or ambigous when its description does not contain some attributes which may be sufficient to recognize it, if they are added to its description. The explanation differs whether the object is justified or not:

a- If the object is justified, then the system uses its most similar training example, to extract their common attributes. The system searches the minimal conjunction of those attributes that allows to justify the object.

b- If the object is refuted or ambigous, the system can exhibit a general objection or builds a contextual objection from its nearest training example.

This process answers to question (i). It also works when using training counter-examples, to give negative answers to the user. For example, an object can be refuted when it verifies 'few' regularities, and is more similar to training counter-examples, or it can be justified when it verifies 'enough' regularities and is less similar to training counter-examples. The expert can use this process to have an answer to questions like (ii). The system may memorize or learn the explanation as a good one, and in the other case, it may modify its acquired knowledge in such a way that it will not provide the same explanation in the future. The user can also revise the description of the object, by adding or removing some attributes.

Remark 1: If an object is not objected, then it becomes a potential example in an empirical reasoning. General objections are a means to define the plausible definition domain of regularities.

Remark 2 : LEGAL is based on Galois lattice. The *main advantage of Galois Lattice*[2] is its exhaustiveness although this has the disadvantage of considerably slowing down the system when dealing with a large amount of data. To avoid rote learning, we use some heuristics to build only a join-semi lattice of regularities, as it is shown in figure 4.

Notation: Downwards, we use **regularity** to designate a valid and pseudo-coherent regularity.

[2] Galois Lattice is a mathematical framework, and has proved to be the best support of our learning hypothesis (Mephu & Sallantin, 1993). Referring to the notion of version spaces defined by Mitchell (1982), Galois Lattice can be viewed as the largest exploratory space of regularities, due to its exhaustiveness.

Conceptual labels	**Algorithm**
Examples	$o_1, o_2, o_3, ..., o_m$
Fact	binary attributes describing examples
Regularity	$\wedge_j$ $fact_j$, $1 \leq j \leq n$ **Relevant attributes** are those which appears at least in one built regularity
Hypothesis	(regularity, set of examples) **Galois connection** : an hypothesis (A,O) is characterize as follows : (i) A is the subset of all attributes that holds for all objects of O, (ii) and O is the subset of all objects that verifies all attributes in A. **Selection criterion** : • A regularity is **valid** if it holds for "enough" examples • It is **pseudo-coherent** if it holds for "few" counter-examples
Concept formulation	Join Semi Lattice of hypothesis Hypothesis are **ordered** by the subhypothesis-superhypothesis relation ≤, i.e. an hypothesis (A_1,O_1) is a subhypothesis of (A_2,O_2) iff $O_1 \subsetneq O_2$ (i.e $A_1 \supsetneq A_2$)
Empirical proof	% {regularity} Principle of **majority vote** onto the set of valid and pseudo-coherent regularities • An object o_i is an example if it verifies enough regularities. o_i is justified. • o_i is *not an example* if it verifies 'few' regularities. o_i is refuted. • Otherwise, o_i is *ambiguous*.
Analogical proof	The set of {regularity} verified by X & an example Y, are nearly the same Principle of decision confirmation : • The *object* o_x *is an example* if: - it verifies 'enough' regularities and is similar to 'enough' examples; - or it verifies 'few' regularities and is similar to a particular example*. • It is*not* if it verifies 'few' regularities and isn't similar to a particular example; • Otherwise, this object is *ambiguous*. ***A particular example** is an initial example which is not similar to initial others.
General objections	{ $\neg(\vee\ fact_k)$ } such that $\neg(\vee\ fact_k) \Rightarrow$ not example A built general objection is the *negation of a disjunction* of some more relevant attributes, such that if it holds for an object then this object cannot verify 'enough' regularities, and is empirically refuted. An object **is objected** if it verifies an objection.
Contextual objections	$\neg(\wedge\ fact_k)$ with X $\Rightarrow$ not as the example Y Considering an initial example Y, a built contextual objection on an object X is the *negation of a conjunction* of some more relevant attributes, such that if it holds for X, then X cannot be similar to Y.

Figure 3 : LEGAL formalism to abstract data.

Column 2 gives the specifications of our symbolic empirical single inductive system LEGAL [Liquiere & Mephu, 1990]. These specifications were chosen in order to allow a coopeartive control. A critic of decision or argumentation of the system (empirical or analogical proof - contextual or general objections) by the expert give rise to a revision of the knowledge.

In practice, the terms "enough" and "few" are defined through some thresholds chosen by the expert or determined by the system, and respectively called the *validation* and the *pseudo-coherence thresholds* when building regularities, the *justification* and the *refutation threshold* for empirical decision, and the *similarity* threshold for analogical decision.

Remark 3 : Methods of machine learning are strongly biased toward symbolic data representations. In some problems, not all the data are available before some decisions must be made. Regularities are strongly dependent on the learning context and the principle of majority vote in decisions can be inadequate due to the dependence between regularities (Mephu & al. 1991). Consequently several errors may of necessity appear in decisions (see figure 5). To avoid this shortcoming, we have implemented an analogical decision method based on numerical taxonomy for its analytic capability, to control LEGAL decisions. We define a context-sensitive similarity between an object and the training objects, ***through the way they interact onto the set of regularities.*** A similarity measure is termed ***context-free***, if the similarity between o_1 and o_2 is independent of A_1's and A_2's rela-tionship to other objects being clustered. *Context-sensitive* measures of similarity have also been developed in which the similarity of two objects is dependent on their relation to additional objects. For example, if we assume that integers are "objects", then using a context-sensitive similarity mea-sure, the integers 1 and 9 would be considered more similar when considered within the range 1 to 100 than when consi-dered within the range 1 to 10 (Fisher & Langley, 1986).

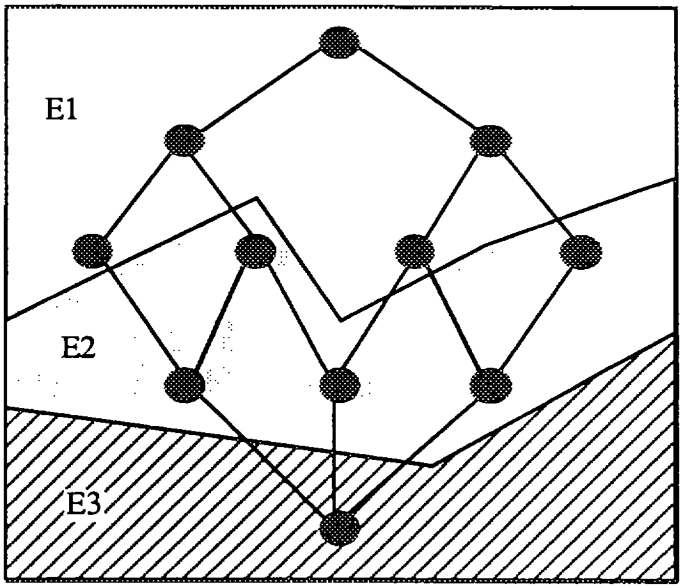

Figure 4 : Regularities Space with LEGAL. Despite the heuristics, the Galois connection is not altered in valid regularities. E1 ∪ E2 = space examined by LEGAL; E3 = space unexplored; E1 = join-semi lattice built.

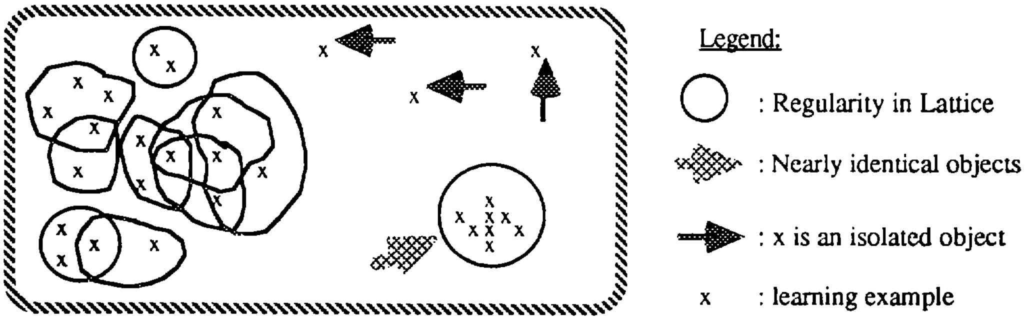

Figure 5 : Examples Space with LEGAL.
A like-circle corresponds to a regularity in LEGAL. Objects o_x in a circle verify the corresponding regularity.
LEGAL is able to point out some particular examples, and this provides more information to the expert. In fact, this may be an outcome of the incompleteness of the description language, of the set of examples, of the concept formulation, or of the knowledge structure bias. It is thus necessary to control learning and decision steps.

Prediction of splice junction sites

We describe here our application. Biology is more and more used in AI as an application domain. We haved chosen it for two reasons : the first relies on the difficulty to clusterize sequences into exclusives classes, and the second is the importance of argumentation for biologists. We focus our interest on the prediction of primate donor and acceptor sites from the DeoxyriboNucleic Acid (DNA) sequence. Let us give some basic definitions which are based on that of Li (1990) and Chan (1992).

Basic definitions

A *gene* is coded by its nucleotidic sequence (DNA sequence), or its RNA strands (RiboNucleic Acid) or its protein sequence (the very building blocks of life). These sequences are divided into groups: amino-acid and nucleotide sequences. We are concerned by the latter group.

Nucleotide sequence....RNA and DNA sequences are nucleotide sequences. The basic building blocks of human genetics are nucleotides. There are four different kinds of nucleotides in DNA: adenine (A), cytosine (C), guanine (G), and thymine (T). A DNA segment can be represented as a sequence of symbols, where each symbol denotes one of the four nucleotides. For example, GAGAGCTT may be a segment of eight nucleotides of a DNA sequence.

Intron and Exon. In general there are some segments in DNA sequencxes which do not encode protein information. These segments are called ***introns***, and are sliced off before *translation*, the process of decoding information on a RNA to generate proteins. Before translation begins, the regions that encode protein information, *exons*, are spliced together after the introns are removed.

Donor and Acceptor sites. All known splice junctions are divided into *acceptor* sites (the boundary between an intron and an exon : IE or 3' sites) and ***donor*** sites (between an exon and an intron : EI or 5' sites).

Data sets

Data are provided from the GenBank database 64.1, which contains the already annotated primate sequences. The problem is to recognize the boundaries between exons and introns in DNA sequences. There is two subtasks: recognizing EI sites, and recognizing IE sites.

Cl.	**instance name**	**sequence segment**
EI,	ATRINS-DONOR-521,	CCAGCTGCATCACAGGAGGCCAGCGAGCAG**GT**CTGTTCCAAGGGCCTTCGAGCCAGTCTG
EI,	ATRINS-DONOR-905,	AGACCCGCCGGGAGGCGGAGGACCTGCAGG**GT**GAGCCCCACCGCCCCTCCGTGCCCCCGC
EI,	BABAPOE-DONOR-30,	GAGGTGAAGGACGTCCTTCCCCAGGAGCCG**GT**GAGAAGCGCAGTCGGGGGCACGGGGATG
...		
EI,	TARHBD-DONOR-468,	GGAAGATGTTGGTGGTGAGGCCCTGGGCAG**GT**CAGTATCATGGCTATGAGGCAGGCTTAA
EI,	TARHBD-DONOR-817,	AAGCTGCATGTGGATCCTGAGAACTTCAGG**GT**AAGTCTAGGAGATGTTCTTCTTTGTCCT
IE,	ATRINS-ACCEPTOR-701,	TTCAGCGGCCTCAGCCTGCCTGTCTCCC**AG**GTCTCTGTCCTTCCACCATGGCCCTGTGGA
IE,	ATRINS-ACCEPTOR-1678,	GGACCTGCTCTGCGTGGCTCGCCCTGGC**AG**TGGGGCAGGTGGAGCTGGGTGGGGGCTCTA
IE,	BABAPOE-ACCEPTOR-801,	GCGGTTGATTGACAGTTTCTCCTTCCCC**AG**ACTGGCCAATCACAGGCAGGAAGATGAAGG
...		
IE,	TARHBD-ACCEPTOR-594,	CTGTCCTGTGGGTTCCTCTCACCCTCTC**AG**GTTGCTGGTCGTCTACCCATGGACCCAGAG
IE,	TARHBD-ACCEPTOR-1884,	CATATGTATCTTTTTACCTTTTCCCAAC**AG**CTCCTGGGCAACGTGCTGGTGTGTGTGCTG
N,	AGMKPNRSB-NEG-1,	CAAAAGAACAAAGCTGGAGGCATCACGCTACCTGACTTCAAACTATACTACAAGGCTACA
N,	AGMORS12A-NEG-181,	AGGGAGGTGTCTGATTGGTCCAGCTTAGTCCATGTCCCTACCCTGAACAGGGGCATGGGG
...		
N,	HUMARMA-NEG-961,	CACTGAGTTGATTTTAGCAGAGAAACGTGGTGACCTGACAAGAGAGAATGTGAACCAGTG
N,	HUMZNF8-NEG-661,	CAGGACAAACCCTACAAATGTACTGACTGTGGGAAGTCGTTTAACCATAACGCACACCTC
N,	LEMHBDPS-NEG-1441,	TTCACCCCACAGGTGCAGGCTGCCTATC**AG**AAGGTGGTGGCTGGTGTGGCTAATGCCCTG
...		
N,	ORARGIT-NEG-241,	TCTCGGGGGCGGCCGGCGCGGCGGGGAGCG**GT**CCCCGGCCGCGGCCCCGACGTGTGTGTC
N,	TARHBB-NEG-541,	ATTCTACTTAGTAAACATAATTTCTTGTGCTAGATAACCAAATTAAGAAAACCAAAACAA
N,	TARHBD-NEG-1981,	AGGCTGCCTATCAGAAGGTGGTGGCTGGTGTGGCTGCTGCTCTGGCTCACAAGTACCATT
		False EI sites
N,		GGCCCCCACCTGGTGGAAGCCCTCTACCTG**GT**GTGCGGGGAGCGAGGTTTCTTCTACGCA
N,		GTTCTAATCATTTCACCATTTTTGTTATTC**GT**TTTAAAACATCTATCTGGAGGCAGGACA
N,		...
N,		ATCAATAAGCTCCTAGTCCAGACGCCATGG**GT**CATTTCACAGAGGAGGACAAGGCTACTA
N,	ORARGIT-NEG-241,	TCTCGGGGGCGGCCGGCGCGGCGGGGAGCG**GT**CCCCGGCCGCGGCCCCGACGTGTGTGTC
		False IE sites
N,		CCAACAGCACCAATATCTTCTTCTCCCC**AG**TGAGCATCGCTACAGCCTTTGCAATGCTCT
N,		GACCTGCAGCACCGGCTGGACGAGGCCG**AG**CAGATCGCCCTCAAGGGCGGCAAGAAGCAG
N,		...
N,		GTTCGTGGGGGCCACTTTGCGGCCTTTG**AG**GAGCCGGAGCTGCTCGCCCAGGACATCCGC
N,	LEMHBDPS-NEG-1441,	TTCACCCCACAGGTGCAGGCTGCCTATC**AG**AAGGTGGTGGCTGGTGTGGCTAATGCCCTG

Figure 7 : Segments nucleotides of true and false donor sites.

"Cl." indicates the class (one of n, ei or ie) of the instance. The data set examples are taken from GenBank 64.1. These data as well as results obtained by Towell, Noordewier and Shavlik are available in the ics.uci.edu UNIX machine. They can be copied by network using an ftp command, and 'anonymous' as login. Data are in the directory "Pub/Machine-Learning-databases/Molecular-Biology".

The dataset[3] contains 767 EI sites (25%), 768 IE sites (25%) and 1655 segments which were neither EI nor IE sites. The examples of the latter subset is referred to as N sites. All the examples of these 3 subsets are sequences segments extracted symmetrically around donor and acceptor sites, starting at position -30 and ending at position +30. Each sequence segments contains 60 nucleotides. Other characters D, N, S, and R indicate ambiguity among the standard characters. They respectively corresponds to (A or G or T), (A or G or C or T), (C or G), (A or G).

Remark 4 : It appears clearly in this dataset that EI and IE examples respectively often have the apparent consensus "GT" at positions 31-32 and "AG" at positions 29-30 (see table 4). Over the 1655 examples of the N subset, there are 8 (.48%) segments with the two consensus, 119 (7.19%) segments with only "AG", 90 (5.44%) with only "GT", and 1438 (86.89%) with no consensus. So using this data set can alter the error rates because there are too many N examples which would necessarily be well predicted due to the absence of the consensus. All these N examples would have been easily well recognized by our method as false sites (F.EI or F.IE). To avoid this shortcoming, we have extracted 4 data sets of IE (384+384), false IE (60+59), EI (384+383) and false EI (45+45) sites, where each site has the

[3] This dataset has been developed by Noordewier & al (1991) to help evaluate a "hybrid" learning algorithm (KBANN) that uses examples to inductively refine preexisting knowledge. Using a "ten-fold cross-validation" methodology on 1000 examples randomly selected from the complete set of 3190, the error rates produced by various ML algorithms , are shown in figure 7.

corresponding consensus. Each of the 4 sets was divided into 2 parts, part I used for training of machine learning algorithms included the first half examples as indicates in the previous brackets, and part II used for testing the remaining half examples.

Remark 5: A code is often defined to translate nucleotique segments in such a way that it may analyze in propositional language In our application, we replace each nucleotide by a conjunction of some properties. For example, the following small segment of 3 nucleotides "AGC" can be described by "101010 010110 011001".

Remark 6: The alphabetical order was maintained when dividing the dataset into training and test in order to avoid strong similarities between sequences.

	IE	N		EI
KBANN	08.47	04.62		07.56
BACKPROP	10.75	05.29		05.74
PEBLS	07.55	06.86		08.18
PERCEPTRON	17.41	03.99		16.32
ID3	13.99	08.84		10.58
COBWEB	09.46	11.80		15.04
Nearest Neighbour	09.09	31.11		11.65
	IE	**F.IE**	**F.EI**	**EI**
LEGAL (Empirical)	09.63	10.12	04.44	04.96

Figure 7 : Error rates obtained with some machine learning and Legal.

The experiments of other learning systems run at the university of Wisconsin, sometimes with local implementations of published algorithms (Noordewier, Towell, & Shavlik, 1991).

Results and Discussion

Figure 7 gives some test results, where symbolic methods such as ID3, COBWEB are less better than neural networks approach such as KBANN, BACKPROP, PERCEPTRON.

The dataset used by these methods are different to ours. However a comparison can be made, and it appears that our empirical prediction results are better than thoses of symbolic methods, and as better as thoses of neural networks approach.

The results of others methods are obtained by a tenfold cross-validation" methodology on 1000 examples randomly selected from the complete set of 3190 examples. This random choice can considerably decrease the error rates of false sites predictions since it appears clearly (see remark 4) that there are nearly 86% of false examples which do not have any of the consensus "GT" or "AG", while all the true examples do.

Because we consider only prediction on false sites with one of the consensus, our methodology can be thus more valid and efficient than the other.

In addition, as shown in figure 8, LEGAL provides some argumentation of its decision if it is needed by the expert, and.is more expressive than neural networks for a biologist. This figure illustrates an example of the interaction that can be established with the user in order to explain the results. The example shown by figure 8 concern the donor model. The same process can be done for the acceptor group.

Remark 7: We notice that all the positive examples of the test set which are not recognized by our system, were similar to initial counter-examples. We have not been able to evaluate the results of our analogical reasoning because we had no interaction with an expert-biologist on these data.

Conclusion

The goal of this paper is to describe and to advocate how important the interaction of a machine learning system and a human-expert is in a knowledge acquisition process. We have explained how, by achieving successively 9 levels, we provide a concept formulation which is easy to test and possibly to refine than in the classical and neural networks methods. We think that the best methods will be those where the expert-user gains confidence. Our method is based on the fact that it is important to help the expert shed light on data that influence a decision, and allow to refine knowledge. So LEGAL can be of great help to keep a knowledge base updated.

We have also shown how to control of learned regularities for knowledge acquisition by an analogical method. We believe that this is an efficient palliative to the uncertainty of a decision based only on learned regularity, and a preliminary for an extraction of a plausible explanation. Combinations of strengths can overcome difficulties in domains that are simultaneously incomplete, noisy, and biased, as it has been previously shown by (Rendell, 1989).

However, the major current limitation of the implemented method is that it remains (even so) a simple method of control with no available automatic feedback to the initial data and/or learned (background) knowledge, as it may be the case with a control method.

Contextual objections have to be validated by a biologist for our application. We have not currently validated this second level of control in our application. We are attempting to address this validation in ongoing research with the laboratory of the French Association of Myopathy. Moreover Brunak, Engelbrecht & Knudsen (1991) obtained also best results with a neuronal network approach, using other data sets from human DNA sequence. Such data would also be well suited for LEGAL.

Availability. The LEGAL program is available on request, and free to academic users. For instructions, send an e-mail or post-mail using the above address.

Acknowledgements. Financial support for this work has been made available by the French Association of Myopathy (A.F.M.). The authors thank Jocelyn Nanard, Marc Nanard, Joel Quinqueton, and the anonymous referees for useful comments and suggestions on earlier drafts of this paper.

Conceptual labels	**A good donor site**
767 Examples of 60 nucleotides where 384 for training & 383 for test	CCAGCTGCATCACAGGAGGCCAGCGAGCAG**GT**CTGTTCCAAGGGCCTTCGAGCCAGTCTG AGACCCGCCGGGAGGCGGAGGACCTGCAGG**GT**GAGCCCCACCGCCCCTCCGTGCCCCCGC .... AAGCTGCATGTGGATCCTGAGAACTTCAGG**GT**AAGTCTAGGAGATGTTCTTCTTTGTCCT
90 Counter-Examples where 45 for training & 45 for test	GGCCCCCACCTGGTGGAAGCCCTCTACCTG**GT**GTGCGGGGAGCGAGGTTTCTTCTACGCA .... TCTCGGGGGCGGCCGGCGCGGCGGGGAGCG**GT**CCCCGGCCGCGGCCCCGACGTGTGTGTC
Fact	A, G, C, or T or combination of nucleotides like D, R, S, N at position i, 1 ≤ i ≤ 18

Regularity
Thresholds :
Validity = 200
Pseudo-coherence = 3
	$.............................{}^{AA}_{GG}.\mathbf{GT}.{}^{A}_{C}G.......................$
	...
	$.............................{}^{C}_{G}..\mathbf{GT}^{AAC}_{GCG}.......................$

Hypothesis
Thresholds :
Justification = 15%
Refutation = 15%
Similarity = 37
	$(.............................{}^{AA}_{GG}.\mathbf{GT}.{}^{A}_{C}G......................., \{o_i\})$
	...
	$(.............................{}^{C}_{G}..\mathbf{GT}^{AAC}_{GCG}......................., \{o_j\})$

Concept formulation
 	
Training objects predicted as good sites:
	Lattice of 2282 regularities where 458 are valid and peudo-coherent
	`                 Empirical  Analogical`
	`examples           96.87 %     97.92 %`
	`counter-exples     04.44 %     03.11 %`

Empirical proof
 	
Good sites in empirical test :
`examples        95.04 %`
`counter-exples  04.44 %`
	≥ 15% {regularity} ⇒ example and < 15% {regularity} ⇒ not example
	o_x = "AGCCAGGGCACTCACCAGGCTGCAAGAA$\underline{\text{CA}}$**GT**GC$\underline{\text{T}}$GGGGTAAGAGGGGAGCGGGGGATCC" is not an example as it verifies 4% of regularities. The expert can critic this decision, and change the justification threshold to less than 5%. He can continue for more information

Analogical proof
 	
Good sites in analogical test:
`examples        95.30 %`
`counter-exples  04.44 %`
	Regularities holds by X and an example Y, are nearly the same
	o_x isn't similar to a particular example. It is similar to the initial counter-example :
	"ACTCTGTATTTTGGCCTGAAACCCATAGTG**GT**GCTGCATGGATATGAAGCAGTGAAGGAA".
	Its nearest good initial example is o_y which verifies 16% of regularities
	o_y = "GTGGGCAAGTGCCGAAGCGCAGGCATC$\underline{\text{AAGGT}}$A$\underline{\text{CTGG}}$CCTCCCATCCTCCCCTCCATTCT"

General objections o_x is objected by the last one. If the expert refuted it then he may modify the description language by returning to level 2.	$\neg(...........................G...........................) \Rightarrow$ not exple $\neg(...........................{}^{C}_{T}...........................) \Rightarrow$ not exple ... $\neg(..........................{}^{C}_{G}....{}^{C}_{G}......................) \Rightarrow$ not exple
Contextual objections	o_x is not similar to example o_y because of $\neg(...........................{}^{A}_{G}G...........................)$ Replacing 'CA' by '${}^{A}_{G}G$' in the first object o_x gives two objects o_a and o_b : o_a = "AGCCAGGGCACTCACCAGGCTGCAAGA$\underline{\text{AAGGT}}$G$\underline{\text{CTGG}}$GGTAAGAGGGGAGCGGGGGATCC" which verifies 15% of regularities and is similar to o_y due to underlined nucleotides.

Figure 8 : Data abstraction for donor splice junction sites.

An object o_i verifies the regularity "$.....{}^{AA}_{GG}.GT.{}^{A}_{C}G.....$" if it has (A or G) at position 6 and 7, G and T at respective positions 9 and 10, (A or C) at position 12, and G at position 13.

References.

Birkhoff, G. 1967. *Lattice Theory*. 3rd edition. American Mathematic Society Ed. Providence, RI.

Brunak, S., Engelbrecht, J., and Knudsen S. 1991. Prediction of Human mRNA Donor and Acceptor Sites from the DNA Sequence. *Journal of Molecular Biology* 220: 49-65.

Chan, P. K. 1991. Machine Learning in Molecular Biology Sequence Analysis, Technical Report, CUCS-041-91. Dept. of Computer Science, Columbia Univ., New-York.

Fichant, G., Quinqueton, J., and Gautier C. 1988. Analyse statistique des séquences génomiques. *Biométrie et Données discrètes* 7. Grenoble, 31 Mai. ENSAR.

Ficket, J. 1982. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research* 10: 5303-5318.

Fisher, D., and Langley, P. 1986. Conceptual Clustering and Its Relation to Numerical Taxonomy. In *Artificial Intelligence and Statistics*, Gale, W. A. Ed. Addison-Wesley.

Green, M. R. 1986. Pre-mRNA splicing. *Annual Review Genetics* 20: 671-708.

Lakatos, I. 1984. *Preuves et Refutations*. Hermann Ed.

Lapedes, A., Barnes, C., Burks, C., Farber, R., and Sirotkin, K. 1990. Application of neural networks and other machine learning algorithms to DNA sequence analysis. In Computers and DNA: the Proceedings of the Interface between Computation Science and Nucleic Acid Sequencing Workshop, 157-182. Bell, G., & Marr, T. Eds., Addison-Wesley, Redwood city, CA.

Li, M. 1990. Towards a DNA Sequencing Theory. In Proceedings of the 31st IEEE symposium on Foundation of Computer Science, 125-134.

Liquière, M., and Mephu Nguifo, E. 1990. LEGAL: LEarning with GAlois Lattice. In Proceedings of the fifth Journées Françaises sur l'Apprentissage, 93-113. Lannion.

Mephu Nguifo, E., Chiche, L., Gracy, J., and Sallantin, J. 1991, New methods for the alignment of weakly homologous protein sequences: The (n+1)th method based on symbolic learning. In Proceedings of GBF Prediction and Experiment 3D Structure of Homologous Proteins, 40-42. Germany, September 18-20.

Mephu Nguifo, E., and Sallantin, J. 1993. Cooperative concept acquisition. Forthcoming.

Muggleton, S.; King, R. D.; and Sternberg, M. J. E. 1992. Protein secondary structure prediction using logic-based machine learning. *Protein Engineering* 5(7): 647-657.

Noordewier, M. O., Towell, G. G., and Shavlik, J. W. 1991. Training Knowledge-Based Neural Networks to Recognize Genes in DNA sequences. In *Advances in Neural Information Processing Systems*, 3. M. Kaufmann.

Rendell, L. 1989. A study of an empirical learning for an involved problem. In Proceedings of the International Joint Conference on Artificial Intelligence, 615-620.

Russell, B. 1956. *The Principles of Mathematics*. Allen Unwin Ed., London.

Wille, R. 1992. Concept Lattices & Conceptual Knowledge Systems. *Computer Mathematic Applied*, 23(6-9):493-515.