ELECTRONIC WORKSHOPS IN COMPUTING

Series edited by Professor C.J. van Rijsbergen

Rainer Manthey and Viacheslav Wolfengagen (Eds)

Advances in Databases and Information Systems 1997

Proceedings of the First East-European Symposium on Advances in Databases and Information Systems, (ADBIS'97), St Petersburg, 2-5 September 1997

Presentation of a Methodology to Help the Researcher in his Experimental Process

Michel Sala

Published in collaboration with the
British Computer Society

Presentation of a Methodology to Help the Researcher in his Experimental Process

Michel Sala

LIRMM, UMR 9928 CNRS-Montpellier II

161 rue Ada, 34392 Montpellier Cedex 5

Tel : (33) 04 67 41 86 08

E-mail : sala@lirmm.lirmm.fr

October 18, 1997

Abstract

In this paper we present a methodology to allow a researcher in experimental sciences to receive explanations which will let him to validate or refute the datas of experimentation. In order to do this, we present the chosen methodology then the description of the differents modules used and we finish by the generation of explanations.

Key words : knowledge revision, modelisation, discovery, explanation.

1 Introduction

In the framework of the scientific research, the researcher is confronted to a mass of experimental datas which must be classified, analysed, managed. Those datas come from his own experimentations or from databases of the field. Every new experimentation is confronted to its model of the domain, directly or after a treatment by calculation tools.

For example, in the field of immunogenetics, the model of an immunoglobulin can be traduced by the fact that it must have anchor points which are particular motifs. The proteinic sequence, represented by a chain of letters (in an alphabet if 20), must contain the motifs C then W then C. To position those anchor points, the researcher has to use alignment tools (for example Clustal W [Thompson 94]) which put in correspondence the letters (called amino acids) by simulation mutation events (a letter transforms into another), deletion event (a letter is deleted), or insertion even (a letter inserts) [Dahoff 79] [Risler 88].

In order to take his decision to validate or refuse his data of experimentation, the researcher needs a specific environment which analysis incrementally the datas of experimentation and provides him a whole of informations which will help him to decide. For this, we need differents modules : - calculation tools to treat datas, - a knowledge base to store the model of the researchers field, - databases to research complementary informations.

From those basic elements, we will describe the methodology we used and the differents control modules necessities to the use of such an environment.

2 Methodology

The environment proposed thus makes interact a base of knowledge of the researcher, the computational tools and the databases. This environment provides to the expert explanations on the differences met between bases

and the experiments or the results of the tools. With the reading of these explanations, the researcher rejects the data of experimentation or revises his knowledge by modifying his base of knowledge.

Knowledge of the field of the researcher is described by constraints. All new data of experimentation is confronted with these constraints. If those are violated, an explanation on the data is provided to the researcher who decides to reject the data or to revise his knowledge. In the contrary case, the various computational tools produce results which are then harmonized because of heterogeneity of the tools. If such constraints are violated, a whole of explanations are provided to the researcher on the constraints of the results.

In order to compare the results of the different tools presented with different formalism, we have been led to create a module called *harmonization* which allows to make canonical the results of the calculation tools to make them comparable to others.

In the case where a data of experimentation violates the knowledge base, the researcher needs to position this data according to data already known, stored in databases of the field. To answer to this need, after having interviewed the researcher upon the informations he wishes to obtain, we have developed a module called *navigation module* which will directly select in databases the asked informations.

The process described can be summarized by the following algorithm :

- addition of a new data of experimentation
- analyse of the new data by the base of knowledge
- launching of the computational tools
- analyze results of the computational tools with the base of knowledge
- if violation
- then - navigation in databases
- posting of the explanations
- call to human expertise
- however enquiring new data validates
- then revision by the researcher bases knowledge
- if not rejection news given
- end of treatment

The explanations are provided to the researcher and concern different elements :

- on the data of experimentation having violated at least a constraint,
- on explanations related to the constraints,
- on results of each tool,
- stemming from the module of navigation and giving additional information on the new data coming from the databases of the field.

We are now going to detail the different modules used then the chosen explanation.

3 The architecture of the environment

In this part, we will detail the various modules of the environment :

- computational tools which starting from data of experiments, provide a result in exit, - knowledge of the field of the researcher which is represented in the form of constraints,
- databases of the field which provide a complementary data useful for the researcher.

3.1 Computational tools

We take as working hypotheses that the whole of the tools used refer to the same field and provide in result a comparable whole of information. We can thus consider that they handle a whole of common information, standard with all the tools both in entry and in result.

The whole of the tools is tested on the same whole of experiments (to be able to bring the results closer obtained). The results once standardized, are compared between them according to the constraints on the results stored in the base of the results.

3.2 The base of knowledge of the researcher

All new data of experimentation is analyzed according to the constraints which were specified by the researcher. These constraints can be various natures :

- constraints on the structure : the researcher must indicate the precise structure of the data of experimentation,
- constraints on the values of attributes : with each attribute, the researcher must specify the fields of value : authorized values, intervals,
- constraints inter-attributes of experimentation : the researcher must fix the rules governing the relations inter- attributes,
- constraints on an attribute compared to an element out of the structure of the data of experimentation : the researcher must specify the whole of controls which use data external with experimentation (for example of the databases).

In the same way the constraints on the results of the computational tools are specified in the following way :

- constraints on the structure of the variables entered : used on the data input of the tools (they can be certain fields of the data of experimentation),
- constraints on the variables of entry : fields of values and beach of value associated with each field with the data,
- constraints inter-variables of entry,
- constraints on the structure of the results : variables of standard representation used for the results,
- constraints on the values of the results : fields of values and beach of value of the awaited results,
- constraints on the results of the tools : rules of control on these results.

3.3 Databases of the field

If during the analysis of the results of a tool, a violation of a constraint occurs, we seek various information requested by the researcher in the databases of the field. This various information was selected starting from all the attributes of all the databases of the field directly by the researcher. From this selection, a structure of data is created containing the unit of these attributes.

From this structure, requests of interrogation of the databases must be defined which constitute " the module of navigation " in the databases of the field. It manages the whole of the answers of the requests of each data base in their specific format, harmonizes them and transmits them to the explanatory module.

4 The explanations

The role of explanations is fundamental in the making of the researcher's decision, and must be built from :

- a confrontation between the data of experimentation and the knowledge base of the researcher,
- the harmonized result of differents tools,
- informations proper to tools,
- concatenation and analysis of informations issued from different databases heterogeneous apriori.

4.1 The explanation elements related to the hypothesis of the researcher on the data of experimentation

Then, the hypothesis of the researcher are specified on the form of constraints and an explanation is attached to every constraint [CLA 83] and is expressed in the formalism usually used by the researcher. So in the case of the violation of a constraint, following an experimentation, it is possible to supply the associated explanation.

Figure 1: Dendrogram of the tool MAP123D

4.2 The explanation elements related to the tools

Three types of explanation can be built : upon the results of the tools, upon the harmonized result of a tool, and upon the comparison of the results of different tools.

- Explanations upon the results of the tool :

The explanations upon the results of the tool provide informations concerning the tool itself, its specifications, algorithms and heuristics it uses. The explanations are tool-dependant.

- Explanations upon the harmonized result of a tool :

The harmonization allows to the researcher to compare the results of the tools by using the same representation. The explanation corresponds to the result of the tool translated into the formalism chosen by the researcher. In our experimentation on the classification of immunoglobulins, the selected format is the one of the tool MAP123D [Gracy 93] according to the format pointed out by the figure below.

- Explanations upon the comparison of the results of different tools

After the use of several tools, an explanation is supplied in a synthetic form which will help the researcher to implicate his hypothesis.

4.3 The explanation elements related to the data bases

In the aim to check the hypothesis stored in knowledge bases, the system provides complementary informations stored in databases available on the Internet network. The explanations traduce themselves by the result of a request ; the structure of the result being define apriori by the researcher.

The elementary explanations provided to the researcher are the following ones :

- informations on the data of experimentation having violated at least one constraint,
- informations on the explanations related to the constraints not abided by the knowledge base,
- synthetic informations on the whole of calculation tools,
- informations on the harmonized results of every tool,
- informations on the elementary results of every tool,
- information coming from the navigation module and giving supplementary informations on the new data coming from the databases of the field.

5 The SIGALE System

In this last case, we are in a situation of the type " contributes to the scientific discovery ". Methodology that we will describe was tested in collaboration with the L.I.G.M. (Laboratoire d'ImmunoGénétique Moléculaire de Montpellier) [Sala 94], direct by professors Marie-Paule and Gérard Lefranc, on the alignment of proteinic sequences of immunoglobulins [Lefranc 94] [Frippia 94] and gave rise to system SIGALE [Sala 96].

The specific application we retained is the dynamic classification of the variables areas of the alpha-chains of humans T-cells receptors. To obtain this dynamic classification, we have an initial sample composed of a

whole of proteinic sequences before all classified. Using an algorithm of classification, we obtain a classification in sub- families. From this initial classification, we add cyclically a new proteinic sequence and we analyse the result of the new classification obtained.

The datas used in entry are proteinic sequences, the constraints on those datas concerns the fact that they must be immunoglobulins. The calculation tools used are alignment tools which provide as result an alignment as well as a dendrogram associated. The model of the researcher's domain is represented by a knowledge base.

The datas of experimentation come from the specific database, called IMGT / LIGMDB (according to the theoricall work of [Mougenot 95]). LIGMDB is a database specialised in the processing of informations connected to immunoglobulins sequences and T-cells receptors.

5.1 The functioning of SIGALE

The control module undertakes the control of the whole of chains. From the introduction of a new sequence, an expertise using the knowledge base is ran. If it detects an anomaly, it is indicated to the Explanation module. On the contrary, the whole of tools are activated and the knowledge base checks the results.

If a violation occurs, an explanation is provided to the biologist according to informations collected by the Navigation module and to the results obtained by the knowledge base from the whole of the results of tools.

6 Conclusion

We have proved that it is possible to place a methodology which allows a researcher in experimental sciences to make his knowledge of the field progress. This environment was tested in experiments by the SIGALE system and made it possible to highlight preserved patterns in the "immunoglobulin superfamily" approved by the scientific community into immunogenetics. We proved that it was possible in the field of the applied sciences to install a system of explanations which can help the researcher in his phase of assistance to the discovery to revise its knowledge on the field. Indeed in our experimentation, we had a rejection by the researcher of any explanation provided in textual form. The explanations are started automatically by the system at the convenient period, when an assumption of the researcher is not respected.

References

- [Clay 84] *Inference ART programming tutorial Vol 2, a first look of viewpoints* - Clay B.D. - Inference Corporation, Los Angeles, 1984
- [Dayhoff 79] *Atlas of protein sequence and structure* - Dayhoff M.O., Schwartz R.M. - National biomedical research foundation, p 353, Washington, 1979
- [Frippia 94] *Genomic organisation of 34 kb the human immunoglobulin lambda locus (IGLV) : restriction map and sequences of new VI III genes* - Frippia J.P., Lefranc M.P. - Molecular Biology, Vol 31, n 9, pp 657-670, 1994
- [Gracy 93] *Conception d'un environnement d'apprentissage appliqué à l'étude structurale des protéines* - Gracy J. - Thèse de troisième cycle, univ de Montpellier II, décembre 93
- [Lefranc 94] *LIGM-DB/IMGT : an integrated database of immunoglobulins and T cell receptors* - Lefranc M.P., Giudicelli V., Busin C., Sala M., Mougenot I., Dehais P., Malik A., Chaume D. - Twentieth EMBO Annual Symposium - Genomes and chromosomes, p 85, EMBL, Heidelberg, 3-6 octobre 1994
- [Mougenot 95] *LIGM-DB, un système de gestion intégré des séquences nucléiques des récepteurs d'antigènes et de leurs annotations*

- [Risler 88] *Amino acid substitutions in structurally* - Risler J.L., Delorme M.O., Delacroix H., Henaut A. - J. Mol. Biol. 204, 344, p 1019, 1988
- [Sala 94] *Environnement pour le transfert de connaissances d'Agents Artificiel vers un agent humain - Application à la classification automatique de séquences protéiques* - Sala M. - III ème Rencontres des Jeunes Chercheurs en I.A., Lumigny, 1994
- [Sala 96] *Proposition d'une méthodologie et d 'un environnement d 'apprentissage d'aide à la découverte scientifique - Application à la génétique : l'environnement SIGALE* - Sala M. - Thèse de troisième cycle, univ de Montpellier II, janvier 96
- [Thompson 94] *Clustal W : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties weight matrix choise* - Thompson J.D., Higgins D.G., Gibson T.J. - Nucleic Acids Reseach, Vol 22, pp 4673-4680, 1994