IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2011

Project-Team ORPAILLEUR

Knowledge Discovery guided by Domain Knowledge

IN COLLABORATION WITH: Laboratoire lorrain de recherche en informatique et ses applications (LORIA)

RESEARCH CENTER
Nancy - Grand Est

THEME
Knowledge and Data Representation and Management

Table of contents

1. Members **1**
2. Overall Objectives **2**
2.1. Introduction 2
2.2. Highlights 2
3. Scientific Foundations **2**
3.1. From KDD to KDDK 2
3.2. Methods for Knowledge Discovery guided by Domain Knowledge 3
3.3. Elements on Text Mining 4
3.4. Elements on Knowledge Systems and Semantic Web 4
4. Application Domains **5**
4.1. Life Sciences 5
4.2. Knowledge Management in Medicine 5
4.3. Cooking 6
5. Software **6**
5.1. Generic Symbolic KDD Systems 6
5.1.1. The Coron Platform 6
5.1.2. Orion: Skycube Computation Software 7
5.2. Stochastic systems for knowledge discovery and simulation 7
5.2.1. The CarottAge system 7
5.2.2. The ARPEnTAge system 7
5.2.3. GenExp-LandSiTes: KDD and simulation 8
5.3. KDD in Systems Biology 8
5.3.1. IntelliGO online 8
5.3.2. WAFOBI : KNIME nodes for relational mining of biological data 8
5.3.3. MOdel-driven Data Integration for Mining (MODIM) 8
5.4. Knowledge-Based Systems and Semantic Web Systems 9
5.4.1. The Kasimir System for Decision Knowledge Management 9
5.4.2. Taaable: a system for retrieving and creating new cooking recipes by adaptation 9
6. New Results **10**
6.1. The Mining of Complex Data 10
6.1.1. FCA, RCA, and Pattern Structures 10
6.1.2. Miscellaneous in FCA and Pattern Mining 10
6.1.3. Skylines, sequences and privacy 11
6.1.4. KDDK in Text Mining 11
6.2. KDDK in Life Sciences 12
6.2.1. Ontology-based Functional Classification of Genes 12
6.2.2. Use of Domain Knowledge for Dimension Reduction 12
6.2.3. Mining Agronomical Data with stochastic models 13
6.3. Structural Systems Biology and Docking 13
6.3.1. Accelerating protein docking calculations using graphics processors 14
6.3.2. Eigen-Hex: Modeling protein flexibility during docking 14
6.3.3. 3D-Blast: A new approach for protein structure alignment and clustering 14
6.3.4. KBDOCK: Protein docking using Knowledge-Based approaches 14
6.3.5. V-Dock: scoring protein-protein interactions using Voronoi fingerprints 14
6.3.6. DOVSA: Developing new algorithms for virtual screening 15
6.4. Around the Kasimir research project 15
6.5. Around the Taaable research project 15
7. Contracts and Grants with Industry **16**
7.1. The BioIntelligence Project 16

7.2. The Quaero Project 17
8. Partnerships and Cooperations **18**
8.1. International projects and collaborations 18
8.1.1. Fapemig INRIA Project: Incorporating knowledge models into scalable data mining algorithms 18
8.1.2. Search for anti-HIV drugs acting as entry-blockers 18
8.1.3. International collaborations in Mining complex data 18
8.2. European Initiatives 19
8.3. National Initiatives 20
8.3.1. ANR Kolflow: man-machine collaboration in continuous knowledge-construction flows 20
8.3.2. ANR Trajcan: a study of patient care trajectories 20
8.4. Local initiatives 20
8.4.1. Contrat Plan État Région" (CPER) 20
8.4.2. Other initiatives 21
8.4.2.1. Cancéropole Grand-Est 21
8.4.2.2. BioProLor 21
9. Dissemination **21**
9.1. Scientific Animation 21
9.2. Teaching 22
10. Bibliography **22**

Project-Team ORPAILLEUR

Keywords: Knowledge Discovery, Data Mining, Ontologies, Knowledge Representation, Reasoning

1. Members

Research Scientists

Amedeo Napoli [Team leader, Senior Researcher, CNRS, HdR]
Marie-Dominique Devignes [Junior Researcher, CNRS, HdR]
Bernard Maigret [Senior Researcher (emiritus), CNRS, HdR]
Chedy Raïssi [Junior Researcher, INRIA]
Dave Ritchie [Senior Researcher, INRIA, HdR]
Yannick Toussaint [Junior Researcher, INRIA, HdR]

Faculty Members

Adrien Coulet [Associate Professor, ESIAL Université Henri Poincaré Nancy)]
Nicolas Jay [Associate Professor, Faculté de Médecine Université Henri Poincaré Nancy)]
Florence Le Ber [Professor (ENGEES Strasbourg), HdR]
Bart Lamiroy [Associate Professor (délégation INRIA until July 2011, MdC INPL)]
Jean Lieber [Associate Professor (MdC Université Henri Poincaré Nancy 1), HdR]
Jean-François Mari [Professor (Université de Nancy 2), HdR]
Emmanuel Nauer [Associate Professor (MdC Université Paul Verlaine Metz)]
Malika Smaïl-Tabbone [Associate Professor (MdC Université Henri Poincaré Nancy 1)]

Technical Staff

Inaki Fernandez [Engineer (until August 2011)]
Renaud Grisoni [Engineer]
Laura Infante-Blanco [Engineer (since October 2011)]
Jean-François Kneib [Engineer]
Luis Felipe Melo [Engineer]
Birama NDiayé [Engineer (until July 2011)]

PhD Students

Mehwish Alam [PhD Student (BioIntelligence Grant, since October 2011)]
Yasmine Assess [PhD Student (INCa Grant, Thesis defended in October 2011)]
Isiru Bayissa [PhD Student (BioIntelligence and Lorraine Region Grant)]
Sid-Ahmed Benabderrahmane [PhD Student (INCa Grant, Thesis defended in December 2011)]
Aleksey Buzmakov [PhD Student (BioIntelligence Grant, since November 2011)]
Emmanuel Bresso [PhD Student (Cifre Harmonic Pharma)]
Victor Codocedo [PhD Student (Quaero Grant, since August 2011)]
Julien Cojan [PhD Student (ATER until September 2011, Thesis defended in October 2011)]
Sébastien Da Silva [PhD Student (INRA - INRIA Grant)]
Valmi Dufour-Lussier [PhD Student (MERT Grant)]
Elias Egho [PhD Student (ANR Trajcan Project)]
Emmanuelle Gaillard [Master Student (from March to December 2011)]
Anisah Ghoorah [PhD Student (ANR Contract)]
Mehdi Kaytoue [PhD Student (ATER until June 2011, Thesis defended in April 2011)]
Thomas Meilender [PhD Student (CIFRE, A2ZI Company)]
Julien Stévenot [PhD Student (ANR Kolflow, since October 2011)]
My Thao Tang [PhD Student (ANR Kolflow, since October 2011)]

Post-Doctoral Fellows

Thomas Bourquard [BioIntelligence Grant]
Ioanna Lykourentzou [Ercim/INRIA Grant (since July 2011)]
Lazaros Mavridis [(until July 2011)]
Violeta Pérez-Nueno [Marie Curie Grant]
Lian Shi [(until June 2011)]
Vishwesh Venkatraman [(until July 2011)]

Administrative Assistant

Emmanuelle Deschamps [Secretary]

2. Overall Objectives

2.1. Introduction

Knowledge discovery in databases –hereafter KDD– consists in processing a large volume of data in order to discover knowledge units that are significant and reusable. Assimilating knowledge units to gold nuggets, and databases to lands or rivers to be explored, the KDD process can be likened to the process of searching for gold. This explains the name of the research team: in French "orpailleur" denotes a person who is searching for gold in rivers or mountains. Moreover, the KDD process is iterative, interactive, and generally controlled by an expert of the data domain, called the analyst. The analyst selects and interprets a subset of the extracted units for obtaining knowledge units having a certain plausibility. As a person searching for gold and having a certain knowledge of the task and of the location, the analyst may use his own knowledge but also knowledge on the domain of data for improving the KDD process.

A way for the KDD process to take advantage of domain knowledge is to be in connection with ontologies relative to the domain of data, for making a step towards the notion of *knowledge discovery guided by domain knowledge* or KDDK. In the KDDK process, the extracted knowledge units have still "a life" after the interpretation step: they are represented using a knowledge representation formalism to be integrated within an ontology and reused for problem-solving needs. In this way, knowledge discovery is used for extending and updating existing ontologies, showing that knowledge discovery and knowledge representation are complementary tasks and reifying the notion of KDDK.

2.2. Highlights

This year, the team would like to indicate in this section two kinds of highlights. Firstly, we would like to emphasize the high quality of some publications obtained by the team in high-level conferences and journals, such as CIKM, ICDM, IJCAI, KDD, and Bioinformatics.

Secondly, the application of KDDK process in the domain of Life Sciences made progress in 2011. Fast algorithms for 3D-shape protein classification and docking using polar Fourier correlations on graphics processor units (GPUs) have been published. The GPU-accelerated version of the Hex docking program (http://hex.loria.fr) has had some 4,000 downloads in the last year, and the GPU-powered server (http://hexserver.loria.fr) has performed some 13,000 docking runs for external users.

3. Scientific Foundations

3.1. From KDD to KDDK

Glossary

Knowledge discovery in databases is a process for extracting knowledge units from large databases, units that can be interpreted and reused within knowledge-based systems.

Knowledge discovery in databases is a process for extracting knowledge units from large databases, units that can be interpreted and reused within knowledge-based systems. From an operational point of view, the KDD process is performed within a KDD system including databases, data mining modules, and interfaces for interactions, e.g. editing and visualization. The KDD process is based on three main operations: selection and preparation of the data, data mining, and finally interpretation of the extracted units. The KDDK process –as implemented in the research work of the Orpailleur team– is based on *data mining methods* that are either symbolic or numerical:

- Symbolic methods are based on frequent itemsets search, association rule extraction, and concept lattice design (Formal Concept Analysis and extensions [91]) [105].
- Numerical methods are based on second-order Hidden Markov Models (HMM2, designed for pattern recognition [101]). Hidden Markov Models have good capabilities for locating stationary segments, and are mainly used for mining temporal and spatial data.

The principle summarizing KDDK can be understood as a process going from complex data units to knowledge units being guided by domain knowledge (KDDK or "knowledge with/for knowledge") [98]. Two original aspects can be underlined: (i) the KDD process is guided by domain knowledge, and (ii) the extracted units are embedded within a knowledge representation formalism to be reused in a knowledge-based system for problem solving purposes.

The various instantiations of the KDDK process in the research work of Orpailleur are mainly based on *classification*, considered as a polymorphic process involved in tasks such as modeling, mining, representing, and reasoning. Accordingly, the KDDK process may feed knowledge-based systems to be used for problem-solving activities in application domains, e.g. agronomy, astronomy, biology, chemistry, and medicine, and also for semantic web activities involving text mining, information retrieval, and ontology engineering [78], [79].

3.2. Methods for Knowledge Discovery guided by Domain Knowledge

Glossary

knowledge discovery in databases guided by domain knowledge is a KDD process guided by domain knowledge ; the extracted units are represented within a knowledge representation formalism and embedded within a knowledge-based system.

Classification problems can be formalized by means of a class of individuals (or objects), a class of properties (or attributes), and a binary correspondence between the two classes, indicating for each individual-property pair whether the property applies to the individual or not. The properties may be features that are present or absent, or the values of a property that have been transformed into binary variables. Formal Concept Analysis (FCA) relies on the analysis of such binary tables and may be considered as a symbolic data mining technique to be used for extracting (from a binary database) a set of formal concepts organized within a concept lattice hierarchy [91]. Concept lattices are sometimes also called Galois lattices [81].

The search for frequent itemsets and the extraction of association rules are well-known symbolic data mining methods, related to FCA (actually searching for frequent itemsets may be understood as traversing a concept lattice). Both processes usually produce a large number of items and rules, leading to the associated problems of "mining the sets of extracted items and rules". Some subsets of itemsets, e.g. frequent closed itemsets (FCIs), allow to find interesting subsets of association rules, e.g. informative association rules. This is why several algorithms are needed for mining data depending on specific applications [120], [119].

Among useful patterns extracted from a database, frequent itemsets are usually thought to unfold "regularities" in the data, i.e. they are the witnesses of recurrent phenomena and they are consistent with the expectations of the domain experts. In some situations however, it may be interesting to search for "rare" itemsets, i.e. itemsets that do not occur frequently in the data (contrasting frequent itemsets). These correspond to unexpected phenomena, possibly contradicting beliefs in the domain. In this way, rare itemsets are related to "exceptions"

and thus may convey information of high interest for experts in domains such as biology or medicine [121], [122].

From the numerical point of view, a Hidden Markov Model (HMM2) is a stochastic process aimed at extracting and modeling a stationary distribution of events. These models can be used for data mining purposes, especially for spatial and temporal data as they show good capabilities to locate stationary segments [100]). one special research effort focuses on the study of the application of HMM2 to composite data, both in the temporal and spatial domain, to produce a multi-dimensional classification based on multiple attributes.

3.3. Elements on Text Mining

Glossary

Text mining is a process for extracting knowledge units from large collections of texts, units that can be interpreted and reused within knowledge-based systems.

The objective of a text mining process is to extract new and useful knowledge units in a large set of texts [77], [88]. The text mining process shows specific characteristics due to the fact that texts are complex objects written in natural language. The information in a text is expressed in an informal way, following linguistic rules, making the mining process more complex. To avoid information dispersion, a text mining process has to take into account –as much as possible– paraphrases, ambiguities, specialized vocabulary, and terminology. This is why the preparation of texts for text mining is usually dependent on linguistic resources and methods.

From a KDDK perspective, the text mining process is aimed at extracting new knowledge units from texts with the help of background knowledge encoded within an ontology and which is useful to relate notions present in a text, to guide and to help the text mining process. Text mining is especially useful in the context of semantic web for ontology engineering [85], [84], [83]. In the Orpailleur team, the focus is put on real-world texts in application domains such as astronomy, biology and medicine, using mainly symbolic data mining methods. Accordingly, the text mining process may be involved in a loop used to enrich and to extend linguistic resources. In turn, linguistic and ontological resources can be exploited to guide a "knowledge-based text mining process".

3.4. Elements on Knowledge Systems and Semantic Web

Glossary

Knowledge representation is a process for representing knowledge within an ontology using a knowledge representation formalism, giving knowledge units a syntax and a semantics. Semantic web is based on ontologies and allows search, manipulation, and dissemination of documents on the web by taking into account their contents, i.e. the semantics of the elements included in the documents.

Usually, people try to take advantage of the web by searching for information (navigation, exploration), and by querying documents using search engines (information retrieval). Then people try to analyze the obtained results, a task that may be very difficult and tedious. Semantic web is an attempt for guiding search for information with the help of machines, that are in charge of asking questions, searching for answers, classifying and interpreting the answers. However, a machine may be able to read, understand, and manipulate information on the web, if and only if the knowledge necessary for achieving those tasks is available. This is why ontologies are of main importance with respect to the task of setting up semantic web. Thus, there is a need for representation languages for annotating documents, i.e. describing the content of documents, and giving a semantics to this content. Knowledge representation languages are (the?) good candidates for achieving the task: they have a syntax with an associated semantics, and they can be used for retrieving information, answering queries, and reasoning.

Semantic web constitutes a good platform for experimenting ideas on knowledge representation, reasoning, and KDDK. In particular, the knowledge representation language used for designing ontologies is the OWL language, which is based on description logics (or DL [76]). In OWL, knowledge units are represented within concepts (or classes), with attributes (properties of concepts, or relations, or roles), and individuals. The hierarchical organization of concepts (and relations) relies on a subsumption relation that is a partial ordering. The inference services are based on subsumption, concept and individual classification, two tasks related to "classification-based reasoning". Furthermore, classification-based reasoning can be associated to case-based reasoning (CBR), that relies on three main operations: retrieval, adaptation, and memorization. Given a target problem, retrieval consists in searching for a source (memorized) problem similar to the target problem. Then, the solution of the source problem is adapted to fulfill the constraints attached to the target problem, and possibly memorized for further reuse.

In the trend of semantic web, research work is also carried on semantic wikis which are wikis i.e., web sites for collaborative editing, in which documents can be annotated thanks to semantic annotations and typed relations between wiki pages [54]. Such links provide kind of primitive knowledge units that can be used for guiding information retrieval or knowledge discovery.

4. Application Domains

4.1. Life Sciences

Participants: Yasmine Assess, Sid-Ahmed Benabderrahmane, Thomas Bourquard, Emmanuel Bresso, Marie-Dominique Devignes, Elias Egho, Anisah Ghoorah, Renaud Grisoni, Nicolas Jay, Mehdi Kaytoue, Bernard Maigret, Lazaros Mavridis, Amedeo Napoli, Violeta Pérez-Nueno, Dave Ritchie, Malika Smaïl-Tabbone, Yannick Toussaint, Vishwesh Venkatraman.

Glossary

Knowledge discovery in life sciences is a process for extracting knowledge units from large biological databases, e.g. collection of genes.

One major application domain which is currently investigated by Orpailleur team is related to life sciences, with particular emphasis on biology, medicine, and chemistry. The understanding of biological systems provides complex problems for computer scientists, and, when they exist, solutions bring new research ideas for biologists and for computer scientists as well. Accordingly, the Orpailleur team includes biologists, chemists, and a physician, making Orpailleur a very original EPI at INRIA.

Knowledge discovery is gaining more and more interest and importance in life sciences for mining either homogeneous databases such as protein sequences and structures, or heterogeneous databases for discovering interactions between genes and environment, or between genetic and phenotypic data, especially for public health and pharmacogenomics domains. The latter case appears to be one main challenge in knowledge discovery in biology and involves knowledge discovery from complex data and thus KDDK. The interactions between researchers in biology and researchers in computer science improve not only knowledge about systems in biology, chemistry, and medicine, but knowledge about computer science as well. Solving problems for biologists using KDDK methods involves the design of specific modules that, in turn, leads to adaptations of the KDDK process, especially in the preparation of data and in the interpretation of the extracted units.

4.2. Knowledge Management in Medicine

Participants: Julien Cojan, Nicolas Jay, Jean Lieber, Thomas Meilender, Amedeo Napoli.

The Kasimir research project holds on decision support and knowledge management for the treatment of cancer [97]. This is a multidisciplinary research project in which participate researchers in computer science (Orpailleur), experts in oncology ("Centre Alexis Vautrin" in Vandœuvre-lès-Nancy), Oncolor (a healthcare network in Lorraine involved in oncology), and A2Zi (a company working in Web technologies and involved in several projects in the medical informatics domain, http://www.a2zi.fr/). For a given cancer localization, a treatment is based on a protocol similar to a medical guideline, and is built according to evidence-based medicine principles. For most of the cases (about 70%), a straightforward application of the protocol is sufficient and provides a solution, i.e. a treatment, that can be directly reused. A case out of the 30% remaining cases is "out of the protocol", meaning that either the protocol does not provide a treatment for this case, or the proposed solution raises difficulties, e.g. contraindication, treatment impossibility, etc. For a case "out of the protocol", oncologists try to *adapt* the protocol. Actually, considering the complex case of breast cancer, oncologists discuss such a case during the so-called "breast cancer therapeutic decision meetings", including experts of all specialties in breast oncology, e.g. chemotherapy, radiotherapy, and surgery.

The semantic Web technologies have been used and adapted in the Kasimir project for several years. Currently, technologies of the semantic Wikis are adapted for the management of decision protocols [66] More precisely, the migration from the static HTML site of Oncolor to a semantic wiki (with limited editing rights and unlimited reading rights) is about to be finished. This has consequences on the editorial chain of the published protocols which is more collaborative. A decision tree editor that has been integrated into the wiki and that has an export facility to formalized protocols in OWL DL has also been developed [67].

4.3. Cooking

Participants: Julien Cojan, Valmi Dufour-Lussier, Inaki Fernandez, Emmanuelle Gaillard, Laura Infante-Blanco, Florence Le Ber, Jean Lieber, Amedeo Napoli, Emmanuel Nauer, Yannick Toussaint.

The origin of the Taaable project is the Computer Cooking Contest (CCC). A contestant of the CCC is a system that answers queries of recipes, using a recipe base; if no recipe exactly matches the query, then the system adapts another recipe. Taaable is a case-based reasoning system that uses various technologies used and developed in the Orpailleur team, such as technologies of the semantic web, knowledge discovery techniques, knowledge representation and reasoning techniques, etc. From a research viewpoint it enables to test the scientific results on an application domain that is at the same time simple to understand and raising complex issues, and to study the complementarity of various research domains. Taaable has been at the origin of the project Kolflow of the ANR CONTINT program, whose application domain is WikiTaaable, the semantic wiki of Taaable. It is also used for other projects under submission.

5. Software

5.1. Generic Symbolic KDD Systems

5.1.1. The Coron Platform

Participants: Mehdi Kaytoue [contact person], Amedeo Napoli, Yannick Toussaint.

The Coron platform [118], [95] is a KDD toolkit organized around three main components: (1) Coron-base, (2) AssRuleX, and (3) pre- and post-processing modules. The software was registered at the "Agence pour la Protection des Programmes" (APP) and is freely available[1]. The Coron-base component includes a complete collection of data mining algorithms for extracting itemsets such as frequent itemsets, frequent closed itemsets, frequent generators. In this collection we can find APriori, Close, Pascal, Eclat, Charm, and, as well, original algorithms such as ZART, Snow, Touch, and Talky-G. The Coron-base component contains also algorithms for extracting rare itemsets and rare association rules, e.g. APriori-rare, MRG-EXP, ARIMA, and BTB. AssRuleX

[1] http://coron.loria.fr

generates different sets of association rules (from itemsets), such as minimal non-redundant association rules, generic basis, and informative basis. In addition, the Coron system supports the whole life-cycle of a data mining task and proposes modules for cleaning the input dataset, and for reducing its size if necessary. The Coron toolkit is developed in Java, is operational, and was already used in several research projects.

5.1.2. Orion: Skycube Computation Software

Participant: Chedy Raïssi [contact person].

This program implements the algorithms described in a research paper published last year at VLDB 2010 [113]. The software provides a list of four algorithms discussed in the paper in order to compute skycubes. This is the most efficient –in term of space usage and runtime– implementation for skycube computation (see https://github.com/leander256/Orion).

5.2. Stochastic systems for knowledge discovery and simulation

5.2.1. The CarottAge system

Participants: Florence Le Ber, Jean-François Mari [contact person].

CarottAge [2] is a data mining system, freely available (GPL license) and based on Hidden Markov Models of second order. It provides a synthetic representation of temporal and spatial data. CarottAge is currently used by INRA researchers interested in mining the changes in territories related to the loss of biodiversity (projects ANR BiodivAgrim and ACI Ecoger) and/or water contamination.

In these practical applications, the system aims at building a partition –called the hidden partition– in which the inherent noise of the data is withdrawn as much as possible. The CarottAge system takes into account: (i) the various shapes of the territories that are not represented by square matrices of pixels, (ii) the use of pixels of different size with composite attributes representing the agricultural pieces and their attributes, (iii) the irregular neighborhood relation between those pixels, (iv) the use of shape files to facilitate the interaction with GIS (geographical information system).

CarottAge has been used for mining hydromorphological data. Actually a comparison was performed with three other algorithms classically used for the delineation of river continuums and CarottAge proved to give very interesting results for that purpose [73].

5.2.2. The ARPEnTAge system

Participants: Florence Le Ber, Jean-François Mari [contact person].

ARPEnTAge [3] (for *Analyse de Régularités dans les Paysages: Environnement, Territoires, Agronomie* is a software based on stochastic models (HMM2 and Markov Field) for analyzing spatiotemporal data-bases [73]. ARPEnTAge is built on top of the CarottAge system to fully take into account the spatial dimension of input sequences. It takes as input an array of discrete data in which the columns contain the annual land-uses and the rows are regularly spaced locations of the studied landscape. Displaying tools and the generation of shape files have also been defined.

We model the spatial structure of the landscape by a Markov Random Field (MRF) whose sites are random Land Uses (LUS) located in the parcels. The dynamics of these LUS are modelled by a temporal HMM2. This leads to the definition of a MRF where the underlying mean field is approximated by a HMM2 that processes a Hilbert-Peano fractal curve spanning the image. This MRF is used to segment the landscape into patches, each of them being characterized by a temporal HMM2. The patch labels, together with the geographic coordinates, determine a clustered image of the landscape that can be coded within an ESRI shapefile.

[2] http://www.loria.fr/~jfmari/App/
[3] http://www.loria.fr/~jfmari/App/

ARPEnTAge is freely available (GPL license pending) and is currently used by INRA researchers interested in mining the changes in territories related to the loss of biodiversity (projects ANR BiodivAgrim and ACI Ecoger) and/or water contamination.

5.2.3. GenExp-LandSiTes: KDD and simulation

Participants: Sébastien Da Silva, Florence Le Ber [contact person], Jean-François Mari.

In the framework of the project "Impact des OGM" initiated by the French ministry of research, we have developed a software called GenExp-LandSiTes for simulating bidimensional random landscapes, and then studying the dissemination of vegetable transgenes. The GenExp-LandSiTes system is linked to the CarottAge system, and is based on computational geometry and spatial statistics. The simulated landscapes are given as input for programs such as "Mapod-Maïs" or "GeneSys-Colza" for studying the transgene diffusion. Other landscape models based on tessellation methods are under studies. The last version of GenExp allows an interaction with R and deals with several geographical data formats.

This work is now part of an INRA-INRIA project about landscape modeling, PAYOTE (2009–2011), that gathers eleven research teams of agronomists, ecologists, statisticians, and computer scientists. The PAYOTE project is now focusing on the comparison of various methods for analyzing and building temporal and spatial landscape structures. Sébastien da Silva is preparing his PhD thesis within this framework and is conducted both by Claire Lavigne (DR in ecology, INRA Avignon) and Florence Le Ber [62]. Florence Le Ber is also involved within a new INRA project on virtual landscape modelling.

5.3. KDD in Systems Biology

5.3.1. IntelliGO online

The IntelliGO measure computes semantic similarity between terms from a structured vocabulary (Gene Ontology: GO) and uses these values for computing functional similarity between genes annotated by sets of GO terms [82]. The IntelliGO measure is made available online (http://plateforme-mbi.loria.fr/intelligo/) to be used by members of the community for exploitation and evaluation purposes. It is possible to compute the functional similarity between two genes, the intra-set similarity value in a given set of genes, and the inter-set similarity value for two given sets of genes.

5.3.2. WAFOBI : KNIME nodes for relational mining of biological data

KNIME (for "Konstanz Information Miner") is an open-source visual programming environment for data integration, processing, and analysis. KNIME has been developed using rigorous software engineering practices and is used by professionals in both industry and academia. The KNIME environment includes a rich library of data manipulation tools (import, export) and several mining algorithms which operate on a single data matrix (decision trees, clustering, frequent itemsets, association rules...). The KNIME platform aims at facilitating the data mining experiment settings as many tests are required for tuning the mining algorithms. The evaluation of the mining results is also an important issue and its configuration is made easier.

A position of engineer ("Ingénieur Jeune Diplomé INRIA") was granted to the Orpailleur team to develop some extra KNIME nodes for relational data mining using the ALEPH program (http://www.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph.pl). The developed KNIME nodes include a data preparation node for defining a set of first-order predicates from a set of relation schemas and then a set of facts from the corresponding data tables (learning set). A specific node allows to configure and run the ALEPH program to build a set of rules. Subsequent nodes allow to test the first-order rules on a test set and to perform configurable cross validations. An INRIA APP procedure is currently pending.

5.3.3. MOdel-driven Data Integration for Mining (MODIM)

Participants: Marie-Dominique Devignes [contact person], Birama Ndiayé, Malika Smaïl-Tabbone.

The MODIM software (MOdel-driven Data Integration for Mining) is a user-friendly data integration tool which can be summarized along three functions: (i) building a data model taking into account mining requirements and existing resources; (ii) specifying a workflow for collecting data, leading to the specification of wrappers for populating a target database; (iii) defining views on the data model for identified mining scenarios. A steady-version of the software has been deposited through INRIA APP procedure in December, 2010.

Although MODIM is domain independent, it was used so far for biological data integration in various internal research studies. A poster was presented at the last JOBIM conference (Paris, June 2011). Recently, MODIM was used by colleagues from the LIFL for organizing data about non ribosomal peptide syntheses. Feedback from users led to extensions of the software. The sources can be downloaded at https://gforge.inria.fr/projects/modim/.

5.4. Knowledge-Based Systems and Semantic Web Systems

5.4.1. The Kasimir System for Decision Knowledge Management

Participants: Nicolas Jay, Jean Lieber [contact person], Amedeo Napoli, Thomas Meilender.

The objective of the Kasimir system is decision support and knowledge management for the treatment of cancer. A number of modules have been developed within the Kasimir system for editing of treatment protocols, visualization, and maintenance. Kasimir is developed within a semantic portal, based on OWL. KatexOWL (Kasimir Toolkit for Exploiting OWL Ontologies, http://katexowl.loria.fr) has been developed in a generic way and is applied to Kasimir. In particular, the user interface EdHibou of KatexOWL is used for querying the protocols represented within the Kasimir system.

The software CabamakA (case base mining for adaptation knowledge acquisition) is a module of the Kasimir system. This system performs case base mining for adaptation knowledge acquisition and provides information units to be used for building adaptation rules [123]. Actually, the mining process in CabamakA is implemented thanks to a frequent close itemset extraction module of the Coron platform (see §5.1.1). A semantic wiki for the collaborative edition of decision protocols was developed and is going to be deployed.

5.4.2. Taaable: a system for retrieving and creating new cooking recipes by adaptation

Participants: Julien Cojan, Valmi Dufour-Lussier, Inaki Fernandez, Emmanuelle Gaillard, Laura Infante-Blanco, Florence Le Ber, Jean Lieber, Amedeo Napoli, Emmanuel Nauer [contact person], Yannick Toussaint.

Taaable is a system whose objectives are to retrieve textual cooking recipes and to adapt these retrieved recipes whenever needed. Suppose that someone is looking for a “leek pie” but has only an “onion pie” recipe: how can the onion pie recipe be adapted?

The Taaable system combines principles, methods, and technologies of knowledge engineering, namely case-based reasoning (CBR), ontology engineering, text mining, text annotation, knowledge representation, and hierarchical classification. Ontologies for representing knowledge about the cooking domain, and a terminological base for binding texts and ontology concepts, have been built from textual web resources. These resources are used by an annotation process for building a formal representation of textual recipes. A CBR engine considers each recipe as a case, and uses domain knowledge for reasoning, especially for adapting an existing recipe w.r.t. constraints provided by the user, holding on ingredients and dish types.

The Taaable system is available on line at http://taaable.fr. After being ranked twice second, in the 2008 and 2009 “Computer Cooking Contests” organized during the ICCBR conference, Taaable won the first price and the adaptation challenge, in 2010. In 2011, no contest was organized but the system has, however, been extended by two new features, both concerning knowledge acquisition using FCA [42]. The first feature uses FCA in order to enrich the domain ontology (especially the ingredient hierarchy), making the case retrieval more progressive and more precise [45]. The second feature uses FCA for extracting adaptation knowledge, in order to be able to better adapt a recipe to given constraints [47]. Current ongoing work on the Taaable project also includes formal representation of preparations [63].

6. New Results

6.1. The Mining of Complex Data

Participants: Mehwish Alam, Isiru Bayissa, Thomas Bourquard, Aleksey Buzmakov, Victor Codocedo, Adrien Coulet, Elias Egho, Nicolas Jay, Mehdi Kaytoue, Florence Le Ber, Ioanna Lykourentzou, Luis Felipe Melo, Amedeo Napoli, Chedy Raïssi, Lian Shi, Yannick Toussaint.

Formal concept analysis, together with itemset search and association rule extraction, are suitable symbolic methods for KDDK, that may be used for real-sized applications. Global improvements may be carried on the scope of applicability, the ease of use, the efficiency of the methods, and on the ability to fit evolving situations. Accordingly, the team is working on extensions of such symbolic data mining methods to be applied on complex data such as biological or chemical data or textual documents, involving objects with multi-valued attributes (e.g. domains or intervals), n-ary relations, sequences, trees and graphs.

6.1.1. FCA, RCA, and Pattern Structures

Recent advances in data and knowledge engineering have emphasized the need for Formal Concept Analysis (FCA) tools taking into account structured data. There are a few extensions of FCA for handling contexts involving complex data formats, e.g. graphs or relational data. Among them, Relational Concept Analysis (RCA) is a process for analyzing objects described both by binary and relational attributes [116]. The RCA process takes as input a collection of contexts and of inter-context relations, and yields a set of lattices, one per context, whose concepts are linked by relations. RCA has an important role in KDDK, especially in text mining [85], [84].

Another extension of FCA is based on Pattern Structures (PS) [90], which allows to build a concept lattice from complex data, e.g. nominal, numerical, and interval data. In (major [5]), pattern structures are used for building a concept lattice from intervals, in full compliance with FCA, thus benefiting of the efficiency of FCA algorithms. Actually, the notion of similarity between objects is closely related to these extensions of FCA: two objects are similar as soon as they share the same attributes (binary case) or attributes with similar values or the same description (at least in part). Various results were obtained in the study of the relations existing between FCA with an embedded explicit similarity measure and FCA with pattern structures [48]. Moreover, similarity is not a transitive relation and this lead us to the study of tolerance relations. In addition, a new research perspective is aimed at using frequent itemset search methods for mining interval-based data being guided by pattern structures and biclustering as well [50], [49].

Pattern structures in association with a similarity measure were applied in the field of decision support in agronomy. In this domain, a set of agro-ecological indicators is aimed at helping farmers to improve their agricultural practices by estimating the impact of cultivation practices on the “agrosystem”. The modeling and the assessment of environmental risk require a large number of parameters whose measure is imprecise. The propagation of the imprecision and the different types of imprecision have to be taken into account in the computation of the value of indicators for decision support. Actually, based on pattern structures with a associated similarity measure, this problem has been approached as an information fusion problems with substantial results [34], [35].

6.1.2. Miscellaneous in FCA and Pattern Mining

In the field of medicine, an approach based on a combination of FCA with sequential pattern mining was developed to explore patients care trajectories (PCT) [46]. When PCT are modeled as multidimensional and multilevel sequences [108], the results of a frequent sequential itemsets search feed an FCA step in order to compute interests measures such as concept stability. These measures help the experts to find the most interesting sequential patterns.

In the context of environmental sciences, research work is in concern with the mining of complex hydroecological data with concept lattices. FCA was compared and combined with statistical approaches to deal with multi-valued contexts in hydroecology [31], [27], [39]. Regarding the preparation of agronomical data, we have developed an episode-based analysis about the design of information systems (actually, this work was carried out during the ANR-ADD COPT project between 2005 and 2008). We focused on the experience of persons in charge of building *observatoires*, i.e. information systems, for the monitoring and the management of rural territories [32]. Moreover, Florence Le Ber –as a member of UMR 7517 Lhyges, Strasbourg– is the scientific head of an ANR project named "FRESQUEAU" (2011–2014) dealing with FCA and data mining and hydroecological data (see http://fresqueau.engees.eu/).

For completing the work on itemset search, there is still on-going work on frequent and rare itemset search, for being able to build lattices from very large data and completing the algorithm collection of the Coron platform. This year, results were obtained on the design of an integrated and modular algorithm for searching for closed and generators itemsets, and equivalence classes of itemsets, thus enabling the construction of the associated lattice [56]. This research aspect is also linked to the research carried on within a the PICS CaDoE research project (see Section 8.1.3).

6.1.3. Skylines, sequences and privacy

Pattern discovery is at the core of numerous data mining tasks. Although many methods focus on efficiency in pattern mining, they still suffer from the problem of choosing a threshold that influences the final extraction result. The goal of a study done this current year (2011) is to make the results of pattern mining useful from a user-preference point of view. That is, take into account some domain knowledge to guide the pattern mining process. To this end, we integrate into the pattern discovery process the idea of skyline queries in order to mine *skyline patterns* in a threshold-free manner. This forms the basis for a novel approach to mining skyline patterns. The efficiency of our approach was illustrated over a use case from *chemoinformatics* and we showed that small sets of dominant patterns are produced under various measures that are interesting for chemical engineers and researchers [55].

Sequence data is widely used in many applications. Consequently, mining sequential patterns and other types of knowledge from sequence data has become an important data mining task. The main emphasis has been on developing efficient mining algorithms and effective pattern representation.

However, important fundamental problems still remained open: (i) given a sequence database, can we have an upper bound on the number of sequential patterns in the database? (ii) Is the efficiency of the sequence classifier only based on accuracy? (iii) Do the classifiers need the entire set of extracted patterns or a smaller set with the same expressiveness power?

In three different works on sequences, we study the problem of bounding sequential patterns with the combinatorial complexity of sequences and the problem of sequence classifiers with the constraints of optimizing both accuracy and earliness [53], [46].

Orpailleur is one of the few project-teams working on privacy challenges which are becoming a core issue with different scientific problems in computer science. Privacy-preserving data publication has been studied intensely in the past years. In our recent works, we introduce two different data anonymization methodologies based on different usability scenarios [57], [58].

6.1.4. KDDK in Text Mining

Ontologies help software and human agents to communicate by providing shared and common domain knowledge, and by supporting various tasks, e.g. problem-solving and information retrieval. In practice, building an ontology depends on a number of "ontological resources" having different types: thesaurus, dictionaries, texts, databases, and ontologies themselves. We are currently working on the design of a methodology and the implementation of a system for ontology engineering from heterogeneous ontological resources. This methodology is based on both FCA and RCA, and was previously successfully applied in contexts such as astronomy and biology. At present, an engineer is in charge of implementing a robust system

being guided by the previous research results and preparing the way for some new research directions involving trees and graphs.

In another work in text mining [19], we propose a method based on a syntactic parsing for extracting rich semantic relationships between pairs of entities co-occurring in a single sentence. The method was applied in pharmacogenomics (study of the impact of individual genomic variation on drug responses) and we obtained a resource encoded in RDF that summarizes pharmacogenomics relationships mentioned into roughly 17 million Medline abstracts. This resource appears to be of major interest since it is used to guide human curation of biomedical databases, and to derive new knowledge about drug-drug interactions [92].

6.2. KDDK in Life Sciences

Participants: Mehwish Alam, Yasmine Assess, Sid-Ahmed Benabderrahmane, Emmanuel Bresso, Thomas Bourquard, Adrien Coulet, Sébastien Da Silva, Marie-Dominique Devignes, Anisah Ghoorah, Renaud Grisoni, Mehdi Kaytoue, Jean-François Kneib, Florence Le Ber, Bernard Maigret, Jean-François Mari, Lazaros Mavridis, Amedeo Napoli, Violeta Pérez-Nueno, Dave Ritchie, Malika Smaïl-Tabbone, Vishwesh Venkatraman.

One of the major challenges in the post genomic era consists in analyzing terabytes of biological data stored in hundreds of heterogeneous databases (DBs). The extraction of knowledge units from these large volumes of data would give sense to the present data production effort with respect to domains such as disease understanding, drug discovery, and pharmacogenomics or systems biology. Research reported here addresses these important issues and shows the spreading of KDDK over such domains.

6.2.1. Ontology-based Functional Classification of Genes

Functional classification involves grouping genes according to their molecular functions or the biological processes they participate in. This unsupervised classification task is essential for interpreting gene datasets produced by postgenomic experiments. As the functional annotation of genes is mostly based on the Gene Ontology (GO), many similarity measures using the GO have been described, but few of them have been used for clustering [107]. We have evaluated a functional classification of genes using our previously described IntelliGO semantic similarity measure with the help of reference sets [38]. The IntelliGO measure computes semantic similarity between genes for discovering biological functions shared by genes and takes into account domain knowledge represented in Gene Ontology [82]. The reference sets consist of genes taken from human and yeast KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways and Pfam clans. Hierarchical clustering and heatmap visualization were used to illustrate the advantages of IntelliGO over several other measures. Because genes often belong to more than one reference set, the fuzzy C-means clustering algorithm was then applied to the datasets using IntelliGO. The F-score method was used to estimate the quality of clustering and the optimal number of clusters. The results were compared with those obtained from the state of the art DAVID (Database for Annotation Visualization and Integrated Discovery) functional classification method. Overlap analysis allows to study the matching between clusters and reference sets, and leads us to propose a set-difference method for discovering missing information [38]. The IntelliGO similarity measure, the clustering tool and the reference sets used for the evaluation are available at http://plateforme-mbi.loria.fr/intelligo.

6.2.2. Use of Domain Knowledge for Dimension Reduction

Data complexity is a major challenge for knowledge discovery approaches. High dimensionality of datasets can impair the execution of most data mining programs and/or lead to the production of numerous and complex patterns, improper for interpretation by the supervising expert. Thus, an important research orientation is dimension reduction as part of the data preparation step [93]. Domain knowledge is essential for achieving such dataset modification with minimal loss of information. The Life Sciences constitute a suitable domain for testing knowledge-guided approaches for dimension reduction because of the continuous increase in the number of both complex datasets and bio-ontologies. Most of these bio-ontologies are used for annotating biological objects leading to high-dimensional datasets. We propose a new approach for reducing dimensions

in a dataset by exploiting semantic relationships between terms of an ontology structured as a rooted directed acyclic graph [40]. Term clustering is performed thanks to the IntelliGO similarity measure and the term clusters are further used as descriptors for data representation. The technique was applied to a set of drugs associated with their side effects collected from the SIDER database. Terms describing side effects belong to the MedDRA terminology. The hierarchical clustering of about 1,200 MedDRA terms into an optimal collection of 112 term clusters led to a reduced data representation. Two data mining experiments were conducted to illustrate the advantage of using such reduced data representation.

Results obtained in the frame of the collaborative Grand Challenge project (see previous report 2009 and 2010) have been published this year. We have designed the HIV-PDI (Protein-Drug Interactions) resource as a decision making tool to propose alternative antiretroviral drugs (ARVs) for personalized antiretroviral treatment [22]. The HIV-PDI is an integrated database in which sequence mutations of viral proteins can be mapped onto three-dimensional structural interactions between these proteins and ARVs. Thus, critical loss of interactions leading to resistance can be detected and serve as indicators for proposing appropriate ARVs escaping the resistance. As a first step, the HIV-PDI was populated with data relating to HIV protease: clinical information on patients, resistance to ARVs treatments, HIV protease structures and mutations, ARV drugs and their 3D interactions with HIV protease models. Possible queries include protein, drug and treatment conditions, coupled with dedicated tools for visualization/analysis of 3D Protein-Drug interactions. Case-studies demonstrate the capabilities of the HIV-PDI resource for retrieving information associated with patients and for analyzing structural data relating proteins and ligands [23].

6.2.3. Mining Agronomical Data with stochastic models

In the framework of agricultural landscape data mining, we have developed an original approach combining two methods used separately so far: the identification of explicit farmer decision rules through on-farm surveys methods and the identification of landscape stochastic regularities through data-mining of the mosaic of agricultural parcels, following preceding work [96]. This approach was assessed in a study on the Niort plain (West of France) database. In this database, provided by the CEBC (UPR CNRS), the land use occupations of the fields covering a $400km^2$ area are recorded during 12 years. It results a segmentation of the landscape, based on both its spatial and temporal organization and partly explained by generic farmer decision rules. This consistency between results points out that the two modelling methods interact and may be combined for land-use modelling at landscape scale and for understanding the driving forces of spatial organization. Based on farm surveys, we were able to retrieve and measure changes in land use occupation and link some farmer decision and spatiotemporal regularities that were observed in the landscapes.

6.3. Structural Systems Biology and Docking

Participants: Thomas Bourquard, Marie-Dominique Devignes, Anisah Ghoorah, Bernard Maigret, Lazaros Mavridis, Violeta Pérez-Nueno, Dave Ritchie, Malika Smaïl-Tabbone, Vishwesh Venkatraman.

Structural systems biology aims to describe and analyze the many components and interactions within living cells in terms of their three-dimensional (3D) molecular structures. Much of our work in this area has been funded by the ANR Chaires d'Excellence project entitled "High Performance Algorithms for Structural Systems Biology" (HPASSB) which was awarded to Dave Ritchie (January 2009 – September 2011). A related follow-on ANR project entitled "Polynomial Expansions of Protein Structures and Interactions" (PEPSI) has recently started (November 2011). The HPASSB project complements existing competencies in the Orpailleur team represented by Marie-Dominique Devignes (CR CNRS) who is coordinating the MBI project (Modelling Biomolecules and their Interactions, http://bioinfo.loria.fr), Malika Smaïl-Tabbone (MCU Nancy University) who is working on data integration and relational data-mining approaches, and Bernard Maigret (DR CNRS) who has an extensive experience of molecular dynamics and virtual screening. We are currently developing advanced computing techniques for molecular shape representation, protein-protein docking, protein-ligand docking, high-throughput virtual drug screening, and knowledge discovery in databases dedicated to protein-protein interactions. The PEPSI project is a collaboration with Sergei Grudinin at INRIA Grenoble (project Nano-D) and Valentin Gordeliy at the Institut de Biologie Structurale in Grenoble. This new project will

involve developing further the above techniques and using them to help solve the structures of large molecular systems experimentally.

6.3.1. Accelerating protein docking calculations using graphics processors

We have recently adapted the *Hex* protein docking software to use modern graphics processors (GPUs) to carry out the expensive FFT part of a docking calculation [115]. Compared to using a single conventional central processor (CPU), a high-end GPU gives a speed-up of 45 or more. This software is publicly available at http://hex.loria.fr. A public GPU-powered server has also been created (http://hexserver.loria.fr) [99]. These advances have facilitated further work on modeling the assembly of multi-component molecular structures using a particle swarm optimization technique [69].

6.3.2. Eigen-Hex: Modeling protein flexibility during docking

Although the *Hex* protein docking software can often make reasonably good predictions about how two proteins might fit together, a major limitation of many current algorithms, including *Hex*, is that that they assume that proteins are rigid objects. In fact, proteins can be highly flexible, and the internal conformations of their atoms often change on going from the unbound forms in the free proteins to the bound conformations in the complex. We have developed a novel approach to model such flexibility using a principal component analysis (PCA) technique to identify and predict the main atomic motions during a docking calculation. Our approach gives better results than rigid body docking, although the flexible docking problem is still by no means solved. A journal article describing this work has been submitted.

6.3.3. 3D-Blast: A new approach for protein structure alignment and clustering

We recently developed a new sequence-independent protein structure alignment approach called 3D-Blast [102], which exploits the spherical polar Fourier (SPF) correlation technique used in the *Hex* protein docking software [114]. This approach recently performed very well in a blind shape comparison experiment organized by Orpailleur as part of Eurographics Workshop on 3D Object Retrieval [103]. The utility of this approach has been demonstrated by clustering subsets of the CATH protein structure classification database [106] for each of the four main CATH fold types, and by searching the entire CATH database of some 12,000 structures using several protein structures as queries. Overall, the automatic SPF clustering approach agrees very well with the expert-curated CATH classification, and ROC-plot analysis of database searches show that the approach has very high precision and recall. We recently proposed that the 3D-Blast approach could ultimately provide a novel way to enumerate and index protein fold space (major [7]).

6.3.4. KBDOCK: Protein docking using Knowledge-Based approaches

Protein docking is the difficult computational task of predicting how a pair of three-dimensional protein structures come together to form a complex. Historically, there has been considerable interest in developing *ab initio* docking algorithms such as the *Hex* docking program developed by Dave Ritchie. However, as structural genomics initiatives continue to populate the space of protein 3D structures, and as several on-line databases of protein interactions have recently become available, using structural database systems to perform docking by homology will become an increasingly powerful approach to predicting protein interactions. In order to explore such possibilities, Anisah Ghoorah has recently developed the KDBOCK system as part of her doctoral thesis project. KDBOCK combines residue contact information from the 3DID database [117] with the Pfam protein domain family classification [89] together with coordinate data from the Protein Data Bank [86] in order to describe and analyze all known protein-protein interactions for which the 3D structures are available. In a recent publication [24] we demonstrated the utility of this approach for template-based docking using 73 complexes from the Protein Docking Benchmark [94]. KBDOCK is available at http://kbdock.loria.fr.

6.3.5. V-Dock: scoring protein-protein interactions using Voronoi fingerprints

There is growing interest in using machine learning techniques to analyze and populate protein-protein interaction (PPI) networks [104]. The aim of this project is to investigate the use of Voronoi fingerprints [16] as a way to distinguish cognate and non-cognate pairs of protein-protein interfaces. In collaboration with colleagues in the INRIA AMIB and INRA Bios teams, we recently applied our Voronoi fingerprint representation (V-Dock) to re-score rigid body docking predictions from *Hex* [60], and we demonstrated that it could be used to improve the ranking of 7 out of 9 docking targets from the CAPRI protein docking experiment [60]. This approach was also used to predict the stability of engineered protein structures for another recent CAPRI target [21].

6.3.6. DOVSA: Developing new algorithms for virtual screening

In 2010, Violeta Pérez-Nueno joined the Orpailleur team thanks to a Marie Curie Intra-European Fellowship (IEF) award to develop new virtual screening algorithms (DOVSA). The aim of this project is to advance the state of the art in computational virtual drug screening by developing a novel consensus shape clustering approach based on spherical harmonic (SH) shape representations [110]. The main disease target in this project is the acquired immune deficiency syndrome (AIDS), caused by the human immuno-deficiency virus (HIV) [109]. However, the approach will be quite generic and will be broadly applicable to many other diseases. So far, good progress has been made on calculating and clustering spherical harmonic "consensus shapes" which represent rather well the essential features of groups of active molecules [30]. Recent progress on this project has been presented orally at the 5th Journée Nationale de Chémoinformatique in Cabourg, the 9th International Conference on Chemical Structures in Noordwijkerhout, and at 3rd International Conference on Drug Discovery and Therapy in Dubai. A review of the state of the art in drug promiscuity was also recently published [29].

6.4. Around the Kasimir research project

Participants: Nicolas Jay, Jean Lieber, Bart Lamiroy, Amedeo Napoli, Thomas Meilender.

This special research project involves researchers working around the Kasimir project and Bart Lamiroy who was attached to the Orpailleur Team during his "INRIA délégation" (2010-2011) and at the same time was a visiting scientist at Lehigh University, USA. The background of Bart Lamiroy is in document and image analysis. Recently he was interested in investigating the application of KDDK to numerical and structural data including document images. The objective is to extend mining tools towards complex and semi-structured multi-media data on the one hand, and to associate image analysis with KDDK techniques on the other hand.

The main research direction which is followed at the moment is in concern with the Kasimir project. Actually, oncology protocols are mainly documented and represented in diagram formats. The classification and CBR techniques used in the Kasimir project require that the ontologies and decision protocols have to be represented in OWL. Based on previous work, we started modeling the mapping of visual features in diagram charts with semantics of the medical domain ontology. The mapping between the visual ontology and the domain ontology should guide a more complete extraction of the protocols from the diagrams for completing the domain ontology of the Kasimir system.

Moreover, during his stay at Lehigh University, Bart Lamiroy developed a new approach for recovering useful information within image data. By recording a wide range of "provenance information" related to complex image analysis processes, the DAE platform (http://dae.cse.lehigh.edu) provides a large set of metadata that can be used by KDDK methods. For example, this allows the correlation and combination of numerical and symbolic aspects, e.g. relating image aspects and domain symbolic representations (within domain ontologies). This work bridges the gap between formal knowledge representation and signal-based pattern recognition and offers a robust experimental environment for further application of KDDK on image data.

6.5. Around the Taaable research project

Participants: Julien Cojan, Valmi Dufour-Lussier, Inaki Fernandez, Emmanuelle Gaillard, Laura Infante-Blanco, Florence Le Ber, Jean Lieber, Amedeo Napoli, Emmanuel Nauer, Yannick Toussaint.

The Taaable project (http://taaable.fr) has been originally created as a challenger of the Computer Cooking Contest (CCC, organized during the ICCBR Conference). A candidate to this contest is a system whose goal is to solve cooking problems on the basis of a recipe book (common to all candidates), where each recipe is a shallow XML document with an important plain text part. The size of the recipe book (about 1500 recipes) prevents from a manual indexing of recipes: this indexing is performed using semi-automatic techniques.

After being ranked twice second, in the 2008 and 2009 CCCs organized during the ICCBR conference, Taaable won the first price and the adaptation challenge, in 2010 (note that no contest was organized in 2011). Beyond its participation to the CCCs, the Taaable project aims at federating various research themes: case-based reasoning, information retrieval, knowledge acquisition and extraction, knowledge representation, minimal change theory, ontology engineering, semantic wikis, text-mining, etc.

The most important original features of this version are:

A module for refining the domain ontology for improving the case retrieval. In Taaable, the retrieval of similar cases is based on a query generalization using an ontology of the cooking domain. In order to make the case retrieval more progressive and more precise, a enrichment of the domain ontology, and especially the ingredient hierarchy, has been studied and implemented [42]. The refinement process consists in inserting intermediate classes into the initial hierarchy of the system for better distinguishing classes that were initially not distinguishable. In order to introduce new classes into the initial hierarchy, the initial classes of ingredients have been characterized with additional properties. These additional properties are cooking actions that can be applied to ingredients and that have been extracted from the texts of recipes. FCA has been used on these new properties for restructuring the initial hierarchy.

A module for computing adaptation knowledge. Adaptation knowledge discovery has been performed for better adapting cooking recipes to user constraints. This paper extends the approach proposed in 2009 [80] for extracting this kind of adaptation knowledge. The adaptation knowledge comes from the interpretation of closed itemsets whose items correspond to the ingredients that have to be removed, kept, or added. An original approach focusing on a restrictive binary context building and on a specific ranking based on the form of the closed itemsets has been proposed [47].

Several theoretical studies have been carried out that should be applied to some future versions of Taaable:

- The representation of preparations in temporal representation formalisms [63].
- An algorithm for adapting cases defined in the expressive description logic $\mathcal{ALC}$ [43], [11].
- The study of the relations between adaptation based on belief revision and other approaches to adaptation [61], [11].
- The study of the extension of the domain ontology to make the retrieval step of a case-based reasoning system more accurate [42].
- The study of adaptation knowledge discovery based on variation of ingredients between pairs of recipes [42].

7. Contracts and Grants with Industry

7.1. The BioIntelligence Project

Participants: Mehwish Alam, Isiru Bayissa, Aleksey Buzmakov, Adrien Coulet, Marie-Dominique Devignes, Mehdi Kaytoue, Luis Felipe Melo, Amedeo Napoli [contact person], Chedy Raïssi, Malika Smaïl-Tabbone.

The objective of the "BioIntelligence" project is to design an integrated framework for the discovery and the development of new biological products. This framework takes into account all phases of the development of a product, from molecular to industrial aspects, and is intended to be used in life science industry (pharmacy, medicine, cosmetics, etc.). The framework has to propose various tools and activities such as: (1) a platform for searching and analyzing biological information (heterogeneous data, documents, knowledge sources, etc.), (2) knowledge-based models and process for simulation and biology in silico, (3) the management of all activities related to the discovery of new products in collaboration with the industrial laboratories (collaborative work, industrial process management, quality, certification). The "BioIntelligence" project is led by "Dassault Systèmes" and involves industrial partners such as Sanofi Aventis, Laboratoires Pierre Fabre, Ipsen, Servier, Bayer Crops, and two academics, Inserm and Inria. An annual meeting of the project usually takes place in Sophia-Antipolis at the beginning of July.

Three thesis related to "BioIntelligence" are beginning in the Orpailleur team. A first one is in concern with ontology re-engineering in the domain of biology. The objective is consider the content of the BioPortal ontologies and to design formal contexts with which we will be able to build a concept lattice, to be used as a support for an ontology schema. The formal concept is built according to external resources such as Wikipedia and domain knowledge as well.

A second thesis is related to the study of possible combination of mining methods on biological data. The mining methods which are considered here are based on FCA and RCA, itemset and association rule extraction, and inductive logic programming. These methods have their own strengths and provide different special capabilities for extending domain ontologies. A particular attention will be paid to the integration of heterogeneous biological data and the management of a large volume of biological data while being guided by domain knowledge lying in ontologies (linking data and knowledge units). Practical experiments will be led on biological data (clinical trials data and cohort data) also in accordance with ontologies lying at the NCBO BioPortal.

A third thesis is based on an extension of FCA involving Pattern Structures on Graphs. The idea is to be able to extend the formalism of pattern structures to graphs and to apply the resulting framework on molecular structures. In this way, it will be possible to classify molecular structures and reactions by their content. This will help practitioners in information retrieval tasks involving molecular structures or the search for particular reactions.

7.2. The Quaero Project

Participants: Victor Codocedo [contact person], Amedeo Napoli.

The Quaero project (http://www.quaero.org) is a program aimed at promoting research and industrial innovation on technologies for automatic analysis and classification of multimedia and multilingual documents. The partners collaborate on research and the realization of advanced demonstrators and prototypes of innovating applications and services for access and usage of multimedia information, such as spoken language, images, video and music.

In this framework, the Orpailleur team participates in the task called "Formal Representation of Knowledge for Guiding Recommendation", whose objectives are to define methods and algorithms for building a "discovery engine" guided by domain knowledge and able to recommend a user some content to visualize. Such a discovery engine has to extend capabilities of usual recommender systems with a number of capabilities, e.g. to select among a huge amount of items (e.g. movie, video, music) those which are of interest for a user according to a given profile. In addition, the discovery engine should take into account contextual information that can be of interest such as news, space location, moment of the day, actual weather and weather forecast, etc. This contextual information changes within time and extracted information has to be continuously updated. Finally, the system has be able to justify or explain the recommendations.

A thesis takes place in the context of the Quaero project. At the moment, document annotation is especially studied for enhancing recommendation but also information retrieval. Information retrieval guided by domain knowledge can be used for selecting resources of interest for these two tasks. Then knowledge discovery based

on Formal Concept Analysis can be used for extracting patterns of interest w.r.t. the context and for enriching the domain and contextual knowledge base.

Finally, the discovery process has to be able to act as a classifier and as an inference engine at the same time for reasoning and classifying elements for recommendation and retrieval.

8. Partnerships and Cooperations

8.1. International projects and collaborations

8.1.1. Fapemig INRIA Project: Incorporating knowledge models into scalable data mining algorithms

Participants: Mehdi Kaytoue, Amedeo Napoli [contact person], Chedy Raïssi.

This Fapemig – INRIA research project involves researchers at Universidade Federal de Minas Gerais in Belo Horizonte –a group led by Prof. Wagner Meira– and the Orpailleur team at INRIA Nancy Grand Est. In this project we are interested in the mining of large amount of data and we target two relevant application scenarios where such issue may be observed. The first one is text mining, i.e. extracting knowledge from texts and document categorization. The second application scenario is graph mining, i.e. determining relationship-based patterns and use these relations to perform classification tasks. In both cases, the computational complexity is large either because the high dimensionality of the data or the complexity of the patterns to be mined.

One strategy to ease the execution of such data mining tasks is to use existing knowledge to restrict the search space and to assess the quality of the patterns found. This existing knowledge may be formalized in ontologies but also in other ways whose study is a research issue in this project. Once we are able to build knowledge models, we need to determine how to use such knowledge models, which is a second major research issue in this project. In particular, we want to design and evaluate mechanisms that allow the exploitation of existing knowledge for sake of improving data mining algorithms.

Finally, the computational complexity of the algorithms remains a major issue and we intend to address it through parallel algorithms. Data mining algorithms, in general, represent a challenge for sake of parallelization because they are irregular and intensive in terms of both computing and communication. Accordingly, in a first joint work, we developed a new parallel algorithm to build skycubes based on the Anthill framework developed at UFMG. The paper was presented in a local Brazilian Conference and an extended journal version will appear in a 2012 special issue of the International Journal of Parallel Programming.

8.1.2. Search for anti-HIV drugs acting as entry-blockers

Participants: Thomas Bourquard, Marie-Dominique Devignes, Anisah Ghoorah, Lazaros Mavridis, Violeta Pérez-Nueno, Dave Ritchie, Malika Smaïl-Tabbone, Vishwesh Venkatraman.

In collaboration with computational chemistry colleagues at the University of Bari and the Institut Chimique de Saria (IQS) in Barcelona, Dave Ritchie has published reviews of the state of *in silico* protein structure modeling and *virtual drug screening* techniques for the CCR5 [87], and CXCR4 [111], entry-blocking molecules. As there now exist several hundred such entry-blockers, there is considerable interest in the chemoinformatics community in how best to use knowledge of known drug molecules to develop new and more potent new drug candidates [112]. The spherical harmonic clustering approach developed by Dave Ritchie and Violeta Pérez-Nueno was recently used successfully in a virtual screening study at the IQS to discover new high-affinity ligands for CXCR4 [109].

8.1.3. International collaborations in Mining complex data

Participants: Isiru Bayissa, Adrien Coulet, Mehdi Kaytoue, Amedeo Napoli, Chedy Raïssi.

A first collaboration involves "Université du Québec à Montréal" (UQAM) in Montréal with Prof. Petko Valtchev and Laboratoire LIRMM in Montpellier with Prof. Marianne Huchard. This collaboration is supported by a CNRS PICS project (2011-2014), which is called "Concept Analysis driving Ontology Engineering" and abbreviated in "CAdOE". The research work within this project is aimed at defining and implementing a semi-automatic methodology supporting ontology engineering based on the joint use of Formal Concept Analysis (FCA) and Relational Concept Analysis (RCA). At the moment, some elements of this methodology are existing and were used in text mining [85], [84]. However, the first methodology should be completed and improved, especially regarding the applicability on complex data and the interoperability with knowledge representation modules. This year, some publications were already obtained and some others are in preparation for next year [36], [56], [75].

A second collaboration involves Sergei Kusnetsov at Higher School of Economics in Moscow (HSE). Mehdi Kaytoue and Amedeo Napoli visited HSE laboratory in July 2010 granted by the Poncelet Laboratory in Moscow, a joint CNRS – INRIA laboratory. This visit was the occasion of preparing a number of publications, among which a publication in a first-rank conference in Artificial Intelligence (major [5]), together with some other important publications [49], [33], [48]. This shows that the collaboration is on-going and that there is still a substantial research work to be done. This year, Amedeo Napoli visited HSE laboratory in June 2011 while Sergei Kuznetsov visited Loria in October 2011.

A third collaboration –a PHC Zenon project– exists with Florent Domenach, associated professor at the University of Nicosia in Cyprus. This project is entitled "Knowledge Discovery for Complex Data in Formal and Relational Concept Analysis" (KD4CD) and is aimed at studying and combining different types of classification process in the framework of FCA. These processes can be based on Galois connections but also on the so-called "overhangings", i.e. a kind of generalization of closure systems. Moreover, another interest is put on consensus theory where the objective is to find the better classification of a set of abjects according to a quality measure (this could be applied to ontologies). This year, there were two visits from France to Cyprus in May and December 2011 while there was one visit from Cyprus to France in October 2011.

8.2. European Initiatives

8.2.1. FP7 Projet: DOVSA

- Title: Development of Virtual Screening Algorithms: Exploring Multiple Ligand Binding Modes Using Spherical Harmonic Consensus Clustering.
- Type: PEOPLE.
- Instrument: Marie Curie Intra-European Fellowships for Career Development (IEF).
- Duration: July 2010 – July 2012.
- Coordinator: INRIA Nancy Grand-Est (France).
- Others partners: None.
- Abstract (see also Section 6.3.6 of this document):

  This project will advance the state of the art in virtual drug screening by developing novel spherical harmonic-based consensus clustering algorithms. The main disease that will be targeted in this project is the acquired immune deficiency syndrome (AIDS), caused by the human immunodeficiency virus (HIV). However, the approach will be quite generic and will be broadly applicable to many other diseases. The approach will be tested and validated using 40 well-known drug targets from the DUD dataset. It will then be used to screen the French Chimiothèque Nationale library of some 36000 compounds for novel ligands which will bind the CCR5 co-receptor and hence block HIV infection. A small list of candidate entry-blocking compounds will be sent to Barcelona for experimental testing. By extending the SH-based consensus clustering technique, this project will provide a generic tool to help deal with cases where multiple ligands may be associated with multiple pocket sub-sites or which may bind multiple targets, and it will help to find new HIV entry-blocking compounds.

8.3. National Initiatives

8.3.1. ANR Kolflow: man-machine collaboration in continuous knowledge-construction flows

Participants: Jean Lieber [contact person], Amedeo Napoli, Emmanuel Nauer, Julien Stévenot, Yannick Toussaint.

Kolflow (http://kolflow.univ-nantes.fr/) is a 3-years basic research project taking place from February 2011 to July 2014, funded by French National Agency for Research (ANR), program ANR CONTINT. The aim of the project is investigation on man-machine collaboration in continuous knowledge-construction flows. Kolflow partners are Edelweiss (INRIA Sophia Antipolis), GDD (LINA Nantes), Silex (LIRIS Lyon), Orpailleur, and Score (LORIA).

8.3.2. ANR Trajcan: a study of patient care trajectories

Participants: Elias Egho, Nicolas Jay [contact person], Amedeo Napoli, Chedy Raïssi.

Since 30 years, many patient classification systems (PCS) have been developed. These systems aim at classifying care episodes into groups according to different patient characteristics. In France, the so-called "Programme de Médicalisation des Systèmes d'Information" (PMSI) is a national wide PCS in use in every hospital. It systematically collects data about millions of hospitalizations. Though it is used for funding purposes, it includes useful knowledge for other public health domains such as epidemiology or health care planning.

The objective of the Trajcan project is to represent and analyze "patient care trajectories" (patient suffering from cancer limited to breast, colon, rectum, and lung cancers) and the associated healthcares. The data are related to patients receiving hospital cares in the "Bourgogne" region and using data from the PMSI. Such an analysis involves various data, e.g. type of cancer, number of visits, type of stays, hospitalization services and therapies used, and demographic factors, i.e. age, gender, place of residence.

One thesis is currently carried out on this subject whose objective is to design a knowledge discovery system working on multidimensional and sequential data for characterizing Patient Care Trajectories (PCT). This thesis combines knowledge discovery and knowledge representation methods for improving the definition of patient care trajectories as temporal objects (sequential data mining). The overall objective id to provide in decision support for improving healthcare in detecting for example typical or exceptional trajectories for planning with precision healthcare for a given population. In order to discover groups of patients showing similar health condition, treatments or journeys through the healthcare system, PCT are mined with multilevel and multidimensional sequential itemsets search, using external knowledge on hospitals, medical procedures and diagnoses. FCA capabilities for dealing with large amounts of data and for filtering (with a measure such as stability) are then used as a post-processing step for selecting the most interesting patterns [46].

8.4. Local initiatives

8.4.1. Contrat Plan État Région" (CPER)

The links between the Regional Administration and LORIA are materialized through an administrative contract called "Contrat Plan État Région" (CPER) running from 2007 to 2013. The associated scientific program is called "Modélisations, informations et systèmes numériques" (MISN) and includes two tracks in which the Orpailleur team is involved.

- "Modeling Bio-molecules and their Interactions" (MBI).

  This project is coordinated by Marie-Dominique Devignes (http://bioinfo.loria.fr) and the general objective is to study how domain knowledge can be taken into account for improving modeling of biomolecules and their interactions, and how, in sequence, this guides the modeling of biological systems. Six scientific projects are currently under development and involve collaborations with computer scientists, and people working either in biology or chemistry.

An INRIA experimental research platform is currently developed in the framework of MBI (http://bioinfo.loria.fr/Plateforme%20MBI). This platform is aimed at sharing data and computing resources. Its specific features are relative to biomolecules modeling, classification, and to data integration for data mining. In parallel with the bioinformatics platforms in Strasbourg, Reims, Lille, and Nancy-INIST, it constitutes the North-East node of RENABI ("Réseau National des Plateformes Bioinformatiques").

- "Traitement Automatique des Langues et des Connaissances" (TALC).

  TALC has to be understood as "Automatic Processing of Languages and Knowledge" and the general objective is to study the relations existing between knowledge discovery, knowledge representation, reasoning, and natural language processing. In this framework, the Orpailleur team plays an important role as the research themes are closely related to those of the team. Actually, research projects are currently under development on knowledge management and decision support in the large involving in particular the Kasimir and the Taaable systems.

8.4.2. Other initiatives

8.4.2.1. Cancéropole Grand-Est

A collaboration with the "Laboratoire de Bioinformatique et Génomique Intégratives (LBGI)" at IGBMC Strasbourg involves a thesis funded by INCa ("Institut National du Cancer") with a bipartite direction. This thesis is considered as one research operation within the annual meeting of "Canceropole Grand-Est".

8.4.2.2. BioProLor

The Orpailleur team is member of the BioProLor consortium composed of 5 enterprises and 7 academic research teams. This consortium is funded for 2 years (2010-2012) by the AME ("Agence pour la Mobilisation Economique"). The objective of BioProLor is the design of a production filière for compounds with high added-value which originate from plants in Lorraine. The Orpailleur team and the associated start-up "Harmonic Pharma" are in charge of the computational aspects of this research work.

In addition, a CIFRE contract was set up with Harmonic Pharma for funding the thesis of Emmanuel Bresso on the following subject: "Organisation et exploitation des connaissances sur les réseaux d'interactions biomoléculaires pour l'identification de gènes candidats et la caractérisation de profils pharmacologiques et effets secondaires de principes actifs".

9. Dissemination

9.1. Scientific Animation

- The scientific animation in the Orpailleur team is based on two seminars, the Team Seminar and the BINGO seminar. The Team Seminar is held at least twice a month and is used either for general presentations of people in the team or for inviting external researchers for general interest. The BINGO seminar is held also at least twice a month and is used for more specific presentations focusing on biological, chemical, and medical topics. Actually, both seminars are active and are useful instruments for researchers in the team.
- Members of the Orpailleur team are all involved, as members or as head persons, in various national research groups (mainly GDR CNRS I3 and BIM).
- The members of the Orpailleur team are involved in the organization of conferences, as members of conference program committees (ECAI, PKDD, ICFCA ...), as members of editorial boards, and finally in the organization of journal special issues.

- This year, the team was deeply involved in the organization of CLA 2011, the 8th International Conference on Concept Lattices and their Applications, held at LORIA between October 17th and October 20th (see http://cla2011.loria.fr/). Amedeo Napoli was the president of the organization committee and one of the two chairman of the Conference.

9.2. Teaching

- The members of the Orpailleur team are involved in teaching at all levels of teaching in the universities of Nancy, especially in Nancy Université including “Université Henri Poincaré Nancy-1”, “Université de Nancy-2”, “Institut Polytechnique de Lorraine”. Actually, most of the members of the Orpailleur team are employed on university positions.
- The members of the Orpailleur team are also involved in student supervision, at all university levels, from under-graduate until post-graduate students.
- Finally, the members of the Orpailleur team are involved in HDR and thesis defenses, being thesis referees or thesis committee members.

10. Bibliography

Major publications by the team in recent years

[1] J. COJAN, J. LIEBER. *An Algorithm for Adapting Cases Represented in ALC*, in "22nd International Joint Conference on Artificial Intelligence - IJCAI 2011", Barcelone, Spain, July 2011, http://hal.inria.fr/inria-00584103/en.

[2] V. DUFOUR-LUSSIER, J. LIEBER, E. NAUER, Y. TOUSSAINT. *Improving case retrieval by enrichment of the domain ontology*, in "19th International Conference on Case Based Reasoning - ICCBR’2011", London, United Kingdom, 2011, http://hal.inria.fr/inria-00617621/en.

[3] A. GHOORAH, M.-D. DEVIGNES, M. SMAÏL-TABBONE, D. RITCHIE. *Spatial clustering of protein binding sites for template based protein docking*, in "Bioinformatics", August 2011 [*DOI :* 10.1093/BIOINFORMATICS/BTR493], http://hal.inria.fr/inria-00617921/en.

[4] C. GRAC, A. BRAUD, F. LE BER, M. TRÉMOLIÈRES. *Un système d’information pour le suivi et l’évaluation de la qualité des cours d’eau – Application à l’hydro-écorégion de la plaine d’Alsace*, in "RSTI - Ingénierie des Systèmes d’Information", 2011, vol. 16, p. 9-30, http://hal.archives-ouvertes.fr/hal-00601991/en/.

[5] M. KAYTOUE, S. O. KUZNETSOV, A. NAPOLI. *Revisiting Numerical Pattern Mining with Formal Concept Analysis*, in "Twenty second International Joint Conference on Artificial Intelligence - IJCAI 2011", Barcelona, Spain, 2011, http://hal.inria.fr/inria-00584371/en.

[6] J.-F. MARI, F. LE BER, E.-G. LAZRAK, M. BENOÎT, C. ENG, A. THIBESSARD, P. LEBLOND. *Using Markov Models to Mine Temporal and Spatial Data*, in "New Fundamental Technologies in Data Mining", K. FUNATSU, K. HASEGAWA (editors), Intech, 2011, p. 561–584, http://hal.inria.fr/inria-00566801/en.

[7] L. MAVRIDIS, A. GHOORAH, V. VENKATRAMAN, D. RITCHIE. *Representing and comparing protein folds and fold families using 3D shape-density representations*, in "Proteins", November 2011 [*DOI :* 10.1002/PROT.23218], http://hal.inria.fr/hal-00641815/en.

[8] C. RAÏSSI, J. PEI. *Towards Bounding Sequential Patterns*, in "17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD-2011", San Diego, United States, C. APTÉ, J. GHOSH, P. SMYTH (editors), ACM, August 2011, ISBN : 978-1-4503-0813-7 [*DOI :* 10.1145/2020408.2020612], http://hal.inria.fr/inria-00623550/en.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[9] Y. ASSESS. *Conception par modélisation et criblage in solico d'inhibiteurs du récepteur c-Met*, Université Henri Poincaré - Nancy I, October 2011.

[10] S. BENABDERRAHMANE. *Prise en compte des connaissances du domaine dans l'analyse transcriptomique : Similarité sémantique, classification fonctionnelle et profils flous. Application au cancer colorectal.*, Université Henri Poincaré - Nancy I, December 2011, http://tel.archives-ouvertes.fr/tel-00653169/en/.

[11] J. COJAN. *Application de la théorie de la révision des connaissances au raisonnement à partir de cas*, Université Henri Poincaré - Nancy I, October 2011, http://hal.inria.fr/tel-00646841/en.

[12] M. KAYTOUE. *Traitement de données numériques par analyse formelle de concepts et structures de patrons*, Université Henri Poincaré - Nancy I, April 2011, http://hal.inria.fr/tel-00599168/en.

[13] D. RITCHIE. *Algorithmes Haute-Performance pour la Reconnaissance de Formes Moléculaires*, Université Henri Poincaré - Nancy I, April 2011, Habilitation à Diriger des Recherches, http://hal.inria.fr/tel-00587962/en.

[14] Y. TOUSSAINT. *Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances*, Université Henri Poincaré - Nancy I, November 2011, Habilitation à diriger des recherches (HDR).

Articles in International Peer-Reviewed Journal

[15] A. K. R. ABADIO, E. S. KIOSHIMA, M. M. TEIXEIRA, N. F. MARTINS, B. MAIGRET, M. S. S. FELIPE. *Comparative genomics allowed the identification of drug targets against human fungal pathogens.*, in "BMC Genomics", January 2011, vol. 12, 75 [*DOI :* 10.1186/1471-2164-12-75], http://hal.inria.fr/inria-00610630/en.

[16] T. BOURQUARD, J. BERNAUER, J. AZÉ, A. POUPON. *A collaborative filtering approach for protein-protein docking scoring functions.*, in "PLoS ONE", 2011, vol. 6, n^{o} 4, e18541 [*DOI :* 10.1371/JOURNAL.PONE.0018541], http://hal.inria.fr/inria-00625000/en.

[17] M. CHAVENT, B. LÉVY, M. KRONE, K. BIDMON, J.-P. NOMINÉ, T. ERTL, M. BAADEN. *GPU-powered tools boost molecular visualization.*, in "Briefings in Bioinformatics", November 2011, vol. 12, n^{o} 6, p. 689-701 [*DOI :* 10.1093/BIB/BBQ089], http://hal.inria.fr/hal-00645161/en.

[18] M. CHAVENT, A. VANEL, A. TEK, B. LÉVY, S. ROBERT, B. RAFFIN, M. BAADEN. *GPU-accelerated atom and dynamic bond visualization using hyperballs: a unified algorithm for balls, sticks, and hyperboloids.*, in "Journal of Computational Chemistry", October 2011, vol. 32, n^{o} 13, p. 2924-35 [*DOI :* 10.1002/JCC.21861], http://hal.inria.fr/hal-00645162/en.

[19] A. COULET, Y. GARTEN, M. DUMONTIER, R. B. ALTMAN, M. A. MUSEN, N. H. SHAH. *Integration and publication of heterogeneous text-mined relationships on the Semantic Web*, in "Journal of Biomedical Semantics", May 2011, vol. 2, n^{o} S2, S10, http://hal.inria.fr/hal-00585215/en.

[20] C. ENG, A. THIBESSARD, M. DANIELSEN, T. B. RASMUSSEN, J.-F. MARI, P. LEBLOND. *In silico prediction of horizontal gene transfer in Streptococcus thermophilus*, in "Archives of Microbiology", January 2011, vol. 193, n^{o} 4, p. 287-297 [*DOI :* 10.1007/s00203-010-0671-8], http://hal.inria.fr/inria-00569081/en.

[21] S. J. FLEISHMAN, T. A. WHITEHEAD, E.-M. STRAUCH, J. E. CORN, S. QIN, H.-X. ZHOU, J. C. MITCHELL, O. N. A. DEMERDASH, M. TAKEDA-SHITAKA, G. TERASHI, I. H. MOAL, X. LI, P. A. BATES, M. ZACHARIAS, H. PARK, J.-S. KO, H. LEE, C. SEOK, T. BOURQUARD, J. BERNAUER, A. POUPON, J. AZÉ, S. SONER, S. K. OVALI, P. OZBEK, N. B. TAL, T. HALILOGLU, H. HWANG, T. VREVEN, B. G. PIERCE, Z. WENG, L. PÉREZ-CANO, C. PONS, J. FERNÁNDEZ-RECIO, F. JIANG, F. YANG, X. GONG, L. CAO, X. XU, B. LIU, P. WANG, C. LI, C. WANG, C. H. ROBERT, M. GUHAROY, S. LIU, Y. HUANG, L. LI, D. GUO, Y. CHEN, Y. XIAO, N. LONDON, Z. ITZHAKI, O. SCHUELER-FURMAN, Y. INBAR, V. PATAPOV, M. COHEN, G. SCHREIBER, Y. TSUCHIYA, E. KANAMORI, D. M. STANDLEY, H. NAKAMURA, K. KINOSHITA, C. M. DRIGGERS, R. G. HALL, J. L. MORGAN, V. L. HSU, J. ZHAN, Y. YANG, Y. ZHOU, P. L. KASTRITIS, A. M. J. J. BONVIN, W. ZHANG, C. J. CAMACHO, K. P. KILAMBI, A. SIRCAR, J. J. GRAY, M. OHUE, N. UCHIKOGA, Y. MATSUZAKI, T. ISHIDA, Y. AKIYAMA, R. KHASHAN, S. BUSH, D. FOUCHES, A. TROPSHA, J. ESQUIVEL-RODRÍGUEZ, D. KIHARA, P. B. STRANGES, R. JACAK, B. KUHLMAN, S.-Y. HUANG, X. ZOU, S. J. WODAK, J. JANIN, D. BAKER. *Community-Wide Assessment of Protein-Interface Modeling Suggests Improvements to Design Methodology.*, in "Journal of Molecular Biology", September 2011 [*DOI :* 10.1016/J.JMB.2011.09.031], http://hal.inria.fr/inria-00637848/en.

[22] L. GHEMTIO, M. SMAÏL-TABBONE, A. DJIKENG, M.-D. DEVIGNES, L. KEMINSE, P. KELBERT, J. FOKAM, B. MAIGRET, O. OUWE-MISSI-OUKEM-BOYER. *HIV-PDI: A Protein-Drug Interaction Resource for Structural Analyses of HIV Drug Resistance: 1. Concepts and Associated Database*, in "Journal of Health & Medical Informatics", 2011, vol. 2, n^{o} 1, 1000104 [*DOI :* 10.4172/2157-7420.1000104], http://hal.inria.fr/hal-00642539/en.

[23] L. GHEMTIO, M. SOUCHET, A. DJIKENG, L. KEMINSE, P. KELBERT, D. RITCHIE, B. MAIGRET, O. OUWE-MISSI-OUKEM-BOYER. *HIV-PDI: A Protein Drug Interaction Resource for Structural Analyses of HIV Drug Resistance: 2. Examples of Use and Proof-of-Concept*, in "Journal of Health & Medical Informatics", 2011, vol. 2, n^{o} 1, 1000105 [*DOI :* 10.4172/2157-7420.1000105], http://hal.inria.fr/hal-00642567/en.

[24] A. GHOORAH, M.-D. DEVIGNES, M. SMAÏL-TABBONE, D. RITCHIE. *Spatial clustering of protein binding sites for template based protein docking*, in "Bioinformatics", August 2011 [*DOI :* 10.1093/BIOINFORMATICS/BTR493], http://hal.inria.fr/inria-00617921/en.

[25] C. JONQUET, P. LEPENDU, S. FALCONER, A. COULET, N. F. NOY, M. A. MUSEN, N. H. SHAH. *NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources*, in "Journal of Web Semantics", September 2011, vol. 9, n^{o} 3, p. 316-324 [*DOI :* 10.1016/J.WEBSEM.2011.06.005], http://hal.inria.fr/lirmm-00622155/en.

[26] M. KAYTOUE, S. O. KUZNETSOV, A. NAPOLI, S. DUPLESSIS. *Mining gene expression data with pattern structures in formal concept analysis*, in "Information Sciences", 2011, vol. 181, n^{o} 10, p. 1989-2001, http://hal.inria.fr/hal-00541100/en.

[27] S. MARTIN, A. BERTAUX, F. LE BER, E. MAILLARD, G. IMFELD. *Seasonal Changes of Macroinvertebrate Communities in a Stormwater Wetland Collecting Pesticide Runoff From a Vineyard Catchment (Alsace, France)*, in "Archives of Environmental Contamination and Toxicology", 2011, 13 [*DOI :* 10.1007/s00244-011-9687-6], http://hal.inria.fr/hal-00607741/en.

[28] L. MAVRIDIS, A. GHOORAH, V. VENKATRAMAN, D. RITCHIE. *Representing and comparing protein folds and fold families using 3D shape-density representations*, in "Proteins", November 2011 [*DOI :* 10.1002/PROT.23218], http://hal.inria.fr/hal-00641815/en.

[29] V. PÉREZ-NUENO, D. RITCHIE. *Identifying and characterizing promiscuous targets: Implications for virtual screening*, in "Expert Opinion on Drug Discovery", November 2011 [*DOI :* 10.1517/17460441.2011.632406], http://hal.inria.fr/hal-00641835/en.

[30] V. PÉREZ-NUENO, D. RITCHIE. *Using Consensus-Shape Clustering To Identify Promiscuous Ligands and Protein Targets and To Choose the Right Query for Shape-Based Virtual Screening*, in "Journal of Chemical Information and Modeling", May 2011, vol. 51, n^{o} 6, p. 1233-1248 [*DOI :* 10.1021/CI100492R], http://hal.inria.fr/inria-00617922/en.

Articles in National Peer-Reviewed Journal

[31] C. GRAC, A. BRAUD, F. LE BER, M. TRÉMOLIÈRES. *Un système d'information pour le suivi et l'évaluation de la qualité des cours d'eau – Application à l'hydro-écorégion de la plaine d'Alsace*, in "RSTI - Ingénierie des Systèmes d'Information", 2011, vol. 16, p. 9-30, http://hal.inria.fr/hal-00601991/en.

[32] F. LE BER, S. NOGRY, C. BRASSAC, M. BENOÎT. *Capitalisation d'expériences pour la mise en place d'observatoires de pratiques agricoles*, in "Revue internationale de Géomatique", 2011, vol. 21, p. 99–118, http://hal.inria.fr/hal-00576238/en.

International Conferences with Proceedings

[33] P. AGARWAL, M. KAYTOUE, S. KUZNETSOV, A. NAPOLI, G. POLAILLON. *Symbolic Galois Lattices with Pattern Structures*, in "Thirteenth International Conference on Rough Sets, Fuzzy Sets and Granular Computing - RSFDGrC-2011", Moscou, Russian Federation, S. O. KUZNETSOV, D. SLEZAK, D. H. HEPTING, B. G. MIRKIN (editors), Lecture Notes in Computer Science, Springer-Verlag, 2011, vol. 6743, p. 191-198 [*DOI :* 10.1007/978-3-642-21881-1_31], http://hal.inria.fr/hal-00631473/en.

[34] Z. ASSAGHIR, M. KAYTOUE, W. MEIRA, J. VILLERD. *Extracting Decision Trees from Interval Pattern Concept Lattices*, in "The Eighth International Conference on Concept Lattices and their Applications - CLA 2011", Nancy, France, A. NAPOLI, V. VYCHODIL (editors), INRIA Nancy Grand Est - LORIA, 2011, http://hal.inria.fr/hal-00640938/en.

[35] Z. ASSAGHIR, A. NAPOLI, M. KAYTOUE, D. DUBOIS, H. PRADE. *Numerical information fusion: Lattice of answers with supporting arguments*, in "23rd IEEE International Conference on Tools with Artificial Intelligence - ICTAI 2011", Boca-Raton, United States, October 2011, p. 621-628, http://hal.inria.fr/hal-00646484/en.

[36] Z. AZMEH, M. HUCHARD, A. NAPOLI, M. ROUANE-HACENE, P. VALTCHEV. *Querying Relational Concept Lattices*, in "CLA'11: The 8th International Conference on Concept Lattices and their Applications", France, 2011, p. 377-392, http://hal.inria.fr/lirmm-00646409/en.

[37] J. AZÉ, T. BOURQUARD, S. HAMEL, A. POUPON, D. RITCHIE. *Using Kendall-Tau Meta-Bagging to Improve Protein-Protein Docking Predictions*, in "PRIB 2011", DELFT, Netherlands, M. LOOG, L. WESSELS, M. J. REINDERS, D. DE RIDDER (editors), 2011, p. 284-295, http://hal.inria.fr/inria-00628038/en.

[38] S. BENABDERRAHMANE, M.-D. DEVIGNES, M. SMAÏL-TABBONE, A. NAPOLI, O. POCH. *Ontology-based functional classification of genes: evaluation with reference sets and overlap analysis*, in "The second workshop on Integrative Data in Systems Biology held in conjunction with the IEEE BIBM 2011 conference", Atlanta, United States, Z. ZHAO (editor), IEEE Computer Society, 2011, http://hal.inria.fr/hal-00644438/en.

[39] A. BRAUD, C. NICA, C. GRAC, F. LE BER. *A lattice-based query system for assessing the quality of hydro-ecosystems*, in "CLA 2011", Nancy, France, A. NAPOLI, V. VYCHODIL (editors), INRIA NGE et LORIA, 2011, p. 265-277, http://hal.inria.fr/hal-00640048/en.

[40] E. BRESSO, S. BENABDERRAHMANE, M. SMAÏL-TABBONE, G. MARCHETTI, A. S. KARABOGA, M. SOUCHET, A. NAPOLI, M.-D. DEVIGNES. *Use of domain knowledge for dimension reduction: application to mining of drug side effects*, in "International Conference on Knowledge Discovery and Information Retrieval - KDIR 2011", Paris, France, A. FRED (editor), INSTICC, SciTePress Digital Library, 2011, 8, http://hal.inria.fr/hal-00642520/en.

[41] V. CODOCEDO, C. TARAMASCO, H. ASTUDILLO. *Cheating to achieve Formal Concept Analysis over a large formal context*, in "The Eighth International Conference on Concept Lattices and their Applications - CLA 2011", Nancy, France, A. NAPOLI, V. VYCHODIL (editors), INRIA Nancy Grand Est - LORIA, 2011, p. 349–362.

[42] J. COJAN, V. DUFOUR-LUSSIER, E. GAILLARD, J. LIEBER, E. NAUER, Y. TOUSSAINT. *Knowledge extraction for improving case retrieval and recipe adaptation*, in "Computer Cooking Contest Workshop", London, United Kingdom, September 2011, http://hal.inria.fr/hal-00646717/en.

[43] J. COJAN, J. LIEBER. *An Algorithm for Adapting Cases Represented in ALC*, in "22nd International Joint Conference on Artificial Intelligence - IJCAI 2011", Barcelone, Spain, July 2011, http://hal.inria.fr/inria-00584103/en.

[44] V. DUFOUR-LUSSIER, B. GUILLAUME, G. PERRIER. *Parsing Coordination Extragrammatically*, in "5th Language & Technology Conference - LTC'11", Poznan, Poland, 2011, http://hal.inria.fr/hal-00639929/en.

[45] V. DUFOUR-LUSSIER, J. LIEBER, E. NAUER, Y. TOUSSAINT. *Improving case retrieval by enrichment of the domain ontology*, in "19th International Conference on Case Based Reasoning - ICCBR'2011", London, United Kingdom, 2011, http://hal.inria.fr/inria-00617621/en.

[46] E. EGHO, N. JAY, C. RAÏSSI, A. NAPOLI. *A FCA-based analysis of sequential care trajectories*, in "The Eighth International Conference on Concept Lattices and their Applications - CLA 2011", Nancy, France, A. NAPOLI, V. VYCHODIL (editors), INRIA Nancy Grand Est - LORIA, October 2011, http://hal.inria.fr/hal-00641649/en.

[47] E. GAILLARD, J. LIEBER, E. NAUER. *Adaptation knowledge discovery for cooking using closed itemset extraction*, in "The Eighth International Conference on Concept Lattices and their Applications - CLA 2011", Nancy, France, October 2011, http://hal.inria.fr/hal-00646732/en.

[48] M. KAYTOUE, S. O. KUZNETSOV, J. MACKO, W. MEIRA, A. NAPOLI. *Mining Biclusters of Similar Values with Triadic Concept Analysis*, in "The Eighth International Conference on Concept Lattices and their Applications - CLA 2011", Nancy, France, A. NAPOLI, V. VYCHODIL (editors), INRIA Nancy Grand Est - LORIA, 2011, http://hal.inria.fr/hal-00640873/en.

[49] M. KAYTOUE, S. O. KUZNETSOV, A. NAPOLI. *Biclustering Numerical Data in Formal Concept Analysis*, in "9th International Conference on Formal Concept Analysis - ICFCA 2011", Nicosia, Cyprus, P. VALTCHEV, R. JÄSCHKE (editors), Lecture Notes in Computer Science, Springer, 2011, vol. 6628, p. 135-150 [*DOI :* 10.1007/978-3-642-20514-9], http://hal.inria.fr/inria-00600203/en.

[50] M. KAYTOUE, S. O. KUZNETSOV, A. NAPOLI. *Revisiting Numerical Pattern Mining with Formal Concept Analysis*, in "Twenty second International Joint Conference on Artificial Intelligence - IJCAI 2011", Barcelona, Spain, 2011, http://hal.inria.fr/inria-00584371/en.

[51] B. LAMIROY, D. LOPRESTI, H. KORTH, J. HEFLIN. *How Carefully Designed Open Resource Sharing Can Help and Expand Document Analysis Research*, in "Document Recognition and Retrieval XVIII - DRR 2011", San Francisco, United States, G. AGAM, C. VIARD-GAUDIN (editors), SPIE, January 2011, vol. 7874, ISBN : 9780819484116 [*DOI :* 10.1117/12.876483], http://hal.inria.fr/inria-00537035/en.

[52] D. LOPRESTI, B. LAMIROY. *Document Analysis Research in the Year 2021*, in "Twenty-fourth International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2011)", Syracuse, NY, United States, Lecture Notes in Computer Science, Springer, July 2011, http://hal.inria.fr/inria-00570000/en.

[53] C. RAÏSSI, J. PEI. *Towards Bounding Sequential Patterns*, in "17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD-2011", San Diego, United States, C. APTÉ, J. GHOSH, P. SMYTH (editors), ACM, August 2011, ISBN : 978-1-4503-0813-7 [*DOI :* 10.1145/2020408.2020612], http://hal.inria.fr/inria-00623550/en.

[54] L. SHI, Y. TOUSSAINT, A. NAPOLI, A. BLANSCHÉ. *Mining for Reengineering: an Application to Semantic Wikis using Formal and Relational Concept Analysis*, in "The Semanic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011", Heraklion, Crete, Greece, G. ANTONIOU, M. GROBELNIK, E. SIMPERL, B. PARSIA, D. PLEXOUSAKIS, J. PAN, P. DE LEENHEE (editors), Lecture Notes in Computer Science, Springer, 2011, vol. 6644, p. 421–435, http://hal.inria.fr/hal-00646450/en.

[55] A. SOULET, C. RAÏSSI, M. PLANTEVIT, B. CRÉMILLEUX. *Mining Dominant Patterns in the Sky*, in "The 11th IEEE International Conference on Data Mining - ICDM 2011", Vancouver, B.C, Canada, IEEE, December 2011, http://hal.inria.fr/inria-00623566/en.

[56] L. SZATHMARY, P. VALTCHEV, A. NAPOLI, R. GODIN, A. BOC, V. MAKARENKOV. *Fast Mining of Iceberg Lattices: A Modular Approach Using Generators*, in "The Eighth International Conference on Concept Lattices and their Applications - CLA 2011", Nancy, France, A. NAPOLI, V. VYCHODIL (editors), INRIA Nancy Grand Est - LORIA, October 2011, http://hal.inria.fr/hal-00640898/en.

[57] M. XUE, P. KARRAS, C. RAÏSSI, H. K. PUNG. *Utility-Driven Anonymization in Data Publishing*, in "20th ACM Conference on Information and Knowledge Management - CIKM 2011", Glasgow, United Kingdom, ACM, October 2011, This is a "to appear" publication, http://hal.inria.fr/inria-00623578/en.

[58] M. XUE, P. PAPADIMITRIOU, C. RAÏSSI, P. KALNIS, H. K. PUNG. *Distributed Privacy Preserving Data Collection*, in "16th International Conference on Database Systems for Advanced Applications - DASFAA 2011", Hong Kong, China, J. X. YU, M. H. KIM, R. UNLAND (editors), Lecture Notes in Computer Science, springer, 2011, vol. 6587, p. 93-107 [*DOI :* 10.1007/978-3-642-20149-3_9], http://hal.inria.fr/inria-00610951/en.

National Conferences with Proceeding

[59] S. BENABDERRAHMANE, M.-D. DEVIGNES, M. SMAÏL-TABBONE, O. POCH, A. NAPOLI, W. RAFFELSBERGER, D. GUENOT, N. HOAN, E. GUERIN. *Benchmarking a new semantic similarity measure using fuzzy clustering and reference sets: Application to cancer expression data*, in "11ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances - EGC 2011", Brest, France, January 2011, http://hal.inria.fr/inria-00617692/en.

[60] T. BOURQUARD, J. AZÉ, A. POUPON, D. RITCHIE. *Protein-protein docking based on shape complementarity and Voronoi fingerprint*, in "Journées Ouvertes Biologie Informatique Mathématiques", Paris, France, E. BARILLOT, C. FROIDEVAUX, EDUARDO PC. ROCHA (editors), Institut Pasteur, July 2011, p. 9-16, http://hal.inria.fr/inria-00613186/en.

[61] J. COJAN, J. LIEBER. *Adaptation par révision et adaptation différentielle : comparaison de deux approches de l'adaptation*, in "19ème atelier Français de Raisonnement à Partir de Cas - Rapc2011", Chambéry, France, Fadi Badra and Amélie Cordier, May 2011, http://hal.inria.fr/inria-00595400/en.

[62] S. DA SILVA, C. LAVIGNE, F. LE BER. *Analyse de la structure des haies dans les vergers pour la définition de paysages mieux adaptés contre les bioagresseurs*, in "SAGEO 2011 - International Conference on Spatial Analysis and GEOmatics Conférence internationale de Géomatique et d'Analyse Spatiale Au sein de la 25e Conférence Internationale de Cartographie", Paris, France, 2011, p. 1-4, http://hal.inria.fr/hal-00615301/en.

[63] V. DUFOUR-LUSSIER, F. LE BER, J. LIEBER. *Quels formalismes temporels pour représenter des connaissances extraites de textes de recettes de cuisine ?*, in "Représentation et raisonnement sur le temps et l'espace", Chambéry, France, S. LABORIE, F. LE BER (editors), May 2011, http://hal.inria.fr/inria-00634735/en.

[64] M. KAYTOUE, S. O. KUZNETSOV, A. NAPOLI. *Revisiting Numerical Pattern Mining with Formal Concept Analysis*, in "Journées Nationales de l'Intelligence Artificielle Fondamentale", Lyon, France, 2011, http://hal.inria.fr/inria-00600222/en.

[65] M. KAYTOUE, S. KUZNETSOV, A. NAPOLI, G. POLAILLON. *Symbolic Data Analysis and Formal Concept Analysis*, in "XVIIIeme Rencontres de la Société Francophone de Classification - SFC 2011", Orléans, France, R. EMILION, G. CLEUZIOU (editors), MAPMO - LIFO Orléans, 2011, http://hal.inria.fr/hal-00646457/en.

[66] T. MEILENDER, N. JAY, J. LIEBER, F. PALOMARES. *Les moteurs de wikis sémantiques : un état de l'art*, in "Extraction et gestion des connaissances (EGC'2011)", Brest, France, 2011, p. 575-580, http://hal.inria.fr/hal-00573821/en.

[67] T. MEILENDER, N. JAY, J. LIEBER, F. PALOMARES. *Édition sémantique d'arbres de décision pour l'oncologie avec KcatoS*, in "1ère édition du Symposium sur l'Ingénierie de l'Information Médicale - SIIM 2011", Toulouse, France, June 2011, http://hal.inria.fr/hal-00646843/en.

[68] M. PLANTEVIT, C. RAÏSSI, B. CRÉMILLEUX. *Motifs séquentiels δ-libres*, in "Extraction et gestion des connaissances (EGC'2011)", Brest, France, Hermann-Éditions, January 2011, http://hal.inria.fr/hal-00653579/en.

[69] V. VENKATRAMAN, D. RITCHIE. *Predicting Multicomponent Protein Assemblies Using an Ant Colony Approach*, in "International Conference on Swarm Intelligence", Cergy, France, June 2011, http://hal.inria.fr/inria-00619204/en.

Conferences without Proceedings

[70] E.-G. LAZRAK, N. SCHALLER, J.-F. MARI. *Extraction de connaissances agronomiques par fouille des voisinages entre occupations du sol*, in "Atelier en marge d'EGC 2011", Brest, France, January 2011, http://hal.inria.fr/inria-00560098/en.

Scientific Books (or Scientific Book chapters)

[71] A. COULET, M. SMAÏL-TABBONE, A. NAPOLI, M.-D. DEVIGNES. *Ontology-based knowledge discovery in pharmacogenomics.*, in "Software Tools and Algorithms for Biological Systems", H. R. ARABNIA, Q.-N. TRAN (editors), Advances in Experimental Medicine and Biology, Springer, 2011, vol. 696, p. 357-66 [*DOI :* 10.1007/978-1-4419-7046-6_36], http://hal.inria.fr/inria-00585072/en.

[72] M. KAYTOUE, S. DUPLESSIS, A. NAPOLI. *Toward the Discovery of Itemsets with Significant Variations in Gene Expression Matrices*, in "Classification and Multivariate Analysis for Complex Data Structures", B. FICHET, D. PICCOLO, R. VERDE, M. VICHI (editors), Studies in Classification, Data Analysis, and Knowledge Organization, Springer Berlin Heidelberg, 2011, p. 463–471 [*DOI :* 10.1007/978-3-642-13312-1_49], http://hal.inria.fr/inria-00600233/en.

[73] J.-F. MARI, F. LE BER, E.-G. LAZRAK, M. BENOÎT, C. ENG, A. THIBESSARD, P. LEBLOND. *Using Markov Models to Mine Temporal and Spatial Data*, in "New Fundamental Technologies in Data Mining", K. FUNATSU, K. HASEGAWA (editors), Intech, 2011, p. 561–584, http://hal.inria.fr/inria-00566801/en.

Books or Proceedings Editing

[74] A. NAPOLI, V. VYCHODIL (editors). *The Eighth International Conference on Concept Lattices and their Applications - CLA 2011*, INRIA Nancy Grand Est - LORIA, 2011.

Other Publications

[75] M. HUCHARD, A. NAPOLI, M. ROUANE-HACENE, P. VALTCHEV. *A gentle introduction to Relational Concept Analysis, Tutorial ICFCA 2011*, May 2011, http://hal.inria.fr/lirmm-00616275/en.

References in notes

[76] F. BAADER, D. CALVANESE, D. MCGUINNESS, D. NARDI, P. PATEL-SCHNEIDER (editors). *The Description Logic Handbook*, Cambridge University Press, Cambridge, UK, 2003.

[77] P. BUITELAAR, P. CIMIANO, B. MAGNINI (editors). *Ontology Learning from Text: Methods, Evaluation and Applications*, Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam, 2005.

[78] P. HITZLER, M. KRÖTSCH, S. RUDOLPH (editors). *Foundations of Semantic Web Technologies*, CRC Press, Boca raton (FL), 2009.

[79] S. STAAB, R. STUDER (editors). *Handbook on Ontologies (Second Edition)*, Springer, Berlin, 2009.

[80] F. BADRA, A. CORDIER, J. LIEBER. *Opportunistic Adaptation Knowledge Discovery*, in "8th International Conference on Case-Based Reasoning - ICCBR 2009", Seattle, United States, L. MCGINTY, D. C. WILSON (editors), Lecture Notes in Computer Science, Springer, July 2009, vol. 5650, p. 60-74, The original publication is available at www.springerlink.com [*DOI :* 10.1007/978-3-642-02998-1_6], http://hal.inria.fr/inria-00437693/en.

[81] M. BARBUT, B. MONJARDET. *Ordre et classification – Algèbre et combinatoire (2 tomes)*, Hachette, Paris, 1970.

[82] S. BENABDERRAHMANE, M. SMAÏL-TABBONE, O. POCH, A. NAPOLI, M.-D. DEVIGNES. *IntelliGO: a new vector-based semantic similarity measure including annotation origin*, in "BMC Bioinformatics", December 2010, vol. 11, n^{O} 1, 588 [*DOI :* 10.1186/1471-2105-11-588], http://www.biomedcentral.com/1471-2105/11/588/abstract, http://hal.inria.fr/inria-00543910/en.

[83] R. BENDAOUD, A. NAPOLI, Y. TOUSSAINT. *A proposal for an Interactive Ontology Design Process based on Formal Concept Analysis*, in "Formal Ontology in Information Systems – Proceedings of the Fifth International Conference (FOIS 2008)", Amsterdam, C. ESCHENBACH, M. GRÜNINGER (editors), Frontiers in Artificial Intelligence and Applications, IOS Press, 2008, p. 311–323.

[84] R. BENDAOUD, A. NAPOLI, Y. TOUSSAINT. *Formal Concept Analysis: A unified framework for building and refining ontologies*, in "Knowledge Engineering: Practice and Patterns - Proceedings of the 16th International Conference EKAW", A. GANGEMI, J. EUZENAT (editors), Lecture Notes in Computer Science 5268, 2008, p. 156–171.

[85] R. BENDAOUD, Y. TOUSSAINT, A. NAPOLI. *PACTOLE: A methodology and a system for semi-automatically enriching an ontology from a collection of texts*, in "Proceedings of the 16th International Conference on Conceptual Structures, ICCS 2008, Toulouse, France", P. EKLUND, O. HAEMMERLÉ (editors), Lecture Notes in Computer Science 5113, 2008, p. 203–216.

[86] H. M. BERMAN, T. BATTISTUZ, T. N. BHAT, W. F. BLUHM, P. E. BOURNE, K. BURKHARDT, L. IYPE, S. JAIN, P. FAGAN, J. MARVIN, D. PADILLA, V. RAVICHANDRAN, B. SCHNEIDER, N. THANKI, H. WEISSIG, J. D. WESTBROOK, C. ZARDECKI. *The Protein Data Bank*, in "Acta Crystallographica Section D-Biological Crystallography", 2002, vol. 58, p. 899–907.

[87] A. CARRIERI, V. PÉREZ-NUENO, A. FANO, C. PISTONE, D. RITCHIE, J. TEIXIDÓ. *Biological Profiling of Anti-HIV Agents and Insight into CCR5 Antagonist Binding Using in silico Techniques*, in "ChemMedChem", 2009, vol. 4, p. 1153–1163, http://dx.doi.org/10.1002/cmdc.200900101.

[88] P. CIMIANO, A. HOTHO, S. STAAB. *Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis*, in "Journal of Artificial Intelligence Research", 2005, vol. 24, p. 305–339.

[89] R. D. FINN, J. MISTRY, J. TATE, P. COGGILL, A. HEGER, J. E. POLLINGTON, O. L. GAVIN, P. GUNASEKARAN, G. CERIC, K. FORSLUND, L. HOLM, E. L. L. SONNHAMMER, S. R. EDDY, A. BATEMAN. *The Pfam protein families database*, in "Nucleic Acids Research", 2010, vol. 38, p. D211–D222.

[90] B. GANTER, S. O. KUZNETSOV. *Pattern Structures and Their Projections*, in "Conceptual Structures: Broadening the Base, Proceedings of the 9th International Conference on Conceptual Structures, ICCS 2001,

Stanford, CA", H. DELUGACH, G. STUMME (editors), Lecture Notes in Computer Science 2120, Springer, 2001, p. 129–142.

[91] B. GANTER, R. WILLE. *Formal Concept Analysis*, Springer, Berlin, 1999.

[92] Y. GARTEN. *Text mining the scientific literature to identify pharmacogenomic interactions*, Stanford University, USA, Dec 2010.

[93] I. GUYON, A. ELISSEEFF. *An Introduction to Variable and Feature Selection*, in "Journal of Machine Learning Research", 2003, vol. 3, n^{o} 7-8, p. 1157–1182.

[94] H. HWANG, T. VREVEN, J. JANIN, Z. WENG. *Protein-protein docking benchmark version 4.0.*, in "Proteins: Structure Function and Bioinformatics", 2010, vol. 78, n^{o} 15, p. 3111–3114.

[95] M. KAYTOUE, F. MARCUOLA, A. NAPOLI, L. SZATHMARY, J. VILLERD. *The Coron System*, in "8th International Conference on Formal Concept Analsis (ICFCA) - Supplementary Proceedings", L. BOUMEDJOUT, P. VALTCHEV, L. KWUIDA, B. SERTKAYA (editors), 2010, p. 55–58.

[96] E.-G. LAZRAK, M. BENOÎT, J.-F. MARI. *Time-Space Dependencies in Land-Use Successions at Agricultural Landscape Scales*, in "International Conference on Integrative Landscape Modelling", Montpellier France, UMR LISAH, 02 2010, http://hal.inria.fr/inria-00482890/en/.

[97] J. LIEBER, M. D'AQUIN, F. BADRA, A. NAPOLI. *Modeling adaptation of breast cancer treatment decision protocols in the KASIMIR project*, in "Applied Intelligence", 2008, vol. 28, n^{o} 3, p. 261–274.

[98] J. LIEBER, A. NAPOLI, L. SZATHMARY, Y. TOUSSAINT. *First Elements on Knowledge Discovery guided by Domain Knowledge (KDDK)*, in "Concept Lattices and Their Applications (CLA 06)", S. B. YAHIA, E. M. NGUIFO, R. BELOHLAVEK (editors), Lecture Notes in Artificial Intelligence 4923, Springer, Berlin, 2008, p. 22–41.

[99] G. MACINDOE, L. MAVRIDIS, V. VENKATRAMAN, M.-D. DEVIGNES, D. RITCHIE. *HexServer: an FFT-based protein docking server powered by graphics processors*, in "Nucleic Acids Research", May 2010, vol. 38, p. W445-W449 [*DOI :* 10.1093/NAR/GKQ311], http://hal.inria.fr/inria-00522712/en.

[100] J.-F. MARI, J.-P. HATON, A. KRIOUILE. *Automatic Word Recognition Based on Second-Order Hidden Markov Models*, in "IEEE Transactions on Speech and Audio Processing", 1997, vol. 5, p. 22 – 25.

[101] J.-F. MARI, F. LE BER. *Temporal and Spatial Data Mining with Second-Order Hidden Models*, in "Soft Computing", 2006, vol. 10, n^{o} 5, p. 406–414.

[102] L. MAVRIDIS, D. RITCHIE. *3D-blast: 3D protein structure alignment, comparison, and classification using spherical polar Fourier correlations*, in "Pacific Symposium on Biocomputing 2010", United States Hawaii, World Scientific Publishing, January 2010, p. 281–292 [*DOI :* 10.1142/9789814295291_0030], http://hal.inria.fr/inria-00434263/en.

[103] L. MAVRIDIS, V. VENKATRAMAN, D. RITCHIE, H. MORIKAWA, R. ANDONOV, A. CORNU, N. MALOD-DOGNIN, J. NICOLAS, M. TEMERINAC-OTT, M. REISERT, H. BURKHARDT, A. AXENOPOULOS, P.

DARAS. *SHREC'10 Track: Protein Models*, in "Eurographics Workshop on 3D Object Retrieval - 3DOR 2010", Sweden Norrköping, 2010, http://hal.inria.fr/inria-00536680/en.

[104] R. MOSCA, C. PONS, J. FERNANDEZ-RECIO, P. ALOY. *Pushing Structural Information into the Yeast Interactome by High-Throughput Protein Docking Experiments*, in "PLoS Computational Biology", 2009, vol. 5, n° 8, e1000490.

[105] A. NAPOLI. *A smooth introduction to symbolic methods for knowledge discovery*, in "Handbook of Categorization in Cognitive Science", H. COHEN, C. LEFEBVRE (editors), Elsevier, Amsterdam, 2005, p. 913–933.

[106] C. A. ORENGO, A. D. MICHINE, S. JONES, D. T. JONES, M. B. SWINDELLS, J. M. THORNTON. *CATH - A Hierarchic Classification of Protein Domain Structures*, in "Structure", 1997, vol. 5, n° 8, p. 1093–1108.

[107] C. PESQUITA, D. FARIA, A. O. FALCÃO, P. LORD, F. M. COUTO. *Semantic Similarity in Biomedical Ontologies*, in "PLoS Comput Biol", 2009, vol. 5.

[108] M. PLANTEVIT, A. LAURENT, D. LAURENT, M. TEISSEIRE, Y. W. CHOONG. *Mining multidimensional and multilevel sequential patterns*, in "ACM Trans. Knowl. Discov. Data", 2010, vol. 4, n° 1, p. 1–37, http://doi.acm.org/10.1145/1644873.1644877.

[109] V. PÉREZ-NUENO, S. PETTERSSON, D. RITCHIE, J. BORRELL, J. TEIXIDÓ. *Discovery of Novel HIV Entry Inhibitors for the CXCR4 Receptor by Prospective Virtual Screening*, in "Journal of chemical information and modeling", Apr 2009, vol. 49, n° 4, p. 810-823 [*DOI :* 10.1021/CI800468Q], http://hal.inria.fr/inria-00434261/en.

[110] V. PÉREZ-NUENO, D. RITCHIE, J. BORRELL, J. TEIXIDÓ. *Clustering and Classifying Diverse HIV Entry Inhibitors Using a Novel Consensus Shape-Based Virtual Screening Approach: Further Evidence for Multiple Binding Sites within the CCR5 Extracellular Pocket*, in "Journal of Chemical Information and Modeling", 2008, vol. 48, n° 11, p. 2146–2165.

[111] V. PÉREZ-NUENO, D. RITCHIE. *Applying in silico Tools to the Discovery of Novel CXCR4 Inhibitors*, in "Drug Development Research", 2011, vol. 72, p. 95–111 [*DOI :* 10.1002/DDR.20406], http://hal.inria.fr/inria-00550645/en/.

[112] V. PÉREZ-NUENO, D. RITCHIE, O. RABAL, R. PASCUAL, J. BORRELL, J. TEIXIDÓ. *Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand-receptor docking*, in "Journal of Chemical Information and Modeling", 2008, vol. 48, n° 3, p. 509–533.

[113] C. RAÏSSI, J. PEI, T. KISTER. *Computing Closed Skycubes*, in "Proceedings of the VLDB Endowment", September 2010, vol. 3, n° 1, p. 838–847, http://hal.inria.fr/inria-00610923/en.

[114] D. RITCHIE, G. KEMP. *Protein Docking Using Spherical Polar Fourier Correlations*, in "Proteins: Structure, Function and Genetics", 2000, vol. 39, n° 2, p. 178–194.

[115] D. RITCHIE, V. VENKATRAMAN. *Ultra-fast FFT protein docking on graphics processors*, in "Bioinformatics", 2010, vol. 26, n° 19, p. 2398–2405 [*DOI :* 10.1093/BIOINFORMATICS/BTQ444], http://hal.inria.fr/inria-00537988/en/.

[116] M. ROUANE-HACENE, M. HUCHARD, A. NAPOLI, P. VALTCHEV. *A proposal for combining Formal Concept Analysis and description Logics for mining relational data*, in "Proceedings of the 5th International Conference on Formal Concept Analysis (ICFCA 2007), Clermont-Ferrand", S. O. KUZNETSOV, S. SCHMIDT (editors), LNAI 4390, Springer, Berlin, 2007, p. 51–65.

[117] A. STEIN, A. CEOL, P. ALOY. *3did: identification and classification of domain-based interactions of known three-dimensional structure*, in "Nucleic Acids Research", 2010, vol. 39, p. D718–D723.

[118] L. SZATHMARY. *Symbolic Data Mining Methods with the Coron Platform*, Université Henri Poincaré (Nancy 1), 2006.

[119] L. SZATHMARY, P. VALTCHEV, A. NAPOLI, R. GODIN. *Constructing Iceberg Lattices from Frequent Closures Using Generators*, in "Discovery Science", J.-F. BOULICAUT, M. BERTHOD, T. HORVÁTH (editors), Lecture Notes in Computer Science 5255, Springer, Berlin, 2008, p. 136–147.

[120] L. SZATHMARY, P. VALTCHEV, A. NAPOLI, R. GODIN. *Efficient Vertical Mining of Frequent Closures and Generators*, in "Proceedings of the 8th International Symposium on Intelligent Data Analysis (IDA-2009), Lyon, France", N. ADAMS, J.-F. BOULICAUT, C. ROBARDET, A. SIEBES (editors), Lecture Notes in Computer Science 5772, Springer, Berlin, 2009, p. 393–404.

[121] L. SZATHMARY, P. VALTCHEV, A. NAPOLI. *Finding Minimal Rare Itemsets and Rare Association Rules*, in "Proceedings of the 4th International Conference on Knowledge Science, Engineering and Management (KSEM-2010), Belfast, Northern Ireland, UK", Y. BI, M.-A. WILLIAMS (editors), Lecture Notes in Artificial Intelligence 6291, Springer, Berlin, 2010, p. 16–27.

[122] L. SZATHMARY, P. VALTCHEV, A. NAPOLI. *Generating Rare Association Rules Using the Minimal Rare Itemsets Family*, in "International Journal of Software and Informatics", 2010, vol. 4, n^{o} 3, p. 219–238.

[123] M. D'AQUIN, F. BADRA, S. LAFROGNE, J. LIEBER, A. NAPOLI, L. SZATHMARY. *Case Base Mining for Adaptation Knowledge Acquisition*, in "Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)", M. M. VELOSO (editor), Morgan Kaufmann, 2007, p. 750–755.