Computer Science & IT with/for Biology

CSBio Reader:
Extended abstracts of the interdisciplinary seminar series

C. Maria Keet and
Enrico Franconi (eds.)

Freie Universität Bozen
Libera Università di Bolzano
Free University of Bozen · Bolzano

Computer Science & IT with/for Biology

CSBio Reader:
Extended abstracts of the interdisciplinary seminar series

This reader contains the extended abstracts of the seminars organised for the "Computer Science and IT with/for Biology" Seminar Series, held at the Faculty of Computer Science, Free University of Bozen-Bolzano, from October to December 2005. Slides of the presentations are available online at: www.inf.unibz.it/krdb/biology.

C. Maria Keet and Enrico Franconi (eds.)
KRDB Research Centre
Faculty of Computer Science
Free University of Bozen-Bolzano
Piazza Domenicani 3
I-39100 Italy

Preface

THE vibrant and emerging research area of 'doing research and engineering in the subject domain of biology and the applied biosciences' comprises one or more (sub-) disciplines of computer sciences and information technology that can be mixed with any of the (sub-) disciplines in biology, ecology, and applied biosciences (such as medicine and agriculture). Depending on the emphasis, this combination tends to favour one or more of the following terms to indicate the type of activity: Computational Biology, Systems Biology, Bioinformatics, *In Silico* Biology, Ecoinformatics, (Bio)Medical Informatics, and bio-ontologies, among others. But what exactly is the breadth and depth of these relatively new fields, and what are its characterstic activities? What is, or can be, used from mathematics to advance biology at a faster pace? What type of problems do bioscientists perceive that need to be solved? Is engineering only a supportive discipline for biology? If not, where and how is biology pushing the frontiers of computer science and IT? How did, and does, the combination of computer science & biology lead to landmark achievements – and which ones are considered to be achievements?

Against this background, the KRDB Research Centre of Faculty of Computer Science at the Free University of Bozen-Bolzano aimed to present and form new expertises and professional profiles who can answer the growing demands of the biosciences and ultimately our societies in the area of using both theoretical and applied aspects of computer science and engineering, thereby contributing to pushing the frontiers in computer science as well as (applied) biology. To this end, it has organized the "CS & IT with/for biology" Seminar Series. The aim of the seminars was to provide a broad spectrum of achievements, opportunities, and challenges on using/combining computer science with/for biology, highlighting diverse foci and approaches traversing biology (sub-) disciplines and applied bioscience and a wide range of computer science approaches. This coverage goes from basic biosciences, such as genetics & cellular processes and larger systems in ecology, and the applied biosciences medicine and agriculture, to CS/IT fields of ontology/ies, logics, NLP, database integration, and software development.

This reader contains the extended abstracts of the invited speakers, offering both a summary of the seminar as well as additional references to give useful pointers to key publications, the most recent resaerch output, and 'hot' topics.

The first chapter in this reader provides a general overview of historical aspects and current characteristics of the rather flexible interpretation that was given to biology & informatics – and the more recent diversification into multiple niche areas. It can aid novices in the field to grasp some of the more, and less, active research activities and 'insiders' to have ample material for discussion. From this introduction, we first take a step back before going into details, by looking at some ethical considerations, as described by Heiner Fangerau. Within a short time span, many new possibilities are (or seem) just around the corner: stem cell research and personalised medicine to name just two; but who benefits, and is a regrouping of the human world population into certain groups with genetic predispositions for particular diseases – technologically not impossible – actually desirable and beneficial for the society at large? Which biases are 'built in' when we do our literature research?

The subsequent chapters go into some detail, both with regard to the technological and computer science aspects as (applied) biology. In chapter 3 Alberto Policriti introduces mathematical modelling for systems biology, with automata and pi-calulus in particular. These topics are relevant for *in silico* simulations of cellular processes and the mathematical complexities of the outstanding problems, i.e. modelling biological knowledge requires new solutions from mathematicians. The next chapter by Marco Roos, on the other hand, takes a case-based approach: biologists desire to understand better e.g. Huntington's Disease and histones, and to achieve this, they need a computer infrastructure to enable them to do their research. A regrouping of this requirement with technological support has resulted in the initiation of a virtual laboratory for e-science. Marie-Paule Lefranc has taken a yet different path (in chapter 5), where demands from biology, immunogenetics in this case, are combined with the latest developments in computer science, such that her laboratory belongs not only to the 'early adopters' of technology over the past 15 years, but also can use it effectively to discover biologically meaningful new information: bio & info in synergy.

The infamous biological data explosion that has occurred over the past 10 years may be well-know, it's 'consequently' disconnected software tools and databases is known in considerably less detail. Apart from the obvious data integration issues between databases and linking database and analysis tools, one first needs to be able to find what is there, and then for the biologist to find what s/he needs. This is a central topic of Sarah Cohen-Boulakia's contribution: what are biologists actually looking for, and how can we, automatically, find the relevant software resources? The issue of finding the right information is addressed from an entirely different angle and context by Werner Ceusters in chapter 7. Advances made in the subdiscipline of natural language understanding can help processing electronic health records, annotated with an ontology, to mine that data and discover new patterns in the patient's treatment and history with as aim to improve biomedicine. Last, with Aldo Gangemi we take a closer look at the usefulness of task and action ontologies for software development in agriculture, with the UN Food and Agriculture Organisation (FAO) among the beneficiaries.

While the topics do not cover all aspects of CS&IT with/for (applied) biology, we hope it will give you some insight in its multifaceted aspects, ranging from applied mathematics and philosophy to software engineering, from core to applied biology, and from enabling information technology to successful combination of bio-info and biology-driven computer science.

Maria Keet and Enrico Franconi
Italy, December 2005

Table of Contents

Current characteristics and historical perspective of Computer Science and IT with/for biology

C. Maria Keet

KRDB Research Centre, Faculty of Computer Science, Free University of Bozen-Bolzano, Italy
keet@inf.unibz.it

Abstract. Although the emerging discipline(s) involved in combining Computer Science and Information Technology with biology may seem a new development, several historical aspects already can be identified. These, with its past and present characteristics, will be presented and discussed. Together, they provide a general introduction covering the breadth of the research topics of CS&IT with/for biology and its related applied life sciences, and offer a background framework to place the subsequent specialist seminars in its appropriate context.

1 Introduction

From an outsider's perspective, any combination of 'something computer science or IT' that refers to 'something bio' then must be bioinformatics. People who actually do combine one with the other, enable one by support of the other, or find new problems to solve induced by another discipline, beg to differ. This may simply be by using a new fancy label to (re)group 'old' activities, as well as comprise really new developments. An incomplete list of terms is: Computational Biology, Biomedical Engineering, Biocomputing, (Meta)Genomics, Proteomics, Metabolomics, Climate modelling, Bioinformatics, Agricultural Informatics, Theoretical Biology, GIS, Ecoinformatics, Nanotechnology, Computational Chemistry, Environmental engineering, Medical Informatics, *in silico* (molecular) Biology, Bio-ontologies, Theoretical Ecology, Mathematics and Biology, and Systems Biology. Here we provide some structure, based on a historical perspective and current characteristics, to give an introductory overview, which is neither complete nor has the final word. Nevertheless, it provides a start for exploring the dynamic, evolving area of mixing computer science, IT, basic biology, and the applied life sciences.

2 Historical aspects

2.1 Before the mid-1990s

Although Systems Biology is considered a new, hot, topic, in a slightly different form it already emerged in the 1930s [17]. A commonality is the 'systems view' [15], but in the 1930s the emphasis was more on non-equilibrium dynamics, pathways and later also self-organisation [17]. Around the 1950s, biomechanics and biomedical engineering was added, which looked into e.g. developing protheses and related mechanical devices for improving medicine. Its counterpart for the core biological disciplines (see Table

1) is the development, and in particular its deployment, of analysis machines for X-ray cristallography, liquid chromatography and so forth; most notably for long-term impact is the former, through which the DNA helix structure could be discovered by Watson and Crick. From a computer science perspective, the emphasis was on hardware and firmware development. This started to change in the 1970s with, for instance, the emergence of the first climate simulation models running on supercomputers – pushing the boundaries of both hardware and software – and agricultural informatics, which resulted in introduction of IT in this applied life sciences area, for instance to model crop rotation, fodder composition, and managing greenhouses with rule-based systems (AI). Around the same time (in the 1980s in particular), medical informatics was having a high conjecture with research into and implementation of decision support systems. In the late 1980s and early 1990s the first molecular biology databases (for genes and proteins) saw the light. Hence, a gradual shift in emphasis, but not necessarily in overall activity, had taken place from hardware & firmware toward software-fucussed aspects. Then, around mid-1990s, the 'big explosion' took place. To place the rapidly unfolding, evolving events and activities in context, a few notes have to be made on the scientific enterprise.

2.2 The past 10 years

Fig. 1 depicts the standard, iterative, experimental research cycle, which can be followed both clockwise and counter-clockwise. Up to the mid-1990s, many theories existed that were suggested plausible explanations about little aspects of nature, but hardly tested because of the laborious work to test it or the impossibility to test it with available means, i.e. a relative block between steps 4 and 1. Put differently: the theories were as "thruthlike" [8] as possible, but it was not clear how close their thruthlikeness was (and still is) to the truth – considering the pursuit of truth about nature to be the main goal of doing research. With Moore's law, cheaper computers, and the Internet, quickly many new possibilities became available to *finally* scale-up labwork, do previously un-doable data analysis, and disseminate research results, which resulted in the now infamous data-explosion induced problems. These new pos-

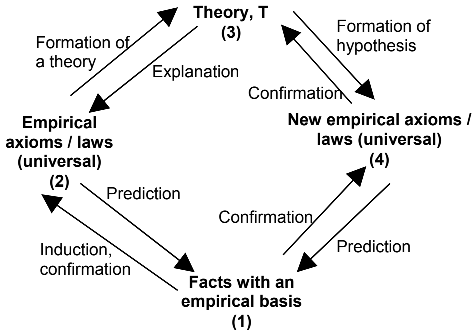


Fig. 1. The standard cycle of the experimental research process.

sibilities translated into a mushrooming of bio-databases, gene sequencers and gene sequence comparison algorithms, and data analysis tools. This, in turn, created new demands from (molecular) biologists, like data(base) integration. For example, the mouse-researchers had developed their own software systems, but so had done the fruitfly [28] and the yeast genome researchers [42] – as well as other research communities who studied what has come to be known as 'model organisms' (*C. elegans* (a type of worm) [18], *Arabidopsis* (thale cress) [41], among others). To link, or even integrate, their systems, a structured controlled vocabulary was deemed necessary, and thus the Gene Ontology Consortium saw the light in 1998 [7] [12]. While certainly not a perfect ontology, if it can be called an ontology at all, the community-effort approach has been very successful and resulted in many bio-ontologies spin-offs and similar approaches (e.g. [38] [14]) that together at least ease annotation and linking database entries across largely autonomous databases and their analysis tools.

Separate from these activities are the developments in data mining, and currently in particular pattern finding in large data sets, grids, and workflow systems. A recent addition is computational linguistics, for both information retrieval of the immense amount of literature [40] and text mining to automatically find e.g. pathway information [2]. Also, one has generally moved from research in gene alignment algorithms to structure and function prediction of proteins. More generally, there is a move from the reductionist 'the genes have the answers to understand nature' to a realisation that nature is more complex and that answers may be found higher up in the "omics planes" [16]. Hence, the new systems biology. To make it more challenging for applied mathematicians, computer scientists and IT engineers, biologists do not want only the customary static models, but rather play, *in silico*, with simulations of cells and organisms alike. A still relatively unchartered area is metagenomics [3], which tries to understand the system as a whole *in vivo* instead of its individual components.

2.3 A few social aspects

A last, contentious, note has to be made regarding the social aspects of research. Apart from the scientifically and technologically 'what is hot and what is not', there are 'old' and new activities, one morphing into the other and new activities created that require new labels for various (strategic) reasons. Bioinformatics used to be anything that combines biology and informatics, moved to *molecular* biology and informatics, and at the time of writing leans more toward actually being only for the final purpose of supporting (human) medicine. Debatable, but systems biology may be seen as a type of high-throughput cell physiology. Counter-actions are the repositioning of informatics & ecology into ecoinformatics, and the distinction between engineering-focussed bioinformatics and the applied mathematics-focussed computational (systems) biology. Quite separate strands of investigation that mix informatics and (applied) biology without cross-fertilization with bioinformatics and computational biology are e.g. nanotechnology and environmental engineering. New terms and its effective marketing can bring new funding opportunities – if its characteristic activities can be called a new research discipline retrospectively remains to be seen.

3 Characteristics

Liberally interpreted, one can combine any of the research areas listed in the left-hand column of Table 1 with one or more listed in the right-hand column. Note that neither each combination is being pursued actively, nor that each existing combination has its own (named) niche in academe, but likely some activity is going on. At first, one might be inclined to think that combining one (sub)discipline with the other results in *inter*disciplinarity, and to tackle the issues requires a full-fledged interdisciplinary *team*. To quote Eddy: “An interdisciplinary team is a committee in which members identify themselves as an expert in something else besides the actual scientific problem at hand, and abdicate responsibility for the majority of the work because it’s not their field.” [4]. This disciplinary approach to a situation that favours actually a *combination* of competencies, is another factor, aside from the data explosion, contributes to the current ’mess’. There are *many* bio-databases, which may be topical (one or two granularity levels, GOLD, HGVBase), species specific (FlyBase [28], AceDB [18]), context (Bad Bug Book [19]), or primary source (TIGR [48]) versus many boutique databases. The yearly inventory published in the journal Nucleic Acids Research [6] is large, and still only partial and other databases of databases exist [39] [30]. Second, there are many single-issue software tools, mainly for data analysis of database content, visualisation, and upcoming simulation software. Both the databases and analysis tools have a high degree of autonomy of development and maintenance of IT tools. Most tools, however, are poorly maintained and some databases are better maintained than others, partly due to end-of-project effects and maintained on a voluntary basis by an interested individual. Third, sub-optimal data management can be identified: data is ill structured, reliability of data is becoming an issue, data redundancy is not adequately addressed, neither is data compatibility. Fourth, research

CS & IT	**(Applied) biology**
Hardware/Firmware	*Core sciences – Biology*
- Robotics	- Microbiology (bacteriology, fungi)
- Grid computing & supercomputing	- Plant sciences
- Analyzers	- Animal sciences
Software	(nematology, ornithology, ethology, ...)
- Neural networks	- Taxonomy
- Workflows	*Core sciences – Molecular Biology*
- Software engineering	- Biochemistry, enzymology
- Databases	- Cell physiology, systems biology
(CM, DB devel., integration, temporal DB)	- Genomics, proteomics, metabolomics
- Distributed processing	*Ecology & - environmental sciences*
- Graphics & visualisation, HCI	- theoretical and experimental ecology
- Computational linguistics	(trophic levels, nutrient cycles, niche)
- Knowledge based systems, ontologies	- Climatology
- Data mining	- System biology
	Applied Biosciences
	- Biomedicine, agriculture, food science

Table 1. An incomplete list of (sub)disciplines in CS, IT, and (applied) biology.

results are reported in many different journals and conferences, as for ‘new’ combinations of activities there are no readily available specialised outlets. This brings forth another issue: where and how to find the right, desired, information? While there are many problems to solve, it must be mentioned that some tools are really useful. To complicate separating the wheat from the chaff, everybody (claims to) develop(s) the ultimate best solution. There is a flurry of things going on, but a) one cannot easily see the forest for the trees on what is there, and what is useful, and b) there is an awakening realisation that the bioinformatics tools at present do not do quite what biologists had in mind the tools would do in helping them to do their research more effectively (but there are also moving targets). It is not uncommon to encounter the ‘jumping on the bandwagon’ feeling, there is ample funding, many (over-)optimistic promises, while playing down, or even ignoring, ethical considerations. IT can enable finding many new insights in nature, but do we really want to know certain things, and if yes, how should it be used?

Apart from these issues, the research activities do require from the scientist to know something about another discipline too; but how to educate such students and researchers? Generally, there is a lag between research output and catching up of education to train competent scientists. And what exactly does ‘the bioinformatician’ need to know anyway? Is s/he a biologist who can program a little, a computer scientist who googles for biological information? Does a BSc in discipline a with an MSc in discipline b suffice? This does not enable a student to internalise different running paradigms, research methodologies, and cultures. Then there are interdisciplinary approaches that are CS/IT-based (the toy example) versus bio-based (enabling technologies), theory versus experimental labwork, and technology push/pull mechanisms. No less important are differences in the use of knowledge (e.g. hierarchical, object-oriented versus associative, networked, knowledge) and ‘the nature’ of the knowledge (certainty vs. good-enough, conjectures, change), and biasism vs. epistemological realism.

Bearing this in mind, then maybe the CS&IT with/for (applied) biology requires interdisciplinary *people* who are ‘multilingual’, fulfill a bridge function between disciplines and are capable of applying some methodology a from discipline x to problem b from discipline y and vice versa – and scale up and merge such new approaches to “inventing new ways to look at the world” [4] alike fixing the biologist’s radio [11] and beyond. An alternative view is that the present ‘anarchy’ in combining disciplines actually is the usual, and fruitful, type of activities that in 10-15 years leads to well-established disciplines. Hence, the new and rearranged sets of activities that are categorised under bio-ontologies, computational biology, etc. are maybe *ante*disciplinary [4] and will develop into a/several discipline(s)? Conversely, one can ask: will/does/did the combination of CS, IT, and biology result in one, or more, new disciplines, and does a named set of a few characteristic activities really make a (new) (sub)discipline? Regardless, for the time being one has the liberty to reconfigure and invent, be creative without being placed in a disciplinary box, and explore utterly unchartered areas to discover and solve a near-limitless range of topics.

4 Examples

More concretely, we mention a few examples of *integrative* approaches, which do not cover the broad range but serve as illustration. One can find more specific projects and references to relevant background material in subsequent chapters.

An exciting new area are the simulations – ranging from developing computer simulations of cells to whole organisms (developing a Virtual Human is a task for the upcoming years). To appreciate what has been achieved, one may like to compare achievements of representing the cholera toxin [13], i.e. its schematic molecule interactions and proteins at different levels of granularity (whilst ensuring having your model grounded in reality with real laboratory data), with more recent simulations for education [45] [47], simulations that feed on real data, and videos of real things at (sub)cellular level [49].

Sequence comparisons, where the protein level receives most attention at present, groups together genetics, biochemistry, data mining, mathematics, visualisation, databases, and Web-access. In applied sciences one has e.g. medical, agricultural, and environmental engineering. The former has its main outlets (and categorisation of themes) in MEDINFO, AMIA, and Journal of Biomedical Informatics, among others. An concrete example of the latter is OntoWEDSS [1], which is a decision support system for managing active sludge in wastewater treatment plants. It combines into one coherent whole the raw data collection, databases, case and rule based systems, an ontology, and a user interface. At the level of ecosystems, the Knowledge Network for Biocomplexity is established for improving ecological and environmental research on biocomplexity [35]. It involves ecological metadata, storage resource broker, distributed data management, data integration, quality assurance, hypothesis modelling for ecological research, and provides visualisation tools. From an engineering point of view, two of the many software development projects are Bio-Linux [21] [5] that contains 60 software packages for data management, and BioBrew with about 223 active groups of bio-software developers.

What these examples indicate is that there is also *intra*disciplinary science: solving complex problems requires expertise from more than one sub-discipline in computer science as well as cross-boundary, systems, research in biology, i.e. both horizontal and vertical cooperation and integration.

5 Concluding remarks

Research areas involved in combining Computer Science and IT with biology can be ‘old’ activities (with or without new terms), where old goes back to about the 1930s concurrent with the early developments in computer science, and in other cases really are new, where this does not only comprise applying new technologies to enable biology and its applied live sciences, but also can be more alike informatics *with* biology where each one pushes the boundaries of the other. Present characteristics show a rather chaotic picture, that nevertheless at the same time facilitates exciting developments in unchartered fields.

This extended abstract served as a brief introductory overview to give a background framework to better place the subsequent seminars in their appropriate context. Several aspects are not touched upon or only cursorily, like difficulties of modelling biological data characteristics, 'requirements' for the type of person suitable for the job, and a problem analysis to solve outstanding issues. Furthermore, inevitably some bias is built in in this abstract and it serves the reader to consult other sources to get a more comprehensive picture of this multifaceted dynamic area. No one has a monopoly on wisdom, but a healthy thirst for knowledge and peeking over the disciplinary fence can provide the researcher necessary ingredients to make a significant contribution to advance our understanding of the world around us.

References

1. Ceccaroni, L., Cortés, U. and Sànchez-Marrè, M. OntoWEDSS: augmenting environmental decision- support systems with ontologies. *Environmental Modelling & Software*, 2004, 19 (9): 785-797.
2. Daraselia, N., Egorov, S., Yazhuk, A., Novichkova, S., Yuryev, A., Mazo, I. Extracting Protein Function Information from MEDLINE Using a Full-Sentence Parser. *Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics.* 2004. pp11-18.
3. DeLong, E.F. Microbial community genomics in the ocean. *Nature Reviews Microbiology*, 2005, 3:459-469.
4. Eddy, S.R. "Antedisciplinary" Science. *PLoS Computational Biology*, 2005, 1(1): e6.
5. Field, D., Tiwari, B., Snape, J. Bioinformatics and Data Management Support for Environmental Genomics. *PLoS Biology*, 2005, 3(8): e297.
6. Galperin, M.Y. The Molecular Biology Database Collection: 2005 update. *Nucleic Acids Research*, 2005, 33: D5-D24.
7. Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 2000, 25: 25-29.
8. Johansson, I. Bioinformatics and Biological Reality. *Journal of Biomedical Informatics*, 2005 (forthcoming). preprint: http://hem.passagen.se/ijohansson/information4.PDF.
9. Keet, C.M. Biological Data and Conceptual Modelling Methods. *Journal of Conceptual Modeling*, Issue 29, 2003. http://www.inconcept.com/jcm.
10. Kitano, H. Computational systems biology. *Nature*, 2002, 420: 206-210.
11. Lazebnik, Y. Can a Biologist Fix a Radio? – or, What I Learned while Studying Apoptosis. *Cancer Cell*, 2002, 2: 179-182.
12. Lewis S.E. Gene Ontology: looking backwards and forwards. *Genome Biology*, 2005, 6: 103.
13. Merritt, E.A., Sarfaty, S., Akker, F. van den, L'Hoir, C., Martial, J.A., Hol, W.G.J. Crystal structure of choleratoxinB-pentamer bound to receptor GM1 pentasaccharide. *Protein Science*, 1994, 3: 166-175.
14. Plant Ontology Consortium. The Plant Ontology Consortium and Plant Ontologies. *Comparative and Functional Genomics*, 2002, 3:137-142.
15. Richmond, B. *Systems thinking - four key questions*. High Performance Systems. 1991.
16. Toyoda, T. and Wada, A. Omic space: coordinate-based integration and analysis of genomic phenomic interactions. *Bioinformatics*, 2004, 20(11): 1759-65.
17. Westerhoff, H.V., and Palsson, B.O. The evolution of molecular biology into systems biology. *Nature Biotechnology*, 2004, **22**(10): 1249-1252.

Some related websites, organisations, and projects

18. A Caenorhabditis elegans DataBase: http://www.acedb.org/.
19. Bad Bug Book. http://www.cfsan.fda.gov/~mow/intro.html.
20. BioBrew: http://bioinformatics.org/project/?group_id=273

21. Bio-Linux: http://www.biolinux.org/.
22. Biopattern: http://www.biopattern.org/index.html.
23. E-cell project. http://www.ndsu.nodak.edu/instruct/mcclean/vc/wwwic-vc5.htm.
24. European Bioinformatics Institute: http://www.ebi.ac.uk.
25. European Ethics Network: http://www.eureth.net.
26. European Federation for Information Technology in Agriculture, Food and the Environment: http://www.efita.net/.
27. European Molecular Biology Laboratory: http://www.embl.org/.
28. Fruitfly: http://www.fruitfly.org.
29. Gene Ontology Consortium: http://www.geneontology.org.
30. Infobiogen. http://www.infobiogen.fr/deambulum/tab.php?page=vue&lg=en.
31. InfoBioMed: http//www.infobiomed.org.
32. Institute for Formal Ontology and Medical Information Science: http://www.ifomis.org.
33. International Society for Computational Biology: http://www.iscb.org/.
34. International Medical Informatics Association: http://www.imia.org.
35. Knowledge Network for Biocomplexity: http://knb.ecoinformatics.org/index.jsp.
36. Marine genomics: http://www.marine-genomics-europe.org.
37. Mouse Genome Database: http://www.informatics.jax.org.
38. Open Biological Ontologies. http://obo.sourceforge.net.
39. Pathway Resource List: http://cbio.mskcc.org/prl/index.php.
40. TREC Genomics: http://ir.ohsu.edu/genomics/.
41. The Arabidopsis Information Resource: http://www.arabidopsis.org/.
42. Saccharomyces Genome Database: http://www.yeastgenome.org.
43. Science Environment for Ecological Knowledge: http://seek.ecoinformatics.org.
44. Semantic Mining in Biomedicine: http://www.semanticmining.org.
45. Simulation of actin: http://cellix.imolbio.oeaw.ac.at/Videotour/video_tour_5.html.
46. Society for Mathematical Biology: http://www.smb.org/.
47. Translation: the movie:
http://vcell.ndsu.nodak.edu/∼christjo/vcell/animationSite/index.htm.
48. The Institute of Genomic Research. http://www.tigr.org.
49. Video of sub-cellular events:
http://cellix.imolbio.oeaw.ac.at/Videotour/video_tour.html.

Finding Bioethical Literature - Databases and Databiases

Heiner Fangerau

Institute for the History of Medicine, Heinrich-Heine-University Düsseldorf
heiner.fangerau@uni-duesseldorf.de

Abstract. The selection of suitable databases for finding medical ethics literature is often difficult. There exists a great variety of electronic bibliographies listing medical ethical literature. These databases follow different quality standards. The quality standards applied by the databases as well as their selection of articles to be indexed have a strong effect on the users research results. Recent studies could show that there seems to be a regional bias in the most popular databases favouring US American periodicals compared to European literature on Medical Ethics. Especially this regional bias or other language or national biases can have an influence on ethicists research results. Examples from hot ethical issues like stem cell research, its ethical and legal evaluation and its representation in electronic bibliographies will be displayed to show of which data-biases researchers in ethics have to be aware. Solutions to overcome these biases will be presented.

1 Ethical Pluralism

Public debates often reveal that great differences in the understanding of life may exist between scientists and the public. Current negative reactions to scientific findings, like the hot debates about human stem cell research and therapeutic cloning, stem from such fundamental differences. During these debates it is often forgotten that researchers not necessarily perform their research without any consideration of its ethical implications. They may just apply different ethical principles than their contemporaries. Scientists also not necessarily share common, harmonised values[1]. Too often the existence of diverse moral concepts within societies is neglected. The opposite of this neglect is true: within pluralistic societies (even in our globalised world) the existence of a variety of moral norms seem to be common [2] [3].

It is one of the tasks of research in bioethics to detect and analyse moral values involved in bio-medical debates and to discuss the life sciences and their potential impact on different societies. As an academic discipline bioethics as “applied ethics” is a branch of moral theory and philosophy. Research in bioethics e.g. helps to understand varying ideas concerning stem cell research in different countries or the great variety of legislation in this field in Europe [4]. The diversity in opinion is not to be moaned. It may be used to reconcile research goals. This again may lead to a theoretical diversity that itself may open the mind for new applications of research findings in other contexts. However, scientific compromising to answer public or scientific ethical demands is in itself not ethically unproblematic, e.g. if the postponement of scientific findings results in the death of people that otherwise could have been saved. Nevertheless, in the case of stem cell research scientific compromising is considered by some authors as a font of creativity [5] [6].

[1] Different ethical theories are described in high quality in book length e.g. by Hinman [1].

2 Literature searches and databases[2]

Research in the field of medical ethics usually requires extensive literature searches. Before technical development made possible the establishment of larger electronic literature databases, scientists had to rely on extensive printed bibliographies from multiple disciplines. Because the bibliography selected for the search limited which articles were found and which were not [8], the researcher often had to search more than one bibliography to get an adequate overview of the relevant literature. Thus, finding suitable articles was often complicated and time-consuming.

The possibilities of finding literature in the field of medical ethics have been widely extended by the computer sciences: various computer based bibliographic databases are available to assist scientists and other users in their search for the required literature. This development sufficiently accelerates the process of searching for the documents in question. Nevertheless, the database selected for research - as with the classical printed bibliography - still determines which literature will be found. According to the considerations above, the origin of the database determines the ethical orientation of the literature it holds. For example, most of the existing literature databanks seem to show a regional bias. They index literature from their point of origin to a greater extent than literature from other regions [9]. Albeit this bias might not affect biomedical research too much, it might be relevant to research in the field of biomedical ethics for several reasons, including the fact that a possible preference for a region or language by a bibliographic database could result in cultural distortion of the facts illustrated in the literature, or a particular emphasis on issues of regional or local importance. In the study quoted above the journal coverage of international periodicals on medical ethics for different electronic bibliographies was ascertained [7]. The questions asked were:

- which literature databases index the highest number of periodicals dealing with medical ethical questions?
- do different databases show decisive regional or language preferences in their indexing practice?

A suitable resource for the clarification of these two questions is the database *ulrichsweb.com*, which is accessible via the Internet (www.ulrichsweb.com). Altogether, 290 periodicals could be identified which according to “Ulrich’s” explicitly publish articles about medical ethical issues. 284 different “Abstracting and Indexing Services” index at least one of the 290 periodicals. 117 periodicals (40.3%) are not abstracted or indexed in any bibliography. The top ten bibliographic databases that collect the highest number of periodicals publishing on medical ethical issues are:

1. Current Contents (http://www.isinet.com) (indexes 66 periodicals out of 290 found in “Ulrich’s”, which means that the coverage is 22.8%),
2. MEDLINE (http://www.nlm.nih.gov) (N=64/ 22.1%),
3. Research Alert (http://www.isinet.com) (N=54/ 18.6%),
4. Social Science Citation Index (http://www.isinet.com) (N=54/ 18.6%),
5. EMBASE (http://www.excerptamedica.com) (N=51/ 17.6%),
6. AgeLine (http://research.aarp.org/ageline/home.html) (N=50/ 17.2%),

[2] For a more detailed description see Fangerau [7].

7. CINAHL (http://www.cinahl.com/) (N=42/ 14.5%),
8. E-psyche (http://www.e-psyche.net) (N=39/ 13.5%),
9. Sociological Abstracts (http://www.csa.com) (N=38/ 13.1%) and
10. Family Index (http://www.famindx.com) (N=36/ 12.4%).

The degree of coverage of the individual databases is rather lower than might have been expected by a user searching for medical ethics literature. The maximum coverage is 22.8% (66 out of 290 possible periodicals in Current Contents).

Only users extending their search from a single to several databases can reach a higher degree of coverage. However, even here, as there is overlap within the databases concerning the indexed periodicals, the degree of coverage that can be reached by using all the ten databases mentioned can only be raised up to a maximum of 45.2% (Fig. 1). The periodicals identified in “Ulrich’s” for this analysis are published in 24 different countries, with a clear regional predominance of US-American literature. 152 periodicals are published in the USA (52.4%), 95 in Europe (32.8%). In Europe British (N=38), Dutch (N=25), German (N=8), Italian (N=7) and French (N=7) periodicals predominate.

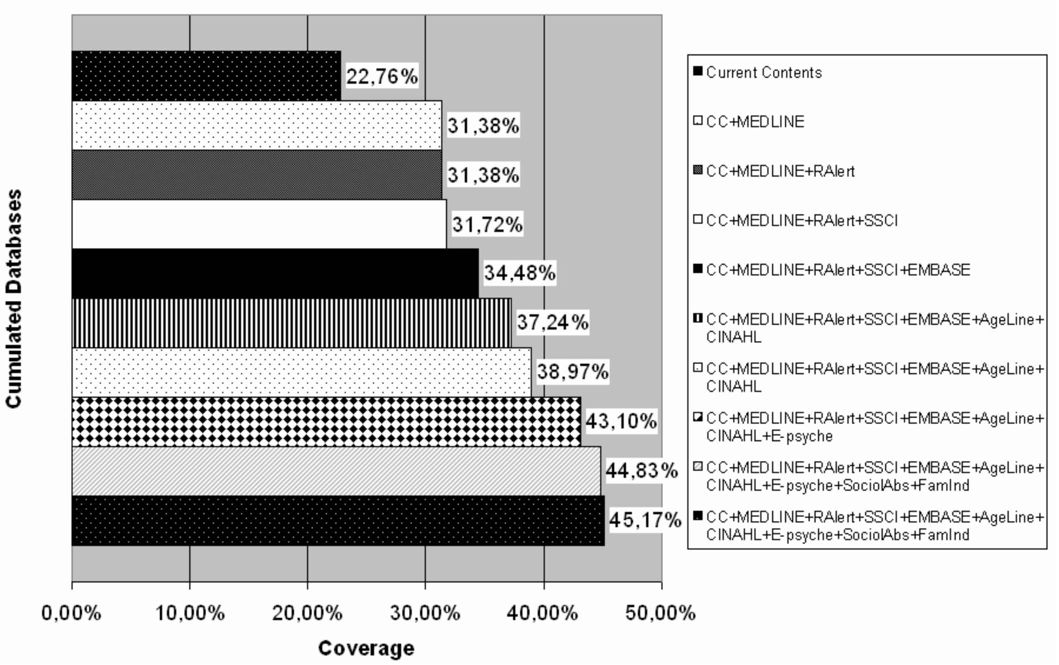


Fig. 1. Coverage of international medical ethical journals by several databases in combination.

The representation of European and US-American periodicals in the examined bibliographic databases were compared. Determining the proportion of the European and US-American periodicals indexed in the different databases serves as a tool for assessing a possible regional preference for either of the regions in each of the bibliographic databases (USA/Europe rate). A very low number of Asian journals found and the lack of African journals may suggest that "Ulrich's Periodicals" itself might show a regional bias. Therefore, the rate occurring in Ulrich's Periodicals of 152/95 (USA/Europe quotient 1.6) had to serve as a base point for a normal value with which a preference can be evaluated. For the eight databases with the highest degree of coverage described above the regional preference for either Europe or the USA is shown in Fig. 2. A comparison of the rates shows that none of the top ten databases index more European periodicals than it would correspond to their share in all periodicals indexed in Ulrich's. Only EMBASE comes close. All the other databases show a tendency to represent a majority of US-American literature.

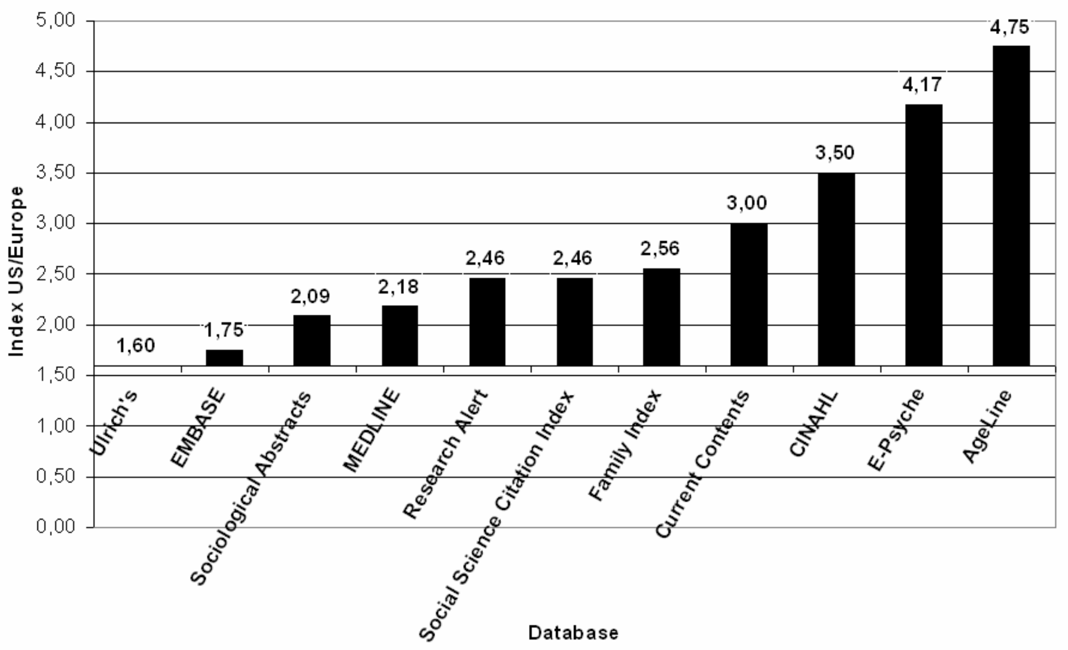


Fig. 2. Regional preferences.

3 Conclusions

Knowledge of using bibliographic databases assists researchers in finding literature on which they can build their research hypotheses. Because a database determines what the user is finding the user has to know the gaps, thematic emphases and indexing preferences of the different databases. The survey shows that the medical ethics literature is represented insufficiently in the most popular existing bibliographic databases. On the assumption that Ulrich's Periodicals might themselves show a regional preference in their indexing practice, the result of the analysis of regional preferences for US-American or European literature in bibliographic databases has to be compared with the "USA/Europe quotient" of Ulrich's Periodicals of 1.6. All of the top ten bibliographic databases show a higher "USA/Europe quotient" (1,75 EMBASE - 4,75

AgeLine). This result makes a predominance of US-American literature in the indexing practice of the ten analysed databases visible. It suggests a "publication bias" that by all means should be taken into account when searching for medical ethics literature. Otherwise a European discussion for example might be reflected insufficiently or an "African" point of view might be neglected at all.

Responding to the need for an increased European contribution to the international discussion on ethics in medicine and biotechnology, some of Europe's leading bioethics institutions have joined forces to establish the international network "EURETHNET". In a project funded by the European Commission (EC) between 2002 and 2005 partners from several European countries have come together to develop an information network and knowledge base in the field of ethics in medicine and biotechnology [10]. Via www.eureth.net their services and a European literature database are accessible. However, the future of the project is insecure as the EC funding stopped. As a consequence, the information bias may be consolidated.

References

1. Hinman, L.M.: Ethics : a pluralistic approach to moral theory. Thomson, Belmont, CA (2002)
2. Engelhardt, H.T.: Bioethics in the third millennium: some critical anticipations. Kennedy Inst Ethics J 9 (1999) 225-243
3. Turner, L.: Bioethics in pluralistic societies. Med Health Care Philos 7 (2004) 201-208
4. Holm, S.: Stem cell transplantation and ethics: a European overview. Fetal Diagn Ther 19 (2004) 113-118
5. Gilbert, D.M.: The future of human embryonic stem cell research: addressing ethical conflict with responsible scientific research. Med Sci Monit 10 (2004) RA99-103
6. Holden, C., Vogel, G.: CELL BIOLOGY: A Technical Fix For an Ethical Bind? Science 306 (2004) 2174-2176
7. Fangerau, H.: Finding European bioethical literature: an evaluation of the leading abstracting and indexing services. J Med Ethics 30 (2004) 299-303
8. Marsh, S.S.: Bibliography of Bioethics and Index Medicus: comparison of coverage, publication delay, and ease of recall for journal articles on bioethics. Bulletin of the Medical Library Association 75 (1987) 248-252
9. Obst, O.: Datenbanken auf dem Prfstand: Ist Medline eine Luftnummer? AGMB aktuell (2000) 24-27
10. Fangerau, H., Simon, A., Wiesemann, C.: Improving information systems in Europe: EURETHNET. Med Health Care Philos 6 (2003) 67-69

Formal Tools for Systems Biology

Alberto Policriti

Dipartimento di Matematica e Informatica, Università di Udine, Italy
policrit@dimi.uniud.it

Abstract. In this talk we will illustrate the idea of exploiting formal models and languages used in software formal verification, with as aim the design of innovative systems to be applied in Systems Biology. We will start by briefly describing the Simpathica tool (Simulation of Pathways and Integrated Concurrent Analysis), integrating mathematical and logical approaches for the study biochemical networks. We will continue with an overview of other possibilities to apply formal tools to cell biology. We conclude with addressing some algorithmic and expressivity issues, problems related with the use of standard/hybrid automata, and logical languages in the context of Systems Biology.

Systems Biology is becoming very popular ([19, 20]) as it is widely recognized that, in biology, the identification and classification of "emergent behaviors" is not an easy task that must be tackled with powerful tools. There is an array of possible approaches to the search, classification, and analysis of such behaviors, which are closely interrelated and highly dynamic, as the quest for applications is rapidly increasing in size and differentiating in type. The needs arise for more and more sophisticated and mathematically well founded computational tools capable of analyzing the models that are and will be at the core of *system biology.* Such computational models should be implemented in software packages faithfully while exploiting the potential trade-offs among usability, accuracy, and scalability dealing with large amounts of data.

In general, Bioinformatics tools focus on creating a finely detailed and "mechanistic" picture of biology at the cellular level by combining the part-lists (genes, regulatory sequences, other objects from an annotated genome, and known metabolic pathways), with observations of both transcriptional states of a cell (using microarrays) and translational states of the cell (using proteomics tools). Attempts to provide pictures of biological behaviors as comprehensive and systematic as possible are undergoing and concurrent and reactive models are playing a central role in many such proposals (see, for example, [25, 4]).

The work described in this presentation is part of a much larger project still in progress, and thus only provides a partial and evolving picture of a new paradigm for computational biology.

Consider the following scenario. A biologist is trying to test a set of hypotheses against a corpus of data produced in very different ways by several *in vitro*, *in vivo*, and *in silico* experiments. The system the biologist is considering may be a piece of a *pathway* for a given organism. The biologist can access the following pieces of information:

- raw data stored somewhere about the temporal evolution of the biological system; this data may have been previously collected by *observing* an in vivo or an in vitro system, or by *simulating* the system in silico;

– some mathematical model of the biological system[1].

The biologist will want to formulate *queries* about the evolution encoded in the data sets. For example, he/she may ask: *will the system reach a "steady state"?*, or *will an increase in the level of a certain protein activate the transcription of another?* Clearly the set of numerical *traces* of very complex systems rapidly becomes unwieldy to wade through for increasingly larger numbers of variables.

Eventually, many of these models will be available in large public databases (e.g. [6, 16–18, 27, 21]) and it is not inconceivable to foresee a biologist to test some hypotheses *in silico* before setting up expensive wet-lab experiments. The biologist will mix and match several models and raw data coming from the public databases and will produce large datasets to be analyzed.

To address this problem, we have proposed a set of theoretical and practical tools, XS-systems and Simpathica, that allow the biologist to formulate such queries in a simple way [2, 4, 5]. The computational tool Simpathica derives its expressiveness, flexibility, and power by integrating in a novel manner many commonly available tools from numerical analysis, symbolic computation, temporal logic, model-checking, and visualization. In particular, an *automaton-based* semantics of the temporal evolution of complex biochemical reactions starting from their representations as sets of differential equations is introduced. Then *propositional temporal logic* is used to qualitatively reason about the systems. When we speak of "qualitative reasoning," as in the preceding sentence, we do not intend to describe an abstract reasoning process devoid of all quantitative information—rather, we focus on the relation among several basic properties (each described by an atomic proposition), where each one may involve some quantitative information, e.g., "property of a protein concentration reaching half of its initial value."

In [3] we continue our research on the computational models at the core of our approach. We bring in several techniques from the fields of Verification, Logic and Control Theory, while maintaining a trade off between the need to manipulate large sets of incomplete data and the requirements arising from the needs to provide a mathematically well founded system.

In particular, we propose the use of *hybrid automata* together with the notions of *bisimulation* and *collapsing*. Hybrid automata are equipped with states embodying time-flow, initial and final conditions, and therefore allow maintenance of more information about the differential equations (S-system, in this particular case) we use to model the change in the involved quantities. The use of the notion of bisimulation in the definition of the *projection operation* (restrictions to a subset of "interesting" variables) provides a way to introduce *reduced* automata satisfying the same formulae as the initial ones. Notice that the idea behind and potential of this notion of bisimulation can be exploited just as fruitfully here as in the context of standard automata. Finally, the notion of collapsing, we introduce, serves a dual purpose: first, it provides a natural approach for qualitative reasoning on the automata extracted from

[1] We note that simulating a system *in silico* actually requires a mathematical model. However, we want to consider the case when such mathematical model is unavailable to both the biologist and the software system.

the analysis of traces summarizing the behavior of biomolecules; second, it tames the otherwise unruly complexity of the automata in terms of their size as a function of the levels of approximation allowed.

A survey on the different approaches for modeling and simulating genetic regulatory systems can be found in [13]: the author takes into consideration different mathematical methods (including ordinary and partial differential equations, qualitative differential equations and others) and evaluates their relative strengths and weaknesses.

The problem of constructing an automaton from a given mathematical model of a general dynamical system has been previously considered in the literature. In particular, it has been investigated by Brockett in [7]: our approach in [4] is certainly more focused, since it deals with specific mathematical models (i.e. S-systems). Here we move farther away from purely discrete models, and adapt hybrid automata to describe the underlying biochemical behavior instead of standard automata. Consequently, we are able to take advantage of the continuous component of hybrid automata for allowing quantitative information in addition to qualitative reasoning.

The use of hybrid automata for the modeling and simulation of biomolecular networks has been proposed also by Alur et al. in [1] and by Chabrier et al. in [8]. In [1] the discrete component of an hybrid automaton is used to switch between two different behaviors (models) of the considered biological system, (for example) depending on the concentration of the involved molecules. The hybrid automaton is then implemented in Charon. In our case, the continuous component is used to model the permanence on a given state depending on the values of the involved variables (reactants), and the discrete component is used for enabling the transition to another state. Moreover, we do not only model the biological systems, but we also query them using temporal logics. A similar approach is considered in [8], where a variant of Euler's method is applied in order to obtain a symbolic representation of the system. Then the authors show how to use symbolic model checkers, such as NuSMV [9] and DMC [14], to study the system.

The notion of concurrency can be explicitly used in modeling biochemical systems by representig the involved reactants as communicating processes running in parallel [25]. (In our case this kind of concurrency becomes implicit since in all the states of the automaton representing an S-system, the values of all the reactants, and their evolutions are represented.) The approach has been described and extensively studied/applied in many papers (see, for example, [26, 11, 10, 12]). A particularly interesting line of researches in this field, is the work done in order to provide simple and expressive bioinformatics tools executing a stochastic version of the so-called π-calculus (see [24]) suitable for introducing a quantitative ingredient in the methodology. The effort for developing such tools includes a specification of graphical language for the stochastic π-calculus ([23]), a definition of a correct abstract machine underlying the implementation ([22]), and the use of stochastic-discrete simulation (Monte Carlo) algorithms for the simulation of the network of concurrent processes exploiting an analogy with molecular dynamics ([15]).

References

1. R. Alur, C. Belta, F. Ivancic, V. Kumar, M. Mintz, G. J. Pappas, H. Rubin, and J. Schug. Hybrid Modeling and Simulation of Biomolecular Networks. In *Hybrid Systems: Computation and Control*, volume 2034 of *LNCS*, pages 19–32. Springer-Verlag, 2001.
2. M. Antoniotti, F. C. Park, A. Policriti, N. Ugel, and B. Mishra. Foundations of a Query and Simulation System for the Modeling of Biochemical and Biological Processes. In *Proc. of the Pacific Symposium of Biocomputing (PSB'03)*, 2003.
3. M. Antoniotti, C. Piazza, A. Policriti, M. Simeoni, and B. Mishra. Taming the Complexity of Biochemical Models through Bisimulation and Collapsing: Theory and Practice. *Theoretical Computer Science*, 325(1):45–67, 2004.
4. M. Antoniotti, A. Policriti, N. Ugel, and B. Mishra. XS-systems: extended S-systems and algebraic differential automata for modeling cellular behaviour. In *Proc. of Int. Conference on High Performance Computing (HiPC'02)*, 2002.
5. M. Antoniotti, A. Policriti, N. Ugel, and B. Mishra. Model Building and Model Checking for Biological Processes. *Cell Biochemistry and Biophysics*, 2003. To appear.
6. U. S. Bhalla. Data Base of Quatitative Cellular Signaling (DOQCS). Web site at `http://doqcs.ncbs.res.in/`, 2001.
7. R. W. Brockett. Dynamical Systems and their Associated Automata. In *Systems and Networks: Mathematical Theory and Applications*, volume 77. Akademie-Verlag, 1994.
8. N. Chabrier and F. Fages. Symbolic Model Checking of Biochemical Networks. In C. Priami, editor, *Computational Methods in Systems Biology (CMSB'03)*, volume 2602 of *LNCS*, pp149–162. Springer-Verlag, 2003.
9. A. Cimatti, E. M. Clarke, E. Giunchiglia, F. Giunchiglia, M. Pistore, M. Roveri, R. Sebastiani, and A. Tacchella. NuSMV 2: An Opensource Tool for Symbolic Model Checking. In E. Brinksma and K. G. Larsen, editors, *Int. Conf. on Computer Aided Verification (CAV'02)*, volume 2404 of *LNCS*, pages 359–364. Springer-Verlag, 2003.
10. M. Curti, P. Degano, and C. T. Baldari. Casual pi-calculus for Biochemical Modelling. In C. Priami, editor, *Computational Methods in Systems Biology (CMSB'03)*, volume 2602 of *LNCS*, pages 21–33. Springer-Verlag, 2003.
11. M. Curti, P. Degano, C. Priami, and C. T. Baldari. Casual π-calculus for Biochemical Modelling. DIT 02, University of Trento, 2002.
12. V. Danos and C. Laneve. Graphs for Core Molecular Biology. In C. Priami, editor, *Computational Methods in Systems Biology (CMSB'03)*, volume 2602 of *LNCS*, pp34–46. Springer-Verlag, 2003.
13. H. de Jong. Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. DIT 4032, Inria, 2000.
14. G. Delzanno and A. Podelski. DMC User Guide. 2000.
15. D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81, 1977.
16. P. D. Karp, M. Riley, S. Paley, and A. Pellegrini-Toole. The MetaCyc Database. *Nucleic Acid Research*, 30(1):59, 2002.
17. P. D. Karp, M. Riley, M. Saier, and S. Paley A. Pellegrini-Toole. The EcoCyc Database. *Nucleic Acids Research*, 30(1):56, 2002.
18. KEGG Database. `http://www.genome.ad.jp/kegg/`.
19. H. Kitano. Systems Biology: an Overview. *Science*, 295:1662–1664, March 2002.
20. P. Nurse. Understandig cells. *Nature*, 24, 2003.
21. PathDB Database. `http://www.ncgr.org/pathdb/`.
22. A. Phillips and L. Cardelli. A correct abstract machine for the stochastic pi-calculus. *ENTCS*, Bioconcur'04, 2004.
23. A. Phillips and L. Cardelli. A Graphical Presentation for the Stochastic Pi-calculus. *ENTCS*, Bioconcur'05, 2005.
24. C. Priami. Stochastic π-calculus. *The Computer Journal*, 38, 1995.
25. A. Regev and E. Shapiro. Cellular Abstactions: Cells as Computations. *Nature*, 419, 2002.
26. A. Regev, W. Silverman, and E. Shapiro. Representation and Simulation of Biochemical Processes using the π-calculus Process Algebra. In *Proc. of the Pacific Symposium of Biocomputing (PSB'01)*, pages 459–470, 2003.
27. WIT Database. `http://wit.mcs.anl.gov/WIT2/`.

Towards a virtual laboratory for integrative bioinformatics research

Marco Roos[1], Han Rauwerda[1], M. Scott Marshall[1], Lennart Post[1,2], Márcia Inda[1], Christiaan Henkel[1], and Timo Breit[1]

[1] Integrative Bioinformatics Unit, Faculty of Science, University of Amsterdam, The Netherlands
[2] Nuclear Organisation Group, Faculty of Science, University of Amsterdam, The Netherlands
{roos, rauwerda, marshall, lpost, inda, chenkel, breit}@science.uva.nl

Abstract. From a biological point of view, the motivation for the development of a virtual laboratory (VL) for integrative bioinformatics research is the need for computer support for experimentation on multifaceted biological problems that require the involvement of experts from different disciplines and incorporation of knowledge and data from all relevant facets. The envisioned VL supports interactive experimentation and multidisciplinary collaboration. The experimental research cycle is empowered by a 'semantic framework' based on Semantic Web technology. It enables integration of data and knowledge, including the data provenance from intermediate steps in the VL. Our approach is 'case study'-driven in the sense that investigating a biological case is performed concurrently with the development of bioinformatics methodology. Biological cases include, among others, transcription regulation and the 'histone code', and Huntington's disease. We will present the first stage of an incremental workflow that performs knowledge-based integration of the histone code to transcription factor biding sites.

1 Introduction

With the advent of 'omics' technologies, data generation is no longer a major bottleneck for life science [1]. Data are produced by a constantly growing number of new high throughput and/or genome-wide techniques for every level of cellular biology: genomics (DNA), transcriptomics (RNA), proteomics (protein), metabolomics (metabolite) and phenomics (phenotype). With nanotechnology and lab-on-a-chip applications on the horizon, the end of this development is not yet in sight. In addition, the sum of results from past experiments that were more limited in scope also represents a formidable amount of information, stored in databases and/or described in articles and text-books and often locked in the minds of human experts. One of the still existing bottlenecks when designing and performing (computational) experiments is the ability to take into account the multifaceted nature of biological problems and the amount of background knowledge involved. In general, 'wet' laboratory experiments produce data that give only partial information. For instance, typical micro array experiments give information on RNA abundance only. However, this data is often used as a direct indicator of gene expression even though one could argue that factors such as RNA turnover should be taken into account. Biologists generally need to speculate beyond the scope of their experiments in order to interpret their results in a biological context. This requires substantial background knowledge. Moreover, each facet of a biological problem, each type of laboratory experiment, and each type of analysis require specific expertise. For a truly integrative approach one would like to make use of all available expertise and take data and knowledge from all appropriate facets into account. We investigate enabling such an approach by incorporating information technology that allows the performance of computational experiments with these elements in a virtual laboratory for integrative bioinformatics (an 'e-bioscience laboratory', i.e. a laboratory for *enhanced* biological science). We address the general steps of experimentation: information analysis, hypothesis generation, and design and execution of (computational) experiments. Our ultimate goal is to have an environment that allows us to gain new insight into biology by the integration of human expertise, and the integration of machine-readable data and knowledge. In collaboration with domain experts (biologists and computer scientists) we are investigating and developing two aspects of a virtual laboratory for integrative bioinformatics. An 'interactive and creative environment' addresses the human aspect of an integrative bioinformatics approach. The key objective is to enable interactive information analysis, hypothesis formation, experiment design, and computational experimentation within a multidisciplinary team of scientists, of which the members can be at multiple locations. A 'general computational layer' addresses the information science aspect of an integrative approach for biological research. It encompasses technology for demanding computation, and technology to employ biological knowledge in computations. We apply a case-study driven approach. Case studies have two equally important objectives: provide scenarios for technological development, and produce results relevant in the biological domain. We consider this approach essential to the success of the development of a virtual laboratory for e-bioscience. Our main case studies are a number of micro array analysis, gene expression regulation, and Huntington's Disease. Currently, we focus on human interaction, knowledge-based data integration, and computational modules for transcriptome analysis.

2 Virtual laboratories for e-science

Virtual Laboratories (VLs) in general are "electronic workspaces for distance collaboration and experimentation in research or other creative activity, to generate and deliver results using distributed information and communication technologies" [2]. High-performance networking and high-throughput computing in a grid environment make VLs possible. However, for the scientist from the application domain (the end-user in a VL), sharing and reuse of resources, i.e. data, information, knowledge, methodologies and equipment, is only practical when the technical complexity of ICT is hidden or abstracted. To this end, a VL is set up as a three-layered structure: the Grid-VL layer, the generic VL-layer and the application specific VL-layer. The Grid-VL layer offers the tools and infrastructure for resource-access and management. The generic VL-layer provides the generic methods, tools and infrastructure to use Grid technology in order to meet the demands from the applications [3]. It holds methods and techniques that are applicable to more than one application domain. Examples of these are middleware to support interactive high-performance computing, methodology for collaborative information management, workflow processing tools and visualization tools. The specific VL-layer provides the interaction between the application domain and the generic VL-layer. With respect to software development, levels of maintenance and certification and the frequency of the release cycle may differ from layer to layer and can be assigned to different groups. Workflow management in VL-e [4] is considered a central element for all its applications, because it is the basis for reproducible computational experimentation. We are particularly interested in the association of data and machine-readable knowledge in workflows.

3 Towards a virtual laboratory for integrative bioinformatics

3.1 General approach

We distinguish four phases of the general experimental research cycle: information analysis, hypothesis formation, experiment design, and (computational) experiment (after which results are fed back into the loop). In line with our case-study driven approach, we step through these phases to obtain results relevant to the domain of the case study, and at the same time to investigate new methodology to enable an integrative approach based on the virtual laboratory concept. We think this approach will facilitate acceptability of the VL concept in the life science domain. For the development of the VL for integrative bioinformatics, we will discuss the two different points of view: human interaction and information technology. The developments from both sides will eventually fit together to form a unified system.

Human point of view: An Interactive and Creative Environment (ICE)
In a VL for integrative bioinformatics, we require an ‘interactive and creative environment’ (ICE) where scientists from different disciplines, perhaps in remote locations, can explore a specific biological problem domain. For instance micro array studies benefit from the collaboration between experts from different fields [5]. The micro array experts in our group can not also be experts on each biological case addressed by micro arrays in addition to being highly qualified statisticians. Similarly, the biology domain experts can not be expected to be micro array experts or experts in other fields as well. Henceforth, a team of scientists should collaborate on such biological studies. The process by which to solve the biological questions is a creative and explorative one, because it is often not possible or even desirable to fully predefine the steps of a micro array analysis. In general, an empirical, explorative approach is an important trait of biological research, which a VL should support. For developing the ICE, we take a direct approach, focusing on making bioinformatics tools and data available, enabling ‘quick and dirty’ data integration, and enabling the visualisation of results quickly. In principle, all results are produced for human evaluation. The physical appearance will be that of a multi-display setup operated by a human operator [6]. GRID networking technology enables remote collaboration [7]. Sophisticated linking of the elements, the input and output, is not the primary objective. This is the objective of the ‘general computational and semantic layer’.

Information technology point of view: The General Computational and Semantic Layer
The biological problems addressed in a VL for integrative bioinformatics require

methodologies for computational experiments and integration and retrieval of data and knowledge. We speak of compiling a 'knowledge-space' that is available for exploration and computational experimentation [1]. We use the term 'semantic framework' as a reference to all semantic technology required. We have defined a general approach to iteratively increment machine-readable knowledge, driven by case studies. We start by designing a 'crude' experiment for a case study. Next, we identify the key terms, which we use to search for data and ontologies, using search engines such as Swoogle [8] and Google. If we find an ontology that contains the key terms, we have to evaluate it: apart from its quality, does it have the necessary level of granularity, and does it take the point of view we need? If so, we will port it to our knowledge database for which we currently use Sesame [9]. If not, we will build an experimental ontology appropriate to the case study in OWL and RDFS. We will add concepts or small scale models as needed. It is not our aim to build community ontologies. We also look for or build models that represent the knowledge behind our hypothesis and experiment. Next, we make connections between the data and the models using RDF: we 'semantically annotate' our data. If we link multiple data sources to models that share concepts, we effectively integrate data. We have chosen Semantic Web technology because it is a W3C recommended international set of standards, but most importantly because of its flexibility that fits an experimental science such as biology, where models are never expected to be 100% correct or complete. Furthermore, we think that the experimenter in a VL should have control over, and be responsible for, connecting the resources of choice. The 'decoupled nature' of RDF allows this in principle, which makes it an obvious choice for 'omics' research [10]. As a result of application by life scientists, knowledge models evolve, and new areas important to experimental biology are investigated. We welcome such developments, especially those that will enable us to capture the uncertainty inherent in experimental results, together with the methods by which evidence was obtained. Karp, for instance, presented the application of an ontology to capture the scientific evidence that supports the information within a database [11].

3.2 Semantic Data Integration

We present an approach for data integration based on the application of formalized knowledge in the form of RDF(S) and OWL. Next to the formalization of knowledge, we formalize the steps of our experiments by applying the component-based computing paradigm in terms of web-services and workflow. After identification of the necessary models and data, the data needs to be imported. The first step is data acquisition, for which we currently build specific web services for each data source. A more general approach for data discovery and retrieval is under study. The following step is conversion to RDF. Most bioinformatics data is available as tables from relational databases, and not readily as RDF. The mapping is performed in two steps. In the first step we make a 'flat' translation from the data structures to RDF (similar to [12]), and little domain semantics is added. The RDF is stored in Sesame. In the next step, the critical step that adds domain knowledge to the data, the RDF data is linked to the knowledge models encoded in OWL/RDFS. Instances of data elements (RDF) will be referenced by instances of concepts of the knowledge models. The data and related models are now available for further experimentation, for instance through the Sesame Query Language. Although the Sesame query language could be used to perform data integration (cf. [12]), reminiscent of table joins in relational database, we prefer to exploit the semantic models. When different data sets are connected to different concepts, one or more relations between these concepts could be discovered by reasoning or by data mining, or they can be manually defined by domain experts. The latter is a form of manual integration and is equivalent to connecting two datasets to one model. In any case, our approach demands more from a Semantic Web query language. Indeed, in our experience the Sesame RDF Query Language (SeRQL [9]) may currently be insufficient for highly sophisticated queries (for a survey of alternatives see [13]).

4 Example: The histone code and gene expression regulation

Histones are proteins that pack DNA into higher order structures and influence processes such as transcription, repair and replication of DNA. They have been implied in diseases such as cancer [14] or Huntington's Disease [15]. Through specific chemical modifications of their structure, they form a 'histone code' across the genome [16] [17]. The aim of this study is to unravel the relationship between the histone code, DNA sequence, and gene expression regulation. The first crude experiment is to relate the distribution of one modified histone across the genome to transcription factor binding sites (TFBSs, sequences that are important for gene expression regulation; transcription factors are proteins that bind to DNA – sometimes via another protein – to regulate the transcription of a nearby gene). We have built a small histone ontology (HistOn) using the OWL plugin in Protégé. After evaluating existing bio-ontologies, we concluded that there were none that fulfilled our requirements sufficiently. The Gene Ontology (GO), although containing a number of entries related to histone modifications, contains only is-a and part-of relationships, lacking the flexibility that we anticipate we will need. In essence, we found no ontologies that contained concepts and relationships specifically concerned with epigenetics. A practical issue was that we wanted to be able to experiment with the OWL models themselves in the context of this case study. Concepts from the HistOn were used to formally describe the hypothesized relationship between histone binding and transcription factor binding sites. The two required data sets were found at the UCSC genome browser: histone abundance as a function of genome location, and genome locations of transcription factor binding sites. In the model made for this experiment the concept of modified histones was linked to transcription factor binding sites through the common concept of genome location. The raw data imported from UCSC (tab delimited files with columns chromosome number, start position, end position), were translated into RDF using the java tool 'Mapper', which uses an XML file that describes the desired translation from one format to another. The RDF output is stored in Sesame. The integration is completed through linking the RDF data to RDF instances of RDFS/OWL concepts from our models. Using the Sesame query language, the collection of transcription factor binding sites of which the genomic location overlaps with that of the histones can be retrieved, although limitations of the language requires substantial knowledge of the RDF graph to accomplish this. For the future, further integration is planned, such as with the transcription factors themselves (of which the data contain links to e.g. Gene Ontology concepts), genome sequence, and gene expression. In addition, we are evaluating computational options that exploit the integrated models and data to unravel the relationships between the histone code, DNA sequence, and gene expression regulation. We contemplate having a number of options, such as reasoning and text/data mining [18], available in the VL to combine in computational experiments.

5 Conclusions and future work

In the seminar a case-study driven approach was discussed for the development of methodology that enables an integrative bioinformatics approach. An integrative approach is needed to address the multifaceted nature of biology and the amount of data and knowledge involved in biological research. We aim for a virtual laboratory that enables integration of data and knowledge by humans and machine. This first is captured by the concept of the ICE, the second by the concept of a computational layer that supports the application of machine-readable knowledge. Ultimately, developments for each of these layers should merge to provide one powerful system for performing *enhanced* biological science (e-bioscience). The envisioned integrative approach is based on the application of ontologies to incrementally build a knowledge space, and workflow to design and run experiments. The knowledge space will be based our own experimental models and models that increasingly become available in the biomedical domain. We intend to extend the computational layer with services to semantically annotate features in biological data, encompassing for instance transcriptome patterns (e.g. ridges [19] [20]), and features in microscope images reflecting DNA-regions of interests, (e.g. again ridges; Goetze et al., unpublished data). These examples are related to Histone features through their common relationship with gene expression regulation. Hence, we investigate the building of a 'semantic web' for gene expression, and in particular its potential to enable an integrative approach in a Virtual Laboratory. We gained experience with Semantic Web technology (RDF, RDFS, and OWL) for the development of a data integration method based on captured knowledge about histones. For multidisciplinary teams working in the ICE on biological cases, our approach will mean that an extra layer, the knowledge layer, becomes available for exploration and experimentation. For the future, we hope to employ a combination of services that make use of this layer (e.g. reasoning) in addition to other bioinformatics data analysis methods (for some examples see references in [18]).

References

1. Rauwerda, H., Roos, M., Hertzberger, B. O., Breit, T. M.: The promise of a Virtual Lab in Drug Discovery. submitted to Drug Discovery Today (2005)
2. Vary, J. P.: Report of the Expert Meeting on Virtual Laboratories. International Institute of Theoretical and Applied Physics (IITAP) and UNESCO, CII-00/WS/01 Ames, Iowa (2000)
3. Afsarmanesh, H., Belleman, R. G., Belloum, A. s. Z.: VLAM-G: A Grid-based virtual laboratory. Scientific Programming 10 (2002) 173-181

4. Zhao, Z., Belloum, A., Wibisono, A., Terpstra, F., Boer, P. T. d., Sloot, P., Hertzberger, B.: Scientific workflow management: between generality and applicability. In: Proc. International Workshop on Grid and Peer-to-Peer based Workflows in conjunction with the 5th International Conference on Quality Software (2005) 357-364
5. Van 't Veer, L. J., De Jong, D.: The microarray way to tailored cancer treatment. Nat Med 8 (2002) 13-4
6. Leigh, J., Johnson, A., Park, K., Nayak, A., Singh, R., Chowdhry, V.: Amplified Collaboration Environments. In: Proc. VizGrid Symposium (2002)
7. Foster, I., Kesselman, C., Tuecke, S.: The Anatomy of the Grid. Int. J. High-Performance Comput. Applic. 15 (2001) 200222.
8. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: a search and metadata engine for the semantic web. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management. ACM Press (2004) 652-659
9. Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. 2342 edn. (2002)
10. Wang, X., Gorlitsky, R., Almeida, J. S.: From XML to RDF: how semantic web technologies will change the design of 'omic' standards. Nat Biotechnol 23 (2005) 1099-103
11. Karp, P. D., Paley, S., Krieger, C. J., Zhang, P.: An evidence ontology for use in pathway/genome databases. Pac Symp Biocomput (2004) 190-201
12. Cheung, K. H., Yip, K. Y., Smith, A., Deknikker, R., Masiar, A., Gerstein, M.: YeastHub: a semantic web use case for integrating data in the life sciences domain. Bioinformatics 21 Suppl 1 (2005) i85-i96
13. Haase, P., Broekstra, J., Eberhart, A., Volz, R.: A Comparison of RDF Query Languages. 3298 edn. (2004)
14. Santos-Rosa, H., Caldas, C.: Chromatin modifier enzymes, the histone code and cancer. Eur J Cancer 41 (2005) 2381-402
15. Steffan, J. S., Bodai, L., Pallos, J., Poelman, M., McCampbell, A., Apostol, B. L., Kazantsev, A., Schmidt, E., Zhu, Y. Z., Greenwald, M., Kurokawa, R., Housman, D. E., Jackson, G. R., Marsh, J. L., Thompson, L. M.: Histone deacetylase inhibitors arrest polyglutamine-dependent neurodegeneration in Drosophila. Nature 413 (2001) 739-43
16. Peterson, C. L., Laniel, M. A.: Histones and histone modifications. Curr Biol 14 (2004) R546-51
17. Strahl, B. D., Allis, C. D.: The language of covalent histone modifications. Nature 403 (2000) 41-5
18. Mukherjea, S.: Information retrieval and knowledge discovery utilising a biomedical Semantic Web. Brief Bioinform 6 (2005) 252-62
19. Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M. C., van Asperen, R., Boon, K., Voute, P. A., Heisterkamp, S., van Kampen, A., Versteeg, R.: The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science 291 (2001) 1289-92
20. Versteeg, R., van Schaik, B. D., van Batenburg, M. F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H. J., van Kampen, A. H.: The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res 13 (2003) 1998-2004

Further reading

21. Antoniou, G., Van Harmelen, F.: A semantic Web primer. MIT Press, Cambridge, Mass. (2004)
22. Cannataro, M., Talia, D.: Semantics and Knowledge Grids: Building the Next-Generation Grid. IEEE Intelligent Systems 19 (2004) 56-63
23. Costa, C.G., Laskey, K.B., Laskey, K.J., Pool, M.: ISWC2005 Notes: Uncertainty Reasoning for the Semantic Web 3. In: Proc. Workshop on Uncertainty Reasoning for the Semantic Web (URSW), part of the 4th International Semantic Web Conference (ISWC) (2005)
24. Hendler, J.: Science and the semantic web. Science 299 (2003) 520-1
25. Morris, R. W., Bean, C. A., Farber, G. K., Gallahan, D., Jakobsson, E., Liu, Y., Lyster, P. M., Peng, G. C., Roberts, F. S., Twery, M., Whitmarsh, J., Skinner, K.: Digital biology: an emerging and promising discipline. Trends Biotechnol 23 (2005) 113-7

26. Schreiber, G.: Knowledge engineering and management: the CommonKADS methodology. MIT Press, Cambridge, Mass. (2000)
27. Searls, D. B.: Data integration: challenges for drug discovery. Nat Rev Drug Discov 4 (2005) 45-58
28. Stevens, R. D., Lord, P. W., McEntire, R., Butler, J. A.: The Eighth Annual Bio-Ontologies Meeting. http://bio-ontologies.man.ac.uk [on line] (2005)

A standardized and dynamic approach for immunogenetics and immunoinformatics: IMGT-Choreography based on the IMGT-ONTOLOGY concepts

Marie-Paule Lefranc

Laboratoire d'ImmunoGénétique Moléculaire, Université Montpellier II, Institut Universitaire de France, Institut de Génétique Humaine, IGH, UPR CNRS 1142, France
lefranc@ligm.igh.cnrs.fr

Abstract. IMGT® , the international ImMunoGeneTics information system® (http://imgt.cines.fr), created in 1989 at Montpellier, France, is a high quality integrated resource specialized in (i) the immunoglobulins, T cell receptors, major histocompatibility complex of human and other vertebrates, (ii) proteins that belong to the immunoglobulin superfamily (IgSF) and to the MHC superfamily (MhcSF), and (iii) related proteins of the immune system (RPI) of any species. IMGT® contains five databases, ten specific interactive tools and 8,000 HTML pages of synthesis and knowledge. Standardization for genome, genetics, proteome and 3D structure data is based on IMGT-ONTOLOGY, the first ontology in immunogenetics. IMGT-ONTOLOGY allows immunogenetics knowledge management by immunoinformatics. IMGT-ONTOLOGY concepts are available for the biologists and IMGT users in the IMGT Scientific chart, and for the computing scientists in IMGT-ML. IMGT-ML includes an XML Schema for each IMGT-ONTOLOGY concept and is used by IMGT Web services to exchange IMGT data. This is the first step towards the implementation of IMGT- Choreography, the bioinformatics process of complex immunogenetics knowledge. IMGT® is freely available at http://imgt.cines.fr.

1 Introduction

The number of genomics, genetics, three-dimensional (3D) and functional data published in the immunogenetics field is growing exponentially, and involves fundamental, clinical, veterinary and pharmaceutical research. The number of potential protein forms of the antigen receptors, immunoglobulins (IG) and T cell receptors (TR) is almost unlimited. The potential repertoire of each individual is estimated to comprise about 10^{12} different IG (or antibodies) and TR, and the limiting factor is only the number of B and T cells that an organism is genetically programmed to produce. This huge diversity is inherent to the particularly complex and unique molecular synthesis and genetics of the antigen receptor chains. This includes biological mechanisms such as DNA molecular rearrangements in multiple loci (three for IG and four for TR in humans) located on different chromosomes (four in humans), nucleotide deletions and insertions at the rearrangement junctions (or N-diversity), and somatic hypermutations in the IG loci (see FactsBooks [1] [2] for reviews).

IMGT® , the international ImMunoGeneTics information system® (http://imgt.cines.fr) [3] [4], was created in 1989, by the Laboratoire d'ImmunoGénétique Moléculaire (LIGM) (Université Montpellier II and CNRS) at Montpellier, France, in order to standardize and manage the complexity of the immunogenetics data. Fifteen years later, IMGT® is recognized as the international reference in immunogenetics and immunoinformatics. IMGT® is a high quality integrated knowledge resource, specialized in (i) the IG, TR, major histocompatibility complex (MHC) of human and other vertebrates, (ii) proteins that belong to the immunoglobulin superfamily (IgSF) and to the MHC superfamily (MhcSF), and (iii) related proteins of the immune systems (RPI) of any species. IMGT® provides a common access to standardized data from genome, proteome, genetics and 3D structures for the IG, TR, MHC, IgSF, MhcSF and RPI [3] [4]. The IMGT® information system consists of databases, tools and Web resources [3]. Databases include four sequence databases, one genome database and one three-dimensional (3D) structure database. Interactive tools are provided for sequence, genome and 3D structure analysis. Web resources ("IMGT Marie-Paule page") comprise 8,000 HTML pages of synthesis and knowledge (IMGT Repertoire, IMGT Scientific chart, IMGT Education, IMGT Index), and external links (IMGT Bloc-notes and IMGT other accesses) [4]. Despite the heterogeneity of these different components, all data in the IMGT® information system are expertly annotated. The accuracy, the consistency and the integration of the IMGT® data, as well as the coherence between the different IMGT® components (databases, tools and Web resources) are based on IMGT-ONTOLOGY [5], the first ontology in immunogenetics and immunoinformatics [5]. IMGT-ONTOLOGY provides a semantic specification of the terms to be used in the domain, and thus, allows the management of immunogenetics knowledge for all vertebrate species.

IMGT-ONTOLOGY concepts are available, for the biologists and IMGT® users, in the IMGT Scientific chart [3] [4], and have been formalized, for the computing scientists, in IMGT-ML [6] [7] which uses XML (Extensible Markup Language) Schema. The IMGT Scientific chart (for biologist agents) and IMGT-ML (for computing agents) are the foundations of the IMGT® data and knowledge management system. In order to extract knowledge from IMGT® standardized immunogenetics data, three main IMGT® biological approaches have been developed: genomics, genetics and structural approaches. On the computer side, this required the modelling of the analysis of the IMGT® components in relation with the concepts. The development of IMGT® Web services using IMGT-ML will allow any IMGT® component to be automatically queried and to achieve a higher level of interoperability within IMGT® and with other information systems. This is the first step towards the implementation of IMGT-Choreography, which corresponds to the process of complex immunogenetics knowledge and to the connection of treatments performed by the IMGT® component Web services.

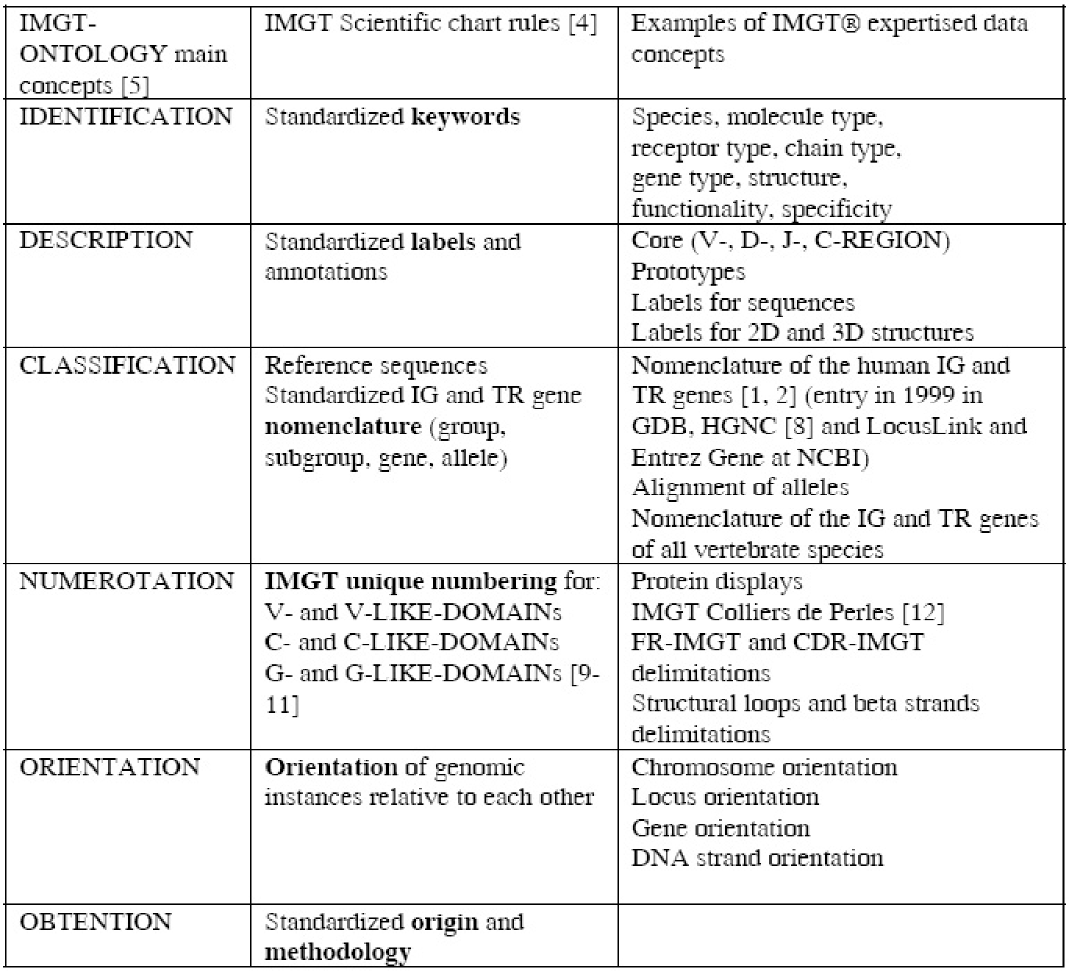
2 IMGT-ONTOLOGY

2.1 IMGT Scientific chart

The IMGT Scientific chart [4] comprises the controlled vocabulary and the annotation rules necessary for the immunogenetics data identification, description, classification and numbering, and for knowledge management in the IMGT® information system.

All IMGT® data are expertly annotated according to the IMGT Scientific chart rules. Standardized keywords, labels and annotation rules, standardized IG and TR gene nomenclature, the IMGT unique numbering, and standardized origin/methodology were defined, respectively, based on the six main concepts of IMGT-ONTOLOGY [5] (Table 1). The IMGT Scientific chart is available as a section of the IMGT® Web resources (IMGT Marie-Paule page). These HTML pages are devoted to biologists, IMGT® users and IMGT® annotators. Examples of IMGT® expertised data concepts derived from the IMGT Scientific chart rules are shown in Table 1.

Table 1. IMGT-ONTOLOGY main concepts, IMGT Scientific chart rules, and examples of IMGT® expertised data concepts.



IMGT-ONTOLOGY main concepts [5]	IMGT Scientific chart rules [4]	Examples of IMGT® expertised data concepts
IDENTIFICATION	Standardized **keywords**	Species, molecule type, receptor type, chain type, gene type, structure, functionality, specificity
DESCRIPTION	Standardized **labels** and annotations	Core (V-, D-, J-, C-REGION) Prototypes Labels for sequences Labels for 2D and 3D structures
CLASSIFICATION	Reference sequences Standardized IG and TR gene **nomenclature** (group, subgroup, gene, allele)	Nomenclature of the human IG and TR genes [1, 2] (entry in 1999 in GDB, HGNC [8] and LocusLink and Entrez Gene at NCBI) Alignment of alleles Nomenclature of the IG and TR genes of all vertebrate species
NUMEROTATION	**IMGT unique numbering** for: V- and V-LIKE-DOMAINs C- and C-LIKE-DOMAINs G- and G-LIKE-DOMAINs [9-11]	Protein displays IMGT Colliers de Perles [12] FR-IMGT and CDR-IMGT delimitations Structural loops and beta strands delimitations
ORIENTATION	**Orientation** of genomic instances relative to each other	Chromosome orientation Locus orientation Gene orientation DNA strand orientation
OBTENTION	Standardized **origin** and **methodology**	

2.2 IMGT-ML

IMGT-ML [6] [7] represents the specification of the main IMGT-ONTOLOGY concepts [5], formalized through an in-house defined mark-up language, based on the

Extensible Markup Language (XML) and constrained through XML Schema. IMGT-ML includes, for each IMGT- ONTOLOGY concept, an XML Schema [6] [7].

- IDENTIFICATION: the “identification” tag, composed of one or more “partIdent” tags, each of them introducing, as attribute, the molecule type (DNA, cDNA ...), the configuration (germline, rearranged ...), gene type (variable, diversity, constant or junction), species, functionality, etc.
- CLASSIFICATION: the “classification” tag, composed of one or more “group”, “subgroup”, “gene”, “allele” tags. The “classification” tag contains the “collection” tag in order to formalize loci with their genes.
- DESCRIPTION: “description” and “annotation” tags gather sequence features with their labels and qualifiers.
- NUMEROTATION: “numerotation” tags introduce “nucSystem” and “proSystem” tags for, nucleotide sequences and amino acid sequences, respectively, within a frame, according to a standardized numbering with gaps and mutations.
- OBTENTION: at the moment the formalization of the “obtention” concept is in progress.

In addition to IMGT-ONTOLOGY tags, tags for factual data, sequences and knowledge have been developed. These tags aggregate IMGT-ONTOLOGY tags, sequence metadata tags (for date, external database references, keywords ...) and literature reference tags. XML is useful both internally for the integration of data and externally for sharing data with other information systems.

3 The IMGT-Choreography biological approaches

Three major IMGT® biological approaches, genomics, genetics and structural approaches, have been selected for the modelling of interactions between the IMGT® components (databases, tools and Web resources). Databases and tools are shown in Figure 1.

Although the IMGT® genome, sequence and 3D structure databases, the IMGT® analysis tools and the IMGT Repertoire Web resources, were initially implemented for the IG, TR and MHC of human and other vertebrates, data and knowledge management standardization has now been extended to the proteins of the immunoglobulin superfamily (IgSF) [13], to the proteins of the MHC superfamily (MhcSF) [14], and to the related proteins of the immune system (RPI) of any species (IMGT Repertoire (RPI)). Thus, standardization in IMGT® contributed to data enhancement of the system and new expertised data concepts were readily incorporated. The IMGT® components in the three IMGT® biological approaches are described in the next sections.

3.1 IMGT® genomics approach

The IMGT® genomics approach is gene-centered and mainly orientated towards the study of the genes within their loci and on the chromosomes (Table 2). Genomic data are managed in IMGT/GENE-DB, which is the comprehensive IMGT® genome database [15]. In November 2005, IMGT/GENE-DB contained 1,377 IG and TR genes and

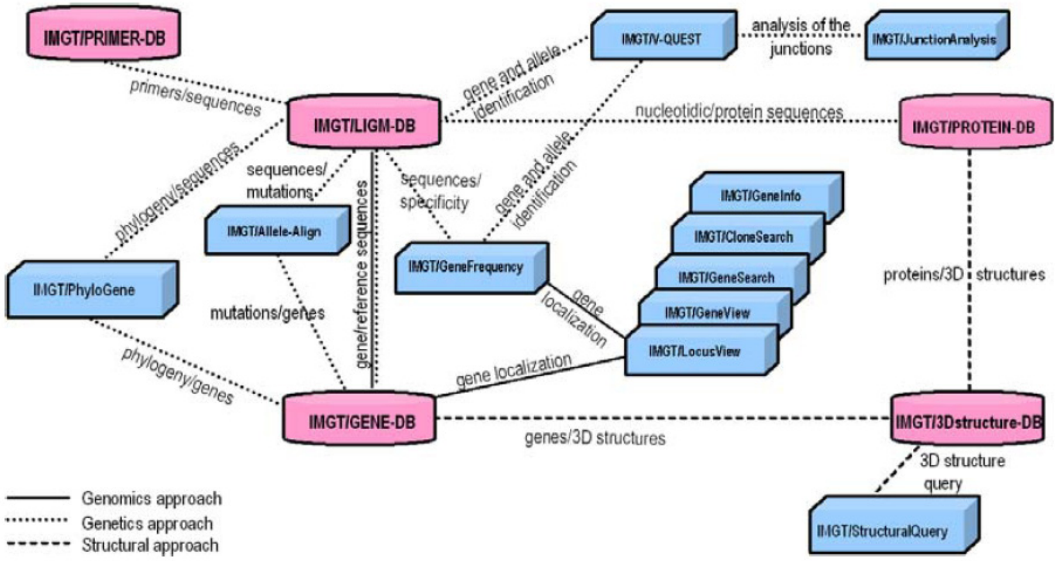


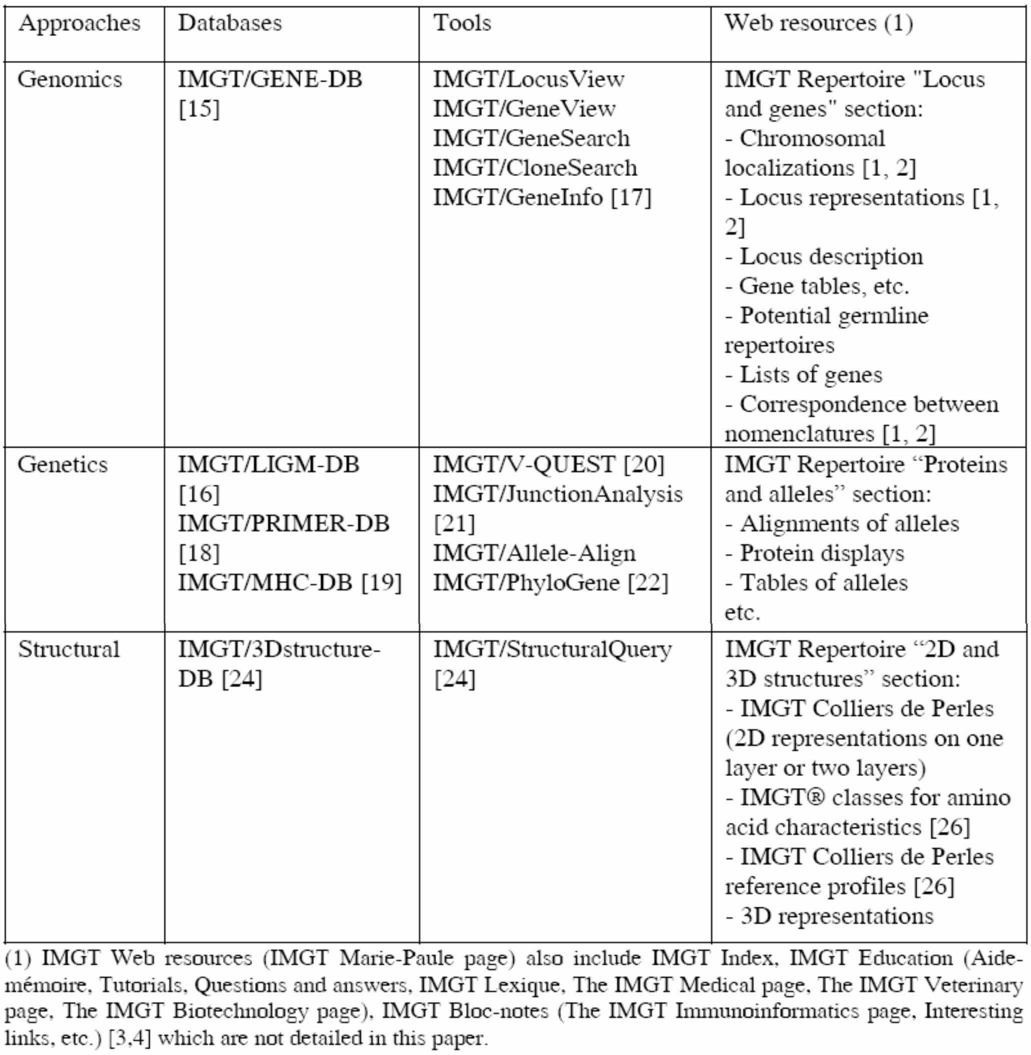
Fig. 1. IMGT-Choreography. Examples of interactions between the IMGT® databases and tools following the three main IMGT® biological approaches: genomics, genetics and structural approaches. The corresponding IMGT Repertoire Web resources (not shown) are described in table 2.

2,207 alleles from human and mouse IG and TR genes. Based on the IMGT® CLASSIFICATION concept, all the human IMGT® gene names [1] [2], approved by the HUGO Nomenclature Committee HGNC in 1999, are available in IMGT/GENE-DB [15] and in Entrez Gene at NCBI (USA). All the mouse IMGT gene and allele names and the corresponding IMGT reference sequences were provided to Mouse Genome Informatics MGI Mouse Genome Database MGD in July 2002 and were presented by IMGT® at the 19th International Mouse Genome Conference IMGC 2005, in Strasbourg, France. IMGT/GENE-DB interacts dynamically with IMGT/LIGM-DB [16] to download and display human and mouse gene-related sequence data. This is the first example of an interaction between IMGT® databases using the CLASSIFICATION concept. The IMGT® genome analysis tools manage the locus organization and gene location and provide the display of physical maps for the human and mouse IG, TR and MHC loci. They allow to view genes in a locus, to search for genes in a locus based on IMGT® gene names, functionality or localization on the chromosome, to provide information on the clones that were used to build the locus contigs (accession numbers are from IMGT/LIGM-DB, gene names from IMGT/GENE- DB), or to display information on the human and mouse IG and TR potential rearrangements.

3.2 IMGT® genetics approach

The IMGT genetics approach refers to the study of the genes in relation with their sequence polymorphisms and mutations, their expression, their specificity and their evolution (Table 2). The IMGT® genetics approach heavily relies on the DESCRIPTION concept (and particularly on the V-, D-, J- and C-REGION core concepts for

Table 2. IMGT-Choreography approaches and IMGT® databases, tools and Web resources



Approaches	Databases	Tools	Web resources (1)
Genomics	IMGT/GENE-DB [15]	IMGT/LocusView IMGT/GeneView IMGT/GeneSearch IMGT/CloneSearch IMGT/GeneInfo [17]	IMGT Repertoire "Locus and genes" section: - Chromosomal localizations [1, 2] - Locus representations [1, 2] - Locus description - Gene tables, etc. - Potential germline repertoires - Lists of genes - Correspondence between nomenclatures [1, 2]
Genetics	IMGT/LIGM-DB [16] IMGT/PRIMER-DB [18] IMGT/MHC-DB [19]	IMGT/V-QUEST [20] IMGT/JunctionAnalysis [21] IMGT/Allele-Align IMGT/PhyloGene [22]	IMGT Repertoire "Proteins and alleles" section: - Alignments of alleles - Protein displays - Tables of alleles etc.
Structural	IMGT/3Dstructure-DB [24]	IMGT/StructuralQuery [24]	IMGT Repertoire "2D and 3D structures" section: - IMGT Colliers de Perles (2D representations on one layer or two layers) - IMGT® classes for amino acid characteristics [26] - IMGT Colliers de Perles reference profiles [26] - 3D representations

(1) IMGT Web resources (IMGT Marie-Paule page) also include IMGT Index, IMGT Education (Aide-mémoire, Tutorials, Questions and answers, IMGT Lexique, The IMGT Medical page, The IMGT Veterinary page, The IMGT Biotechnology page), IMGT Bloc-notes (The IMGT Immunoinformatics page, Interesting links, etc.) [3,4] which are not detailed in this paper.

the IG and TR), on the CLASSIFICATION concept (gene and allele concepts) and on the NUMEROTATION concept (IMGT unique numbering [9]-[11]).

IMGT/LIGM-DB is the comprehensive IMGT® database of IG and TR nucleotide sequences from human and other vertebrate species, created in 1989 by LIGM, Montpellier, France, on the Web since July 1995 [16]. The IMGT/LIGM-DB annotations (gene and allele name assignment, labels) allow data retrieval not only from IMGT/LIGM-DB, but also from other IMGT® databases. As an example, the IMGT/GENE-DB entries provide the IMGT/LIGM-DB accession numbers of the IG and TR cDNA sequences which contain a given V, D, J or C gene. Standardized information on oligonucleotides (or Primers) and combinations of primers (Sets, Couples) for IG and TR are managed in IMGT/PRIMER-DB [17], the IMGT® oligonucleotide database on the Web since February 2002. IMGT/MHC-DB [18] hosted at EBI comprises IMGT/HLA for human MHC (or HLA) and IMGT/MHC-NHP for MHC of non-human primates.

The IMGT® tools for the genetics approach comprise IMGT/V-QUEST [20], for the identification of the V, D and J genes and of their mutations, IMGT/ JunctionAnalysis [21] for the analysis of the V-J and V-D-J junctions which confer the antigen receptor specificity, IMGT/Allele-Align for the detection of polymorphisms, and IMGT/Phylogene [22] for gene evolution analyses. IMGT/V-QUEST (V-QUEry and STandardization) is an integrated software for IG and TR [20]. This tool, easy to use, analyses an input IG or TR germline or rearranged variable nucleotide sequence. IMGT/V-QUEST results comprise the identification of the V, D and J genes and alleles and the nucleotide alignments by comparison with sequences from the IMGT reference directory, the delimitations of the FR-IMGT and CDR-IMGT based on the IMGT unique numbering, the protein translation of the input sequence, the identification of the JUNCTION, and the two-dimensional (2D) IMGT Collier de Perles representation of the V- REGION. IMGT/JunctionAnalysis [21] is a tool, complementary to IMGT/V-QUEST, which provides a thorough analysis of the V-J and V-D-J junction of IG and TR rearranged genes. Several hundreds of junction sequences can be analysed simultaneously. The automatic annotation of rearranged human and mouse cDNA sequences in IMGT/LIGM-DB is performed by IMGT/Automat [23], an internal Java tool which implements IMGT/V-QUEST and IMGT/ JunctionAnalysis.

3.3 IMGT® structural approach

The IMGT® structural approach refers to the study of the 2D and 3D structures of the IG, TR, MHC and RPI, and to the antigen- or ligand-binding characteristics in relationship with the protein functions, polymorphisms and evolution (Table 2). The structural approach relies on the CLASSIFICATION concept (IMGT® gene and allele names), DESCRIPTION concept (receptor and chain description, domain delimitations), and NUMEROTATION concept (amino acid positions according to the IMGT unique numbering [9]-[11]). Structural and functional domains of the IG and TR chains comprise the variable domain or V-DOMAIN (9-strand beta-sandwich) which corresponds to the V-J-REGION or V-D-J-REGION and is encoded by two or three genes [1] [2], the constant domain or C-DOMAIN (7-strand beta-sandwich), and, for the MHC chains, the groove domain or G-DOMAIN (4 beta-strand and one alpha-helix). The IMGT unique numbering has been initially defined for the V-DOMAINs of the IG and TR and for the V-LIKE-DOMAINs of IgSF proteins other than IG and TR [9]. It has been extended to the C-DOMAINs of the IG and TR and to the C-LIKE-DOMAINs of IgSF proteins other than IG and TR [10]. More recently, the IMGT unique numbering has also been defined for the groove domain (G-DOMAIN) of the MHC class I and II chains, and for the G-LIKE-DOMAINs of MhcSF proteins other than MHC [11].

Structural data are compiled and annotated in IMGT/3Dstructure-DB, the IMGT® 3D structure database, on the Web since November 2001 [24]. IMGT/3D-structure-DB comprises IG, TR, MHC and RPI with known 3D structures. Coordinate files extracted from the Protein Data Bank (PDB) [25] are renumbered according to the standardized IMGT unique numbering [9]-[11]. The IMGT/3Dstructure-DB cards provide IMGT® annotations (assignment of IMGT® genes and alleles, IMGT® chain and domain labels, IMGT Colliers de Perles on one layer and two layers), downloadable renumbered IMGT/3Dstructure-DB flat files, vizualisation tools and external links. IMGT/3Dstructure-DB residue cards provide detailed information on the inter- and intra-domain contacts of each residue position.

The IMGT/StructuralQuery tool [24] analyses the intramolecular interactions for the V- DOMAINs. The contacts are described per domain (intra- and inter-domain contacts) and annotated in term of IMGT® labels (chains, domain), positions (IMGT unique numbering), backbone or side-chain implication. IMGT/StructuralQuery allows to retrieve the IMGT/3Dstructure-DB entries, based on specific structural characteristics: phi and psi angles, accessible surface area (ASA), amino acid type, distance in angstrom between amino acids, CDR- IMGT lengths [24].

In order to appropriately analyse the amino acid resemblances and differences between IG, TR, MHC and RPI chains, eleven IMGT® classes were defined for the 'chemical characteristics' amino acid properties and used to set up IMGT Colliers de Perles reference profiles [26]. The IMGT Colliers de Perles reference profiles allow to easily compare amino acid properties at each position whatever the domain, the chain, the receptor or the species. The IG and TR variable and constant domains represent a privileged situation for the analysis of amino acid properties in relation with 3D structures, by the conservation of their 3D structure despite divergent amino acid sequences, and by the considerable amount of genomic (IMGT Repertoire), structural (IMGT/3Dstructure-DB) and functional data available. These data are not only useful to study mutations and allele polymorphisms, but are also needed to establish correlations between amino acids in the protein sequences and 3D structures and to determine amino acids potentially involved in the immunogenicity.

4 The IMGT-Choreography informatics approaches

4.1 IMGT tool diamonds

In order to enhance the interoperability between the IMGT® components, IMGT® tools were analysed for input and output parameters, performed tasks and accompanying databases (IMGT reference directories). Graphical diamond-shaped representations, designated as "IMGT tool diamonds" [4] (Fig. 2) were developed to obtain tool profiles and to compare the state of the art of each tool in relation with the IMGT ontological concepts. Each IMGT tool diamond is composed of 16 modules and each module comprises 4 facets: *left*: input parameters, *bottom*: task, *top*: IMGT reference directory and *right*: output parameters [4]. For a given module, each facet acts as a Boolean switch and indicates whether input parameters are necessary or not, whether a task is performed or not, whether an expertised IMGT reference directory is needed or not, and whether output parameters are provided or not, respectively.

The four modules at the core of the IMGT tool diamond (red) correspond to the major concepts of the tool supported by specific tasks. The 12 outer modules correspond to concepts usually shared with other tools: those of the west pole (blue) correspond to the gene configuration (germline, rearranged or not defined), those of the north pole (orange) to the functionality of the germline sequences (Functional (F), Open Reading Frame (ORF), Pseudogenes (P)), those of the south pole (yellow) to the functionality of the rearranged sequences (productive, unproductive) (IDENTIFICATION concept), and those of the east pole (green) include the labels (DESCRIPTION concept), IMGT unique numbering (NUMEROTATION concept) and the localization and the orientation (ORIENTATION concept).

The IMGT tool diamonds are particularly useful for the IMGT® Web service developers, as they allow to control and to enhance the coherence inside and between the IMGT® tools in the frame of IMGT-Choreography. Indeed, the comparison of two IMGT tool diamonds allows to identify the modules and their “switched on” facets (in colour in the graph), and then, to analyse the expertised IMGT® concepts that are involved and are relevant to both tools. Thus, in the example in Figure 2, three modules are directly relevant to both the IMGT/V-QUEST [20] and IMGT/JunctionAnalysis [21] tools, as the output parameters of each of them (right facet, circled in Fig. 3A) are the necessary input parameters (left facet, circled in Fig. 3B) of the corresponding IMGT/JunctionAnalysis modules (Fig. 2) [4].

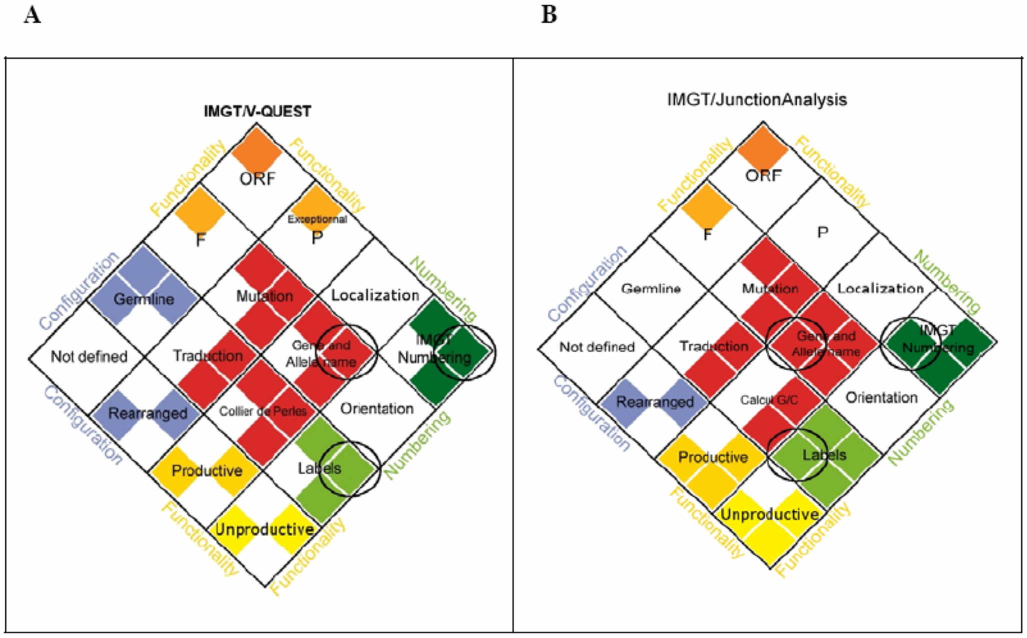


Fig. 2. IMGT tool diamond profiles of two sequence analysis tools, (A) IMGT/V-QUEST [20] and (B) IMGT/JunctionAnalysis [21]. Three modules were selected for analysis: “Gene and allele name” (core, red), “IMGT numbering” and “Labels” (east pole, green). The output parameters of these three IMGT/V-QUEST modules (right facet, circled in (A)) are the necessary input parameters of the corresponding IMGT/JunctionAnalysis modules (left facet, circled in (B)). Note that, in contrast to IMGT/JunctionAnalysis, IMGT/V-QUEST does not require input parameters for these modules (empty facets) (from [4]).

4.2 IMGT® Web services

Web services have been chosen as the means to create dynamic interactions between IMGT® databases and tools. The choice of the Web services to be developed in priority is based on the major existing or potential "conversation nodes" detected in the IMGT biological approaches or with the IMGT tool diamonds.

The Web Service paradigm considers as service any application accessible over Internet fulfilling the requirements of interoperability, weak-coupling and platform independence between applications by making extensive use of open standards, based for example on XML, and existing networking protocols. Precisely, Service Oriented Architectures (SOA) use the Web Services Description Language (WSDL) for the description of new services, the Simple Object Access Protocol (SOAP) ensures communication between services, and the Universal Description, Discovery and Integration (UDDI) protocol enables applications to quickly, easily, and dynamically find and use Web services over the Internet. However, this framework does not specify the underlying semantics of communications. IMGT-SOA introduces a semantic layer by imposing that messages, that are exchanged between service providers and consumers, be encoded using valid IMGT-ML streams. IMGT-ML can be seen as a kind of Rosetta stone since it extends the ease of interconnection between IMGT® Web services. IMGT-ML is the unique language used for both services inputs and outputs, the output of a IMGT® Web service being used as an input for any other relevant Web service. Clients and providers for these services can be written using any SOAP-capable programming language (i.e the SOAP::lite) development library for Perl or webMethods Glue for JAVA) thus facilitating the conversion of legacy applications to services. IMGT® Web services are developed using the JAVA programming language and deployed using the Apache Axis Web services development framework. Apache Axis is an implementation of the SOAP submission to W3C.

The IMGT/LIGM-DB Web service is the first Web service currently developed and implemented with Axis. It includes the "queryKnowledge" and "querySeqData" services [4]. The queryKnowledge service provides the lists of instances for the IMGT-ONTOLOGY concepts, for example the list of chain types, functionalities, specificities defined in the IDENTIFICATION concept, the lists of groups and subgroups defined in the CLASSIFICATION concept, or the list of labels defined in the DESCRIPTION concept. The querySeqData service allows the retrieval of any sequence related data, identified, classified, described according to the IMGT® concepts, such as the nucleotide sequence, the description labels, the literature references, the metadata, etc. The querySeqData input has the form of an incomplete IMGT-ML data entry. The given values are used as criteria to query the database. The result is then a list of data entries, in IMGT-ML format, sharing these given values [4]. Other Web services are developed to automatically query IMGT databases and tools.

5 Conclusions

IMGT-Choreography has for goal to combine and join the IMGT® database queries and analysis tools [27], and thus, to enhance the dynamic interactions between the

IMGT® components to answer complex biological and clinical requests. IMGT-Choreography is based on the Web service architecture paradigm. It orchestrates dynamic procedure calls between databases querying and analysis tools. It will allow any IMGT® component to be automatically queried and to achieve a higher level of interoperability within IMGT® and with other information systems. Conversations between Web services are expressed using IMGT-ML language both for queries and result fetches. This ensures semantic consistency between exchanged messages as IMGT-ML is an XML Schema formalization of the IMGT-ONTOLOGY concepts. In order to keep only significant approaches, a rigorous analysis of the scientific standards, of the biologist requests and of the clinician needs has been undertaken in the three main biological approaches: genomics, genetics and structural approaches. The detailed interactions between IMGT® components are currently being carefully modelled in Unified Modeling LanguageTM (UML) [28].

Since July 1995, IMGT® has been available on the Web at http://imgt.cines.fr. IMGT® has an exceptional response with more than 140,000 requests a month. The information is of much value to clinicians and biological scientists in general. IMGT® databases, tools and Web resources are extensively queried and used by scientists from both academic and industrial laboratories, who are equally distributed between the United States (one-third), Europe (one-third) and the remaining world (one-third). IMGT-Choreography will further increase the IMGT® leadership in immunogenetics and immunoinformatics. IMGT® is used in very diverse domains: (i) fundamental and medical research (repertoire analysis of the IG antibody recognition sites and of the TR recognition sites in normal and pathological situations such as autoimmune diseases, infectious diseases, AIDS, leukemias, lymphomas, myelomas), (ii) veterinary research (IG and TR repertoires in farm and wild life species), (iii) genome diversity and genome evolution studies of the adaptive immune responses, (iv) structural evolution of the IgSF and MhcSF proteins, (v) biotechnology related to antibody engineering (single chain Fragment variable (scFv), phage displays, combinatorial libraries, chimeric, humanized and human antibodies), (vi) diagnostics (clonalities, detection and follow-up of residual diseases) and (vii) therapeutical approaches (grafts, immunotherapy, vaccinology). Owing to its high quality and data distribution based on IMGT-ONTOLOGY, IMGT® has an important role to play in the development of immunogenetics Web services. The design of IMGT-Choreography and the creation of dynamic interactions between the IMGT® databases and tools, using the Web services and IMGT-ML, represent novel and major developments of IMGT® , the international reference in immunogenetics and immunoinformatics.

Citing IMGT®

If you use IMGT® databases, tools and/or Web resources, please cite [3] and this article, and quote the IMGT® Home page URL address, http://imgt.cines.fr.

Acknowledgments

I thank Véronique Giudicelli, Denys Chaume, Patrice Duroux, Chantal Ginestoux, Joumana Jabado-Michaloud, Géraldine Folch, Elodie Duprat, Quentin Kaas and Gé-rard Lefranc for helpful discussion.

I am grateful to the IMGT® team for its expertise and constant motivation and especially to our curators for their hard work and enthusiasm. IMGT® is a registered mark of the Centre National de la Recherche Scientifique (CNRS). IMGT® has received the National Bioinformatics Platform RIO label since 2001 (CNRS, INSERM, CEA, INRA). IMGT® was funded in part by the BIOMED1 (BIOCT930038), Biotechnology BIOTECH2 (BIO4CT960037) and 5th PCRDT Quality of Life and Management of Living Resources programmes (QLG2-2000- 01287) programmes of the European Union (EU). IMGT-ML was developed in collaboration with the EU Online Research Information Environment for the Life Sciences, ORIEL project (IST- 2001-32688). IMGT® is currently supported by the CNRS, the Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche (MENESR) (Université Montpellier II Plan Pluri-Formation, Institut Universitaire de France, ACI-IMPBIO IMP82-2004), GIS AGENAE, Réseau National des Génopoles, Association pour la Recherche sur le Cancer (ARC) and the Région Languedoc-Roussillon BIOSTIC-LR2004.

References

1. Lefranc, M.-P. and Lefranc, G. (2001). The Immunoglobulin FactsBook. Academic Press, London, UK, 458 pages.
2. Lefranc, M.-P. and Lefranc, G. (2001). The T cell receptor FactsBook. Academic Press, London, UK, 398 pages.
3. Lefranc M.-P., Giudicelli V., Kaas Q., Duprat E., Jabado-Michaloud J., Scaviner D., Ginestoux C., Clment O., Chaume D., Lefranc G. (2005) IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res, 33, D593-D597.
4. Lefranc M.-P., Clément, O., Kaas, Q., Duprat, E., Chastellan, P., Coelho, I., Combres, K., Ginestoux, C., Giudicelli, V., Chaume, D. and Lefranc, G. (2005). IMGT-Choreography for Immunogenetics and Immunoinformatics. E pub In Silico Biology 5 0006 http://www.bioinfo.de/isb/2004/05/0006/ 24 December 2004.
5. Giudicelli, V. and Lefranc, M.-P. (1999). Ontology for Immunogenetics: the IMGT-ONTOLOGY. Bioinformatics 12, 1047-1054.
6. Chaume, D., Giudicelli, V. and Lefranc, M.-P. (2001). IMGT-ML a language for IMGT-ONTOLOGY and IMGT/LIGM-DB data. In: CORBA and XML: Towards a bioinformatics integrated network environment, Proceedings of NETTAB 2001, Network tools and applications in biology, 71-75.
7. Chaume, D., Giudicelli, V., Combres, K. and Lefranc, M.-P. (2003). IMGT-ONTOLOGY and IMGT-ML for Immunogenetics and immunoinformatics. In: Abstract book of the Sequence databases and Ontologies satellite event, European Congress in Computational Biology ECCB'2003, September 27-30, Paris, France, pp. 22-23.
8. Wain, H.M., Bruford, E.A., Lovering, R.C., Lush, M.J., Wright, M.W. and Povey, S. (2002). Guidelines for human gene nomenclature. Genomics 79, 464-470.
9. Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V. and Lefranc, G. (2003). IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev. Comp. Immunol. 27, 55-77.
10. Lefranc, M.-P., Pommié, C., Kaas, Q., Duprat, E., Bosc, N., Guiraudou, D., Jean C., Ruiz M., Da Piedade, I., Rouard, M., Foulquier, E., Thouvenin, V. and Lefranc, G. (2005). IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. Dev. Comp. Immunol. 29, 185-203.
11. Lefranc, M.-P., Duprat, E., Kaas, Q., Tranne, M., Thiriot, A. and Lefranc, G. (2005) IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. Dev. Comp. Immunol. 29, 917-938.
12. Ruiz, M. and Lefranc, M.-P. (2002). IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics 53, 857-883.
13. Williams, A.F. and Barclay A.N. (1988). The immunoglobulin family: domains for cell surface recognition. Annu. Rev. Immunol. 6, 381-405.
14. Maenaka K. and Jones E.Y. (1999). MHC superfamily structure and the immune system. Curr. Opin. Struct. Biol. 9, 745-753.

15. Giudicelli, V., Chaume, D. and Lefranc, M.-P. (2005). IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res. 33, D256-D261.
16. Giudicelli, V. Duroux, P., Ginestoux, C. Folch, G., Jabado-Michaloud, J., Chaume, D. and Lefranc, M.-P. (2006). IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res. 34, (in press)
17. Baum, T.P., Pasqual, N., Thuderoz, F., Hierle, V., Chaume, D., Lefranc, M.-P., Jouvin-Marche, E., Marche, P.,N. and Demongeot, J. (2004). IMGT/GeneInfo: enhancing V(D)J recombination database accessibility. Nucleic Acids Res. 32, D51-D54.
18. Folch G., Bertrand J., Lemaitre M. and Lefranc M.-P.(2004). IMGT/PRIMER-DB. In: Database listing. Galperin M.Y. (ed.). The Molecular Biology Database Collection: 2004 update, Nucleic Acids Res. 32, D3-D22.
19. Robinson, J., Waller, M.J., Parham, P., de Groot, N., Bontrop, R., Kennedy, L.J., Stoehr, P. and Marsh, S.G. (2003). IMGT/HLA and IMGT/MHC sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res. 31, 311-314.
20. Giudicelli, V., Chaume, D. and Lefranc, M.-P. (2004). IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. Nucleic Acids Res., 32, W435-W440.
21. Yousfi Monod, M., Giudicelli, V., Chaume, D. and Lefranc, M.-P. (2004). IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. Bioinformatics, 20, I379-I385.
22. Elemento, O. and Lefranc, M.-P. (2003). IMGT/PhyloGene: an on-line tool for comparative analysis of immunoglobulin and T cell receptor genes. Dev. Comp. Immunol. 27, 763-779.
23. Giudicelli, V., Protat, C. and Lefranc, M.-P. (2003). The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat. In: Proceedings of the European Conference on Computational Biology ECCB'2003, September 27-30, Paris, France, DKB-31, pp. 103-104.
24. Kaas, Q., Ruiz, M. and Lefranc, M.-P. (2004). IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. Nucleic Acids Res. 32, D208-D210.
25. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N.and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. 28, 235-242.
26. Pommié, C., Sabatier, S., Lefranc, G. and Lefranc, M.-P. (2004). IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. J. Mol. Recognit. 17, 17-32.
27. Chaume D., Giudicelli V., Combres K., Ginestoux C. and Lefranc M.-P. (2005). IMGT-Choreography: processing of complex immunogenetics knowledge, CMSB 2004, Paris May 26-28 2004, Lecture Notes in Computer Science Springer-Verlag GmbH, ISSN: 0302-9743, vol 3082/2005, pp. 73-84.
28. Cranefield, S. (2001). UML and the Semantic Web. In: Proceedings of SWWS'01, The first Semantic Web Working Symposium, Stanford University, California, US.

Web sites

Apache Axis: http://ws.apache.org/axis/
Extensible Markup Language (XML): http://www.w3.org/XML/
IMGT, the international ImMunoGeneTics information system: http://imgt.cines.fr
IMGT-ML: http://imgt.cines.fr/textes/IMGTindex/IMGT-ML.html
NCBI: http://www.ncbi.nlm.nih.gov/
Simple Object Access Protocol (SOAP): http://www.w3.org/TR/soap/
SOAP::lite: http://www.soaplite.com/
Unified Modeling LanguageTM (UML): http://www.uml.org/
Universal Description, Discovery and Integration (UDDI) protocol:
http://www.uddi.org/about.html
Web Services Description Language (WSDL): http://www.w3.org/TR/wsdl
XML Schema (W3C consortium): http://www.w3.org/XML/Schema

BioGuide: Supporting the Scientist during Data Sources Querying

Sarah Cohen-Boulakia

LRI, CNRS UMR 8023, Université Paris-Sud, Orsay, France
cohen@lri.fr

Abstract. Life sciences are continuously evolving so that the number and size of new sources providing specialized information in biological sciences have augmented significantly in the last few years, as well as the number of tools required to carry out bioinformatics tasks. As a consequence, scientists are increasingly confronted with the problem of selecting appropriate sources and tools. To address this problem, we have designed BioGuide[2], a user-centric framework that helps scientists choose sources and tools according to their *preferences* and *strategy.* BioGuide allows the user to specify his/her query through a user-friendly visual interface.
Availability: http://www.lri.fr/∼cohen/bioguide/bioguide.html.

1 Answering Scientists Requirements

1.1 Introduction

The number and size of new biological data sources[1] together with the number of tools available for analysing this data have increased exponentially in the last few years, to a point where it is unrealistic to expect scientists to be aware of all of them. However, as these sources and tools are often complementary, focused on different objects and reflecting various experts' points of view, scientists should not limit themselves to the sources they already know well, and thus have to face the problem of selecting sources and tools when interpreting their data. The diversity of sources and tools available makes it difficult to perform this selection without assistance.

A questionnaire was developed based on lists of user requirements (cf. [9], [10] and [6]). After interviewing scientists working in various domains, we found that they expressed *preferences* concerning the sources queried and the tools used. Moreover, this study emphasized the fact that the process of querying itself – the *strategy* – varies from one scientist to another. In response to these findings, we introduced BioGuide [2] which assists the scientist with data searches within sources and takes into account his/her strategy and preferences. BioGuide provides information concerning the sequences of sources to be consulted and the tools to be used: the *paths* between sources to be followed.

1.2 Need for transparent queries

BioGuide aims at supporting users in the specification of their queries. Our study of how scientists consider the query process revealed that from a question expressed in

[1] See the annual Nucleic Acids Research database issue (January).

natural language, they first identify the underlying biological entities and the relationships between them. For instance, in question *"On which chromosome is the BAC of my CGH array located?"*, the underlying entities are Chromosome and Bac and the underlying relationship is *isOn*. In BioGuide, the user is supported in this task by a graphical representation of the biological domain, represented through the **entities graph** (Fig.1), in which nodes are biological entities and labeled edges are biological relationships between them (relationships are symmetric). This graph models biological knowledge (e.g. *proteins are encoded by genes*) as well as knowledge about tools (e.g. *proteins and genes may be similar*).

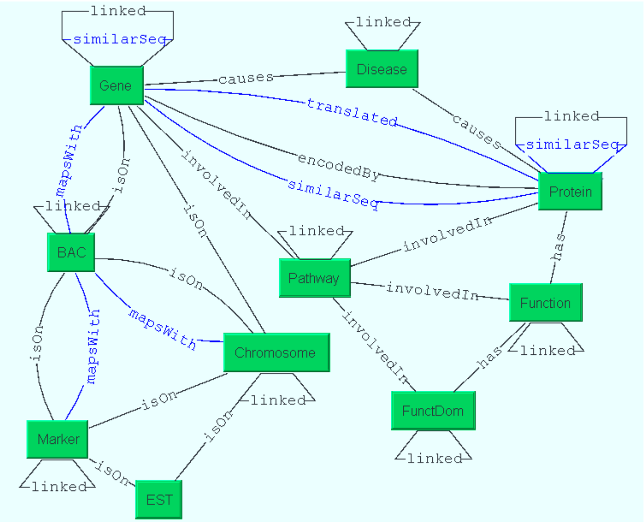


Fig. 1. Entities graph fragment

Scientists can make use of this graph to build BioGuide *transparent queries* (without named sources and tools) by selecting entities and, possibly, relationships between these entities.

1.3 Need for expressing preferences

Answers to questionnaires have also revealed that users punctually need to cite some sources or tools and have preferences on the **kind** of sources to be queried (e.g. access only *reliable sources*). In BioGuide, they are supported in this task by a graphical representation of sources, offered by the **sources-entities graph** (Fig. 2), in which each node represents an entity in a source and arrows indicate the links between two entities (in the same source or in another). Labels on arrows specify the kind of link: cross-reference (*CrossRef*), internal link (*Internal*)– links between entities in the same source – and tools (e.g. *Blast*). For instance, the link $GenBank\_BAC \xrightarrow{Blast} EMBL\_Gene$ means that the *GenBank* source provides a *Blast* tool which can be used to compare the BACs contained in *GenBank* and the genes from EMBL.

Using the sources-entities graph, scientists can thus complete their *transparent query* by an *extended query* in which they (possibly) specify the sources and tools to access or to avoid. This graph is also used to visualize which sources-entities contain

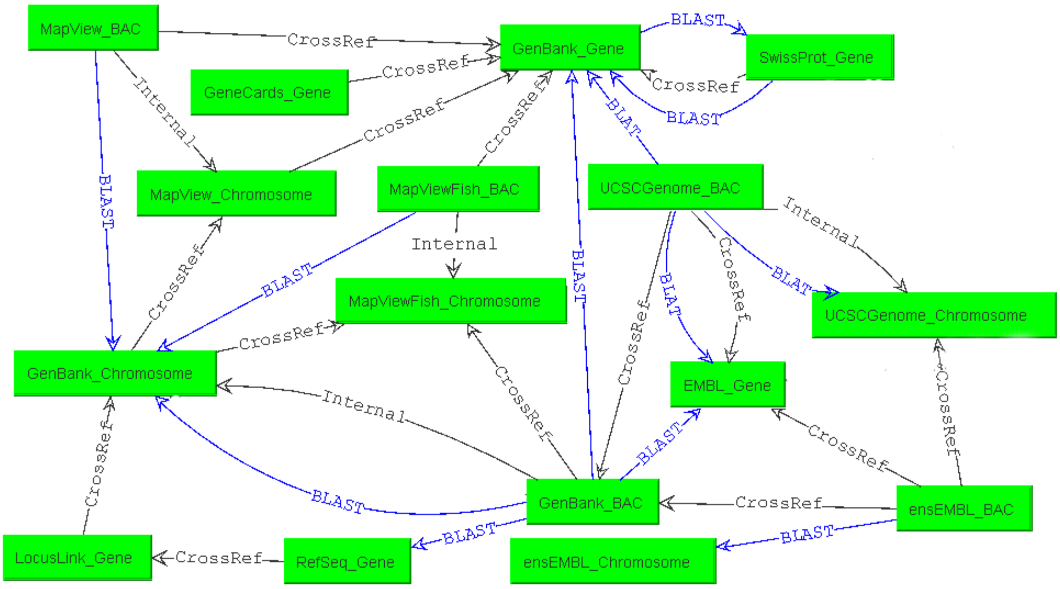


Fig. 2. Sources-entities graph fragment

a given entity and which links achieve a given relationship: the two graphs are *in correspondence* with each other.

1.4 Need for multiple querying strategies

Last but least, interviews revealed that each scientist follows paths between sources and queries the sources by first considering each biological entity for which information was sought and then by linking information about entities by means of cross-references or tools. Since information is collected entity by entity, each entity is treated exactly once.

However, the scientists differed considerably in other aspects of querying, in particular whether or not (i) they followed an order on the entities searched, (ii) they were willing to explore other entities, and (iii) they were willing to visit a source more than once. We term these querying criteria *Ordered*, *OnlyGivenEntities* and *SourceOnceForAll*, respectively, and call the combination of criteria the querying **strategy**.

BioGuide allows the users to express the strategy they want to follow.

2 Ongoing and Further Work

The biological significance of the results obtained with BioGuide has been shown through the task of positioning genomic BAC clones on the draft of the human genome sequence [3] [2]. Moreover, we have newly developed a module [5] to enable the use of BioGuide on top of the well-known SRS platform in order to automatically retrieve instances corresponding to the paths generated by BioGuide.

Furthermore, we have recently defined a complete RDF representation of BioGuide and introduced XPR[4], an RDF path language extending FSL[2]. We plan to study the semantics of this language by following [1]. Finally, BioGuide has been compared with other "paths-based" proposals such as [8] and [7].

References

1. De Bruijn, J., Franconi E., Tessaris, S.: Logical Reconstruction of RDF and Ontology Languages, *Proc. of Workshop on Principles and Practice of Semantic Web Reasoning (PPSWR)*, 65-71 (2005)
2. Cohen-Boulakia, S., Davidson, D., Froidevaux, C.: A User-centric Framework for Accessing Biological Sources and Tools. *Proc. of Data Integration for the Life Sciences (DILS'05)*, Springer-Verlag, Lecture Notes in Computer Science (LNCS) series, LNBI **3615**, 3-18 (2005).
3. Cohen-Boulakia, S., Lair, S., Stransky, N., Graziani, S., Radvanyi, F., Barillot, E., Froidevaux, C.: Selecting biomedical data sources according to user preferences, *Bioinformatics*, **20**, i86-i93 (2004).
4. Cohen-Boulakia, S., Pietriga, E., Froidevaux, C.: Selecting Biological Data Sources and Tools with XPR, a Path Language for RDF, Proc. of Pacific Symposium on Biocomputing (PSB'06) (To Appear), (2006).
5. Cohen-Boulakia, S., Froidevaux, C.: Generating SRS Queries according to User Preferences and Strategy, Technical Report.
6. De Santis, L., Scannapieco, M., Catarci, T.: Trusting Data Quality in Cooperative Information Systems, *Proc. of CoopIS/DOA/ODBASE 2003*, 354-369 (2003).
7. Lacroix, L., Parekh, K., Vidal M., Cardenas, M., Marquez, M.: BioNavigation: Selecting Optimum Paths Through Biological Resources to Evaluate Ontological Navigational Queries, *Proc. of Data Integration for the Life Sciences (DILS'05)*, Springer-Verlag, Lecture Notes in Computer Science (LNCS) series, LNBI **3615**, 275-283 (2005)
8. Mork, P., Halevy, A., Tarczy-Hornoch, P.: A model for data integration systems of biomedical data applied to online genetic databases, *Proc. of AMIA Symposium*, 473-477 (2001).
9. Naumann, F., Leser, U., Freytag, J.C.: Quality-driven Integration of Heterogenous Information Systems, *Proc. of Int. Conf. Very Large DataBases (VLDB)*, 447-458 (1999).
10. Zhao, J., Wroe, C., Goble, C., Stevens, R., Quan, D., Greenwood, M.: Using Semantic Web Technologies for Representing e-Science Provenance *Proc. of Semantic Web Conference (ISWC2004)*, 92-106 (2004).

[2] http://www.w3.org/2005/04/fresnel-info/fsl/

Requirements for natural language understanding in referent-tracking based electronic health records

Werner Ceusters[1,4] and Barry Smith[2,3,4]

[1] European Centre for Ontological Research, Saarbrücken, Germany
[2] Institute for Formal Ontology and Medical Information Science, Saarbrcken, Germany
[3] Department of Philosophy, University at Buffalo, NY, USA
[4] National Center for Biomedical Ontology, University at Buffalo, NY, USA
werner.ceusters@ecor.uni-saarland.de, phismith@buffalo.edu

Abstract. Most electronic patient records contain identifiers to uniquely identify entities such as the patient, the physician, and the healthcare facility. None, however, contains thus far identifiers that uniquely identify the particular disorders patients have, the symptoms they experienced, the actual treatments that have been applied, and so forth. Referent tracking has been introduced as a paradigm to make this also a standard procedure. In this talk, we discuss how natural language understanding can contribute to this.

1 Introduction

Electronic health records (EHRs) consist primarily of descriptions of a patient's medical condition, the treatments administered, and the outcomes obtained. These descriptions are about concrete entities in reality: for example about the particular pain that the particular patient John experienced in his chest on this specific day; or about the particular pacemaker – with its specific serial number assigned to it by its manufacturer – that was implanted in John during the particular surgical procedure that started at a precise moment in time on a certain day.

The descriptions contained in current EHRs contain very few explicit references to such entities. This lack of explicit reference is usually a minor problem for human interpreters, but it makes an accurate understanding of EHR data nearly impossible for machines. This is because reference resolution in running text (still the most common format for descriptions in EHRs) is one of the hardest problems in natural language understanding [1]. But even those EHR systems which incorporate data in more structured formats, for example by resorting to controlled vocabularies, terminologies or ontologies, are in no better shape in this respect. This is because the terms or codes contained in the latter are used simply as an alternative to what would otherwise have been registered by means of general terms in natural language. By picking a code from such a system and then registering that code in an EHR, one refers generically to *some* instance of the class represented by the code. It is still left at best only partially, and indirectly, specified which particular instance is intended in concrete reality.

We therefore argue that, where coding systems and terminologies provide rich vocabularies to describe the entities that exist in reality using general terms, there is a need also for an associated mechanism for expressing *what those descriptions are about*, i.e. the particular entities in reality to which they refer. Such a mechanism is indispensable at the interface where coding systems meet the clinical record if we are to gain maximal advantage from coding efforts and from so-called formal representations and descriptions in EHR systems [2].

2 Referent tracking

Drawing upon our experience in EHR research and standardisation [3] [4] [5], and also from our philosophical research on universals and particulars [6], we introduced referent tracking as a paradigm under which it will become possible to refer explicitly to all of the concrete individual entities relevant to the accurate description of each patient's condition, therapies, and outcomes through the assignment of unique identifiers [7] [8]. Such an identifier is called a IUI (pronounced to rhyme with 'CUI' as used in the UMLS [9]), for 'Instance Unique Identifier'. This means that, not only does each patient receive a IUI, but so also does the particular fracture he is suffering from, the particular bone that is fractured, and even, if the clinician finds this important, the particular pain the patient is experiencing in a certain time period, or the particular document in which the pain is first recorded.

As such, referent tracking goes much further than current practices, under which entities are uniquely identified only when they belong to a restricted range of entity-types, including human beings (the patient himself, the physicians involved), buildings (the hospital in which the patient is treated), certain instruments and devices, and so forth. Moreover, where the majority of entities uniquely identified under current schemes are outside the patient (physicians, instruments, wards, X-ray images), referent tracking extends the facility of unique identification also to the patient's body parts, the specific diseases he has suffered from, the symptoms he has exhibited, and so forth. It goes beyond established approaches also in the degree to which it takes seriously the notion of 'uniqueness'. For where, in many EHR systems, patients and physicians are uniquely identified only relative to some local context (for example of the hospital in which a given EHR-system is used), referent tracking aims for *global* uniqueness.

Note that IUIs refer to the real entities themselves out there in reality, and not to data about these entities. IUIs are the means whereby those constellations of particular entities (tokens, instances) in reality that are relevant to clinical care can be represented in an EHR in the same direct way in which the corresponding classes (types, universals) are already represented by means of clinical coding systems.

Thus IUIs are also *not the entities themselves*. This might seem obvious, but use-mention confusions ('Swimming is healthy and is a concept included in our terminology') – in which an entity in reality and its representation are confounded together – are abundantly present in the literature on knowledge representation in general and on concept-based terminology systems in particular [10].

3 Statements in referent tracking systems

In [8] we explored ways in which the referent tracking paradigm can be implemented in the healthcare environment. Our hypothesis is that, once the right infrastructure is in place, the burden on clinicians and nurses (or on whomever is assigned the task of registering patient data) will be not significantly greater than under existing strategies for data entry – but that the benefits, in terms of semantic interoperability of computer systems and also in terms of patient management, cost containment, epidemiology and disease control, as well as for the advance of science in the domain of biomedicine, can be enormous.

The purpose of a referent tracking system (RTS) is, as its name suggests, to keep track of *referents*. Referents are entities that exist in reality, i.e. in the spatio-temporal world that surrounds us. Most referents are *particulars*, examples being a copy of the manual in which this paper is published, and its authors. Other referents are *universals*, examples being *journal*, *manual*, *person*, H_2O, and so forth.

An RTS will primarily contain information about particulars. The users who enter this information will be required to employ IUIs in order to assure explicit reference to the particulars about which they are providing information. Thus the information that is currently captured in the EHR by means of sentences such as: "this patient has a left elbow fracture", would in the future be conveyed by means of descriptions such as "#IUI-5089 is located in #IUI-7120", together with associated information to the effect that "IUI-7120" refers to the patient under scrutiny, and "IUI-5089" to a particular fracture in patient #IUI-7120 (and not to some similar left elbow fracture from which he suffered earlier). The RTS must correspondingly contain information relating particulars to universals, such as "IUI-5089 **instance_of** fracture" (where 'fracture' might be replaced by a unique identifier pointing to the representation of the universal *fracture* in an ontology) [6].

4 Natural language processing for referent tracking

Of course, EHR systems that endorse the referent tracking paradigm should have mechanisms to capture such information in an easy and intuitive way, including mechanisms to translate generic statements into the intended concrete form, a form which may itself be operative primarily behind the scenes, so that the IUIs themselves remain invisible to the human user. One could indeed imagine that natural language processing (NLP) software will one day be in a position to replace in a reliable fashion the generic terms in a sentence ('John's mother', 'John's pacemaker') with corresponding IUIs for the particulars thereby denoted, with manual support in flagged problematic cases. This corresponds on the level of particulars to what users already expect from EHR systems on the level of universals in supporting entry of codes or terms from coding systems.

The requirements for such natural language analysers are thus:

- To distinguish in clinical narrative the words and phrases that refer to either particulars or universals or (as will commonly be the case) both;
- to identify what specific particulars (among those already described in the RTS) are referred to by the terms and phrases that denote particulars, and what universals (in terms of the foundational ontology and associated domain ontologies linked to the RTS) are referred to by the terms and phrases that denote universals;

- to identify those particulars and universals not thus far recognised within the RTS or the associated ontologies respectively, and to incorporate them appropriately into the system;
- to create adequate linguistic ontologies and link these to foundational ontologies in a way that will enable the machine to infer from the semantic (as understood in linguistics) relationships between the words and phrases in a source text what the ontological relationships are that obtain between the particulars and universals referred to;
- to represent the results of this analysis in such a way that semi-automated population of the RTS is achievable.

References

1. Popescu-Belis A and Lalanne D. Reference resolution over a restricted domain: References to documents. ACL 2004 Workshop on Reference Resolution and its Applications, Barcelona, Spain, July 2004, 71-78.
2. Smith B, Ceusters W. An Ontology-Based Methodology for the Migration of Biomedical Terminologies to Electronic Health Records. AMIA 2005, October 22-26, Washington DC.
3. Zanstra P, Rector A, Ceusters W, de Vries Robbé P. Coding systems and classifications in healthcare: the link to the record. International Journal of Medical Informatics 1998; 48: 103-109.
4. Ceusters W. Language, medical terminologies and structured electronic patient records: how to escape the Bermuda Triangle. In De Moor G, De Clercq E (eds.) Proc MIC, 2000, 7-14.
5. Smith B, Ceusters W. Towards industrial-strength philosophy. How analytical ontology can help medical informatics. Interdisciplinary Science Reviews, 2003; 28 (2): 106-111.
6. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall CJ, Neuhaus F, Rector A, Rosse C. Relations in biomedical ontologies. Genome Biology, 2005 2005, 6 (5), R46.
7. Ceusters W, Smith B. Tracking referents in electronic healthcare records. Forthcoming in Proc MIE 2005.
8. Ceusters W, Smith B. Strategies for Referent Tracking in Electronic Health Records. Journal of Biomedical Informatics. In press. (draft, presented during the IMIA WG6 workshop Ontology and Biomedical Informatics, Rome, Italy, April 29 - May 1, 2005).
9. Powell T, Srinivasan S, Nelson SJ, Hole WT, Roth L, Olenichev V. Tracking meaning over time in the UMLS Metathesaurus. Proc AMIA Symp. 2002, 622-6.
10. Smith B. Beyond concepts, or: Ontology as reality representation. In: Varzi AC, Vieu L (eds.), Formal Ontology and Information Systems. Proc Third International Conference (FOIS 2004), Amsterdam: IOS Press, 2004, 73-84.

Applying ontology design patterns to practical expertise: roles, tasks and techniques in the agricultural domain

Aldo Gangemi

Laboratory for Applied Ontology, CNR-ISTC, Rome, Italy
aldo.gangemi@istc.cnr.it

Abstract. Ontology design patterns help building a rationale for ontology construction, mapping, and evaluation. They are reusable components, but they also contribute to the formalization of relevant expertise in a domain for some task. Within the Agricultural Ontology Service project at UN agency FAO, some patterns of varied logical types have been applied in order to migrate legacy taxonomies, and to capture the expertise for large information service design. The most challenging representation issues concern roles, tasks, techniques, regulations, warning guidelines, etc., which required a complex design pattern based on a semantics for reified social objects. In the seminar I introduce the issues, methods and models employed in the creation of a large Fishery Ontology. Pointers to papers, resources, and related work are given[1], as well as sample explanations and figures.

1 Introduction

Ontology design patterns help building a rationale for ontology construction, mapping, and evaluation. They are reusable components, but they also contribute to the formalization of relevant expertise in a domain for some task. Within the Agricultural Ontology Service project at UN agency FAO, some patterns of varied logical types have been applied in order to migrate legacy taxonomies, and to capture the expertise for large information service design. The most challenging representation issues concern roles, tasks, techniques, regulations, warning guidelines, etc., which required a complex design pattern based on a semantics for reified social objects. In the seminar I introduce the issues, methods and models employed in the creation of a large Fishery Ontology.

2 Formal ontology

Formal ontology, as a research area in ontology engineering, is a set of methods, techniques and models aimed at building and reusing components in ontology projects. A formal ontology typically contains predicates expressing very general notions, richly axiomatized, and with a broad domain coverage [21] [22] [31].

3 Ontology design patterns

A design pattern [11] for an ontology is a fragment of a formal modelling solution, or of a formal ontology, which has some features that enable the easy reuse of models in

[1] Bold-faced references are selected readers for the mentioned topics.

ontology projects for some use case type [1] [24] [25] [26] [30] [36] [37]. It is possible to distinguish between *logical* and *conceptual* ontology design patterns [1] [25] [30], depending on whether the pattern includes or not a specific vocabulary, besides the primitives in the representation language (e.g. OWL).

3.1 The *Description* ↔ *Situation* pattern

Ontology design patterns are specially useful when the expertise models of a domain include complex and highly interrelated notions, which are also dependent on a rich set of contextual assumptions. Typical examples of complex expertise models include the specification of regulations, diagnoses, methods, designs, projects, information objects, etc. All of these examples present us with the need to match a ground, *situated* knowledge to a way of *describing* that knowledge, e.g. a regulation is devised with reference to (non-)conforming social circumstances, diagnoses focus on a set of systemic conditions, projects anticipate possible realizations of behaviors or artifacts, etc.

A pattern with a very general character has been designed in order to match situated and descriptive knowledge, called *Description* ↔ *Situation* [32] [34] [35]. That pattern is a fragment of the social ontology developed as a plugin to the **DOLCE** *foundational ontology* [32] [6], and it introduces a correspondence between logical and social reification (see below).

The patterns presented here are fragments of the **DOLCE-Lite-Plus** ontology [32] [6], developed within the WonderWeb European project [33], which extends DOLCE with the so-called ontology of *Descriptions and Situations* (D&S), a theory that builds upon the reification of contexts, roles, tasks, parameters, and situations. D&S has been applied in many domains [34] [36] [37] [38] for representing methods, norms, plans, etc.

Basic DOLCE top-level includes the following categories and relations:

- **Endurants** (Objects or Substances) and **Perdurants** (Events, States, or Processes) are distinct categories linked by the relation of *participation* (e.g., a group of people participates in an expedition).
- Endurants are *localized in* space, and get their temporal location from the perdurants they participate in. Perdurants are localized in time, and get their spatial location from the endurants participating in them (this is the so-called *participation pattern* [1]).
- **Qualities** *inhere in* either Endurants (as Physical or Abstract Qualities) or in Events (as Temporal Qualities), and they corresponds to "individualized properties", i.e. they inhere only in a specific entity, e.g. "the color of this red herring", "the depth of the water at this point", etc.
- Each kind of Quality is associated to a **Quality Space** representing the space of the values that qualities can assume (e.g. a metric space).
- Quality Spaces, as all **Abstracts** (the fourth category), are neither in time nor in space.
- **Space** and **Time** are specific quality spaces.
- Different kinds of space and time are admitted (e.g. Galilean vs. Newtonian vs. Anatomical).

- Different endurants or perdurants can be spatio-temporally co-localized, e.g. a fish and the anatomical parts it is made of.
- Relations between instances of a same category are contemplated, e.g.: *part*, *constitution*, *connectedness*, etc.

D&S includes the following categories and relations:

- **Descriptions** and **Situations** are distinct categories linked by the relation of *satisfaction.*
- Descriptions are *social objecs*, and get their spatial location from the agents that are able to conceive them, e.g. a fishery technique, depending on the people who know it (and /or its encoding in some document).
- Descriptions define and use **concepts**, another kind of social objects. Special kinds of concepts are **roles**, which can be *played by* some endurant (e.g. crew, captain), **courses**, which can *sequence* some actions or processes (e.g. a route, a set of instructions for a gear), and **parameters**, which must be *valued by* at least one value in a region (e.g. a high exploitation indicator for a stock, a budget).
- Descriptions also define **figures** like FAO or provide an ordering to **information objects** like a web page).
- Situations are *constructed* entities that are logically dependent on descriptions (they must *satisfy* descriptions), e.g. a fishery situation. The *setting* of a situation is constituted by entities that must either play a role, or be sequenced by a course, or be values for a parameter in that description (see below for examples).

Parameters are *requisites for* either roles or courses (e.g. an exploitation indicator for an aquatic resource). Roles can *target* e.g. a *task* (e.g. a captain can be *obliged* to take a certain route).

4 Re-engineering Fishery Knowledge Organization Systems

In the beginning of 2002 the Food and Agriculture Organization of the United Nations (FAO, in the following) took action in order to enhance the quality of its information and knowledge services related to fishery. The FOS project was designed to the creation, integration and utilization of ontologies for information integration and semantic interoperability in fishery information systems. FOS naturally fitted the wider AOS (Agriculture Ontology Service) long-term programme[2], started by FAO at the end of 2001. The Laboratory for Applied Ontology assisted the FAO in the design and development of FOS [5] [7].

The following resources have been singled out from the fishery information systems considered in the project:

OneFish [15] is a portal for fishery projects and a participatory resource gateway for the fisheries and aquatic research and development sector. It is organized as hierarchical *topic trees* (more than 1,800 topics, regularly increasing), topics have brief summaries, identity codes and attached knowledge objects (documents, web sites, various metadata).

AGROVOC thesaurus [16] has been developed by FAO and the Commission of

[2] http://www.fao.org/aims/aos.jsp

the European Communities in the early 1980s and is used for document indexing and retrieval. AGROVOC contains approximately 2,000 fishery related descriptors out of about 16,000 descriptors.

ASFA thesaurus [17] supports an abstracting and indexing service covering the world's literature on the science, technology, management, and conservation of aquatic resources and environments, including their socio-economic and legal aspects. It consists of more than 6,000 descriptors.

FIGIS [18] is a global network of integrated fisheries information. Presently its thematic sections (*reference tables*) are five: aquatic species, geographic objects, aquatic resources, marine fisheries, and fishing technologies. The tables consist of approximately 300 top-level concepts, with a max depth of 4, about 30,000 *objects* and multilingual support.

4.1 Migration

The sources to be integrated are rather variate under many perspectives (semantic, lexical and structural), then they require a reengineering based on a same *framework of reference.*

Once made clear that different fishery information systems provide different views on the domain, we can apply the paradigm of *ontology integration* [19]. In our perspective, thesauri, topic trees and reference tables can be considered as *informal* schemata that have been conceived in order to query semi-structured or informal databases such as texts, forms and tagged documents (positions on this topic can be found in [8] [9] [10] [12]).

In order to benefit from ontology integration, we must transform informal schemata into *formal* ones. Formality is not enough though, because different views will still be different after formalization. That is why interoperability in FOS needed a common framework for KOS reengineering: a comprehensive set of *reference ontologies* that satisfy the following constraints:

- be (partly) *domain-independent* ontologies that are shared by the legacy KOSes
- be *flexible* enough, so that different views are accomodated in a common context
- be focused on the *core reasoning schemata* for the fishery domain, otherwise the common framework would be too abstract.

In the procedure described in [5], after a common format and an integrated ontology data model have been obtained from the source Terminological DataBases (TDB) [14], an Ontology Representation Language has been chosen. Some tests have been performed at the beginning of the project, and we have decided to take a multi-level approach, maintaining the reengineered ontologies into languages of increasing expressivity (and related reasoning services). RDF(S) [27] has been chosen for the basic level, DAML+OIL [28] (currently OWL-DL [21]) for the middle level, and KIF [29] for the expressive level. The KIF version has been used to carry out ontology learning procedures (see phase 4). The OWL-DL version has been used as the standard language to reason over the SW. The RDF(S) version has been used to maintain a lightweight ontology.

For certain terminological data types, a refinement has been performed at this stage and after alignment (see phase 3). For example, AGROVOC makes no difference between descriptors denoting owl:Classes (e.g. agrovoc:River), and descriptors denoting owl:Individuals (e.g. agrovoc:Amazon). Most individuals have been found in subdomains like geography and institutions.

Translation and refinement have been complemented by transforming the applications of RT and of owl:ObjectProperties lifted from FIGIS into formal owl:Restrictions. RT relationships declare associations between classes, and trasformations to ontology datamodel must clarify what is the intended semantics of those associations. We made some working hypotheses in making these transformations:

- RT is a maximally generic owl:ObjectProperty
- an application (triple) of RT to classes is equivalent to an owl:Restriction
- the resulting owl:Restrictions are inheritable to all the subclasses of the owl:Class to which the restriction pertain, and
- the quantification applicable to owl:Restriction derived from RT application is owl:someValuesFrom (the soundness of this hypothesis is mostly empyrical, but also based on the common sense of thesaurus builders.

4.2 Core Ontology of Fishery

The *Core Ontology of Fishery* (COF) provides the backbone to ontology mapping in FOS. COF has been designed by specializing the **DOLCE-Lite-Plus** ontology introduced at the beginning of this primer [7][3].

For example, the *fishery technique* design pattern from COF (Fig.1) [14] is a specialization of D&S. It represents constraints for the entities involved in techniques for fisheries. It states that a cof:Fishery_technique (which is subClassOf Class(edns:technique partial edns:description) has three possible constraint types: cof:Fishery_task, cof:Fishery_role, and cof:Fishery_parameter. A constraint type has subclasses, e.g.:

Class(cof:Route partial cof:Fishery_task)
Class(cof:Fishing_zone partial cof:Fishery_role)
Class(cof:Budget partial cof:Fishery_parameter).

This is the so-called descriptive section of the pattern. The constraints are used to select the entities whose classes are defined in the ground section of the pattern. These can be *actions* like cof:Expedition or cof:Freezing (sequenced by a fishery task), *objects*: cof:Aquatic_organism or cof:Water_area (playing a fishery role), or *attributes*: cof:Exploitation_indicator or cof:Monetary_value (being values for a fishery parameter). The exemplification in Fig.1 suggests that e.g. a tuna fishery situation must comply to an established technique, in the sense that e.g. *expeditions* (activity) must be carried out across certain steps specified in a *route* (a fishery task):

Class(cof:Expedition partial own:Journey$Journeying)
Class(cof:Route partial
    restriction(edns:sequences allValuesFrom(cof:Expedition)))

or that certain *water areas* (endurants) targeted during the expedition play the role

[3] The ontologies mentioned here are available in various languages and formats from: http://dolce.semanticweb.org and http://www.fao.org/aims/onto_domains.jsp.

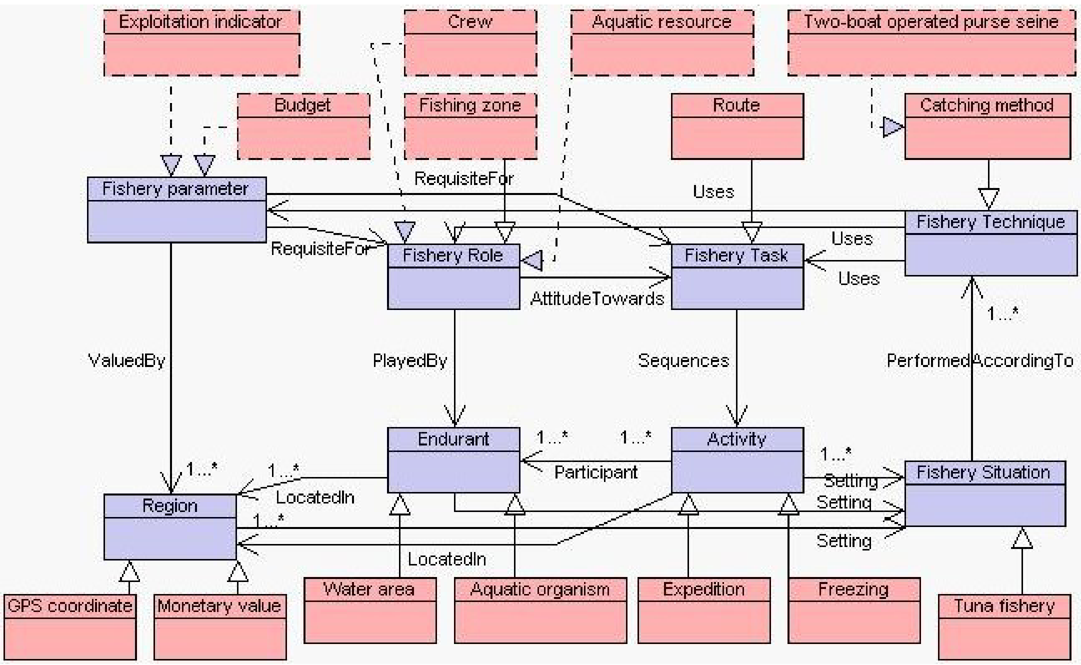


Fig. 1. The *Fishery Technique* ontology design pattern as a specialization of D&S. Dashed boxes are individuals. Topmost and lowest nodes are exemplifications of COF specialization.

of *fishing zone* (a fishery role):

Class(cof:Water_area partial cof:Geographical_object)
Class(cof:Fishing_zone partial
restriction(edns:played_by allValuesFrom(cof:Water_area)))

or that a certain *monetary value* (region of a metric space) is the estimated cost of an expedition with respect to the expected *budget* (fishery parameter):

Class(cof:Monetary_value partial dol:Abstract_region).
Class(cof:Budget partial
restriction(edns:valued_by someValuesFrom(cof:Monetary_value)))

If a set of entities from the ground section of the pattern obeys the constraints provided by the entities in the descriptive section, a cof:Fishery_situation (like a cof:Tuna_fishery) results to be pla:performedAccordingTo some cof:Fishery_technique (e.g. *two-boat operated purse-seine*).

In order to build the COF, we have used TDB top levels, legacy TDB schemata, elicitation from experts, and other ontology design patterns defined elsewhere. In particular, ASFA provided more than 1,600 top-level classes as candidates for the COF, Agrovoc only 83, FIGIS about 400 (including a set of DTDs that control the XML databases of FIGIS). Only about 10% of these candidate classes have been included in the COF, according to the following rationale:

- Some classes are equivalent across sources

- Many classes are not fishery specific, and have been aligned to generic purpose ontologies like WordNet [39] (in the OntoWordNet version [2])
- Many classes have been refined in lower taxonomical positions

The main subdomains represented in the COF as containers for *core* classes of fishery, according to experts' advise, are:

- Biological entities (organisms, anatomy)
- Continental and water areas (geography)
- Ecosystems
- Techniques (capture fishery, aquaculture)
- Vessels and Gears
- Resources, stocks, and management
- Commodities and commercialization
- Institutions and regulations

A summary of the fishery ontology library is depicted in Fig.2.

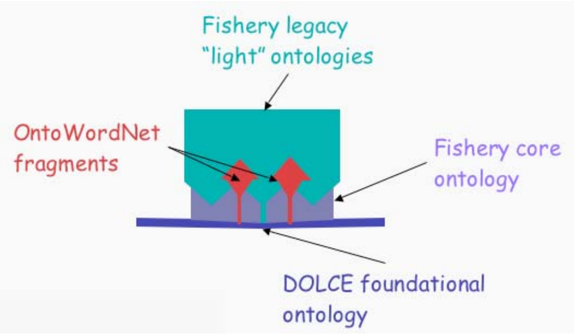


Fig. 2. The FOS ontology library as a "toy house" metaphor. Ground (DOLCE-Lite-Plus), walls (COF), some supporting posts (OntoWordNet), and roof/floors (domain ontologies).

4.3 Mapping

COF has been used (together with other ontologies, especially OntoWordNet) to map the migrated sources. Typical problems have arisen during mapping (alignment, merging). Some of them are mentioned here.

Consistency

For example, the class asfa:Trap_fishing has originally two superclasses: asfa:Catching_methods and asfa:Fishing. From the alignment, we know that asfa:Catching_methods is (transitively) rdfs:subClassOf dlp:object (methods are conceptualized as static objects used for e.g. planning purpose), while asfa:Fishing is (transitively) rdfs:subClassOf dlp:activity. But in DOLCE-Lite-Plus it holds: (disjointClasses dol:object edns:activity), then the attribution of two superclasses to asfa:Trap_fishing leads to inconsistency after the alignment to COF and DOLCE-Lite-Plus (see below for inconsistency management).

BT polysemy

After migration, possible polysemy of the BroaderThan relation became apparent. Decisions for dubious cases have been taken by using some heuristics from foundational or core (*reference*) ontologies. For example, after the translation of the source terminologies, it holds that Class(agrovoc:Blood_Cells partial agrovoc:Blood) (because BT is mapped to rdfs:subClassOf). This is inconsistent on the grounds of a biomedical core ontology (e.g. ON9, which includes the formalization of the UMLS Semantic Network [20] [8]). ON9 contains the following axioms:

```
Class(on9:Blood_cell partial on9:Cell)
Class(on9:Blood partial on9:Tissue)
(DisjointClasses on9:Cell on9:Tissue)
Class(on9:Cell partial
    restriction(on9:finer_grain_component_of someValuesFrom(on9:Tissue)))
```

Therefore, on the basis of ON9, we can conclude that the original BT is polysemous, since a cell cannot be a tissue (the two classes are disjoint), and that the intended meaning of BT could be in this case:

```
Class(agrovoc:Blood_Cells partial
    restriction(dlp:finer_grain_component_of someValuesFrom(agrovoc:Blood)))
```

Emergent polysemy

Alignment generates a lot of potentially redundant ontology elements, because classes (as well as individuals and properties) from different domain ontologies may have the same intended meaning, for example: agrovoc:Trawlers, figis:Trawlers, asfa:Trawlers, or may even show false similarities.

If we had no taxonomic structure, and if class names corresponded 1:1 to intended meaning, the solution would be straightforward: just merge homonym classes into one. Unfortunately, this is not the case, since equivalent classes across ontologies have heterogeneous positions, and since names have a m:n mapping to intended meanings. Heterogeneous position may lead to multiple meanings for the same name across different ontologies (*emergent polysemy*):

```
Class(agrovoc:Dredgers partial agrovoc:Ships)
Class(asfa:Dredgers partial asfa:Work_platforms)
```

AGROVOC's one is a class of fishing vessels, while ASFA's is a class of fishing platforms, while vessels and platforms are disjoint classes.

Emergent synonymy

Another case of m:n mapping shows multiple names for the same meaning across different ontologies (*emergent synonymy*), e.g. asfa:Ships and figis:Non-fishing_vessels have the same intended meaning (according to experts).

Validation and exploitation

Current tools (e.g. [41] [42] [43], a summary in [44]) for bulk merging of ontologies mostly use similarity of the names of class pairs. This technique is appropriate only to the case of emergent polysemy. Moreover, *validation* must be done on the basis of the similarity of superclasses, annotations, and other hints, which require reasoning according to the "components" of intended meanings. These components are mostly represented in reference ontologies. E.g. the *minimal conceptual difference* between a "ship" and a "platform" grounds on notions that do not exist in fishery domain ontologies, but can be encoded in COF or other core ontologies.

In order to solve the validation problem, and to treat emergent synonymy, we have adopted ONIONS [20], which contains several methods for ontology merging. E.g. a (semi-automatic) method splits a domain into finer subdomains. This job is facilitated by reusing the subject trees existing in oneFish and AGROVOC. Another (mostly intellectual) method looks at existing glosses (or elicits new ones), which can be used to learn those minimal conceptual differences (see [14] for examples). Still another adopts relation learning from texts (see [3] for examples in a biochemical domain).

Validation also depends on the intended exploitation of the ontologies [4] [31].

In order to decide on possible exploitation, the reference persons of existing service platforms for Fishery have been interviewed. The OneFish responsible has indicated a list of query patterns (types with examples, [44]) that has been used to define a preliminary taxonomy of query types. Moreover, FAO-GILW has made a questionnaire, and sent it out to final users of fishery IR services, in order to learn what *pull* recommendations should be implemented.

A taxonomy of elementary query types, partly inspired by interviews and questionnaires has been sketched which distinguishes between data, document, and within-document searches. This study enabled us to design and realize a prototype for information retrieval services (synonyms, multilingual access, query expansion, terminology brokering, semantic navigation of bibliographical metadata, ontology navigation), and a mock-up for distributed database querying services. Details on the applications are contained in the documents downloadable from [14].

Several tools have been used for ontology building or exploitation. Making a detailed assessment of the many tools we have considered, and describing the set of functionalities that we want to find in ideal tools is largely besides the scope of this paper. We just mention here some of the Semantic Web and Knowledge Representation tools that we have used: KAON [45], Loom+Ontosaurus [46], OilEd [47], FaCT++ [48], RACER [49], OWL Validator [50], Protégé [51], OCML [52].

References

1. **Gangemi, A. Ontology Design Patterns for Semantic Web Content. Y. Gil et al., Proceedings of ISWC2005, Springer (2005).**
2. Gangemi, A., Navigli, R., Velardi, P.: The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet, Meersman R, et al. (eds.), Proceedings of ODBASE03 Conference, Springer (2003).
3. **Ciaramita M., Gangemi A., Ratsch E., Saric J., and Rojas I., (2005), Unsupervised Learning of Semantic Relations between Concepts of a Molecular Biology Ontology, in Proceedings of the 19th International Joint Conference on Artificial Intelligence.**
4. **Gangemi A., Catenacci C., Ciaramita M., and Lehmann J., (2005), Ontology evaluation: A review of methods and an integrated model for the quality diagnostic task. Technical Report available at http://www.loa-cnr.it/Publications.html.**

5. **Gangemi, A., Fisseha, F., Keizer, J., Lehmann, J., Liang, A. Pettman, I., Sini, M., Taconet, M., A Core Ontology of Fishery and its Use in the Fishery Ontology Service Project, in Gangemi A. and Borgo S. (eds.), Workshop on Core Ontologies in Ontology Engineering, CEUR-WS, vol.64, http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-118/ (2004).**
6. **The DOLCE-Lite-Plus Ontology Library in OWL, http://dolce.semanticweb.org.**
7. **The Fishery ontology repository, http://www.fao.org/aims/onto_domains.jsp.**
8. Pisanelli, D.M., Gangemi, A., Steve, G.: An Ontological Analysis of the UMLS Metathesaurus, J. of American Medical Informatics Association, 5 (1998).
9. Wielinga, R., Schreiber, G., Wielemaker, J, Sandberg, J.A.C.: From Thesaurus to Ontology. Proceedings of the First KCAP Conference, ACM Press, New York (2001).
10. Hahn, U., Schulz, S.: Turning Lead into Gold? Feeding a Formal Knowledge Base with Informal Conceptual Knowledge. In Proceedings of 13th EKAW Conference, Springer, Berlin (2002).
11. Alexander, C.: The Timeless way of building. Oxford University Press, New York (1979).
12. Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S.: Reengineering Thesauri for New Applications: AGROVOC Example, Journal Of Digital Information, 4 (2004).
13. Gangemi, A., Fisseha, F., Pettman, I., Pisanelli, D.M., Taconet, M., Keizer, J.: A Formal Ontological Framework for Semantic Interoperability in the Fishery Domain, in Euzenat J, et al. (eds.), Workshop on Ontologies and Semantic Interoperability, CEUR-WS, vol.64, http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-64/ (2002).
14. Gangemi, A., et al.: FOS Project Final Report, http://www.fao.org/aims/onto_domains.jsp (2004).
15. http://www.onefish.org/global/index.jsp
16. http://www.fao.org/agrovoc/
17. http://www.fao.org/fi/asfa/asfa.asp
18. http://www.fao.org/figis/servlet/FiRefServlet?ds=staticXML&xml=webapps/figis/wwwroot/fi/figis/index.xml&xsl=webapps/figis/staticXML/format/webpage.xsl
19. Calvanese, D., De Giacomo, G., Lenzerini, M.: A Framework for Ontology Integration. Proceedings of the First International Semantic Web Symposium (SWWS) (2001).
20. Gangemi, A., Pisanelli, D.M., Steve, G.: An overview of the ONIONS project. Data & Knowledge Engineering, 31 (1999).
21. McGuinness D.L., van Harmelen F. (eds.): Owl web ontology language overview, W3C Recommendation, (February 2004), http://www.w3c.org/TR/owl-features/ .
22. Guarino, N. and Welty, C.: Evaluating Ontological Decisions with OntoClean. Communications of the ACM 45(2) (2002).
23. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A.: Restructuring WordNet's Top-Level, AI Magazine, Fall (2003).
24. Reich, J.R.: Ontological Design Patterns: Modelling the Metadata of Molecular Biological Ontologies, Information and Knowledge. In DEXA 2000 (2000).
25. **Rector, A.L., Rogers, J. Patterns, Properties and Minimizing Commitment: Reconstruction of the GALEN Upper Ontology in OWL, in [8] (2004).**
26. Svatek V.: Design Patterns for Semantic Web Ontologies: Motivation and Discussion. In: 7th Conference on Business Information Systems, Poznan. (2004).
27. http://www.w3.org/TR/rdf-schema/
28. http://www.daml.org/language/
29. http://cl.tamu.edu/
30. **Semantic Web Best Practices and Deployment Working Group, Task Force on Ontology Engineering Patterns. Description of work, archives, W3C Notes and recommendations available from http://www.w3.org/2001/sw/BestPractices/OEP/ (2004-5).**
31. Gruninger, M., and Fox, M.S.: The Role of Competency Questions in Enterprise Engineering. Proceedings of the IFIP WG5.7 Workshop on Benchmarking – Theory and Practice, Trondheim, Norway (1994).
32. **Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., WonderWeb Deliverable D18: The WonderWeb Library of Foundational Ontologies, http://wonderweb.semanticweb.org (2003).**
33. **http://wonderweb.semanticweb.org**

34. **Gangemi, A., Mika, P.: Understanding the Semantic Web through Descriptions and Situations. In Meersman, R., et al. (eds.), Proceedings of ODBASE03 Conference, Springer, Berlin (2003).**
35. **Masolo, C., Vieu, L., Bottazzi, E., Catenacci, C., Ferrario, R., Gangemi, A., Guarino, N.: Social Roles and their Descriptions, to appear in Welty, C. and Dubois, D. (eds.), Proceedings of the Ninth KR Conference (2004).**
36. **Gangemi, A., Catenacci, C., Battaglia, M.: The Inflammation Ontology Design Pattern: an Exercise in Building a Core Biomedical Ontology with Descriptions and Situations. In Pisanelli D and Smith B, Biomedical Ontologies, IOS Press (2004).**
37. Gangemi, A., Prisco, A., Sagri, M.T., Steve, G., Tiscornia, D.: Some ontological tools to support legal regulatory compliance, in Jarrar, M. et al. (eds.), Proceedings of the WORM03 Workshop at ODBASE03 Conference, Springer, Berlin (2003).
38. Oberle, D., Mika, P., Gangemi, A., Sabou, M.: Foundations for service ontologies: Aligning OWL-S to DOLCE, to appear in Staab S and Patel-Schneider P (eds.), Proceedings of the World Wide Web Conference (WWW2004), (2004).
39. Fellbaum, C., ed.: WordNet - An electronic lexical database. MIT Press (1998).
40. Noy, N., Musen, M. A.: PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. National Conference on Artificial Intelligence, Austin Texas USA (2000).
41. McGuinnes, D. L., Fikes, R. et al.: The Chimaera Ontology Enviroment. National Conference on Artificial Intelligence (AAA2000, Austin,Texas USA (2000).
42. Stumme, G., Maedche, A.: FCA-MERGE: Bottom-Up Merging of Ontologies, in Nebel, B. (ed.), Proc. of the 17th IJCAI Conference, Morgan Kauffman, San Francisco (2001).
43. Gomez-Pérez, A., Fernandez-Lopez, M., Corcho, O.: Ontological Engineering, Springer, Berlin (2004).
44. Baron Varley, J.: personal email to FOS project staff on query improvements (2003).
45. Volz, R., Oberle, D., Staab, S., Motik, B.: KAON SERVER - a Semantic Web Management System, Proceedings of WWW12 Conference, Budapest, Hungary (2003).
46. http://www.isi.edu/isd/LOOM/LOOM-HOME.html
47. Bechhofer, S., Horrocks, I., Goble, C., Stevens, R.: OilEd: a Reason-able Ontology Editor for the Semantic Web, in: Proceedings of KI2001, Springer, Vienna (2001).
48. Tsarkov, D., Horrocks, I.: WonderWeb Deliverable D13, Reasoner Prototype, http://wonderweb.semanticweb.org (2003).
49. http://www.cs.concordia.ca/~haarslev/racer/download.html
50. http://phoebus.cs.man.ac.uk:9999/OWL/Validator
51. http://protege.stanford.edu/
52. Motta, E.: Reusable Components for Knowledge Modelling, IOS, Amsterdam (1999).