RAPPORT DE RECHERCHE

L R I

A PRECISE AUTOMATIC EXTRACTION OF TERMINOLOGY IN GENOMICS

MATTE-TAILLIEZ O / ROCHE M / KODRATOFF Y

Unité Mixte de Recherche 8623
CNRS-Université Paris Sud-LRI

12/2002

Rapport de Recherche N° 1344

CNRS – Université de Paris Sud
Centre d'Orsay
LABORATOIRE DE RECHERCHE EN INFORMATIQUE
Bâtiment 650
91405 ORSAY Cedex (France)

A Precise Automatic Extraction of Terminology in Genomics

Oriane Matte-Tailliez [1, 2], Mathieu Roche [1], Yves Kodratoff [1]
email: oriane, roche, yk@lri.fr
1 Equipe Inférence et Apprentissage, Laboratoire de Recherche Informatique, Université de Paris XI, 91405 Orsay Cedex, France.
2 Equipe Bioinformatique des Génomes, Institut de Génétique et Microbiologie, Université de Paris XI, 91405 Orsay Cedex, France.

Summary

Modern scientists tend to be overloaded with new data and new information. Many research fields work on solutions to this dilemma, such as Statistics (for data), and Natural Language Processing (for information written as texts). Two new research fields are even born since some 10 years, namely Data Mining and Text Mining that attempt to regroup and cross-fertilise all methods dealing with this overload. This paper presents how our success in one unavoidable step of Text Mining, namely the detection of the terms terminology, can be applied to texts relative to the eukaryotic DNA-binding proteins.

Introduction

Genomics developed powerful and large scale methods generating heaps of data, leading to a well-known information overload (see, for instance, Andrade and Valencia, 1998; Palakal et al., 2002). The scientific community acknowledges widely that the process of knowledge extraction from texts is greatly improved when an "ontology" of the domain is available, be it in Genomics or in any other domain dealing with large quantities of texts such as many industrial processes. Our results improve one specific and unavoidable step in this process, viz. the gathering of significant terms ("terminology") that will constitute the nodes of the ontology. Our research topic is the eukaryotic DNA-binding proteins because of the importance of these proteins on spatial and dynamical genome organisation (Gilbert et al., 2000). We are thus applying knowledge extraction from biological texts, specialising into the yeast *Saccharomyces cerevisiae* as being the most suitable eukaryotic organism since so much information relative to it has been already gathered (Guelzim et al., 2002).

To achieve this goal, we are developing an original chain of automatic tools starting with the texts, putting them in a more standardised form ("cleaning"), tagging the words, spotting significant sequences of words ("terminology"), and finally structuring the terms in an ontology. This paper reports on interesting results obtained at the stage of terminology. Note that contrary to other authors such as (Andrade and Valencia, 1998; Collier et al., 2001; Ohta et al., 2002;), we start with the untagged texts themselves, not with a set of examples of what a good term might be. Our approach thus belongs to the so-called "unsupervised learning" approach which can be judged by an expert after the learning has taken place, unlike the "supervised learning" approach, such as used by (Collier et al., 2001), where expert knowledge is provided to the learning system before the automatic learning takes place. A recent example of such an unsupervised approach in the parent field of biomedical methodology is given by (Bodenreider, Rindflesch, and Burgun, 2002).

Our goal is the building of an ontology relative to eukaryotic DNA-binding proteins. In order to simplify the automation of the whole process, we try to obtain an ontology which is as much as possible a taxonomy, where a term belongs to one concept only, whenever this is possible. This problem is known for being very difficult since it implies dealing with polysemy, i.e., words can show many different meanings depending on their context. There is no known way to solve this problem completely, even in a language of speciality since polysemy is such an essential component of natural language, as shown to us by our experience with other fields, such as texts in human resources, and scientific publications (Kodratoff, 2001a; 2001b). One of the possible ways of dealing with polysemy supposes being able to extract terms that are long enough (i.e., they are a composition of several words) to have a unique meaning, and this paper describes a powerful method for generating particular, quite long terms, as strongly suggested by the results of (Bodenreider, Burgun, and Rindflesch, 2002) relative to biomedical terminology, on the problem of the particularisation of nouns by adjectives.

The methodology

1 - Gathering the corpus

We queried the National Library of Medicine (NLM)'s Medline (PubMed) database with the keywords *DNA-binding, proteins, yeast* thus obtaining a corpus of 6119 abstracts. The request was done in Jan. 2002 and we obtained so many interesting terms with this corpus that we concentrated on the improvement of our methods rather than on dealing with corpus maintenance. In the conclusion, we shall see how the method we developed will be useful for corpus maintenance also.

1- Cleaning the corpus

At this step, we perform an homogenisation of these textual data in order to obtain the cleaned corpus. This step is very important since it governs the good behaviour of all further linguistic treatments but it is made of a very large number of rules. Besides some seemingly trivial transforms such as suppressing the authors names, data base formatting, etc. we essentially performed two types of cleaning. The vocabulary is not fixed and, for instance, we replaced by "C-term" all occurrences of "carboxy-terminal", "carboxy termini", "carboxyl terminal", "carboxyl termini", "COOH-terminal", "COOH-termini", "C02H-terminal", and "CO2H termini." This operation is performed by some 100 groups of rules. We also replaced the gene aliases by their generic name, as given in ftp://genome-ftp.stanford.edu/yeast/.../registry.genenames.txt , thus generating 1932 groups of rules.

2-Tagging the corpus

Brill's tagger (Brill, 1994) tags automatically words in context. Tagging associates a grammatical label to the words of the corpus. This is far from being a trivial task because Brill's tagger has been trained on a general corpus, not on the topic of biology. It is thus unable to tag properly English of speciality, for instance 70% of the words of our corpus are unknown to the standard version of Brill's tagger. We had to modify its original rules and we developed what we call "GenoBrill," a version of Brill's adapted to molecular biology and Genomics. To that effect we introduced new rules in GenoBrill. As an example of such a rule: if a word ends by "ine," then attribute to it the label noun, which writes as the following rule :

if ($word =~ /ine$/) push @tags, 'NN'.

Even though quite simple, it was also very important to recognise formula within texts, and we introduced a new tag, "formula," besides the classical ones such as "noun", "adjective," etc.

A total of 30 lexical rules and 4 contextual rules were added by which we improved from a recognition rate of 30% to a rate of 85%.

3-Extracting the relevant terms

From this tagged corpus, we extracted the most relevant terms for the field. The importance of the tagging comes from the fact that many words can have different syntactic roles (such as the same word being adjective or noun), and their behaviour in forming terms depends of these roles. This is why we added to Brill a new syntactic form "formula-noun" describing the cases where a formula is followed by a noun.

The relevance measure relies on a measure of association favouring the association of words that are as seldom as possible associated to any other words. For instance, when considering the term "double-stranded-DNA," DNA appears linked to many other words, and this decreases the relevance of the term, while double-stranded appears practically only when followed by DNA, and this increases the relevance of the term. There exist many such measures and we tested the versions of (Church and Hanks, 1990), (Dunning, 1993), (Jacquemin, 1997) and (Daille, 1998). Our main observation is that the precision of the result, as defined immediately below, does not substantially depend on this choice. For theoretical reasons, we finally decided to chose Jacquemin's that combines pure relevance and the number of instances of the terms in the texts, and is defined as follows. Let x and y be two adjacent words appearing as "x y", n(x) and n(y) their total number of occurrences in the texts, and n(x, y) the number of their common occurrences as the sequence "x y". Their mutual information is given by $I(x, y) = - \log_2 (n(x, y) / (n(x)*n(y)))$. Let nmax be the maximum number of common occurrences for any couple of words, and Imax the maximum mutual information, then the relevance of the term x-y two words x and y, A(x, y), is defined by $A(x, y) = 1/2 ((I(x, y) / Imax) + (n(x, y) / nmax))$. This definition is trivially extended to existing terms in order to form terms of length larger than 2.

The precision measure is performed as follows. Generate terms with a relevance measure RelMeas1. This generates several thousands of terms, of which the first 2000 most relevant are chosen. The biologist studies these 2000 terms and classes them in 4 possible categories: term non relevant to biology, general term relevant to general biology, term relevant to Genomics, specific term relevant to a subfield of biology, different from Genomics. The computed global precision is the ratio of the sum of relevant terms (of the three kinds) to the total number of generated terms.

To compare with RelMeas2, generate the terms following another relevance measure, and the expert classes again the best 2000 terms relative to RelMeas2. The best relevance measure is the one showing the highest precision.

Even though the load on the expert is very large, this work was possible because the results of the measures do not differ too much, and the expert had to class only a few new terms she did not class at the preceding try.

All relevance measures yielded a precision ranging around 82%. This rate is slightly higher than the 79% published on a similar task (Rindflesch et al., 1999) which can be explained by our careful cleaning and, over all, the good performance of our GenoBrill.
In order to significantly increase the precision, we added several heuristics to the relevance measure. The simplest one is that the terms already used by the authors are favoured. For instance, if some authors write "double strand" and others "double-strand" then the formation of the term "double-strand" is favoured. Let us call A the relevance measure of a given term, and let n(x, y) be the number of times it appears as a term in the texts, then we compute an actual relevance R(x, y) = A(x, y) * exp(n(x, y)).
Another simple heuristics takes into account the number of different texts where a term appears. When a term appears in very few different texts, then the effective number of occurrences of this term is decreased. Let N be total number of texts, n_i be the number of occurrences of the term in the i-th text, then we compute an effective number of occurrences, n-effect, by:

$$n - effect = \sum_{i=1}^{N} \left[n_i - \sum_{j=0}^{n_i - 1} \frac{j}{10} \right]$$

Finally, our main heuristics is to obtain the terminology in several iterative steps, while the terms (or words of the first iteration) of the (i-1)-th iteration are favoured at the i-th iteration. The value of the relevance is multiplied by a coefficient computed from the iteration (i-1) by:

$$\frac{3}{2} + F(i-1) * k$$

where F(i-1) is the frequency with which the word is included in a term at iteration (i-1), and where k is a parameter bounding the value of the coefficient.
We also set a lower limit on the number of occurrences necessary for a term to be accepted in our list. This number increases in a simple way with the decrease of the relevance rank of the term, and the increasing number of iterations.

The above formulas were built and their coefficients found by a trial-and-error method where improving the precision defined success. After all these trials, the list of terms actually checked by the expert in the final run numbers exactly 1860 terms. The percent of terms acknowledged as significant by the expert is 88.4%.
Since the list of the all the terms we obtained counts to 9014, among which 2054 coming directly from the authors (thus supposed to be significant), this means that our methodology generated 6960 terms, and that we were able to generate approximately 6152 significant terms. The total of significant terms we are working with is thus 6152 + 2054 = 8206.
Some claim (Tsujii, 2001) that around 10^6 terms are necessary to properly map Cell-Signal Pathway. The relatively "small" number of terms we discovered is nearer to the one found in Gene Ontology. This can be explained by the fact that we deal with abstracts only, or alternately, that such claims are exaggerated. At any rate, be it 10^6 or 10^4, automation will be necessary to generate sets of term properly reflecting the literature.

Validation relative to Gene Ontology

Gene Ontology (GO, http:\www.genontology.org) is one of the most significant terminology base in Genomics. This base offered an assistance for genome annotation (ref), and is already established in a number of biological bases. GO displays some 13000 controlled terms (October 2002), that are nodes of the ontology, with a monthly update. GO is a very reliable system because the relevance of the terms is regularly checked, and the terminology is adapted according to evolution in genomic area, thus terminology and ontology are meticulously controlled by experts. In order to validate the sets of terms we generated, we compared our results to the terms included in this widely used existing ontology. The large differences we then observed lead us to spend a significant amount of effort in understanding these differences. GO is visibly built by experts and for experts, not for an automatic building and exploitation, which is our approach. Being built by experts, it lacks the completeness that can be brought by an automatic approach – and we shall see that we are already able to propose more than 5000 new terms to include to their approximately 13 000. This paper will present a small extract of such terms. Being built for experts, its use for automatic knowledge extraction would be very uneasy, but it is already so well-built that it could be used a starting point for such a use. On the other hand, we shall see that an automatic search of all the relevant biological terms useful in genomic sequence annotation would be an interesting improvement.

In order to perform a large scale comparison between our terminology and GO's, we had to transform all its nodes into terms, according to a standardised syntax. Since they contain various signs such as "(", "<", "\," and English words such as "and", "associated to" etc. it is very uneasy to eliminate them systematically (and correctly!) from a list of terms (Colliers et al., 2001). In such a way, we form thus a list of terms we shall call "GO-ourterms." GO-ourterms does not contain some terms that indeed exist in GO, but that are hardly understandable to a non expert as they are presented. For instance, GO: 0007001 labels "chromosome organization and biogenesis (sensu Eukaria)" which point at two terms, at least for a human specialist, "chromosome-organization-sensu-Eukaria" and "biogenesis-sensu-Eukaria". Note that knowledge of the field is necessary to balance in such a way the words around the "and".

As another example, consider GO:0006139 : nucleobase, nucleoside, nucleotide and nucleic acid metabolism. In this case, the "," is not equivalent to a "-" since "nucleobase" and "nucleoside" are terms by themselves and " nucleobase-nucleoside" is not a valid term (at least, it is obviously not intended to be so in GO).

Another problem comes from synonyms of interest for biological texts, or simple syntactic variants, such as our term "initiation-of-transcription" trivially corresponding to G0:0006352 "transcription-initiation." Note however that, since we did not find significant occurrences of the term "transcription-initiation" in the texts, this means that the authors express themselves as our term shows.

In this way, we generated a GO-ourterms containing 13641 terms. It is nevertheless obvious that we introduce here some error since some terms an expert would recognise in GO were not generated by our automatic method. In comparing our results to GO's terminology, we have thus to compute a total rate of error. In order to achieve this goal, we selected among the 1860 terms validated by the expert the 1428 ones that are not in GO-ourterms, and we sampled 100 of them among these 1428. The expert checked then how many of this 100 were erroneously believed not to belong to GO. It happens that this error rate is very low, around 4%. In other words, we can

say that combining the two different sources of error we introduce (one is by generating non significant terms from the texts, the other one by generating non significant terms from GO) is of the order of 15%. Since the total number of terms we generate and that are not in GO-ourterms is 8553, this means that 1283 of them might be errors, that is we generate at least 7270 valid terms that are not in GO. This shows also that GO's way of presenting the terms is quite systematically done: the exceptions we signalled above are indeed exceptional.

Among the 1428 validated terms not in GO-ourterms, we chose a small set of terms illustrating how these terms could fit into GO. This list of concepts (and their relations) are shown in table 1. Our corpus and the results of the various treatments on this corpus are at http://www.lri.fr/ia/genomics/ .

The terms in table 1 illustrate various relationships between our terms and the ones of GO.

Some of these could be easily incorporated in GO. For example "Hsp90-chaperone" is an obvious instance of GO-existing "chaperone" GO:0003754.

Some others are conceptually vey far from any GO concepts, and it would be interesting to create links relating these terms. For instance, "basic-helix-loop-helix-leucine-zipper-motif" is a child of DNA-binding-protein-motif, itself a child of DNA-binding-protein-domain, itself a child of GO existing DNA-binding GO:0003677. Table 1 shows that our terms contain the necessary intermediates.

Some of our terms could be inserted between two GO concepts. For instance, our term meiotic-prophase can be placed between meiosis (GO:0007126) and meiotic-prophase-I (GO:0007128).

Finally, some GO terms, such as DNA-double-strand-break-processing (GO:0000729) could receive a number of obvious parents, such as our terms double-strand-break and DNA-damage.

cellular-response	cellular-differentiation	pseudohyphal-differentiation	
chaperone (GO:0003754)	Hsp90-chaperone		
chromatin-remodeling	**chromatin-remodeling-complex (GO:0016585)**		
chromosome (GO:0005694)	chromosome-structure	arm-of-chromosome	
CRE-binding-proteins	**cAMP-response-element-binding-protein-binding (GO:0008140)**		
DNA-binding (GO:0003677)	DNA-binding-domain		
DNA-binding (GO:0003677)	DNA-binding-protein	activation-domain	glutamine-rich-activation-domain
DNA-binding (GO:0003677)	DNA-binding-protein		
DNA-binding (GO:0003677)	DNA-binding-protein-family	TATA-box-binding-protein	
DNA-binding (GO:0003677)	DNA-binding-site	Gal4-binding-sites	
DNA-binding-protein-domain	basic-zipper		
DNA-binding-protein-domain	DNA-binding-protein-motif	basic-helix-loop-helix-leucine-zipper-motif	
DNA-binding-protein-domain	zinc-finger-domain		
DNA-damage	DNA-damaging-agent		
DNA-damage	double-strand-break	**DNA-double-strand-break-processing (GO:0000729)**	
DNA-double-strand-breaks	illegitimate-recombination		
DNA-recombination (GO:0006310)	***	holliday-junction	
DNA-recombination (GO:0006310)	illegitimate-recombination		
DNA-recombination (GO:0006310)	intrachromosomal-recombination		
DNA-replication-initiation (GO:0006270)	initiation-of-chromosomal-DNA-replication		
DNA-replication-licensing (GO:0030174)	minichromosome-maintenance		
DNA-sequence	promoter-sequence	cis-acting-element	
DNA-sequence	DNA-element	cis-acting-element	

DNA-sequence	DNA-element	MADS-box		
DNA-sequence	repeat-sequence	inverted-repeat		
DNA-sequence	repeat-sequence	telomeric-repeat		
DNA-transposition (GO:0006313)	transposable-element			
double-strand-break-repair (GO:0006302)	DSB-induced-gene-conversion			
double-strand-break-repair (GO:0006302)	Rad51-family			
extragenic-suppressor				
histone-acetyltransferase (GO:0004402)	***	histone-acetyltransferase-1		
mating (GO:0007618)	silent-mating-type			
meiosis (GO:0007126)	early-meiotic-gene			
meiosis (GO:0007126)	meiotic-prophase	**meiotic-prophase-I (GO:0007128)**		
meiosis (GO:0007126)	meiotic-prophase	**meiotic-prophase-II (GO:0007136)**		
osmotic-stress	**response-to-osmotic-stress (GO:0006970)**			
oxidative-stress	**response-to-oxidative-stress (GO:0006979)**			
protein-folding	Hsp90-chaperone			
response-to-heat (GO:0009408)	heat-shock-factor			
response-to-nitogen-starvation (GO:0006995)	pseudohyphal-differentiation			
RNA-modification (GO:0009451)	posttranscriptional-modification			
RNA-polymerase	**DNA-directed-RNA-polymerase (GO:0003899)**			
RNA-polymerase	**RNA-directed-RNA-polymerase (GO:0003968)**			
RNA-splicing (GO:0008380)	***	alternative-splicing		
single-strand-DNA-binding (GO:0003697)	single-stranded-DNA-binding-protein			
spindle (GO:0005819)	mitotic-spindle	**mitotic-spindle-assembly (GO:0007052)**		
spindle (GO:0005819)	mitotic-spindle	**mitotic-spindle-elongation (GO:0000022)**		
spindle (GO:0005819)	mitotic-spindle	**mitotic-spindle-positionning (GO:0018986)**		
telomere (GO:0005696)	***	telomeric-repeat		
transcriptional-activator (GO:0016563)	Gal4-DNA-binding-domain			
transcriptional-activator (GO:0016563)	PACE-binding-protein			
transcription-factor (GO:0003700)	GATA-family			
transcription-factor (GO:0003700)	heat-shock-factor			
transcription-factor-binding-sites	E-box-binding-protein			
transcription-initiation (GO:0006352)	promoter-sequence	TATA-box		
transcription-initiation (GO:0006352)	start-site-of-transcription			
transcription-regulator (GO:0030528)	repressor-activator			
transport-protein	ATP-binding-cassette-transporters			
zinc-finger-protein	zinc-finger-domain	LIM-homeodomain		
zinc-finger-protein	zinc-finger-regions	zinc-finger-domain	zinc-finger-motif	C2H2-zinc-finger-motif
zinc-finger-protein	zinc-finger-regions	zinc-finger-domain	zinc-finger-motif	C4-zinc-finger-motif

Table 1: *A small extract showing how our terms could be inserted into GO*
The terms are ranked by generality. The more to the left, the more general is the term.
Terms belonging to GO are written with bold letters, terms we discovered and that are not in GO are in thin letters.
*We insert a set of *** when we think that intermediate terms would be necessary to clarify the generality relationship.*
When a parent has several children, the parent is repeated.

A side effect of the way GO completes its ontology brings a further validation to our approach. The queried terms are looked upon by GO's keepers, and those worthwhile are added to the ontology. We noticed that most of the terms we queried, and that were not in GO in October 02

are now inserted into it. We do not claim that our query is the reason why it was included, but we claim that it shows that the terms we discovered in the texts are of interest to the Genomics community.

Conclusion

The aim of this paper is showing that an automatic term generation is worthwhile if we want to build accurate specialised ontologies. Our corpus is obviously incomplete but, since we obtained good results from it, it shows that improving text analysis is at least as much important as text gathering. For instance, now that our term generator is shown to work satisfactorily, we can apply it to any new texts on the topic of Genomics, and be able to track the new terms appearing in the literature on an almost daily basis. This is all the most interesting since even if the nomenclature of genes and associated proteins for the well-known organism *Saccharomyces cerevisiae* is already established, we should not forget that the current functional and structural genomic techniques leads to the determination of proteic functions hitherto unknown. It is thus of a great interest to build methods enabling continuous work on the terminology extraction of such a field.
Similarly, applying our methodology to another subfield of biology might ask some more effort in order to build several specific "BioBrill," but once this task is achieved, it becomes very easy to gather terms specific to a specific area of Biology.
That GO be still incomplete is not the point we want to raise here. This is obvious, and GO itself is constantly under improvement. The problem we raise as critics is rather the one of GO's browser : in order to find a term in GO, it is necessary to query it exactly the way it is written, including unexpected comas or hyphens. Instead of using (for instance) Google's browser as a solution, and when our terms are semantically equivalent to some of GO, ours being the ones of the literature, they could be used as equivalent in order to directly improve GO's browser.
GO being a specialised ontology, it contains many highly specialised terms, that is, terms made of five or more words. Since we dealt with summaries only, the author do not use the terms in their full extent, and we miss many of these terms. We are quite aware that we would much improve our terminology by dealing with the full texts, and our next effort will go in that direction. In this way, we would be able to compare our results even to the more complete lists available from E-BioSci, EMBO's initiative at http://www.e-biosci.org to set up a platform that will provide services relating to access and retrieval of digital information in the life sciences, ranging from bibliographic or factual data to published full text.
Inversely, we find many general terms made of two or three words are not included in GO, which is normal since GO is aimed at people that are supposed to know about these concepts. The use of an ontology for information extraction, or any other automatic treatment of the texts, demands completion, and this is another gain of our automated approach.

Bibliography

Andrade M.A. & Valencia A. 1998. "Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families," *Bioinformatics.* **14**:600-607.

Bodenreider O., Burgun A., Rindflesch T. C. "Assessing the consistency of a biomedical terminology through lexical knowledge," *Int. J. Med. Inf.* **67**:85-95, 2002.

Bodenreider O., Rindflesch T. C., Burgun A. "Unsupervised, corpus-based method for extending biomedical terminology," NAACL Workshop NLP in the Biomedical domain, Philadelphia, July 2002.

Brill E. Some Advances in Transformation-Based Part of Speech Tagging. 1994. *AAAI*, **1**:722-727.

Church K. and Hanks P. Word Association Norms, Mutual Information, and Lexicography. 1990. *Computational Linguistics*, **16**: 22-29.

Collier N., Nobata C., Tsujii J. "Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain," *Journal of Terminology 7*, 239-258, 2001.

Daille B., Gaussier E. & Langé J.-M. An Evaluation Of Statistical Scores for Word Association. *J. Ginzburg, Z. Khasidashvili, C. Vogel, J.-J. Levy & E. Vallduvi (eds) The Tbilisi Symposiumon Logic, Language and Computation : Selected Papers, CSLI Publications,* pp. 177-188, 1998.

Gilbert D. R., Schroeder M., van Helden J. "Itnteractive visualization and exploration of relationships between biological objects," *Trends Biotechnol.* **18**: 487-494, 2000.

Guelzim N., Bottani S., Bourgine P., Kepes F. "Topological and causal structure of the yeast transcriptional regulatory network," *Nature Genetics* **31**, 2002, 60-63.

Jacquemin C. Internal publication of Université de Nantes, 1997.

Kodratoff Y. "Comparing Machine Learning and Knowledge Discovery in DataBases: An Application to Knowledge Discovery in Texts," in *Machine Learning and Its Applications* G. Paliouras, V. Karkaletsis, C.D. Spyropoulos (Eds.): LNAI 2049, p. 1-21, Springer Verlag, 2001b.

Kodratoff Y. "On the Induction of "Interesting" Rules," *Noesis*, XXVI, pp. 103-124, 2001a.

Rindflesch T. C. Hunter L., Aronson A. R. Mining molecular binding terms from biomedical text. Proceedings of the 1999 AMIA Fall Symposium, 127-31.

Palakal M., Mukhopadhyay S., Mostafa J., Raje R., N'Cho M., Mishra S. "An intelligent biological information managem,ent system," *Bioinformatics* **18**, 2002, 1283-1288.

Tsujii J. 2001, Chapter 5 of "Joint proposal for a tutorial: Mining the Biomedical Literature," unpublished, available at http://ismb01.cbs.dtu.dk/pdf/prop10.pdf.

RAPPORTS INTERNES AU LRI - ANNEE 2002

N°	Nom	Titre	Nbre de pages	Date parution
1300	COCKAYNE E J FAVARON O MYNHARDT C M	OPEN IRREDUNDANCE AND MAXIMUM DEGREE IN GRAPHS	15 PAGES	01/2002
1301	DENISE A	RAPPORT SCIENTIFIQUE PRESENTE POUR L'OBTENTION D'UNE HABILITATION A DIRIGER DES RECHERCHES	81 PAGES	01/2002
1302	CHEN Y H DATTA A K TIXEUIL S	STABILIZING INTER-DOMAIN ROUTING IN THE INTERNET	31 PAGES	01/2002
1303	DIKS K FRAIGNIAUD P KRANAKIS E PELC A	TREE EXPLORATION WITH LITTLE MEMORY	22 PAGES	01/2002
1304	KEIICHIROU K MARCHE C URBAIN X	TERMINATION OF ASSOCIATIVE-COMMUTATIVE REWRITING USING DEPENDENCY PAIRS CRITERIA	40 PAGES	02/2002
1305	SHU J XIAO E WENREN K	THE ALGEBRAIC CONNECTIVITY, VERTEX CONNECTIVITY AND EDGE CONNECTIVITY OF GRAPHS	11 PAGES	03/2002
1306	LI H SHU J	THE PARTITION OF A STRONG TOURNAMENT	13 PAGES	03/2002
1307	KESNER D	RAPPORT SCIENTIFIQUE PRESENTE POUR L'OBTENTION D'UNE HABILITATION A DIRIGER DES RECHERCHES	74 PAGES	03/2002
1308	FAVARON O HENNING M A	UPPER TOTAL DOMINATION IN CLAW-FREE GRAPHS	14 PAGES	04/2002
1309	BARRIERE L FLOCCHINI P FRAIGNIAUD P SANTORO N	DISTRIBUTED MOBILE COMPUTING WITH INCOMPARABLE LABELS	16 PAGES	04/2002
1310	BARRIERE L FLOCCHINI P FRAIGNIAUD P SANTORO N	ELECTING A LEADER AMONG ANONYMOUS MOBILE AGENTS IN ANONYMOUS NETWORKS WITH SENSE-OF-DIRECTION	20 PAGES	04/2002
1311	BARRIERE L FLOCCHINI P FRAIGNIAUD P SANTORO N	CAPTURE OF AN INTRUDER BY MOBILE AGENTS	16 PAGES	04/2002
1312	ALLARD G AL AGHA K	ANALYSIS OF THE OSSC MECHANISM IN A NON-SYNCHRONOUS TRANSMISSION ENVIRONMENT	12 PAGES	04/2002
1313	FOREST J	A WEAK CALCULUS WITH EXPLICIT OPERATORS FOR PATTERN MATCHING AND SUBSTITUTION	70 PAGES	05/2002
1314	COURANT J	STRONG NORMALIZATION WITH SINGLETON TYPES	19 PAGES	05/2002
1315	COURANT J	EXPLICIT UNIVERSES FOR THE CALCULUS OF CONSTRUCTIONS	21 PAGES	05/2002
1316	KOUIDER M LONC Z	STABILITY NUMBER AND (a,b)-FACTORS IN GRAPHS	12 PAGES	05/2002
1317	URBAIN X	MODULAR AND INCREMENTAL PROOFS OF AC-TERMINATION	20 PAGES	05/2002

RAPPORTS INTERNES AU LRI - ANNEE 2002

N°	Nom	Titre	Nbre de pages	Date parution
1318	THION V	A STRATEGY FOR FREE-VARIABLE TABLEAUX FOR VARIANTS OF QUANTIFIED MODAL LOGICS	12 PAGES	05/2002
1319	LESTIENNES G GAUDEL M C	TESTING PROCESSES FROM FORMAL SPECIFICATIONS WITH INPUTS, OUTPUTS AND DATA TYPES	16 PAGES	05/2002
1320	PENT C SPYRATOS N	UTILISATION DES CONTEXTES EN RECHERCHE D'INFORMATIONS	46 PAGES	05/2002
1321	DELORME C SHU J	UPPER BOUNDS ON THE LENGTH OF THE LONGEST INDUCED CYCLE IN GRAPHS	20 PAGES	05/2002
1322	FLANDRIN E LI H MARCZYK A WOZNIAK M	A NOTE ON A GENERALISATION OF ORE'S CONDITION	8 PAGES	05/2002
1323	BACSO G FAVARON O	INDEPENDENCE, IRREDUNDANCE, DEGREES AND CHROMATIC NUMBER IN GRAPHS	8 PAGES	05/2002
1324	DATTA A K GRADINARIU M KENITZKI A B TIXEUIL S	SELF-STABILIZING WORMHOLE ROUTING ON RING NETWORKS	20 PAGES	06/2002
1325	DELAET S HERAULT T JOHNEN C TIXEUIL S	ACTES DE LA JOURNEE RESEAUX ET ALGORITHMES REPARTIS, 20 JUIN 2002	52 PAGES	06/2002
1326	URBAIN X	MODULAR AND INCREMENTAL AUTOMATED TERMINATION PROOFS	32 PAGES	06/2002
1327	BEAUQUIER J JOHNEN C	ANALYZE OF RANDOMIZED SELF-STABILIZING ALGORITHMS UNDER NON-DETERMINISTIC SCHEDULER CLASSES	18 PAGES	06/2002
1328	LI H SHU J	PARTITIONING A STRONG TOURNAMENT INTO k CYCLES	14 PAGES	07/2002
1329	BOUCHERON S	RAPPORT SCIENTIFIQUE PRESENTE POUR L'OBTENTION D'UNE HABILITATION A DIRIGER DES RECHERCHES	97 PAGES	08/2002
1330	JOHNEN C	OPTIMIZATION OF SERVICE TIME AND MEMORY SPACE IN A SELF-STABILIZING TOKEN CIRCULATION PROTOCOL ON ANONYMOUS UNIDIRECTIONAL RINGS	21 PAGES	09/2002
1331	LI H SHU J	CYCLIC PARTITION OF STRONG TOURNAMENTS	15 PAGES	09/2002
1332	TZITZIKAS Y SPYRATOS N	RESULT FUSION BY MEDIATORS USING VOTING AND UTILITY FUNCTIONS	30 PAGES	09/2002
1333	AL AGHA K	RAPPORT SCIENTIFIQUE PRESENTE POUR L'OBTENTION D'UNE HABILITATION A DIRIGER DES RECHERCHES	63 PAGES	10/2002
1334	ALVAREZ-HAMELIN J I FRAIGNIAUD P	REDUCING PACKET-LOSS BY TAKING LONG RANGE DEPENDENCES INTO ACCOUNT	20 PAGES	10/2002

RAPPORTS INTERNES AU LRI - ANNEE 2002

N°	Nom	Titre	Nbre de pages	Date parution
1335	EGAWA Y ENOMOTO H FAUDREE R J LI H SCHIERMEYER I	TWO-FACTORS EACH COMPONENT OF WHICH CONTAINS A SPECIFIED VERTEX	16 PAGES	10/2002
1336	LI H WOZNIAK M	A NOTE ON GRAPHS CONTAINING ALL TREES OF GIVEN SIZE	10 PAGES	10/2002
1337	ENOMOTO H LI H	PARTITION OF A GRAPH INTO CYCLES AND DEGENERATED CYCLES	10 PAGES	10/2002
1338	BALISTER P N KOSTOCHKA A V LI H SCHELP R H	BALANCED EDGE COLORINGS	20 PAGES	10/2002
1339	HAGGKVIST R LI H	LONG CYCLES IN GRAPHS WITH SOME LARGE DEGREE VERTICES	16 PAGES	10/2002
1340	DRACH-TEMAM N	RAPPORT SCIENTIFIQUE PRESENTE POUR L'OBTENTION D'UNE HABILITATION A DIRIGER DES RECHERCHES	96 PAGES	11/2002
1341	FLANDRIN E LI H SHU J	A SUFFICIENT CONDITION FOR CYCLABILITY IN DIRECTED GRAPHS	18 PAGES	12/2002
1342	HU Z LI H	PARTITION OF A GRAPH INTO CYCLES AND VERTICES	16 PAGES	12/2002
1343	DJELLOUL S KOUIDER M	MINIMUM k-SELF-REPAIRING GRAPHS	16 PAGES	12/2002