

Contents

• Alternative Splicing Special Interest Group Meeting	pag. 1
o About the workshop	pag. 2
o Background and Aims	pag. 2
o Program	pag. 3
o The Complex World of Splicing Regulatory Sequences	pag. 5
o Genome-Wide Identification of Alternative Isoforms Targeted by Nonsense-Mediated mRNA Decay in Drosophila	pag. 7
o Genome-wide Prediction of Novel Alternative Splicing Events	pag. 8
o The consequences of alternative splicing on biological pathways	pag. 10
o Patterns in Alternative Splicing - Interaction between splicing and transcription components	pag. 11
o Pervasive unproductive splicing of SR proteins associated with ultraconserved elements	pag. 13
o Hundreds of new alternatively spliced variants discovered in C.elephas Genome	pag. 15
o Data Models for alternative splicing	pag. 17
o Correlation of ESE prediction programs with in vivo splicing behaviour in three different cell lines of hMLH1 and hMSH2 missense mutations	pag. 19
o Choosing the End: Regulation of Alternative 3' Splice Sites	pag. 21
o Identifying Splicing Regulatory Elements using Correlation with Expression	pag. 23
o Differentiated evolutionary rates in alternative exons and the implications for splicing regulation	pag. 25
o Comparison of Alternative Splicing Structures in Eukaryotes	pag. 27
o HOLLYWOOD: a comparative alternative splicing database for studying mechanisms mediated by exonic splicing enhancers and silencers	pag. 29
o SAGE/MPSS Bioinformatics: going after splicing variants differentially expressed in tumors	pag. 31
o Using the Affymetrix Exon Array to investigate variation in alternative splicing in a human population	pag. 32
o Selection of splicing variant candidates in breast tumor through microarray platform	pag. 34
o Informatics Issues in Affymetrix Human Exon Array Data Analysis	pag. 35
• Bioinformatics Open Source Conference 2006	pag. 39
o Schedule	pag. 41
o Welcome	pag. 42
o Keynotes	pag. 43
o BioPostpres	pag. 43
o QObase: A Graph Database System for Managing and Exploring Gene Ontology	pag. 43
o BioMobv Helping the Crop Scientists	pag. 44
o XRATE: Scheme-y trees, phylo-HMMs and phylo-grammars	pag. 45
o Twease: Searching Medline, One Sentence at a Time	pag. 46
o The ZMap Genome Annotation Viewer	pag. 47
o BioInformants: Biological, Informational Agents on the Internet	pag. 47
o CalIntegrator - An Open Source Translational Informatics Platform to Integrate Clinical Trials and High Throughput Molecular Analysis in Support of Transition to Tailored Therapy	pag. 48
o SHOQUIN - A Large Scale Machine Learning Toolbox for Biological Sequence Analysis	pag. 50
o XMLPipeDB: A Reusable, Open Source Tool Chain for Building Relational Databases from XML	pag. 51
o BioRuby Shell and RubyTools	pag. 52
o Informatics for Evolutionary Science and Synthesis	pag. 53

• JBB 2006: The Joint BioLINK-Bio-Ontologies Meeting	pag. 57
o Program Committee	pag. 58
o Keynote Speakers	pag. 58
o Agenda	pag. 60
o Short Papers	pag. 61
o Poster Abstracts	pag. 62
o Automatically Adapting an NLP Core Engine to the Biology Domain	pag. 66
o Can literature analysis reveal similarities among cellular processes?	pag. 69
o The design of a wiki-based curation system for the Ontology of Functions	pag. 73
o Distributed Representations of Bio-Ontologies for Semantic Web Services	pag. 77
o SpindleViz: A Three Dimensional, Order Theoretical Visualization Environment for the Gene Ontology	pag. 81
o Exploring the Construction and Applications of a Protein Description Corpus	pag. 85
o Improving Biomedical Corpus Annotation Guidelines	pag. 89
o Improving Biomedical Text Categorisation with NLP	pag. 93
o leXML: Towards a Framework for Interoperability of Text Processing Modules to Improve Annotation of Semantic Types in Biomedical Text	pag. 97
o Use of Text Mining for Protein Structure Prediction and Functional Annotation in Lack of Sequence Homology	pag. 101

AS-SIG



ISMB2006 Alternative Splicing
Special Interest Group meeting

August 4-5, 2006

Fortaleza Convention Center
(Convention Center of Ceará)
Fortaleza, Brazil

Organizers:

- Prof. Shoba Ranganathan, Organizing Chair
Macquarie University, Sydney, Australia
- Prof. Sandro de Souza, Co-Chair
Ludwig Insitute of Cancer Research, Sao Paulo, Brazil
- Prof. Roderic Guigo, Co-Chair
Institut Municipal d'Investigacio Medica, Barcelona, Spain

About the Workshop

The organizers of AS-SIG would like to invite you to participate in the second ISMB Special Interest Group meeting on Alternative Splicing, on August 4-5, 2006 at Fortaleza, Brazil. This workshop is scheduled immediately before [ISMB2006](#), Aug. 6-10, 2006 and is jointly sponsored by the state of Sao Paulo, Brazil and [ISCB](#).

AS-SIG 2006 follows on from the highly successful [Alternative Splicing SIG meeting](#) held at the [Pacific Symposium of Biocomputing](#), 2004, and the ISMB2005 [AS-SIG](#) meeting.

Background and Aims

Alternative splicing generates multiple products from a single eukaryotic gene and is a major mechanism responsible for diversity in the transcriptome of higher organisms, using combinations of "genes in pieces" to assemble transcripts.

This is an exciting time in the field of alternative splicing, combining new discoveries from genomics, bioinformatics and molecular biology. Long considered to be an interesting but less common form of regulation, alternative splicing has emerged as a ubiquitous mechanism of regulation, thanks to genome analysis of human and other higher organisms. Whereas the Human Genome Project has produced a net result of 25,000 – 30,000 genes, alternative splicing evidently produces over 100,000 distinct transcript forms. Identifying, quantifying and analyzing the regulation, function and evolution of these forms constitutes a "Human Transcriptome Project", and will require as remarkable and as concerted an effort as the Human Genome Project. Above all, it will require close collaboration between bioinformaticists and experimentalists, to build a community of shared tools, databases, nomenclature and standards that permit everyone to contribute what they do best, while benefiting from what everyone else has done. The AS-SIG aims to establish a permanent forum for bioinformaticists and experimentalists to come discuss collaboratively what needs to be done in transcriptomics.

AS-SIG will address the latest results and questions in this exciting field, and to bring together bioinformaticists and experimentalists, focusing on questions that demand their collaborative inputs.

Besides oral presentation sessions, the workshop will have a panel discussion on standards for data representation.

Program

August 4, 2006 (Day 1)

08.00-09.00 *Registration*

09.00 **Plenary Session - Shoba Ranganathan (Chair)**

09.00-09.10 Welcome - *Shoba Ranganathan*

09.10-10.30 "The Complex World of Splicing Regulatory Sequences"
Gil Ast

10.30-11.00 *Coffee Break*

11.00 **Genomics 1 - Sandro de Souza (Chair)**

11.00-11.30 "Genome-Wide Identification of Alternative Isoforms Targeted by
Nonsense-Mediated mRNA Decay in *Drosophila*"
Steven Brenner

11.30-12.00 "Genome-wide Prediction of Novel Alternative Splicing Events"
Gunnar Rätsch

12.00-13.30 *Lunch Break*

13.30 **Genomics 2 - Sandro de Souza (Chair)**

13.30-14.00 "The consequences of alternative splicing on biological
pathways"
Melissa Cline

14.00-14.30 "Patterns in Alternative Splicing - Interaction between splicing
and transcription components"
Eugene Melamud

14.30-15.00 "Pervasive unproductive splicing of SR proteins associated with
ultraconserved elements"
Steven Brenner

15.00-15.30 "Hundreds of new alternatively spliced variants discovered in *the*
C.elegans genome"
M. Tabish

15.30-16.00 *Coffee Break*

16.00 **Panel Discussion - Win Hide (Chair)**

16.30-17.00 ""Data Models for alternative splicing"
Oliver Hofmann

17.00-18.30 Discussion

August 5, 2006 (Day 2)

- 09.00** **Regulation of Alternative Splicing 1 - Eduardo Eyras (Chair)**
- 09.00-9.30 "Correlation of ESE prediction programs with *in vivo* splicing behaviour in three different cell lines of hMLH1 and hMSH2 missense mutations"
Alessandro Stella
- 09.30-10.00 "Choosing the End: Regulation of Alternative 3' Splice Sites"
Martin Akerman
- 10.00-10.30 "Identifying Splicing Regulatory Elements using Correlation with Expression"
Deopriya Das
- 10.30-11.00** **Coffee Break**
- 11.00** **Regulation of Alternative Splicing 2 - Eduardo Eyras (Chair)**
- 11.00-11.30 "Differentiated evolutionary rates in alternative exons and the implications for splicing regulation"
Eduardo Eyras
- 11.30-12.00 "Comparison of Alternative Splicing Structures in Eukaryotes"
Michael Sammeth
- 12.00-13.30** **Lunch Break**
- 13.30** **Experimental and Computational Platforms - Shoba Ranganathan (Chair)**
- 13.30-14.00 "HOLLYWOOD: a comparative alternative splicing database for studying mechanisms mediated by exonic splicing enhancers and silencers"
Dirk Holste
- 14.00-14.30 "SAGE/MPSS Bioinformatics: going after splicing variants differentially expressed in tumors"
Pedro Galante
- 14.30-15.00 "Using the Affymetrix Exon Array to investigate variation in alternative splicing in a human population"
Jacek Majewski
- 15.00-15.30 To be confirmed
- 15.30-16.00** **Coffee Break**
- 16.00-16.30 "Selection of splicing variant candidates in breast tumor through a microarray platform"
Maria Rangel
- 16.30-17.00 "Informatics Issues in Affymetrix Human Exon Array Data Analysis"
Fan Meng
- 17.00-17.30 Summary
Shoba Ranganathan
- 17.30-18.00** **Summary and Concluding remarks**

The Complex World of Splicing Regulatory Sequences

Gil Ast

Department of Human Genetics and Molecular Medicine, Sackler Faculty of Medicine, Tel-Aviv University, Tel Aviv 69978, Israel.

1. INTRODUCTION

An average human gene consists of 8.8 exons of approximately 130 nucleotides separated by 7.8 introns of ~3000 nucleotides¹. Thus, the mechanism that directs the splicing machinery to locate short exonic sequences that are flanked by long intronic sequences is of a great interest. Studies of the molecular basis of splicing reveal the existence of exonic *cis*-acting regulatory sequences, which bind *trans*-acting factors, and thus, influence splice-site selection. These *cis*-acting elements are relatively short, usually 4-to-18 nucleotides^{2,3}. These motifs are classified as exonic splicing enhancers (ESE) and silencers (ESS), and they are required for the regulation of both constitutive and alternative splicing^{4,5}. Specific binding of splicing regulatory proteins (such as SR proteins) to ESEs and ESSs assists in the placement of the spliceosome on the appropriate splice sites^{6,7}. Alternative splicing is a process in which more than one mRNA is produced from the same precursor messenger RNA by using different 5'ss and/or 3'ss, giving rise to functionally different proteins⁸. Alternative splicing increases the diversity of the transcriptome within and between cells and adds an additional regulatory dimension to the expression pattern of an organism^{9,10}.

The idea underlying a computational search for regulatory sequences using comparative genomics is that selective pressure causes regulatory elements to evolve at a slower rate than less- or non-functional sequences. Therefore, highly conserved sequences in a collection of homologous regions are good candidates for functioning as regulatory elements¹¹. It is estimated that human and mouse diverged 75-130 million years ago¹². Most of the human genes (99%) have a mouse ortholog with a high level of sequence similarity of the exons (88%). However, while orthologous alternatively spliced exons show a very high degree of conservation (94%), constitutively spliced exons show a lower degree of only 89% conservation^{13,14}. Therefore, human-mouse orthologous exons are good model system for such analysis.

We developed a comparative genomics method that identifies putative exonic splicing regulatory (ESR) sequences, based on the combination of two features: (i) the evolutionary conservation of wobble positions between 46,000 human and mouse orthologous exons, since the wobble positions are almost free of coding constraints; and (ii) the analysis of the overabundance of sequence motifs, compared with their random expectation, given by their codon relative frequency. This method resulted in 285 significant ESR motifs. Alternatively spliced exons that are either short in length or contain weak splice sites show the highest conservation level of those ESRs, especially toward the edges of exons. ESRs that are abundant in those subgroups show a different distribution between constitutively and alternatively spliced exons. We validated our putative ESRs experimentally by choosing 10 motifs and placing them in two sub-optimal exons (minigenes). Remarkably, we found distinctions in the inclusion/exclusion levels between the two minigenes mediated by the same ESR candidates — the same ESR sequence can function as an enhancer in one exon and a silencer in the other. We demonstrated the functionality of those ESRs in their natural environment by mutating four ESRs located in three different reporting exons and show their effect on splicing. In addition, we slid two known SR binding sites (SRp40 and SF2/ASF) along the sub-optimal alternatively spliced exon. Strikingly, we observed that those sequences could function as ESE or ESS, depending on their positions along the exon. Elimination those sites by point mutations restored the original splicing level, indicating the specificity of the positional dependent effect. Thus, we suggest that the distinction between ESE and ESS based solely on the sequence of the motif should be refined — the function of exonic splicing regulatory sequences also depends on their positions in the exons and on other spatial effectors.

2. REFERENCES

1. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
2. Cartegni, L., Chew, S.L. & Krainer, A.R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* **3**, 285-98 (2002).
3. Fairbrother, W.G., Yeh, R.F., Sharp, P.A. & Burge, C.B. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007-13 (2002).

4. Blencowe, B.J. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci* **25**, 106-10 (2000).
5. Graveley, B.R. Sorting out the complexity of SR protein functions. *Rna* **6**, 1197-211 (2000).
6. Sanford, J.R., Ellis, J. & Caceres, J.F. Multiple roles of arginine/serine-rich splicing factors in RNA processing. *Biochem Soc Trans* **33**, 443-6 (2005).
7. Singh, R. & Valcarcel, J. Building specificity with nonspecific RNA-binding proteins. *Nat Struct Mol Biol* **12**, 645-53 (2005).
8. Maniatis, T. & Tasic, B. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**, 236-43 (2002).
9. Ast, G. How did alternative splicing evolve? *Nat Rev Genet* **5**, 773-82 (2004).
10. Graveley, B.R. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet* **17**, 100-7 (2001).
11. Blanchette, M. & Tompa, M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res* **12**, 739-48 (2002).
12. Waterston, R.H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).
13. Sorek, R. et al. A non-EST-based method for exon-skipping prediction. *Genome Res* **14**, 1617-23 (2004).
14. Sugnet, C.W., Kent, W.J., Ares, M., Jr. & Haussler, D. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac Symp Biocomput*, 66-77 (2004).

Genome-Wide Identification of Alternative Isoforms Targeted by Nonsense-Mediated mRNA Decay in *Drosophila*

Qi Meng¹, Marco Blanchette¹, Richard E. Green^{1,2}, Kasper D. Hansen¹, Fabian L. Gallusser¹, Jan Rehwinkel³, Elisa Izaurralde³, Sandrine Dudoit¹, Donald C. Rio¹, Steven E. Brenner^{1*}

¹ University of California, Berkeley, USA, ² Max Planck Institute of Evolutionary Anthropology, Leipzig, Germany, ³ European Molecular Biology Laboratory, Heidelberg, Germany

*To whom correspondence should be addressed: brenner@compbio.berkeley.edu

1. INTRODUCTION

Nonsense-mediated mRNA decay (NMD) is a surveillance mechanism conserved from yeast to human that eliminates mRNAs containing premature termination codons (PTCs) before they produce potentially dangerous truncated proteins. Previously thought to only degrade abnormal transcripts, NMD was later discovered to degrade a significant number of natural transcripts generated from alternative splicing or alternative transcription (1), and it is part of the post-transcriptional regulation circuits that control proper gene expression levels (2).

How the NMD machinery recognizes a PTC and initiates the decay of the mRNA is best understood in mammalian cells, where NMD is triggered by any premature stop codon located 50nt or more upstream of the 3'most exon/exon junction of an mRNA. Previously, computational analysis of the human transcripts suggested that one-third of alternate transcripts contain PTCs based on this 50-nt rule. In *Drosophila melanogaster*, PTC recognition occurs independently of the exon-exon boundaries and no rules have been formulated on PTC recognition. Fewer than one ninth of *Drosophila melanogaster* annotated alternative transcripts have stop codons capable of making a truncated version of the full-length protein. We are interested in identifying true NMD targets among these transcripts and analyzing the mRNA features correlated with their NMD status.

2. RESULTS

We knocked down the NMD effectors, UPF1 and UPF2 using RNAi in SL2 cells. Then, we looked for isoforms whose abundance increased when NMD was inhibited. We designed a 44-k custom Agilent oligonucleotide microarray that included virtually all genes with annotated alternative isoforms. The array has extensive exon and junction probe sets covering both alternative and constitutive regions of the alternatively spliced genes. We developed a nonlinear least squares algorithm to deconvolute the array probe data to infer isoform changes upon NMD inhibition. F-statistics were applied to identify isoforms that showed significantly higher fold increase upon NMD inhibition than other isoforms of the same gene. These isoforms are considered NMD targets under our experimental conditions. Our algorithm separates isoform change from gene level change, allowing us to identify minor isoforms that are NMD targets. Ninety percent of our identified NMD targets were not identified as NMD targets by Affymetrix gene expression array (3), presumably because of our array's sensitivity to alternate splice forms.

When NMD is down-regulated, about one tenth of the isoforms that have stop codons capable of making a truncated version of the full-length protein were significantly increased in abundance. We do not see any correlation of the NMD status to the mRNAs features such as the length of the 3'UTR, the length of the ORF, the existence of uORFs or the 50-nt rules, which apply in mammals. Our preliminary results suggest that the mRNA features important for PTC recognition by the NMD machinery in fly are subtle and natural NMD targets are a minor portion of the whole fly transcriptome.

We are using RT-PCR and qPCR to validate array data and to separate direct and secondary effects of NMD inhibition.

3. REFERENCES

- (1). Hillman, R.T., Green, R.E. and Brenner, S.E. 2004. An unappreciated role for RNA surveillance. *Genome Biology* 5:R8. doi:[10.1186/gb-2004-5-2-r8](https://doi.org/10.1186/gb-2004-5-2-r8)
- (2). Lewis, B.P., Green, R.E., Brenner, S.E. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A*. 100:189-92. doi:[10.1073/pnas.0136770100](https://doi.org/10.1073/pnas.0136770100)
- (3). Rehwinkel, J., Letunic, I., Raes, J., Bork, P., Izaurralde, E. 2005. Nonsense-mediated mRNA decay factors act in concert to regulate common mRNA targets. *RNA* 10:1530-44.

Genome-wide Prediction of Novel Alternative Splicing Events

Cheng Soon Ong¹, Uta Schulze¹, Hyunjung Shin¹, Sören Sonnenburg² and Gunnar Rätsch¹

¹ Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany; ² Fraunhofer FIRST, Berlin, Germany

1. INTRODUCTION

While traditional methods for computational recognition of alternative splicing are usually solely based on expressed sequences (ESTs or cDNAs; cf. Gupta et al., 2004, and references therein), more recent techniques try to identify and exploit local sequence features for prediction (Sorek & Ast 2003; Sakai & Maruyama, 2004; Dror, Sorek & Shamir, 2004; Sorek et al., 2004; Hiller et al., 2004). For instance, in Dror, Sorek & Shamir (2004) features like the exon length, its divisibility by three, the length of the flanking introns and the intensity of the poly-pyrimidine tract were utilized. Moreover, conservation patterns to another organism have been taken into account. Those are among the most discriminative features (Sorek & Ast, 2003). However, this is only possible for a fraction of exons in human, as exons are frequently not conserved, making the conservational features unavailable. In our previous work we aimed to design a classifier that accurately distinguishes constitutive from alternatively spliced exons. Our method only used information that is always available and might also be used by the cellular splicing machinery; namely, features derived from the exon and intron lengths and features based on the pre-mRNA sequence (Rätsch, Sonnenburg & Schölkopf, 2005).

We have now extended our work to other alternative splicing events such as intron retention, alternative 5' and 3' (intron) splicing. Furthermore, previously we only considered the model organism *C. elegans*. Now we have included higher organisms such as the zebra-fish, the plant *Arabidopsis thaliana*, the fruit-fly and other nematodes. In this work we show that our method can accurately predict the mentioned types of alternative splicing for the four model organisms. We use our method to complete the inventory of known alternative splicing events with computational predictions. We obtain many new putative alternative splicing events for zebra-fish and *A. thaliana*. Also, we use our classifiers to annotate the genomes of *C. briggsae* and *C. remanei* for which virtually no alternative splicing events are known yet. Our algorithms predict surprisingly high incidences of alternative splicing for *C. briggsae*, which is subject to further testing.

2. DATA AND METHODS

We started by aligning ESTs from dbEST and available cDNAs against the genomes using *blat* with proper postprocessing and quality control. From the remaining transcripts we obtained lists of introns and exons, which were first clustered and later combined into splicing graphs (Heber et al., 2002). As a next step we defined rules for detecting alternative splicing events: intron retention, exon skipping (also exclusive and multiple exon skipping), alternative 5' and 3' splicing. We used these detectors to obtain positive examples for training and testing our algorithms. Negative examples were generated from sites with no evidence of alternative splicing and requiring that every splice site is confirmed at least five times. Regions covered only once by an EST or cDNA (which are therefore the best candidates for yet unknown alternative splicing events), were used as *unlabeled set* which our algorithms have to annotate.

We use support vector machines (SVMs; see e.g. Schölkopf and Smola, 2002) to learn to classify alternative splicing events, using a method similar to the one described in Rätsch, Sonnenburg, Schölkopf (2005). SVMs find a maximum margin separating hyperplane between the positive and negative training in a feature space that is defined via a similarity function, the so-called kernel. We have developed such a kernel that works on DNA sequences and is able to recognize sequence elements at loosely defined positions – for instance relative to the splice site. This kernel turned out to be very well suited for the prediction of alternative splicing, where regulative elements are positioned in the vicinity of the splice sites.

For exon skipping and intron retention we used sequence windows around the two affected splice sites (which are assumed to be known), as well as features describing the length of the introns and exons. Given these features we try to learn to distinguish constitutive from alternatively spliced introns and exons. For alternative 5' and 3' splicing we only used a window around one of the two affected splice sites. Here, we learn to classify whether the splice site is involved in alternative 5' or 3' splicing (note that from this prediction one does not obtain the alternative splice site).

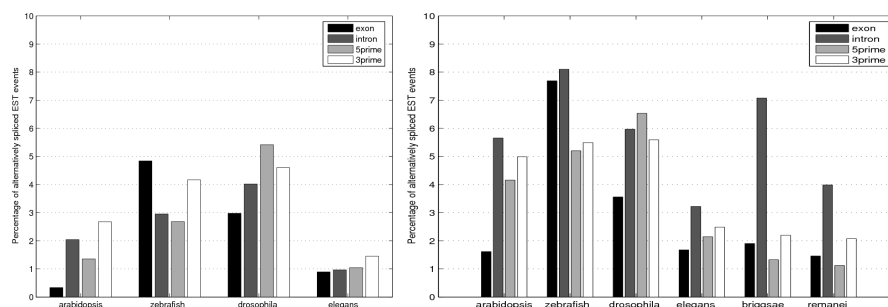


Figure 1: The incidence of several forms of alternative splicing for several model organisms. Left: The frequency based on the current EST databases, Right: The frequency after including the events predicted with our method.

3. RESULTS

After extracting all alternative splicing events from the splicing graphs for the four organisms, we first determined the incidences of these events, e.g. the fraction of intron retention events relative to all known introns. See Figure 1 [left] for a summary. For the two nematodes *C. briggsae* and *C. remanei* there only very few known events of alternative splicing.

In order to apply our learning techniques, we first split the data into a training and evaluation set. We apply the learning method on the training set and measure the performance on the test set. In Table 1 we report the area under the Receiver Operator Curve and the true positive (TP) rate at 1% false positives. We notice that the performance in terms of the TP rate is normally quite high (22%-79%). For *Arabidopsis*, however, the accuracy is generally not as high. The performance for the prediction of intron retention (the most frequent form in *A. thaliana*) is particularly poor. We suspect that our negative set does include quite a few yet unknown intron retention events (in Wang and Brendel, 2006) they have around 20% more events than in our database).

Finally, we combine the known alternative splicing events with our safest predictions (above the threshold for 1% false positives) on the unlabeled set (only one EST confirmation). In Figure 1 [right] we report the combined incidences for all six organisms. For the nematode genomes we used the classifiers trained for *C. elegans*. We observe rather large increases of alternative splicing events for the zebra-fish and *A. thaliana*. Furthermore, we find many strong predictions for *C. briggsae* (in particular intron retention).

	A. thaliana	Zebra-fish	Fruit-fly	C. Elegans
Exon skipping	87.6 (22.4)	95.2 (76.8)	94.3 (43.0)	92.3 (50.0)
Intron retention	83.0 (35.9)	97.0 (78.6)	84.1 (23.9)	93.3 (53.1)
Alternative 5'	81.9 (25.7)	92.8 (62.4)	91.3 (39.3)	81.8 (28.3)
Alternative 3'	83.9 (27.5)	94.3 (50.2)	87.0 (28.0)	90.5 (47.4)

Table 1: Area under receiver operating characteristic (AROC) in percent and the true positives rate at 1% false positives for various organisms.

The consequences of alternative splicing on biological pathways

Melissa Cline*, Benno Schwikowski

Pasteur Institute, 25-28 rue du Docteur Roux, 75015 Paris, France

*To whom correspondence should be addressed: cline@pasteur.fr

1. INTRODUCTION

In this work, we assessed how alternative splicing can modulate biological pathways through changes in protein domain composition. 7085 of 21025 human genes in ENSEMBL[1] encode two or more different proteins, with 3910 genes encoding proteins with different domain compositions. We evaluated how this change in domain composition may modulate biological networks.

First, in regulatory networks, alternative splicing often produces transcription factor proteins with no DNA binding domains [2]. In our observations, 438 of 3910 genes, some protein isoform lacks a zinc finger domain present in others. In well-studied cases [3], when a DNA binding domain is spliced out of a transcription factor protein, it interacts with most of the usual co-factors to form a transcription factor complex, but a complex that cannot bind DNA. Such proteins not only produce no transcription, they absorb low-abundance co-factors; thus, the loss of a DNA binding domain can transform a transcription activator into a transcription repressor.

Additionally, when a domain is spliced out of a protein, certain protein-protein interactions may cease to occur. We evaluated the frequency of this using domain network analysis [4]; given an interaction between two proteins, domain network analysis identifies the protein domains most likely to interact. We evaluated the frequency of this using interaction data from the Rual human interaction dataset [5]. Out of 2450 interactions studied, 490 show evidence suggesting that domain loss through alternative splicing would effectively cancel the interaction.

Most biological network modeling relies on microarray data to assess when a gene is expressed. Domain-level changes in splicing cannot be seen with traditional microarrays, which measure overall expression levels; alternative splicing microarrays offer improvements, but involve greater analysis complexity. Thus, we must consider what measurement platforms are most appropriate for our questions. Furthermore, where a gene's products differ in their interaction patterns, and thus their effect, we must reconsider our assumptions on "gene function".

4. REFERENCES

1. Birney, E., *Ensembl: a genome infrastructure*. Cold Spring Harbor symposia on quantitative biology, 2003. **68**: p. 213-5.
2. Taneri, B., et al., *Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific*. Genome Biol, 2004. **5**(10): p. R75.
3. Grob, T.J., et al., *Human delta Np73 regulates a dominant negative feedback loop for TAp73 and p53*. Cell death and differentiation, 2001. **8**(12): p. 1213-23.
4. Albrecht, M., et al., *Decomposing protein networks into domain-domain interactions*. Bioinformatics, 2005. **21 Suppl 2**: p. ii220-ii221.
5. Rual, J.F., et al., *Towards a proteome-scale map of the human protein-protein interaction network*. Nature, 2005. **437**(7062): p. 1173-8.

Patterns in Alternative Splicing - Interaction between splicing and transcription components.

Eugene Melamud*, John Moulton
Center for Advanced Research in Biotechnology, UMBI, University of Maryland
9600 Gudelsky Drive, Rockville MD, USA
melamud@umbi.umd.edu

INTRODUCTION:

The process of formation of final mRNA transcript involves a complex interaction network between spliceosomal and transcriptional components. Effects of these interactions are poorly understood, yet they have important consequences on types and observed frequency of alternative splicing and alternative transcription events. Furthermore, many transcripts contain multiple alternative events, and it is entirely unclear if these events are coordinated, and if so, how? To help address these questions we have developed a symbolic language for describing differences between transcripts. We have analyzed splicing patterns in human genes and identified statistically significant correlated splicing and transcription events. We find that alternative transcription events are the dominant source of diversity in human transcripts and that there is strong linkage between transcription and splicing events. The majority of splicing patterns involving multiple spliceosomes are not enriched, indicating that there is little communication between adjacent spliceosomal assemblies.

BACKGROUND:

In eukaryotes, there are three major mechanisms that contribute to the diversity of transcripts: alternative splicing, alternative transcription initiation, and alternative transcription termination. Alternative splicing is a consequence of a spliceosome choosing an alternative set of splice junctions during the intron removal process. Alternative 5' end formation is a result of alternative promoter selection by RNA polymerase complex. Alternative 3' end formation is result of selection of the alternative cleavage site by cleavage factors.

Although, traditional view held that both the 3' end formation and splicing are post transcriptional events, we now realize that these processes occur co-transcriptionally [1,2]. The final mRNA transcript is a result of interaction between a number of concurrent processes that involve a multitude of components from transcription and splicing regulatory pathways. The interactions between components can be complex, with alternative promoters causing changes in splicing[3], polymerase processivity causing alternative splicing, splicing factors enhancing polyadenylation[5,6] and even presence of proximal 5' splice sites enhancing transcriptional initiation[7].

The regulation of splicing for small systems with few introns has been worked out in great detail [8], however formation of splicing patterns in large systems remains largely unexplored field. Is there a significant communication between multiple spliceosome assemblies in formation of splicing patterns that involve multiple introns? Furthermore, very little about the strength and the frequency of interaction between transcription and splicing machinery. How strong are these interactions and how frequently do changes in one system propagate to another?

To address these questions we have developed symbolic representation of differences between isoforms. We have extended a formalism typically used to describe alternative splicing events into more detailed categories and introduced symbolism for alternative transcription events. We compiled statistics on splicing patterns in human transcripts and developed a model that simulates formation of random exon patterns. The random model allows us to derive accurate statistics on frequency and correlation between various transcription and splicing events.

Our analysis indicates that the dominant role in diversification of transcripts is alternative transcription events. We observe strong correlation between alternative splicing events and alternative transcription initiation events. With an exception of few examples, we do not see high enrichment in splicing patterns that are generated by multiple spliceosomal assemblies, but splicing patterns produced by a

single spliceosome assembly are highly enriched. Curiously, we find enrichment in patterns that represent modification of both 5' and 3' splice sites within a single spliceosomal assembly.

INTERACTIONS BETWEEN COMPONENTS OF SPLICING AND TRANSCRIPTION MACHINERY

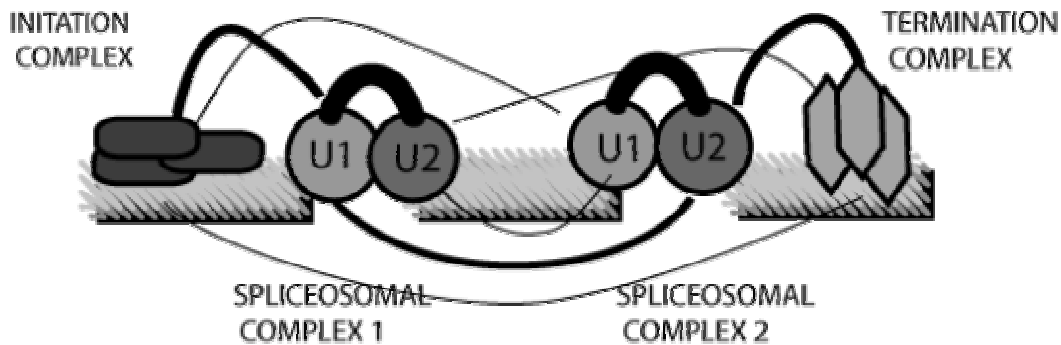


Figure 2: Strength of interaction in formation of alternative transcripts deduced from splicing patterns. Highly enriched interactions are shown as thick lines.

REFERENCES:

1. Maniatis, T. and R. Reed, *An extensive network of coupling among gene expression machines.* Nature, 2002. **416**(6880): p. 499-506.
2. KORNBLIHTT, A.R., et al., *Multiple links between transcription and splicing.* RNA, 2004. **10**(10): p. 1489-1498.
3. Pagani, F., et al., *Promoter Architecture Modulates CFTR Exon 9 Skipping.* J. Biol. Chem., 2003. **278**(3): p. 1511-1517.
4. Nogues, G., M.J. Munoz, and A.R. Kornblihtt, *Influence of Polymerase II Processivity on Alternative Splicing Depends on Splice Site Strength.* J. Biol. Chem., 2003. **278**(52): p. 52166-52171.
5. Vagner, S., C. Vagner, and I.W. Mattaj, *The carboxyl terminus of vertebrate poly(A) polymerase interacts with U2AF 65 to couple 3'-end processing and splicing.* Genes Dev., 2000. **14**(4): p. 403-413.
6. Castelo-Branco, P., et al., *Polypyrimidine Tract Binding Protein Modulates Efficiency of Polyadenylation.* Mol. Cell. Biol., 2004. **24**(10): p. 4174-4183.
7. Furger, A., et al., *Promoter proximal splice sites enhance transcription* Genes Dev., 2002. **16**(21): p. 2792-2799.
8. Black, D.L., *Mechanisms of alternative pre-messenger RNA splicing.* Annu Rev Biochem, 2003. **72**: p. 291-336.

Pervasive unproductive splicing of SR proteins associated with ultraconserved elements

Liana F. Lareau¹, Richard E. Green^{1,3}, Maki Inada², Jordan C. Wengrod¹, Qi Meng², Steven E. Brenner^{1,2*}

Departments of ¹Molecular and Cell Biology and ²Plant and Microbial Biology, University of California, Berkeley, 94720, USA

³Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

*To whom correspondence should be addressed: brenner@compbio.berkeley.edu

1. INTRODUCTION

SR proteins play critical roles in the regulation of pre-mRNA splicing. We show that most human SR protein genes are themselves alternatively spliced, and their alternative splice forms are often targets of a degradation pathway known as nonsense-mediated mRNA decay (NMD). By mining EST databases for sequences corresponding to the 11 human SR protein genes, we have identified premature stop (nonsense) codons in alternative mRNA forms of 10 of the genes. These isoforms are expected to be degraded by NMD rather than producing protein. The frequency with which these genes are alternatively spliced, the similar splice patterns, and the striking conservation imply that the alternative splice forms are functionally important.

We observe two classes of alternative splicing patterns among the SR proteins; in both cases the splicing is likely to decrease production of functional protein by targeting mRNA for decay. For some SR proteins, the minor splice form includes a cassette exon harboring a stop codon (figure 1). For others, the 3' untranslated region (UTR) is alternatively spliced to introduce a splice junction downstream of the normal stop codon, marking that stop codon as premature. Closely-related SR proteins tend to fall within the same class.

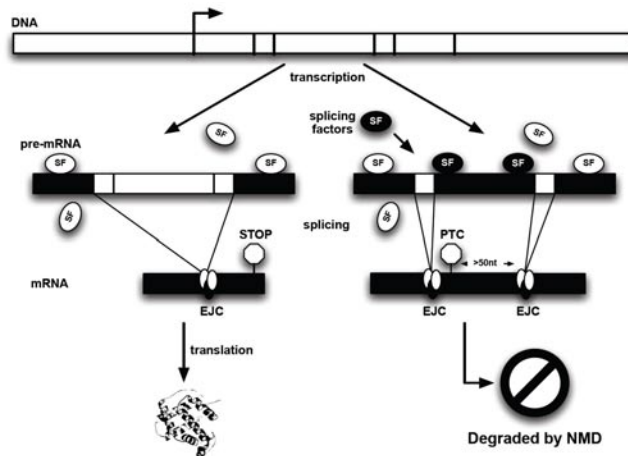


Figure 1 Alternative splicing can include a cassette exon harboring a premature termination codon (PTC), and thus target the mRNA for decay.

Alternative splice forms are not generally conserved between human and mouse. It is striking, then, that the mouse orthologs of human SR proteins exhibit the same unproductive splicing patterns. In at least two cases, the alternative splicing is also conserved between human and the basal chordate *Ciona intestinalis*. Further, the alternatively-spliced 3' UTRs or noncoding cassette exons and flanking introns show remarkable sequence conservation between human and mouse – some are more conserved than the protein-coding exons of the same genes. Five of the genes contain ultraconserved elements, stretches of 100% nucleotide identity between human and mouse extending for at least 200nt (1), in their alternatively spliced regions.

To demonstrate that our predicted SR isoforms are indeed NMD targets, we have established methods to detect specific transcript expression levels in the absence of NMD. After using RNAi to target UPF1,

a key NMD effector, in HeLa cells, we measured the levels of transcripts containing premature stop codons using real-time PCR. Results from the first 5 SR genes display stabilization of predicted NMD target isoforms, in agreement with our predictions.

Some SR proteins have been shown to modulate splicing of their own transcripts. The SR protein SC35 is thought to autoregulate its splicing to produce unstable alternative forms that are likely to be degraded by NMD in order to attenuate the level of SC35 protein in the cell (2). Previously, we showed that a third of all human alternate splice forms may be targets of NMD, and that regulation of expression via degradation may be widespread (3,4). Our analyses suggest that the known cases of regulated alternative splicing of SR proteins may represent a highly conserved mode of gene regulation shared by almost all members of the SR protein family.

2. REFERENCES

1. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S. and Haussler, D. 2004. Ultraconserved elements in the human genome. *Science* 304:1321-1325.
2. Sureau, A., Gattoni, R., Dooghe, Y., Stevenin, J. and Soret, J. 2001. SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. *EMBO J.* 20:1785-1796.
3. Lewis, B.P., Green, R.E. and Brenner, S.E. 2003. Evidence for widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci USA.* 100:189-192. doi:[10.1073/pnas.0136770100](https://doi.org/10.1073/pnas.0136770100)
4. Hillman, R.T., Green, R.E. and Brenner, S.E. 2004. An unappreciated role for RNA surveillance. *Genome Biol* 5:R8. doi:[10.1186/gb-2004-5-2-r8](https://doi.org/10.1186/gb-2004-5-2-r8)

Hundreds of new alternatively spliced variants discovered in *C.elegans* genome

Luv Kashyap, M.Tabish*

AM University, Aligarh, Uttar Pradesh, India

*To whom correspondence should be addressed: m.tabish@lycos.com

INTRODUCTION

The 97 Mb genomic sequence of the eukaryotic, soil dwelling, free living nematode *C. elegans* is complete and was the first multicellular organism to be sequenced(1). Deciphering the biological information from these sequenced genomes is of great importance and use, as the information that we get from *C. elegans* is directly applicable to more complex organisms like humans. 30-40% of *C. elegans* genes share a direct homology with human genes (2). Secondly, it has been shown that human genes replace their *C. elegans* homologs when introduced into transgenic *C. elegans*. Conversely, many *C. elegans* genes can function similarly to mammalian genes. Thus studies in *C.elegans* genes can directly be used for better understanding of human genes. Alternative splicing of pre mRNAs is a powerful and versatile regulatory mechanism that can affect quantitative control of gene expression and functional diversification of proteins (3). It contributes to major developmental decisions and also to fine-tuning of gene function. More than 50% of the genes of recently sequenced eukaryotic genomes (4, 5, 6, 7) are now believed to undergo alternative splicing to generate different transcript and protein isoforms under different developmental, tissue-specific, and disease conditions, thus bringing a new set of challenges to gene prediction programs and the encompassing annotation processes. Alternative splicing is found extensively in all higher eukaryotes with most information from well-studied organisms like *C.elegans*, *Drosophila*, Mouse and Humans (8,9,10,11,12). EST based approach for detection of alternatively spliced variants of a gene, was considered the best approach till now but this approach is successful only in the case of organisms which have extensive EST coverage specially humans. In organisms like *C.elegans* where the EST coverage is limited and only 1% of the total coding genes have EST coverage, detection of alternatively spliced variants is rather difficult. Bioinformatics methods have successfully overcome these limitations, involving computational analysis and experimental verification to detect these spliced variants in *C. elegans* (13), and mouse(14).

RESULTS

Taking motivation from these experiments we did complete analysis of chromosome 1 of *C. elegans* to look for new exons and genes encoded by chromosome 1 using a combination of various gene finding, exon predicting, ORF finding programmes and other bioinformatics tools to predict alternatively spliced transcripts in *C. elegans* genes. We found roughly 120-150 new alternatively spliced variant and exons from chromosome 1 analysis. To experimentally verify our findings we have done RT-PCR experiments for some of the predicted spliced variants of the genes. These new coding sequences, not annotated or identified earlier will not only enhance the available splice data base of *C.elegans* but will also enhance the our knowledge about understanding of the genome structure and evolution of higher eukaryotes specially in context to humans

REFERENCES

1. *C. elegans* Sequencing Consortium 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282, 2012–2018
2. euGenes: Homologous Genes Summary Table August-2005.
<http://eugenesis.org/all/hgsummary.html>
- 3 Lopez AJ. 1998 Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation *Annu Rev Genet.*;32:279-305.
- 4 Mironov, A.A., Fickett, J.W., and Gelfand, M.S. 1999. Frequent alternative splicing of human genes. *Genome Res.* 9: 1288-1293

- 5 Kan, Z., Rouchka, E.C., Gish, W.R., and States, D.J. 2001. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.* 11: 889-900
- 6 Modrek, B., Resch, A., Grasso, C., and Lee, C. 2001. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* 29: 2850-2859
- 7 Zavolan, M., van Nimwegen, E., and Gaasterland, T. 2002. Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome. *Genome Res.* 12: 1377-1385
- 8 Modrek, B. and Lee, C. 2003 Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genetics* 34: 177-180.
9. Black, D. L. 2000 Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* 103:367-370
10. Graveley, B. R.. 2001 Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17:100-107
- 11 .Schmucker, D., J. C. Clemens, H. Shu, C. A. Worby, and J. Xiao *et al.*, 2000 *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* 101:671-684.
12. Zahler, A.M. Alternative splicing in *C. elegans*.
http://www.wormbook.org/chapters/www_altsplicing/altsplicing.html
13. Tabish, M. Clegg, R.A, Rees, H.H. and Fisher, M.J. 1999. Organization and alternative splicing of the *Caenorhabditis elegans* cAMP-dependent protein kinase catalytic-subunit gene (*kin-1*) *Biochem. J.* 339 :209–216
- 14 Tabish M, Ticku MK 2004. Alternate splice variants of mouse NR2B gene *Neurochem Int.* ;44:339-43

Data Models for alternative splicing

Oliver Hofmann and Winston Hide

South African National Bioinformatics Institute, University of the Western Cape, Private Bag X17,
Bellville 7535, South Africa

*To whom correspondence should be addressed: oliver@sanbi.ac.za

1. INTRODUCTION

Transcript isoforms structure is inconsistently described. The problem impacts on the ability to perform effective research on alternative splicing and the integration of different data repositories. With this in mind, we have developed a formal description of alternative splicing (AS) based on existing standards for the annotation and analysis of transcript isoforms.

The formal description is based on a) a controlled vocabulary describing the standard splice events, b) a graph-based representation of a transcriptional unit with all its different transcript variants, (figure 1) and c) a rule-based definition of splicing events represented by the graph (table 1). Transcript isoforms of a given transcriptional unit (or gene) are merged into a single graph, with nodes representing transcription start-, stop-, 3' and 5' splice-sites and edges representing introns or exons.

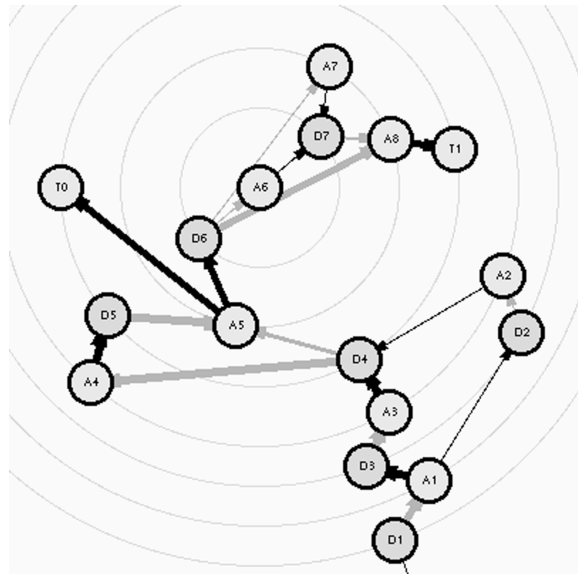


Figure 1: Subgraph for a graphbased representation of a gene. Nodes represent start and end (T) positions of transcripts as well as donor (D) and acceptor (A) splice sites. Edges represent exons (black) or introns (grey). The strength of the edge reflects it's EST support, diverging paths represent different splice patterns. An example of this is the exon A4->D5 which can be skipped by following the path from D4 to A5 instead. Edges are annotated with the gene expression state of the transcripts that were used to generate the graph (e.g edge A4->D5 was found in samples from liver, heart and skin).

The data model is tightly integrated with the eVOC gene expression system (1), enabling us to link transcript isoforms to a given gene expression state. Edges representing exons are be linked to gene expression data using the eVOC system based on supporting ESTs and cDNAs, allowing for the quick identification of subgraphs and paths specific to a particular set of annotatons.

All components of the graph are annotated with Sequence Ontology (2) concepts to facilitate the integration of different data sources. The graph-based formalization of splicing patterns allows us to describe and find splice variants as subgraphs using a graph-based query language independently of the annotation used to initially describe the splicing events.

Finally, all transcripts and alternative splicing events are encoded in XML, utilizing a variant of the community-accepted CHAOS-XML standard (3).

Table 1: Defining an exon-skip (cassette exon) event based on a rule applied to the transcript graph. KIF and OWL definitions kindly provided by Chris Mungall.

Format	Rule
Basic rule	(I <is an> INTRON) and (E <is an> EXON) and (E <is contained in> I)
OWL Web Ontology Language	intersectionOf(Exon hasValuesFrom(containedIn Intron))
Knowledge Interchange Format (KIF)	(<=> (instance-of ?e skipped-exon) (exists (?i) (and (instance-of ?e exon) (instance-of ?i intron) (contained-in ?e ?i))))

2. REFERENCES

1. Kelso J., Visagie J., Theiler G., Christoffels A., Bardien-Kruger S., Smedley D., Otgaar D., Greyling G., Jongeneel V., McCarthy M., Hide T., and Hide W. 2003. eVOC: A Controlled Vocabulary for Gene Expression Data. *Genome Res.* 13: 1222 – 1230.
2. Eilbeck K., Lewis S.E., Mungall C.J., Yanell M., Stein L., Durbin R., and Ashburner M. 2005. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* 6(5): 44
3. Mungall, C. Chaos-XML library. URL: <http://www.fruitfly.org/chaos-xml/>

Correlation of ESE prediction programs with in vivo splicing behaviour in three different cell lines of hMLH1 and hMSH2 missense mutations

Patrizia Lastella¹, Nicoletta Surdo¹, Nicoletta Resta¹, Ginevra Guanti¹, and Alessandro Stella^{1*}
¹ Section of Medical Genetics, Department of Biomedicine in Childhood, University of Bari, Italy.
Policlinico P.zza G.Cesare 11 70124 Bari ITALY

*To whom correspondence should be addressed: alexst@medgene.uniba.it

1. INTRODUCTION

The precision and correctness of intron removal during pre-mRNA splicing are largely dependent on the recognition of several discrete elements some of which, as the splicing donor and acceptor sites, are virtually invariant. Recently, several reports have shown that exonic sequences are able to regulate splicing proficiency, and that nucleotide substitutions in these sequences may lead to abnormal splicing or exon skipping (1,2). Moreover, it has been demonstrated that aberrant splicing can occur as a consequence of mutations that disrupt exonic splicing enhancers (ESEs) or create exonic splicing suppressors (ESSs) (reviewed in 3). Exonic splicing enhancers have been identified on the basis of exon mutations that block splicing, of computational comparison of exon sequences, and of the selection of sequences that activate splicing or that bind to specific regulatory proteins, most notably the SR (serine-arginine rich) proteins.

A recent work (4) has demonstrated that pathogenic missense mutations in the hMLH1 and hMSH2 genes, in contrast to polymorphic variants, tend to colocalize in ESE sequences more frequently than would have been expected. In contrast, non pathogenic polymorphic variants were distributed randomly in relation to the ESE sites. However, we have recently reported that abrogation or decrement of SR protein score matrices does not necessarily lead to a splicing defect (5).

On the basis of all the above evidence, we decided to evaluate the 99 hMLH1/hMSH2 missense mutations listed in the InSIGHT database with three different ESE prediction programs ESEfinder (<http://exon.cshl.org/ESE/>), Rescue-ESE (<http://genes.mit.edu/burgelab/rescue-ese/>), and PESX (<http://cubweb.biology.columbia.edu/pesx/>). We next investigated the consequences of 20 exonic missense mutations with different predicted effects on the splicing proficiency. To further extend our analysis, we assessed the effects of the 20 missense mutations in three different mammalian cell lines.

2. RESULTS AND DISCUSSION

Of the 99 different missense mutations reported in (4), 50 were localized in ESE sequences identified by ESE finder. We analyzed the same mutation data set with RescueESE and PESX which found respectively 40 and 41 mutations as lying in ESE sites. A total of 7 mutations lying in ESE sites were identified by all the three algorithms. Among these, only 2 caused the same type of predicted change (ie both all programs predicted no change, or creation/addition of novel ESE motifs, or disruption of ESE sites). However, in only one case did all the three algorithms equally predicted ESE sites disruption without the concurrent creation of novel ESE sites. We then selected 20 hMLH1 and hMSH2 mutations with different predicted effects on ESEs, and examined the consequences in a splicing assay we already had available (6). The criteria for selection was that mutations should create or abolish one or more ESE sites according to the predictions of at least one algorithm. All the minigene constructs were assembled in the pSPL3 vector.

The results (Figure 1) of the in vivo splicing analysis showed that only 8 of the 20 mutations investigated caused an a significant change in the splicing pattern and 6 lead to a decrease by of at least 50% of the rate of exon inclusion when compared to the wild type allele. Of these 8, four were correctly predicted by ESEfinder, 6 were localized in ESE or ESS sequences identified by PESX which correctly predicted 5 of them, while only 3 lie in ESE sequences recognized by RescueESE that predicted the splicing behaviour of 2 of these 3. Furthermore, when exons were included or skipped completely in our assay they appeared to be insensitive to any change affecting ESE or ESS sequences. The higher PESX prediction power probably relies on the fact that , PESX differently from ESE finder and Rescue-ESE predicts not only exonic enhancer sequences but also those with a suppressor effect. Overall, the results of our splicing assay were comparable to the in vivo splicing profile of hMLH1 and

hMSH2, since all the exons normally presenting alternative splicing showed an inclusion rate ranging from null to 73%, but never complete.

To investigate whether some mutations may show tissue-specific differences, we used the same constructs analyzed in Cos-7 cells to transfect the cervical adenocarcinoma-derived HeLa cell line and the hepatocellular carcinoma cell line Hep-3B. The RT-PCR on RNA extracted from these two cell lines 48 hours after transfection with the different mutated and normal constructs demonstrated a variable level of inclusion for the exons already reported to be alternatively spliced in vivo (exons 2,3,5,6 and 10 for hMSH2, exons 10 and 17 for hMLH1). We next analyzed the consequences on splicing of the mutations that had been able to cause splicing abnormalities when tested in Cos-7 cells. Although for the majority of the mutations analyzed the effects on the splicing proficiency were similar to those observed in Cos-7 cells, for the mutations situated in the alternatively spliced exons there was an evident variability in the level of inclusion.

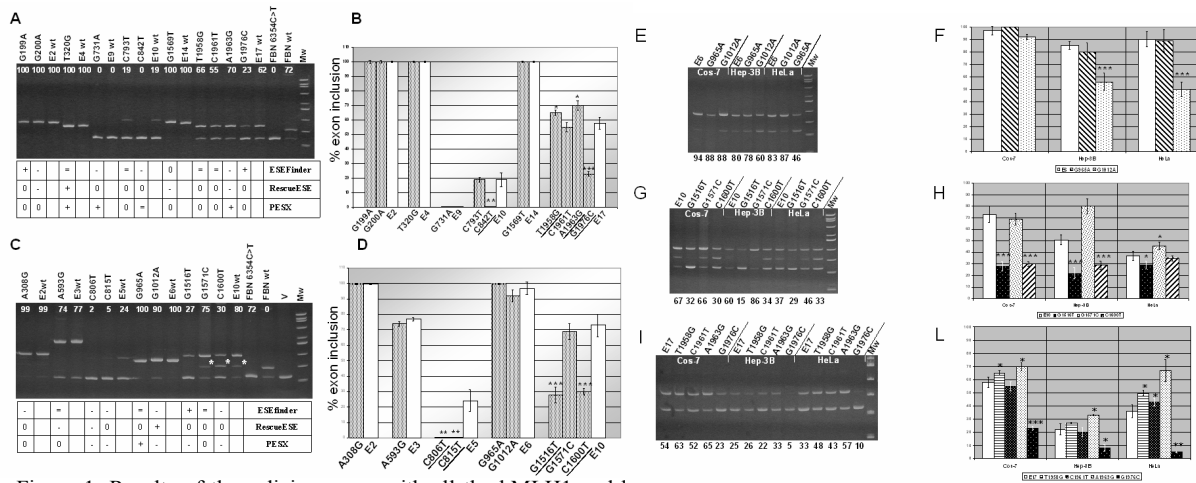


Figure 1. Results of the splicing assay with all the hMLH1 and hMSH2 mutated constructs in Cos-7 cells (A and C), and or hMSH2 exon 6 mutations in Cos-7, Hep3B and HeLa (panel E), hMSH2 exon 10 mutations (panel G) and hMLH1 exon 17 mutations (panel I) in the same three cell lines. A,B,C,D numbering is relative to the nucleotide position in the ORF. Cells were transfected with 1 mg of the indicated mutant minigene variants or the corresponding wild-type exon, RNA was extracted, reverse transcribed and amplified with primers SD6 and SA2. The RT-PCR products were resolved on GeneGel Excel, stained with ethidium bromide and quantitated with an image analyzer. V=vector only; Mw=size standard. The black arrowhead represents the exon skipped product. The percentage of exon inclusion is indicated above each lane. The white asterisks show the splicing product deriving from use of an internal cryptic donor site. A, C Below the gel are reported the predictions for the three algorithms: = no change; + the mutation creates an ESE or abrogates an ESS sequence; - the mutation creates an ESS or abrogates an ESE sequence; 0 the mutations is not localised in and does not create or disrupt any regulatory sequence. (B,D,F,H,L) Graphic representation of the results of the splicing assay. The average of percent exon inclusion is reported in the y-axis and represents the mean of two independent transfections done in triplicate for each construct (x-axis). White bars are used for normal alleles, patterned for mutated constructs (B,D). Mutations within the same exons are grouped together and with their corresponding normal exon. Error bars represent standard deviation. The mutated constructs causing significant differences when data were analysed using Student's *t* test are underlined (*= $P<0.05$, **= $P<0.01$, ***= $P<0.001$)

3. REFERENCES

- Montera M, Piaggio F, Marchese C, Gismondi V, Stella A, Resta N, Varesco L, Guanti G, Mareni C et al 2001 A silent mutation in exon 14 of the APC gene is associated with exon skipping in a FAP family. *J Med Genet.*, 38:863-867.
- Yang Y, Swaminathan S, Martin BK, Sharan SK 2003 Aberrant splicing induced by missense mutations in BRCA1: clues from a humanized mouse model. *Hum Mol Genet.*, 12:2121-2131.
- Pagani F, Baralle FE 2004 Genomic variants in exons and introns: identifying the splicing spoilers. *Nat Rev Genet.*, 5:389-396.
- Gorlov IP, Gorlova OY, Frazier ML, Amos CI 2003 Missense mutations in hMLH1 and hMSH2 are associated with exonic splicing enhancers. *Am J Hum Genet.*, 73:1157-1161
- Lastella P, Resta N, Miccolis I, Quagliarella A, Guanti G, Stella A 2004 Site directed mutagenesis of hMLH1 exonic splicing enhancers does not correlate with splicing disruption. *J Med Genet.*, 41:e72.
- Stella A, Wagner A, Shito K, Lipkin SM, Watson P, Guanti G, Lynch HT, Fodde R, Liu B et al 2001 A nonsense mutation in MLH1 causes exon skipping in three unrelated HNPCC families. *Cancer Res.*, 61:7020-7024.

Choosing the End: Regulation of Alternative 3' Splice Sites

Martin Akerman and Yael Mandel-Gutfreund*

Faculty of Biology, Technion-Israel Institute of Technology, Haifa 32000, Israel

*To whom correspondence should be addressed:

yaelmg@tx.technion.ac.il

1. INTRODUCTION

Alternative splicing (AS) constitutes a major mechanism creating protein diversity in humans. This can result from skipping entire exons or by altering the selection of the splice sites that define the exon borders. Alternative 3' Splice Sites (A3SS) represents ~18.4% of all AS events conserved between human and mouse(1). Half of these events involve the NAGNAG motif at the 3' splice sites (2,3). Though the NAGNAG motif is found frequently in 3' splice sites, only ~4% of these tandem acceptors are confirmed by EST to be alternatively spliced in human and mouse, while in 86% of the cases the proximal splice site is constitutively selected (P) and in 10% the distal splice site is chosen (D). We have previously shown that it is possible to distinguish between sequences that undergo alternative versus constitutive splicing (CS) at the NAGNAG motifs without relying on EST data. Among the features which are characteristic of the AS NAGNAG sites are high sequence conservation of the motif, high conservation of ~30 bp at the intronic regions flanking the 3' splice site and overabundance of *cis*-regulatory elements in the flanking intronic and exonic regions (2).

In this study we compute these and others characteristics for a variety of training sets of alternatively and/or constitutively spliced NAGNAG motifs and show that we can automatically classify them using a Support Vector Machine (SVM). In addition, based on the assumption that tandem acceptors are a subset of all Alternative 3' Splice Site events, we expand our analysis to further cases in which the acceptors sites are spaced from each other and successfully discriminate AS from CS events. Finally, we suggest a possible mechanism of splice site selection both in Alternative and Constitutive Splicing.

2. RESULTS

Applying a machine learning approach we automatically classify three types of splicing events at the NAGNAG motif (fig 1A). As shown we can distinguish AS events from both P and D events with relative high accuracy. Interestingly, a high performance is attained when separating P from D events. Generally, the most discriminative parameters between the groups are splice site composition, namely the sequence of the NAGNAG motif, intronic conservation near the splice site and the strength and relative position of the polypyrimidine tract.

Further we examine A3SS cases in which the acceptor sites are separated at distances varying from 4 to 100 nucleotides. Using similar parameters the SVM succeeds to discriminate between A3SS events and a control set of Constitutive Splicing events including an AG site in varying distances (fig 1B). In addition, we observe that when using a control set of sequences in which the AG which does not serve as a splice site is not evolutionarily conserved, the SVM performance is considerably higher than when the conservation of the AG dinucleotide is not accounted for. The later result could suggest that 1- there is contamination of the CS ESTs within the AS events, 2- In some cases an AS-like regulatory process is required in order to avoid the selection of a proximal splice site, which is usually chosen as the default (4,5).

In order to examine the second assumption we have carried out additional tests comparing subsets of CS sequences to the AS datasets. In a subset of sequences in which the distance between the splice site and the nearest AG site is 4-12 nt, we find that when the AG dinucleotide is evolutionarily conserved, SVM performance considerably drops compared to cases in which the AG dinucleotide is not conserved (fig 1C). We interpret that only in the former group the sequence environment resembles that of Alternative 3' Splice Site events. However, when the AG dinucleotide is located far from the splice site (30-100 nt), the sequence environment at the CS events does not display the characteristic features of AS neither when the AG site is conserved nor when it is not (fig 1D). Overall, our results imply that regulation is necessary in order to avoid the selection of undesired AG sites. Interestingly, we do not observe this regulation when the AG site is not conserved, perhaps because these sites are still evolving. Moreover such regulation seems not to be required when an AG is placed far from the real splice site, presumably upstream the branch point, since this AG will be sequestered in the lariat during the second step of transesterification.

In Conclusion our findings indicate that both the selection between two alternative splice sites and also the recognition of a constitutive AG site from a range of possible splice sites are regulated processes.

3. FIGURES

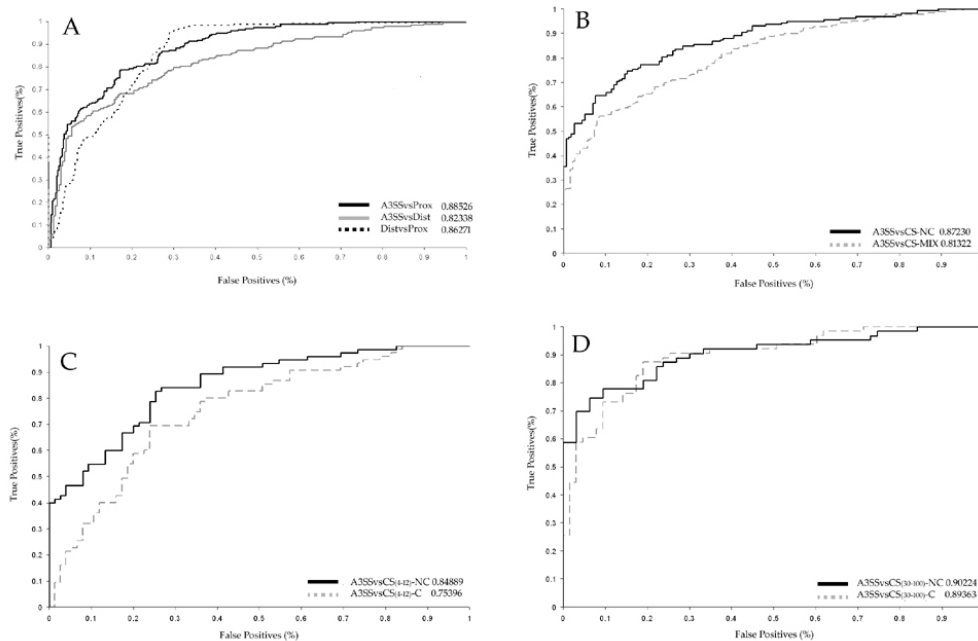


Figure 1: ROC plots for the different SVM runs. (A) Three SVM experiments with the NAGNAG motif :Alternative Splicing versus Proximal, Alternative splicing versus Distal and Distal versus Proximal.(B) Alternative 3' Splice Site events of length 4-100 versus a control set of all constitutive splicing events(A3SSvsCS-MIX) compared to events where the AG dinucleotide is not evolutionarily conserved (A3SSvsCS-NC). (C) Alternative 3' Splice Site events of length 4-12 versus constitutive splicing events in which an AG dinucleotide is conserved (A3SSvsCS-C) or not (A3SSvsCS-NC).(D) Alternative 3' Splice Site events of length 30-100 versus constitutive splicing events in which an AG dinucleotide is conserved (A3SSvsCS-C) or not (A3SSvsCS-NC).

4. REFERENCES

- .1 Sugnet, C.W., Kent, W.J., Ares, M., Jr. and Haussler, D. (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac Symp Biocomput*, 66-77.
- .2 Akerman, M. and Mandel-Gutfreund, Y. (2006) Alternative splicing regulation at tandem 3' splice sites. *Nucleic Acids Res*, **34**, 23-31.
- .3 Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backofen, R. and Platzer, M. (2004) Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat Genet*, **36**, 1255-1257.
- .4 Chua, K. and Reed, R. (2001) An upstream AG determines whether a downstream AG is selected during catalytic step II of splicing. *Mol Cell Biol*, **21**, 1509-1514.
- .5 Smith, C.W., Chu, T.T. and Nadal-Ginard, B. (1993) Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. *Mol Cell Biol*, **13**, 4939-4952.

Identifying Splicing Regulatory Elements using Correlation with Expression

Debopriya Das^{1*}, Tyson A. Clark², Simon Minovitsky³, Inna Dubchak³, John E. Blume², John G. Conboy¹

¹Life Sciences Division and ³Genomics Division, Ernest O. Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

²Affymetrix, Inc. 3420 Central Expressway. Santa Clara, CA 95051, USA

*To whom correspondence should be addressed: ddas@potternexus.lbl.gov

1. INTRODUCTION

Alternative splicing is a versatile modulator of cell-type specific gene expression in mammals, and is a major determinant of cell-type fate and specificity. The complex decision process involving which subset of exons on the primary RNA transcript (henceforth, pre-mRNA) will get spliced into the mature mRNA isoform is mediated by a combination of *cis*-regulatory elements organized across exons and introns, quite analogous to the *cis*-regulation of transcription[1]. Global identification of splicing regulatory elements has been difficult and has been primarily limited to exonic elements[2, 3]. Availability of splicing microarrays[4-6], which interrogate expression levels of exons genome-wide under any particular biological condition, has opened up new possibilities however. One can now apply analogous computational approaches used for dissecting transcriptional regulation[7] to decipher the splicing regulatory elements, with genes replaced by exons and promoters by pre-mRNA regions proximal to the splice sites.

A new set of approaches based on correlation with expression have been particularly successful in identifying *cis*-regulatory elements governing transcription[8-11]. Here, the premise is that gene expression results from integration of multiple signals on to the promoter region, as mediated by binding of *trans*-factors to the *cis*-elements. This implies that the regulatory motif parameters (occurrence frequencies and position weight matrix (PWM) scores) must be correlated with the expression levels across genes under any specific biological condition. Active motifs show statistically significant correlation; inactive motifs do not. Thus, one can identify the motifs functional under the tested condition. Furthermore, expression data from a single test condition and a reference condition are sufficient for the analysis. The choice of reference data depends on the context[12]. In addition, unlike clustering-based approaches, interacting combinations of motifs can be inferred with high confidence[9]. Finally, a recent study shows that such approaches can accurately identify direct targets of *trans*-factors binding to the active motifs, even when motifs are very degenerate[12]. Target identification in such instances has been quite challenging. Thus, one can delineate the transcriptional regulatory networks using correlation with expression. This has proven effective in both lower eukaryotes, e.g. yeast[8, 9, 13, 14], and mammals[12].

2. RESULTS

Here we have applied correlation with expression based approaches on exon microarray data to systematically identify the intronic *cis*-regulatory elements directing the muscle-specific alternative splicing program. From tissue-specific microarray data[15], we identified 56 exons whose expression levels were elevated in heart and skeletal muscle, relative to 14 other tissues. These muscle-enriched exons are highly conserved across vertebrates, relatively short (median size = 84nt) and reside in large genes (median size = 123kb) that are enriched in cytoskeletal functions. We used both linear regression[8] and linear splines[12] to examine whether *cis*-elements in introns adjoining these exons correlate with gene-normalized exon expression in muscle. Candidate regulatory motifs were obtained using enumerative approaches, and included both non-degenerate RNA words and degenerate PWMs. The statistically significant motifs were also examined for over-representation against pre-mRNA sequences around constitutive exons. The latter analysis was repeated for conserved exons in mouse, chicken and frog. In downstream (upstream) 200nt sequence, we identified a total of 35 (27) hexamers that were significantly correlated with expression and also over-represented in human, of which 9 (3) were over-represented in at least one other species. Several of these elements have been previously characterized experimentally as regulators of muscle-specific splicing: UGCAUG, a highly specific binding site for Fox-1 related proteins; UG-rich element, known for binding CELF splicing factors; pyrimidine-rich elements in upstream intron, known for PTB binding. We also identified a novel

branchpoint-like element, ACUAAC, which has been predicted as muscle-specific in another computational study [16]. Similar results were obtained using PWMs. In conclusion, results from correlation with expression are supported by over-representation and phylogenetic analyses, indicating that this is an effective approach for dissecting the *cis*-regulation of alternative splicing.

3. REFERENCES

1. Black, D.L. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 72: 291-336.
2. Fairbrother, W.G., *et al.* 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* 297: 1007-1013.
3. Wang, Z., *et al.* 2004. Systematic identification and analysis of exonic splicing silencers. *Cell* 119: 831-845.
4. Johnson, J.M., *et al.* 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302: 2141-2144.
5. Frey, B.J., *et al.* 2005. Genome-wide analysis of mouse transcripts using exon microarrays and factor graphs. *Nat Genet* 37: 991-996.
6. Clark, T.A., C.W. Sugnet, and M. Ares, Jr. 2002. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* 296: 907-910.
7. Tompa, M., *et al.* 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 23: 137-144.
8. Bussemaker, H.J., H. Li, and E.D. Siggia. 2001. Regulatory element detection using correlation with expression. *Nat Genet* 27: 167-171.
9. Das, D., N. Banerjee, and M.Q. Zhang. 2004. Interacting models of cooperative gene regulation. *Proc Natl Acad Sci U S A* 101: 16234-16239.
10. Conlon, E.M., *et al.* 2003. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci U S A* 100: 3339-3344.
11. Keles, S., M. van der Laan, and M.B. Eisen. 2002 Identification of regulatory elements using a feature selection method. *Bioinformatics* 18: 1167-1175.
12. Das, D., Z. Nahlé, and M.Q. Zhang. 2006. Adaptively inferring human transcriptional subnetworks. *Mol Syst Biol (Nature/EMBO)* 2: doi:10.1038/msb4100067. <http://www.nature.com/msb/journal/v2/n1/synopsis/msb4100067.html>
13. Wang, W., *et al.* 2002. A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* 99: 16893-16898.
14. Wang, W., *et al.* 2005. Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. *Proc Natl Acad Sci U S A* 102: 1998-2003.
15. Clark, T.A., *et al.*, Discovery of tissue-specific exons using comprehensive human microarrays. *Submitted*.
16. Sugnet, C.W., *et al.* 2006. Unusual Intron Conservation near Tissue-Regulated Exons Found by Splicing Microarrays. *PLoS Comput Biol* 2: e4.

Differentiated evolutionary rates in alternative exons and the implications for splicing regulation

Eduardo Eyras* and Mireya Plass

Pompeu Fabra University, Passeig Maritim de la Barceloneta 37-49, E08003, Barcelona, Spain

*To whom correspondence should be addressed: eduardo.eyras@upf.edu

1. INTRODUCTION

Two contradicting properties have been associated to alternative exons: higher sequence conservation (1,2) and higher rate of non-synonymous substitutions (3,4), relative to constitutive exons. In order to clarify this issue, we have performed an analysis of the evolution of alternative and constitutive exons, using a large set of protein coding exons conserved between human and mouse and taking into account the conservation of the transcript exonic structure.

The motivation for our analysis is the following: Splicing regulatory elements are abundant in constitutive and alternative exons and this is likely to exert some selection pressure on the exon sequence at the pre-mRNA level. On the other hand, changes in the regulatory elements are known to affect the splicing pattern of a gene. Thus, a purifying selection at the pre-mRNA would be linked to a constraint on the splicing regulation, and therefore to a conservation of the exonic structure, whereas a relaxation of this selection would be linked to a weaker constraint on the splicing regulation, and to a lack of conservation of the exonic structure. In summary, the conservation of the exonic structure is expected to influence the sequence conservation of alternative and constitutive exons, and consequently, the measurements of the synonymous and non-synonymous substitutions. Accordingly, we separated our set of constitutive and alternative exons into four groups according to whether they were part of a transcript with an exonic structure that is conserved in the orthologous gene or not. Exons in transcripts with conserved exonic structure (CES) are called CES exons, whereas exons in transcripts with non-conserved exonic structure are called non-CES exons. A non-CES exon is such that there is a pattern of splicing of the pre-mRNA, which includes this exon, and which is never the same in the orthologous mRNA whenever the orthologous exon is included.

2. RESULTS

We find evidence for a relation between the lack of conservation of the exonic structure and the weakening of the sequence evolutionary constraints in alternative and constitutive exons. Non-CES exons have higher synonymous (dS) and non-synonymous (dN) substitution rates than CES exons (Figures 1a and 1b). Moreover, alternative non-CES exons are the least constrained in sequence evolution, and at high EST-inclusion levels they are found to be very similar to constitutive exons (Figures 2a and 2b), whereas alternative CES exons have dS values significantly lower than average at all EST-inclusion levels (Figure 2a). At high inclusion levels, alternative and constitutive exons of the same type (CES or non-CES) have indistinguishable dN distributions. However, at high inclusion levels, CES and non-CES exons can still be separated by their dN. We conclude that most of the differences in dN observed between alternative and constitutive exons can be explained by the conservation of the transcript exonic structure. Additionally, low dS values are characteristic of alternative CES exons at all EST-inclusion levels, but not of alternative non-CES exons. These results provide evidence for a selection pressure related to the splicing of the pre-mRNA.

Furthermore, we have also defined a measure of the variation of the arrangement of exonic splicing enhancers (ESE-conservation score) to study the evolution of splicing regulatory sequences. We have used this measure to correlate the changes in the arrangement of ESEs with the divergence of exon and intron sequences. We find a higher conservation in the arrangement of ESEs in constitutive exons compared to alternative ones. Additionally, the sequence conservation at flanking introns remains constant for constitutive exons at all ESE-conservation values, but increases for alternative exons at high ESE-conservation values, indicating a higher density of regulatory signals.

3. FIGURES

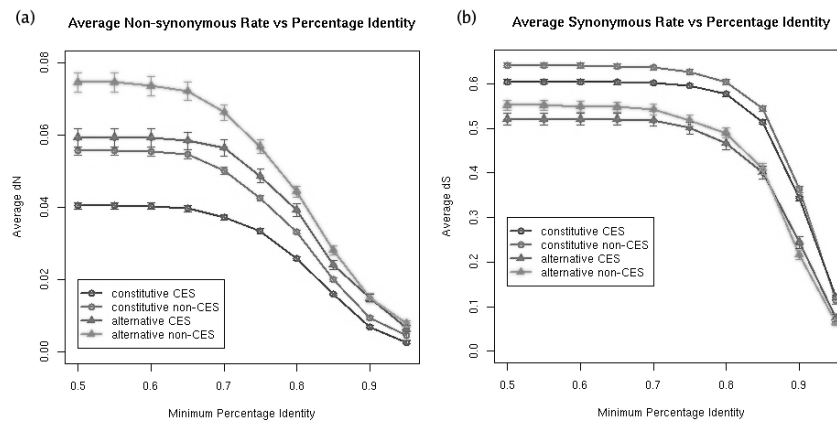


Figure 1. Average (a) non-synonymous (dN) and (b) synonymous (dS) substitution rates (y-axis) for different minimum percentage identity conservation values (x-axis).

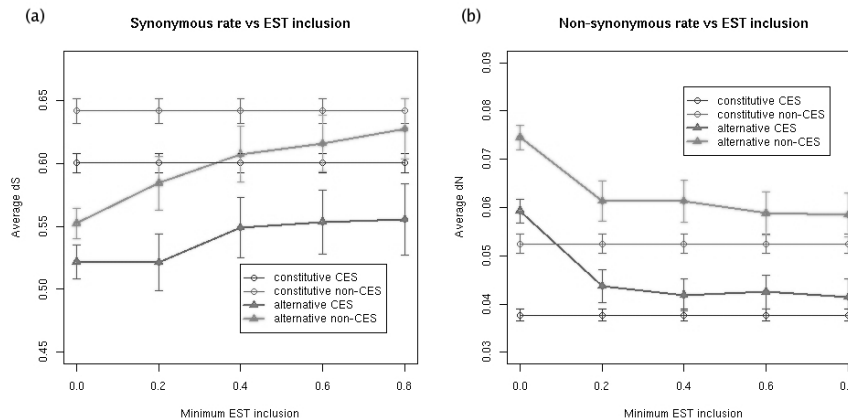


Figure 2. Variation of the (a) synonymous (dS) and (b) non-synonymous (dN) substitution rates for alternative CES and non-CES exons (y-axis) with the EST inclusion level (x-axis). For comparison, we have superimposed the average synonymous rate for all constitutive CES and non-CES exons as a straight line with the corresponding error bars.

4. REFERENCES

1. Xing, Y. and Lee, C. 2005. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc Natl Acad Sci USA*. 102(38):13526-31.
2. Chen, F-C., Wang S-S., Chen, C-K., Li, W-H. and Chuang, T-J. 2006. Alternatively and Constitutively Spliced Exons Are Subject to Different Evolutionary Forces. *Mol. Biol. Evol.* 23:675-682.
3. Sorek, R., Shemesh, R., Cohen, Y., Basechess, O., Ast, G. and Shamir, R. 2004. A non-EST-based method for exon-skipping prediction. *Genome Res.* 14(8):1617-1623.
4. Philipps, D.L., Park, J.W. and Graveley, B.R. 2004. A computational and experimental approach toward a priori identification of alternatively spliced exons. *RNA* 10(12):1838-44.

Comparison of Alternative Splicing Structures in Eukaryotes

M Sammeth*, S Foissac, R Guigó

Genome Bioinformatics Laboratory, Institut Municipal d'Investigació Mèdica, 8003 Barcelona, Spain

*To whom correspondence should be addressed: msammeth@imim.es

1. INTRODUCTION

Little is known about the origin and evolution of alternative splicing (AS). There are, however, several hypotheses that try to give explanations. For instance, it has recently been proposed that AS may have appeared from DNA mutations in constitutive splice sites that gave rise to the suboptimal recognition of exons by the splicing machinery (1), producing alternative exons. Another mechanism proposed for the origin of AS is the exonization of introns. In particular, it has been observed that specific mutations in repeat sequences of type ALU can give rise to weak splice sites, which can be recognized as alternative exons or splice sites (2). A third proposed mechanism is the tandem duplication of exons (3,4). Following this explanation, both variants may keep a similar level of expression but would develop specialized functions. In a summary, the evidence shows that AS may provide an organism with the possibility to explore new protein functions while not disrupting the fitness of the original protein by allowing the addition of novel domains while maintaining the gene fitness intact (5). Accordingly, alternative exons are expected to have more freedom to change in sequence. Indeed, there is evidence that alternative exons have on average a higher aminoacid substitution rate than constitutive exons (6,7). Furthermore, it is well known that only a small fraction of all possible exonic combinations is found in a cell (8) and it is suggested that these pattern may change during evolution. Although recently also first structure catalogs on AS patterns have been published (9), a comparative structural analysis to evaluate their evolution has not yet been described.

2. STRUCTURAL COMPARISON OF AS

We propose a novel approach to assess the evolution of alternative splicing by compare the AS structures across known orthologous genes of different species. Firstly, we introduce a notation system for the formal description of arbitrarily complex AS events (Figure 1). This notation is based on a graph of splice sites (vertices) and splices (edges) and we provide a form to extract the relevant information in a matrix representation for convenient structural analysis (Figure 1b). We describe an AS event in a more general way using the term of splice variations. A *splice variation* is a set of varying splice sites between two conserved flanking splice sites (or the respective transcription start or transcription end), i.e., the description of two or more diverging paths through the splice graph. The sites are as such defined when comparing two or more transcripts (Figure 1a). By this, the information relevant to analyze AS is encoded in a series of splice variations, and the evolution of AS can be studied in terms of the changes in such splice variations across orthologous genes in different species.

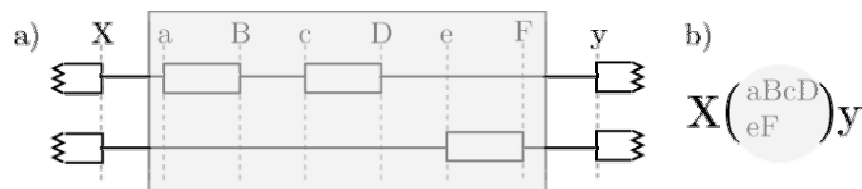


Figure 1 A pairwise splice variation comprising two exon-skipping events and a mutually-exclusive exon event. (a) The exon-intron structure of two transcripts is compared and a conserved donor (X) and acceptor site (y) flanking a pairwise splice variation (highlighted in the grey box) are identified. All splice sites in the splice variation are named alphabetically according to their relative position within the gene and donor sites are assigned capital letters and acceptor sites lowercase letters (top). (b) A non-redundant matrix representation of the splice variation in (a).

The definition of splice variations allows us to model every AS pattern in a very flexible way, without having to focus on either an exon- or an intron-based nomenclature. Furthermore, we can describe novel types of events (e.g., mutually exclusive introns), and complex events that can not be expressed by using a composition of the 5 traditional AS events. Moreover, this description can specify some direction in the variation – splice variations can be distinguished by the relative orientation of the neighboring events (Figure 2). A further benefit of splice variations is that they are sticked to splice sites and therefore the notation is capable of capturing an arbitrary number of exons and introns.

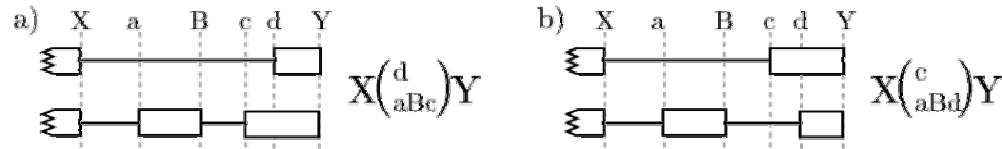


Figure 2 Example of two splice variations capturing the same two AS events (i.e., exon skipping and alternate acceptor), but in a different relative orientation to each other: in (a) the alternate donor is more downstream in the transcript that is skipping an exon, while in (b) it is reversed.

On the graph based notation, we describe novel measures to capture attributes of AS patterns, e.g., the *degree* (a complexity measure). To also describe the difference between 2 AS patterns of corresponding genes (orthologs/paralogs), we set up a distance metrics. To our knowledge, this is the first time that descriptive measures and metrics have been defined for the structural comparison of AS patterns. In the end, the described methods are applied to the genomic data of 10 species (human, mouse, rat, dog, cow, chicken, frog, fish, fruitfly, mosquito) as extracted from the Ensembl database (10).

3. REFERENCES

1. Ast, G. 2004. How did alternative splicing evolve? *Nature Rev. Genetics* 5; 773-782.
2. Sorek, R., Lev-Maor, G., Reznik, M., Dagan, T., Belinky, F., Graur, D. and Ast, G. 2004. Minimal conditions for exonization of intronic sequences; 5' splice site formation in Alu exons: *Mol. Cell.* 14; 221-231.
3. Kondrashov, F.A. and Koonin, E.V. 2001. Origin of alternative splicing by tandem exon duplication: *Hum. Mol. Genet.* 10; 2661-2669.
4. Letunic, I., Copley, R.R. and Bork, P. 2002. Common exon duplication in animals and its role in alternative splicing: *Hum. Mol. Genet.* 11, 1561-1567.
5. Modrek, B. and Lee, C. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss: *Nature Genetics* 34: 177-180.
6. Iida, K. and Akashi, H. 2000. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes: *Gene* 261: 93-105.
7. Xing, Y. and Lee, C. 2005. Assessing the application of Ka/Ks ratio test to alternatively spliced exons: *Bioinformatics* 21: 3701-3703.
8. Graveley, B.R. 2001. Alternative splicing: increasing diversity in the proteomic world: *Trends in Genetics* 17: 100-107.
9. Nagasaki, H., Arita, M., Nishizawa, T., Suwa, M. and Gotoh, O. 2005. Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. *Gene* 364:53-62.
10. Birney, E. et al. 2006. Ensembl 2006: *Nucleic Acids Res.* 1: D556-D561.

HOLLYWOOD: a comparative alternative splicing database for studying mechanisms mediated by exonic splicing enhancers and silencers

Holste D^{1*}, Tung V¹, Huo G², Yeo G³, Faibrother WG¹, Wang Z¹, Shomron N¹, Lim LP⁷, and Burge CB¹

¹Institute of Molecular Pathology, Dr-Bohr-Gasse 7, A-1030, Vienna, Austria

²Department of Biology and ³Department of Computer Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 023193

⁴Salk Institute-CNL, 10010 North Torrey Pines Road, La Jolla, CA 92037

⁵Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02319

⁶Rosetta Inpharmatics LLC, 401 Terry Avenue North, Seattle, WA 98109

*To whom correspondence should be addressed: holste@imp.ac.at

Keywords: cDNA sequences, expressed sequence tags, RNA splicing, exonic splicing enhancers, exonic splicing silencers, relational database

INTRODUCTION

In metazoan genomes, where the majority of genes are transcribed as pre-mRNAs and intervened by introns, the splicing of precursors to mRNAs constitutes a critical mode for the regulation of gene expression (1). RNA splicing is often variable, giving rise to multiple alternatively spliced (AS) mRNA isoforms from a single gene locus, and a significant number of human genes encode proteins by AS pathways [2-4]. The investigation of alternative mRNA isoforms and translated protein products are important to decipher the 'RNA code', and hence to fundamentally understand the control of gene expression in development, tissue specificity, and human disease at the level of RNA processing [2,5-8]. Using the transcript-to-genome alignment systems, one can detect and distinguish AS events in terms of whether mRNA isoforms differ by inclusion or exclusion of an exon, or whether isoforms differ in the usage of a 3' splice site (3'ss) or 5'ss, producing alternative 3'ss exons or alternative 5'ss exons, respectively. These descriptions are not necessarily exclusive, but an exon can make several alternative splice site choices. In addition to biochemical studies, computational identification and analysis of AS events have been conducted, enabled by the abundance of acquisition of millions of different transcripts compiled in sequence databases.

The design of effective bioinformatics tools for investigations of AS poses a challenge [9]. Here, challenges for bioinformatics of AS are the recognition of authentic AS patterns from an average of 200-300 expressed sequence tags (ESTs) for each annotated human gene, the clear annotation and characterization of AS events in different genomes and the linking of such patterns across different lineages, as well as the generation of testable hypotheses for regulatory mechanisms of alternative splice site choices. In an effort to integrate accurate exon and splice site annotation to foster the deeper understanding about splicing regulatory elements and predicted AS events, and to link information about the splicing of orthologous genes in different species, we have developed the HOLLYWOOD system [10]. It was built upon genomic annotation of splicing patterns of known genes derived from spliced alignment of cDNAs and ESTs, and links features such as splice site sequence and strength,

exonic splicing enhancers and silencers, conserved patterns of splicing, and cDNA library information for inferred alternative exons. We describe the background of HOLLYWOOD and show applications by means of case studies. HOLLYWOOD was implemented as a relational database and currently contains comprehensive information for the human and mouse genome. It is accompanied by a web query tool that allows searches for sets of exons with specific splicing characteristics or splicing regulatory element composition, or gives a snapshot of splicing patterns for a specific gene. A streamlined graphical representation of gene splicing patterns is provided, and these patterns can alternatively be layered onto existing information in the UCSC Genome Browser. The database is accessible at <http://hollywood.mit.edu>.

REFERENCES

- (1) Maniatis T. and Reed R. (2002) An extensive network of coupling among gene expression machines. *Nature* 416:499-506
- (2) Lopez A.J. 1998. Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu Rev Genet* 32:279-305
- (3) Black D.L. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 72:291-336.
- (4) Matlin A.J., Clark F. and Smith C.W. 2005. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol* 6:386-398
- (5) Cartegni L., Chew S.L., and Krainer A.R. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* 3:285-298
- (6) Caceres J.F. and Kornblihtt A.R. 2002. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet* 18:186-193
- (7) Venables J.P. 2004. Aberrant and alternative splicing in cancer. *Cancer Res* 64:7647-7654
- (8) Faustino N.A. and Cooper T.A. 2003. Pre-mRNA splicing and human disease. *Genes Dev* 17:419-437
- (9) Modrek B., and C. Lee. 2002. A genomic view of alternative splicing. *Nat Genet* 30:13-19
- (10) Holste D. et al., HOLLYWOOD: a comparative relational database of alternative splicing. *Nucleic Acids Res* 34:D56-62

SAGE/MPSS Bioinformatics: going after splicing variants differentially expressed in tumors

Pedro A.F. Galante^{1,4}, Gustavo S. Guimarães², Gregory Riggins³, Janete Cerutti² and Sandro J. de Souza^{1*}

1. Ludwig Institute for Cancer Research, São Paulo Branch
2. Disciplina de Reumatologia, Universidade Federal de São Paulo
3. Dep. of Neurosurgery, Johns Hopkins University
4. Ph.D Program, Departamento de Bioquímica, Universidade de São Paulo

Last few years have witnessed a dramatic improvement in the development of bioinformatics tools to explore the repository of data from Serial Analysis of Gene Expression (SAGE) experiments. Sage Genie (cgap.nci.nih.gov/SAGE) was developed by us in an attempt to better assign SAGE tags to genes and vice-versa. The approaches developed for SAGE Genie were used here to explore SAGE and MPSS (Massive Parallel Signature Sequencing) data to quantify the expression of splicing variants. This was possible due to the fact that some splicing variants either lack the 3'most NlaIII (for SAGE) or DpnII (for MPSS) sites or gain new 3'most sites. This allowed us to identify cases in which the new tag is differentially expressed in tumors while the original tag is equally expressed. We have identified more than 2,000 variants differentially expressed in human tumors. We presented experimental validation for one variant coding for the COBW protein.

PAFG is supported by FAPESP.

Using the Affymetrix Exon Array to investigate variation in alternative splicing in a human population

Tony Kwan^{1,2}, Rob Sladek¹, Christel Dias¹, Scott Gurd¹, David Serre^{1,2}, Harry Zuzan¹, Thomas J. Hudson^{1,2}, Tyson A. Clark³, Anthony Schweitzer³, Hui Wang³, Michelle K. Staples³, John E. Blume³, and Jacek Majewski^{1,2*}.

¹ Department of Human Genetics, McGill University, Montreal, QC H3A 1A4, Canada

² McGill University and Genome Quebec Innovation Center, Montreal, H3A 1A4 QC, Canada

³ Affymetrix, Inc. Santa Clara, CA 95051, USA

*To whom correspondence should be addressed: jacek.majewski@mcgill.ca

INTRODUCTION

The sequencing and annotation of the human genome revealed an unexpectedly low number (~ 25000) of protein coding gene [1]. Since this number is comparable to the number of genomic loci found in much simpler organisms, such as *C. elegans* and *Arabidopsis*, recent hypotheses postulate that alternative pre-mRNA splicing may be an important mechanism for regulation of gene expression and increasing the complexity of the mammalian transcriptome [2]. Furthermore, variation in alternative splicing among individuals may be responsible for phenotypic diversity and differential susceptibility to genetic disorders [3].

The vast majority of our genome-scale data on alternative splicing has so far been derived from EST analyses. This data is often noisy, incomplete, and much better suited for detection of alternative isoforms across different tissues than across individuals. Microarray-based studies promise to provide new insights into genome-wide splicing patterns. However, large-scale analyses of alternative splicing have generally been carried out using custom chips designed for specific experimental purposes [4], and/or were limited by the amount of splicing-related genome annotation [5]. The recently released Affymetrix Exon Array relies on sets of probes targeted to individual exons, allowing independent measurement of expression levels of over 1.4 million probesets corresponding to approximately 1 million known and predicted human exons.

Here, we show the effectiveness of the Exon Array for investigating variation in alternative splicing among individuals. Our approach is based on comparing lymphoblastoid cell lines derived from the CEPH population [6]. Our experimental setup provides easy access to large amounts of high quality mRNA, as well as detailed, high resolution (1kb) information on individual genotypes from the HapMap project [7]. Although using a single cell type results in a vast underestimate of the total number of splicing differences, it allows us to set a lower limit on the extent of variation within the entire organism.

In our initial approach, we used replicate (15x) lymphoblastoid cell lines from two individuals to select candidate exons with differential levels of inclusion. The results were subsequently subjected to verification using RT-PCR, and the confirmed true positives were tested for genetic effects by following their transmission in families (linkage) and significant genotype association within the CEPH HapMap panel (60 unrelated individuals).

We show that the Exon Array is a valuable discovery tool, allowing us to detect both known (EST-based) and novel (previously unannotated) cases of alternative splicing. We demonstrate that several differences in splicing patterns are further supported by linkage and/or association analyses, suggesting that they have underlying genetic causes. We are currently investigating candidate genetic differences in order to elucidate the mechanisms of splicing variation. We are also extending our microarray coverage to the entire panel of 60 unrelated individuals, which will allow us to assess the amount of splicing variation in the population and identify its underlying genetic basis.

REFERENCES

1. Claverie, J.M., *Gene number. What if there are only 30,000 human genes?* Science, 2001. **291**(5507): p. 1255-7.
2. Kim, H., et al., *Estimating rates of alternative splicing in mammals and invertebrates.* Nat Genet, 2004. **36**(9): p. 915-6.
3. Faustino, N.A. and T.A. Cooper, *Pre-mRNA splicing and human disease.* Genes Dev, 2003. **17**(4): p. 419-37.
4. Ule, J., et al., *Nova regulates brain-specific splicing to shape the synapse.* Nat Genet, 2005. **37**(8): p. 844-52.
5. Johnson, J.M., et al., *Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.* Science, 2003. **302**(5653): p. 2141-4.
6. Cohen, D., I. Chumakov, and J. Weissenbach, *A first-generation physical map of the human genome.* Nature, 1993. **366**(6456): p. 698-701.
7. Altshuler, D., et al., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.

Selection of splicing variant candidates in breast tumor through microarray platform

Rangel MCR, Kirschbaum-Slager NS, Camargo LP, Souza SJ, Brentani HP, Carraro DM
Ludwig Institute for Cancer Research, Rua Antonio Prudente, 109 - 4º andar, São Paulo-SP, Brazil
Hospital do Câncer AC Camargo, Rua Antonio Prudente, 109 - 1º andar, São Paulo-SP, Brazil
*mrangel@ludwig.org.br

1. INTRODUCTION

In the Brazilian female population, breast cancer represents the highest death rate among all cancers. Although some molecular markers have been utilized routinely for the prognosis of this disease, the search for additional markers, that may contribute to increase the global survival and to more effective treatment is necessary.

Current analyses have shown that alternative mRNA splicing (AS) appears in at least 70% of human genes (1), contributing to the wide diversity of transcripts and increasing the proteomic complexity. It is known that some variants generated by AS or by aberrant splicing are preferentially expressed in human tumors (2). In cancer, the importance of genetic diversity generated by alternative splicing resides in the possibility to correlate specific variants with clinical information in order to identify tumor markers (3).

To identify splicing variants associated with breast cancer, 270 tumor-associated exons were selected from the exons of all genes by a computational analysis (4), of which 75 were over expressed in breast tumor. Exons were immobilized on a nylon membrane and hybridized with 27 tumoral and 5 non-neoplastic breast tissues. Samples were manually microdissected and total RNA was amplified in 2 rounds using T7 based methodology using both template switching (TS) and T7 promoter. Five micrograms of amplified RNA was labeled with 60 uCi of $\alpha^{32}\text{P}$ -dCTP and hybridized in duplicate. The generated signal was captured by a Phosphorimager and analyzed by the Array Vision program. Normalization of the data was performed by using the intensity mean of 24 replicates of constitutive genes (GAPDH and ACTB). The T Student test was applied in order to determine differentially expressed exons ($p < 0.05$) comparing different tumor and non-neoplastic cell lines and samples.

Fourteen exons over expressed in breast cancer were selected with differential expression fold ranging from 3.0 to 5.9, displaying significant representativity among the 75 exons previously selected by computational analysis as breast cancer-associated. A SAGE analysis was performed in order to verify if the detected over expression was related with the exon variant or with the tumor over expression of the whole gene. From 7 exons with reliable information at SAGE bank, none of the respective genes showed over expression in breast, supporting that the over expressions are exclusive of their variants. Functional annotation was performed and some of the genes, like BAP1, LIP8, MBTPS1, PRKD2, TRIM 37 and PPP1R8 are likely related with cancer.

2. REFERENCES

1. Johnson JM, Castle J, Garrett-Engle P, et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction *microarrays*. *Science* 2003; 302:2141-4.
2. Venables JP. Aberrant and alternative splicing in cancer. *Cancer Res* 2004; 64:7647-54.
3. Brinkman BM. Splice variants as cancer biomarkers. *Clin Biochem* 2004; 37:584-94.
4. Kirschbaum-Slager N, Parmigiani RB, Camargo AA, de Souza SJ. Identification of human exons over-expressed in tumors through the use of genome and expressed sequence data. *Physiol Genomics* 2005; 21:423-32.

Informatics Issues in Affymetrix Human Exon Array Data Analysis

Manhong Dai¹, Haiming Chen², Melvin McInnis², Margit Burmeister^{1,2}, Marquis Vawter³,
Stanley J. Watson^{1,2}, Fan Meng^{1,2*}

¹Molecular and Behavioral Neuroscience Institute and ²Psychiatry Dept., U. of Michigan, MI 48109, USA; ³ Department of Psychiatry and Human Behavior, U. of California, Irvine, CA 92697, USA

*To whom correspondence should be addressed: mengf@umich.edu

1. INTRODUCTION

The release of the Human Exon 1.0 Sense Transcript Array (HsExon Array) in the second half of 2005 represents a major advance in our ability to study transcriptome-wide splicing changes. According to Affymetrix, over 5 million sense probes on the HsExon array are able to interrogate 1.4 million exons derived from different annotations. Affymetrix groups probe sets into three categories (core, extended and full), from the most conservative to highly exploratory, for maximizing the chance of discovering novel splicing and expression patterns (1). However, there are three main problems in Affymetrix' approach:

1). While the inclusion of all possible exons derived from different methods is a good chip design strategy, it leads to difficulties in the interpretation of analysis results. This is because Affymetrix' gene/transcript/exon definitions are not widely adopted by the research community and none of the three Affymetrix probe set categories is 100% consistent with popular genome annotation databases such as ENSEMBL or Entrez Gene. The use of individual high quality genome annotation databases in the data analysis stage will lead to more straightforward interpretation, with improved statistical power since including annotations from many different sources aggravates the multiple testing problem.

2). About 6.9 % of the probes on the HsExon Array can be aligned to multiple genomic locations. These non-specific probes can influence the expression/splicing level of more than one transcript or exon, leading to mostly false positives and also false negatives under some situations. It is estimated that $\sim 1-(1-0.069)^4 = 25\%$ of the probe sets can potentially be influenced by the non-specific probe problem, for an average probe set of 4. Although not every non-specific probe will affect results, eliminating all non-specific probes from exon probe set definitions is necessary for reliable analysis.

3). Affymetrix' HsExon Array includes 271,710 probes with single genome hit and overlap over 57,000 SNPs with known minor allele frequency (MAF) above 0.1 in the central 15 bp region of the probe (Table 1). The real number of SNPs with MAF>0.1 is likely to be higher based on existing HapMap/ ENCODE data (2).

Table 1. SNPs Represented by Allele-Specific Probes on the Affymetrix HsExon Array

Minor Allele Frequency (MAF)	Number of SNPs represented on the HsExon Array*	Estimated MAF percentage from HapMap/ENCODE	Estimated SNP count on the HsExon Array based on HapMap/ENCODE data
MAF > 0.10	57532 [#] 12918 ^{##} 4472 ^{###}	~41.5%	93007 [#] 20883 ^{##} 7229 ^{###}
MAF > 0.20	40606 [#] 9175 ^{##} 3212 ^{###}	~26.5%	59390 [#] 13419 ^{##} 4698 ^{###}
MAF > 0.30	26353 [#] 5988 ^{##} 2100 ^{###}	~16%	35858 [#] 8148 ^{##} 2857 ^{###}

* Number of SNPs represented by at least one probe ([#]), two probes (^{##}) and three probes (^{###}) on the HsExon Array. Multiple probes often overlap due to the inclusion of different gene/transcript/exon definitions.

Probes with central 15-bp matches are likely to generate very different signals in samples with different genotypes, based on the existing literature (3,4). The presence of such probes with relatively high MAF in a probe set may reduce the sensitivity of detecting real splicing changes due to increased noise level. It may also increase the false positive rate when the sample size is not very large due to the uneven allele distribution. E.g., a MAF=0.2 allele has a 0.0029 probability of displaying $\geq 40\%$ allele frequency difference in a 20 subjects vs. 20 subjects sample set. This means most likely 117 (40606 x 0.0029) or more SNPs will exhibit $\geq 40\%$ frequency difference in these two groups by chance alone, creating ample opportunities for exon-level signal difference given the average size of exon probe set is only 4.

2. MATERIALS AND METHODS

In order to evaluate the impact of non-specific probes and the allele-specific probes on HsExon Array analysis results, we generated 100 sample sets (21 vs. 21) based on HsExon Array CEL files from 42 postmortem brain samples derived from different human subjects. A large number of permutation generated data sets will allow us to examine the effect of different probes in different expression and

genotype combinations using limited amount of experimental data.

We created three types of gene/exon probe sets with probe set size ≥ 3 using the gene/exon boundaries defined in the ENSEMBL CORE database (Version 37) in a procedure described previously (5): 1) raw gene/exon probe sets (32445 genes and 233933 exons) based on probe genomic location 2) specific gene/exon probe sets (29330 genes and 214201 exons) generated by removing non-specific probes in the raw probe sets 3) best gene/exon probe sets (27464 genes and 202517 exons) derived by eliminating known allele-specific probes in the specific probe sets. However, the best probe sets will contain unknown allele-specific probes or non-specific probes due to limitations in current knowledge. We use the gene expression-level normalized exon signal level as an indicator for potential differential alternative splicing events. The R-package SIGGENES (6) is used to estimate the False Discovery Rate (FDR) for each gene expression-level normalized exon signal in every permutation generated data set.

3. RESULTS

We summarized analysis results from 100 permutation generated sample groups in Table 2. It can be seen clearly that the inclusion of non-specific probes in the RAW probe sets dramatically increase the number of gene expression normalized exon signals passing the FDR and fold-change thresholds, particularly at relatively stringent thresholds. As expected, the removal of allele-specific probes increases the sensitivity of detecting potential splicing events by more than 2 fold at $FDR \leq 0.1$ levels (SPECIFIC vs. BEST). The last two columns in Table 2 show the average percentage of the gene expression level normalized exon signals that are also detected by the SPECIFIC and the RAW probe sets. About half of the potential splicing signals cannot be detected in the presence of prob, under the assumption that the BEST probe sets provide more reliable splicing signal estimations.

Table 2. Comparison of Affymetrix Human Exon Array Analysis Results

FDR	Fold Change	RAW probe set	SPECIFIC probe set	BEST Probe set	% BEST in SPECIFIC	%BEST in RAW
≤ 0.05	≥ 1.5 or $\leq 1/1.5$	47 ± 108	2.6 ± 2.6	6.0 ± 12	40 ± 22	30 ± 22
≤ 0.1	≥ 1.5 or $\leq 1/1.5$	90 ± 180	19 ± 28	46 ± 72	42 ± 20	44 ± 21
≤ 0.2	≥ 1.5 or $\leq 1/1.5$	310 ± 510	220 ± 340	230 ± 350	57 ± 21	58 ± 19
≤ 0.05	≥ 1.25 or $\leq 1/1.25$	120 ± 280	3.0 ± 2.7	7.4 ± 16	41 ± 24	30 ± 23
≤ 0.1	≥ 1.25 or $\leq 1/1.25$	240 ± 600	38 ± 70	108 ± 193	38 ± 21	40 ± 21
≤ 0.2	≥ 1.25 or $\leq 1/1.25$	1160 ± 2210	690 ± 1230	830 ± 1540	52 ± 24	53 ± 24

In conclusion, while the Affymetrix Exon Array platform provides the possibility of monitoring potentially splicing changes for over 80% of known exons, it is critical to consider the effect of non-specific and allele-specific probes. Our data suggest that the presence of problematic probes may significantly increase false positives as well as reduce the sensitivity of detecting real alternative splicing events.

Acknowledgement: M. Dai, M. Burmeister, M. Vawter, S. J. Watson and F. Meng are members of the Pritzker Neuropsychiatric Disorders Research Consortium, which is supported by the Pritzker Neuropsychiatric Disorders Research Fund L.L.C.. H. Chen is supported by NIMH K01MH064596 and NARSAD Young Investigator Award 2003-3005.

4. REFERENCES

1. Affymetrix. (2005) Exon Array Computational Tool Software User's Guide <http://www.affymetrix.com/products/arrays/specific/exon.affx>
2. Altshuler, D., Brooks, L.D., Chakravarti, A., Collins, F.S., Daly, M.J. and Donnelly, P. (2005) A haplotype map of the human genome. *Nature*, **437**: 1299-1320.
3. Mei, R., et al. (2003) Probe selection for high-density oligonucleotide arrays. *Proc Natl Acad Sci U S A*, **100**: 11237-11242.
4. Lee, I., Dombkowski, A.A. and Athey, B.D. (2004) Guidelines for incorporating non-perfectly matched oligonucleotides into target-specific hybridization probes for a DNA microarray. *Nucleic Acids Res*, **32**: 681-690.
5. Dai, M., et al. (2005) Evolving gene/ transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*, **33**: e175.
6. Schwender, H. (2005) siggenes. <http://www.bioconductor.org/repository/devel/package/html/siggenes.html>.

Bioinformatics Open Source Conference 2006

Table of Contents

1. Schedule
2. Welcome
3. Abstracts
 - 3.1. Keynotes
 - 3.2. BioPostgres
 - 3.3. GObase: A Graph Database System for Managing and Exploring Gene Ontology
 - 3.4. BioMoby Helping the Crop Scientists
 - 3.5. XRATE: Scheme-y trees, phylo-HMMs and phylo-grammars
 - 3.6. Twease: Searching Medline, One Sentence at a Time
 - 3.7. The ZMap Genome Annotation Viewer
 - 3.8. Bioinformants: Biological, Informational Agents on the Internet
 - 3.9. CaIntegrator - An Open Source Translational Informatics Platform to Integrate Clinical Trials and High Throughput Molecular Analysis in Support of Transition to Tailored Therapy
 - 3.10. SHOGUN - A Large Scale Machine Learning Toolbox for Biological Sequence Analysis
 - 3.11. XMLPipeDB: A Reusable, Open Source Tool Chain for Building Relational Databases from XML
 - 3.12. BioRuby Shell and RubyTools
 - 3.13. Informatics for Evolutionary Science and Synthesis

1. Schedule

Table 1. Schedule

	Friday August 4th	Saturday August 5th
08:45am	call to order	call to order
09:00am	Stott Parker: BioPostgres	KeyNote: Alberto M.R. Davila
09:30am	Ruey-Lung Hsiao: GObase: A Graph Database System for Managing and Exploring Gene Ontology	
10:00am	Martin Senger: BioMoby Helping the Crop Scientists	Subhashree Madhavan: CaIntegrator - An Open Source Translational Informatics Platform to Integrate Clinical Trials and High Throughput Molecular Analysis in Support of Transition to Tailored Therapy
10:30am	Coffee	Coffee
11:00am	Ian Holmes: XRATE: Scheme-y trees, phylo-HMMs and phylo-grammars	Soren Sonnenburg: SHOGUN - A Large Scale Machine Learning Toolbox for Biological Sequence Analysis
11:30am	Matt Wood: Twease: Searching Medline, One Sentence at a Time	Kam Dalquist: XMLPipeDB: A Reusable, Open Source Tool Chain for Building Relational Databases from XML Sources
12:00	Lunch	Lunch
1:30pm	Roy Storey: The ZMap Genome Annotation Viewer	Toshiaki Katayama and Pjotr Prins: BioRuby Shell and BioRuby Tools
2:00pm	Jose Lopez: Bioinformants: Biological, Informational Agents on the Internet	Hilmar Lapp: Informatics for Evolutionary Science and Synthesis
2:30pm	Lightning talks and Demos	Lightning talks and Demos
3:30pm	Coffee:	Coffee:
4:00pm	KeyNote: Amos Bairoch: UniProtKB and bioinformatics open source: where do we go from here?	Lightning talks and Demos
5:00pm	BOFs	BOFs

2. Welcome

Welcome to BOSC 2006. This is the 6th official Bioinformatics Open Source Meeting. We are pleased to announce two exciting keynote speakers this year. Amos Bairoch, of the Swiss Institute of Bioinformatics, will be speaking on Friday afternoon in a talk entitled 'UniProtKB and bioinformatics open source: where do we go from here?'. On Saturday, Alberto M.R. Davila, Department of Biochemistry and Molecular Biology Instituto Oswaldo Cruz - Fiocruz, will be speaking.

We have an impressive group of speakers scheduled each day, focusing on a broad range of development projects representing the breadth of the Open Source Development Community. During the afternoon sessions we have scheduled Lightning Talks and Software demonstrations. Birds of a Feather (BOF) discussions will occur at the end of each day. Please take advantage of this time to attend discussions on specialized topics. If you would like to schedule a BoF see the signup chart that will be available in the mornings.

We hope you enjoy yourself, learn a lot, and most importantly get to know each other and become part of the community of open source development in the life sciences.

Conference Committee

Darin London (chair)	Institute for Genome Science and Policy, Duke University Medical Center
Ewan Birney	European Bioinformatics Institute
Hilmar Lapp	Nescent, Durham, North Carolina
Nomi Harris	University of California, Berkeley
Jason Stajich	TODO JASON

3. Abstracts

3.1. Keynotes

Amos Bairoch, Swiss Institute of Bioinformatics

Alberto M.R. Davila, DBBM, Instituto Oswaldo Cruz, Rio de Janeiro, Brasil

3.2. BioPostgres

D.S. Parker, Ph.D. UCLA

Friday August 4th, 9:00am

BioPostgres is a collection of modules that can be used to extend PostgreSQL, the popular open-source database system, for bioinformatics and computational biology. Each of the BioPostgres modules is intended to be independent, and to be usable separately or in conjunction with the others. Current modules include:

BLASTgres -- Biosequence database extensions
GObase -- GeneOntology (GO) extensions
PostGraph -- Graph database extensions
PostMake -- Data derivation dependency extensions
PostModel -- Model base/data mining extensions

Each extension implements new datatypes with query operators and other tools, providing powerful large-scale query and analysis, quick inclusion of new types of biological information, and integration of diverse existing bioscience data. For example, two important datatypes implemented in the modules above are biosequence locations and graphs, and query operators include BLAST access and graph component extraction.

BioPostgres illustrates how the extensibility of PostgreSQL and open-source development can be harnessed for bioscience data management. Modules are added to a user's PostgreSQL source tree as needed; each module is compiled separately and installed in a loadable library directory. PostgreSQL itself is not recompiled, but instead permits library modules to be dynamically linked into a database -- added "on the fly" -- with appropriate SQL commands.

BioPostgres is being developed at the Center for Computational Biology (CCB) at UCLA. Currently all modules are distributed under the GNU GPL.

Web site: www.biopostgres.org

3.3. *GObase: A Graph Database System for Managing and Exploring Gene Ontology*

Ruey-Lung Hsiao, UCLA

Friday August 4th, 9:30am

The Gene Ontology (GO) defines standard vocabulary for biological terms in three aspects -- molecular functions, biological processes, and cellular components. Its wide acceptance and semantic annotation has led to a wide range of applications -- semantic integration, functional analysis, and microarray gene clustering, to name a few. It represents an important new route for connecting information of different types and has become an essential component in system biology.

The success of the GO underscores the importance of having ways to manage, query and visualize it. Despite this importance, most researchers still use an inefficient way to represent and query the GO. The central data structure of the standard relational database for the GO consists of two tables: term and term2term. The term table keeps track of the basic information, such as accession number and term type, of a particular GO term and term2term represents the relationship between two terms (i.e., edges in the term graph). However, most queries for ontology are recursive in nature (One example: find all the descendants of a particular node), this representation in conjunction with SQL can fare poorly under a variety of performance measures.

In order to resolve this issue, the GO database pre-computes the transitive closure of the graph to keep track of every descendants of the node in the graph. In this way, graph traversal queries can be answered through this table without issuing recursive SQL commands. However, the pre-computation of transitive closure leads to very inefficient usage of storage space by storing a huge number of node pairs. There is a clear and significant need for better support of GO representation.

To address these issues we developed the GObase system, a publicly-available open-source platform for the management, query, and visualization of GO information. GObase is a graph database, implemented as an extension of the PostgreSQL database with a graph datatype. This datatype permits storage and (with addition of new functions) sophisticated query of graph data. The graph representation is both more natural and more efficient for queries than the widely-used GO relational database. There are existing browsers for visualizing GO information, but the GO Term Viewer of GObase provides a more powerful interface, allowing both interactive query and interactive annotation of GO terms. In addition, GObase also links the GO with various other biological information resources.

website: www.go-base.org license: GNU General Public License

3.4. BioMoby Helping the Crop Scientists

Martin Senger, Ph.D, International Rice Research Institute, Philippines

Friday August 4th, 10:00am

The Generation Challenge Programme (GCP) is an international agricultural research consortium, currently numbering 20 agricultural research institutes, working on the characterization of plant genetic resources and the application of comparative genomics, toward crop improvement for the developing world. Given the global dispersion of GCP partners, distributed access to informatics resources is a major challenge. Open source projects, such as BioMoby, are essential for linking GCP components into a coherent information gateway.

The BioMoby is an integration tool helping to connect databases, analysis programs, and other resources into a unified distributed system, linking gene discovery with genetic resource characterization and crop evaluation data. GCP developers added new components there, some of them are presented here.

BioMoby Dashboard is a rich standalone Java application for BioMoby developers, assisting through the whole process of creating, deploying and using Biomoby services. It has a plug-in architecture allowing additional extensions.

BioMoby MoSeS ("Moby Service Support") is a set of code generators that transform information and ontology trees stored in a central service registry into Java source code, helping developers of new services to concentrate only on business logic and not on the protocol and messaging details. Such approach guarantees scalability of the developing process.

BioMoby Environment is an automated way to regularly check if the deployed services are running and producing correct results. In the environment with so many participants, such quality control tool is essential.

BioMoby can cooperate with other integrating tools. A separate project "BioCASE & BioMoby" shows how to use BioMoby to "wrap a wrapper". The BioCASE integrates access to many databases, and BioMoby spreads its data into existing clients and networks.

Links: GCP <http://www.generationcp.org> BioMoby <http://biomoby.org> BioMoby Java projects: http://biomoby.open-bio.org/CVS_CONTENT/moby-live/Java/docs/ BioCASE <http://www.biocase.org>

3.5. XRATE: Scheme-y trees, phylo-HMMs and phylo-grammars

Ian Holmes, Ph.D, Berkeley University

Friday August 4th, 11:00am

A big hit in the past couple of years has been the "phylo-HMM", a multi-sequence HMM employing Felsenstein's pruning algorithm to compute emission scores. (Stochastic grammar extensions to this idea include phylo-GHMMs and phylo-SCFGs.) The phylo-HMM idea was first introduced for genome analysis by Churchill and Felsenstein, and further developed e.g. for RNA structure prediction by Knudsen and Hein. The use of phylo-grammars by Siepel, Pedersen, Bejerano, Haussler et al for gene prediction, evolutionary analysis of rate variation, and other forms of genome annotation has gotten lots of attention in recent years.

Much of the appeal of phylo-grammars is the straight transfer of intuition and expertise from the areas of HMMs and SCFGs. However, the well-known EM algorithms used to train these models (Baum-Welch, Inside-Outside) are a little less straightforward to apply to phylo- grammars. In contrast to (say) Baum-Welch, the phylo-EM algorithm is pretty hairy and not something you'd really want to implement twice.

In the past 4 years we have taken the theory of phylo-EM algorithms from a theoretical treatment (Holmes & Rubin, JMB, 2002) up to a full-blown open-source implementation of a general phylo-grammar prototyping, training and annotation engine (XRATE). Grammars can be specified using a Scheme-like format, "trained" on alignments using phylo-EM, and then used to annotate alignments. The phylo-EM code in our open-source C++ library can also be linked to by external applications (e.g. Jakob Pedersen's EVOFOLD program, which has been used to investigate recently-evolving human ncRNAs). Several developers have contributed full-time to the process, and there is considerable stability, including a battery of automated tests.

XRATE is an easy-to-use Unix app that brings the unrestricted power of phylo-grammars in reach of a first-year grad student or smart undergrad. In a historical aside, XRATE has its roots in a grammar compiler, TELEGRAPH, that was itself based on Ewan Birney's DYNAMITE (and is related to Guy Slater's EXONERATE). TELEGRAPH was presented at BOSC 2000. At BOSC 2006, I'll show how far XRATE has come by giving a tour of its rate-measurement and annotation abilities, accompanied by visualizations of the interesting variety of patterns (covariation, neighbor-dependence, conservation, lineage-specific acceleration, selection...) that can be observed in the mutation rates of genomic features.

3.6. Twease: Searching Medline, One Sentence at a Time

Matt Wood, Ph. D. Cornell University.

Friday August 4th, 11:30am

With 16 million abstracts available in Medline, most searches match more documents than is possible to read. We asked if sentence level searches would be more effective in retrieving articles of interest than the whole abstract method currently used to support most biomedical searches. We present and evaluate a web-based tool to search Medline at the sentence level (available from <http://www.twease.org/>). The tool indexes each sentence of Medline individually and provides features that help correct for the lack of context introduced when searching sentences separately from the rest of the sentences in an abstract. We evaluated Twease with queries assembled from an independently obtained protein-protein interaction dataset (2,789 distinct interactions), as a measure of performance when retrieving abstracts with conjunctive queries of biomedical interest. Experimental results indicate that, on average, the first 25 Twease hits contain as many relevant abstracts as the first 100 PubMed hits. The first 25 Twease hits are also twice more likely to contain a relevant hit than the first 100 PubMed hits. These results indicate that sentence level searches, as implemented in Twease, are a competitive strategy when searching the biomedical literature for articles about multiple concepts (e.g., protein-protein interactions, or disease gene/protein relationships). Because a Twease index can be created directly from a text collection and does not require custom semantic resources, the approach implemented in Twease can be used to index and search any text collection of abstracts or full text articles. Twease implements highly scalable algorithms and approaches that will be discussed during the presentation and is released under the GNU General Public License. The distribution can be downloaded from: <http://icb.med.cornell.edu/crt/twease/index.xml>.

3.7. The ZMap Genome Annotation Viewer

Roy Storey, Sanger Institute, Hinxton, Cambridge

Friday August 4th, 1:30pm

We present a software package, ZMap, that is a visualisation tool for genomic features. The software is a multi-threaded application written in C, utilising the gnome toolkit (GTK2) and draws features on the foocanvas. ZMap accepts input from multiple sources in multiple formats across multiple genomes and is written in way that the addition of further formats is made as trivial as possible. Currently the list of formats includes GFF v2 and DAS1 which may reside in any one of; a file, an acedb instance, a http server. Multiple genomes and their associated features can be displayed in a single view as aligned blocks providing support for comparative annotation.

A wide complement of browser features are implemented in the ZMap GUI. Users may zoom to any resolution in the range from individual bases to a whole chromosome. The display may be split in both vertical and horizontal axis multiple times in a way akin to emacs. The contents of the display are completely copied in order that, as for emacs, the two views are independently scrollable. In this way it is possible to utilise screen space to view both ends of features at a much higher resolution than would otherwise be possible. Users are able to perform operations on the full sequence context, such as reverse complement, print, export and search.

Internally ZMap has a client/server architecture, where the GUI control thread acts as the client making requests to each server that communicates with a unique source. Each server runs within its own thread enabling the graphical thread to remain responsive. Various servers may add to the display at any point during the lifetime of the application. The features are merged with the current context allowing features from multiple sources to be viewed along side each other. As threads are separated from the control interface a user can kill the request in the event of an unresponsive or slow server.

ZMap does not natively include any utility for editing the features which it displays. It does however provide a powerful external interface with which to modify the features which are displayed on the canvas. ZMap utilises the X11 atom mechanism for interprocess communication, via which it is possible to communicate in xml. A perl module X11:XRemote is provided to facilitate ease of integration with perl annotation tools. Using this interface Sanger's production annotation tool otterlace is used to annotate sequence present in the Otter database which in turn updates to the Vega website.

Software License -----

The GNU General Public License (GPL) Version 2

3.8. Bioinformants: Biological, Informational Agents on the Internet

Jose Lopez, Ph.D. University of Tachira, San Cristobal, Tachira, Venezuela

Friday August 4th, 2:00pm

We define a bioinformant as a software agent with specific abilities to serve as an assistant for a scientist. Bioinformants are embedded in a web application which, apart from being the agents' testbed, is also a server/client application to allow access to a conventional GNU/Linux environment from a browser. The goal of the project is two fold. First, we want to develop a platform to recover and analyze genetic information, (re)using existing tools for biologists. But, we also want to integrate newer tools to process biodata. In both cases, we are working to embed agents to deal with applications use, integration and layered learning. This far, we are preparing the release, as free and open source software, of the basic test bed for agents and a knowledge base to support datamining exercises on DNA sequences.

License used for software release: Our software is free under the GNU General Public License as published by the Free Software Foundation, version3.

Related URL: <http://sourceforge.net/projects/simulants>

3.9. CalIntegrator - An Open Source Translational Informatics Platform to Integrate Clinical Trials and High Throughput Molecular Analysis in Support of Transition to Tailored Therapy

Subha Madhavan, Ph.D. , National Cancer Institute

Saturday August 5th, 10:00am

Progress in finding better therapies for cancer treatment is hampered by the lack of opportunity to integrate biomedical data from disparate sources to enable translation of therapies from bench to bedside and back. Hence, a critical factor in the advancement of biomedical research and delivery is the ease with which data from clinical trials can be integrated, redistributed and analyzed both within and across functional domains. Novel biomedical informatics infrastructure and tools are essential for developing individualized patient treatment regimens based on the specific genomic signatures in each patientsed gene expression, Copy number Abnormality (CNA), SNPs, clinical trials data etc.) in a cohesive fashion.

Following are some of the high-level features of the caIntegrator framework:

- N-Tiered Architecture: The caIntegrator framework is implemented using Java 2 Enterprise Edition, a Data Warehouse, and various other open source technologies. It is designed to conform to an n- tiered architecture that includes several layers: A web-based graphical user interface, a business object layer, server components that process the queries and result sets, a data access layer and a data warehouse.
- A Common set of interfaces (APIs) and a domain object model: The domain object model called CGOM (Clinical Genomics Object Model) provides standardized programmatic access to the integrated biomedical data collected in the caIntegrator data system. The model represents data from clinical trials, microarray-based gene expression, SNP genotyping and copy number experiments, and Immunohistochemistry-based protein assays. Clinical domain objects in CGOM allow

access to Clinical trial protocol, treatment arms, patient information, sample histology, clinical observations and assessments. Genomic domain objects allow access to biospecimen information, raw experimental data, in-silico transformation and analyses performed on the raw experimental datasets and biomarker findings. The application's user interface communicates with its caIntegrator-based middle-tier services via domain as well as business objects.

- A real-time analytical service that provides on-the-fly computational analysis capability for caIntegrator applications and currently supports class comparison analysis, principal component analysis and clustering analysis. It is designed to easily incorporate other types of analysis in the future, and scale to provide performance.

- The caIntegrator data system consists of a star schema database, which contains the clinical, and annotation data as dimensions, pre-calculated gene expression copy number data as facts. For performance reasons, normalized gene expression data used by the real time analysis module is stored as R-binary files.

- A plotting interface to allow visualization of genomic data (copy number scatter plots and ideogram) via the webGenome application (<http://webgenome.nci.nih.gov>).

The overall goal of the caIntegrator project is to provide a caBIG-compatible (https://cabig.nci.nih.gov/guidelines_documentation) framework with the infrastructural components needed to develop enterprise level translational applications. One such reference implementation at NCICB is REMBRANDT (Repository of Molecular BRAin Neoplasia DaTa) - <http://rembrandt.nci.nih.gov>. REMBRANDT is a powerful and intuitive informatics system designed to integrate genetic and clinical information from brain tumor clinical trials for improved research, disease diagnosis, and treatment. It provides researchers with the ability to perform ad hoc querying and reporting across multiple data domains, such as Gene Expression, Chromosomal aberrations and Clinical data. Scientists are able to answer basic questions related to a patient or patient population and view the integrated data sets in a variety of contexts. Tools that link data to other annotations such as cellular pathways, gene ontology terms and genomic information are embedded within this system. Two other cancer study applications that are being developed using the caIntegrator framework are:

- I-SPY Breast cancer trial: The primary object of the I SPY Trial is to identify surrogate markers of response to neoadjuvant chemotherapy that are predictive of pathologic remissions and survival in Stage III breast cancer. Goal is to identify Molecular markers and/or MRI results that predict 3-year disease-free survival in these patients.

- CGEMS: Cancer Genetic Markers of Susceptibility (CGEMS) is an NCI project to identify genetic alterations that make people susceptible to prostate and breast cancers. Goal of CGEMS is to analyze the entire genome for most common type of SNPs that are associated with each of these diseases.

caIntegrator knowledge framework offers a paradigm for rapid sharing of information and accelerates the process of analyzing results from various biomedical studies with the ultimate goal to rapidly change routine patient care.

Resources:

CaIntegrator website: <http://caintegrator.nci.nih.gov> REMBRANDT (caIntegrator reference implementation): <http://rembrandt.nci.nih.gov> REMBRANDT application URL: <http://rembrandt-db.nci.nih.gov> CaIntegrator open source license: <http://ncicb.nci.nih.gov/download/caintegratorlicenseagreement.jsp>

3.10. SHOGUN - A Large Scale Machine Learning Toolbox for Biological Sequence Analysis

Soren Sonnenburg, Fraunhofer Institut, Max Planck Society.

Saturday August 5th, 11:00am

We have developed a Machine Learning toolbox, called SHOGUN, which is designed for large scale sequence analysis tasks appearing in computational biology. It features a number of machine learning algorithms such as Support Vector Machines [3] for classification and regression, but also Hidden Markov Models, Multiple Kernel Learning [1, 8], Linear Discriminant Analysis, Linear Programming Machines and Perceptrons. Most of these algorithms are able to deal with several different data types including sparse vectors and sequences.

SHOGUN provides a generic SVM object interfacing to seven different SVM implementations, among them are LibSVM[2] and SVMlight[5], two state-of-the-art SVM implementations. Each of these can be combined with a variety of kernels. The toolbox not only provides efficient implementations of the most common kernels, like the linear, polynomial, Gaussian and sigmoid kernel, but also comes with a number of recently proposed string kernels including the Fischer & TOP kernel [4, 15], the Spectrum kernel [6], the locality improved kernel [16] and the weighted degree kernel [7, 13]. For most of the sequence kernels we have implemented optimized data structures such as tries to speed-up training and evaluation of SVMs.

We have successfully used this toolbox to tackle the following sequence analysis problems: Protein Super Family classification[15], Splice Site Prediction[10, 11], Interpreting the SVM Classifier [12, 8], Splice Form Prediction[7], Alternative Splicing[9] and Promotor Prediction[14]. Some of them come with no less than 10 million training examples, others with 7 billion test examples.

SHOGUN is implemented in C++ and interfaces to MatlabTM, R, Octave and Python/Numarray. The Source Code is freely available at <http://www.fml.mpg.de/raetsch/projects/shogun> under the GNU General Public License, Version 2.

References F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In C. E. Brodley, editor, Twenty-first international conference on Machine learning. ACM, 2004. C.-C. Chang and C.-J. Lin. Libsvm: Introduction and benchmarks. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2000.

3.11. XMLPipeDB: A Reusable, Open Source Tool Chain for Building Relational Databases from XML

Kam Dalquist, Department of Biology, Loyola Marymount University

Saturday August 5th, 11:30am

XMLPipeDB is an open source suite of Java-based tools for automatically building relational databases from an XML schema (XSD). XMLPipeDB provides functionality for managing, querying, importing, and exporting information to and from XML data with minimum manual processing of the data. While its applicability is fairly general, the original motivation for XMLPipeDB was to create a solution for the management of biological data from different sources that are used to create Gene Databases for GenMAPP (Gene Map Annotator and Pathway Profiler), software for viewing and analyzing DNA microarray and other genomic and proteomic data on biological pathways. The creation of Gene Databases for GenMAPP has been difficult because there are a number of different gene ID systems in common usage, necessitating that we relate one set of gene identifiers to the other. Currently, the GenMAPP Gene Databases use the integrated data source from Ensembl for this task. However, this limits the number of species that can be represented in GenMAPP to the mostly animal species supported by Ensembl. Here we report that we have used the XMLPipeDB software tool chain to create relational databases for UniProt and Gene Ontology. In turn, we have used these databases to generate UniProtcentric GenMAPP Gene Databases for *Escherichia coli* and other bacterial species, extending the functionality of GenMAPP to species not currently supported by the GenMAPP.org project team. Moreover, since XMLPipeDB can create the relational databases based solely on the XSD and XML files, it will be more robust to changes in the source files made by the data providers.

XMLPipeDB has the following tools for developers and database designers: the XSD-to-DB application takes a well-formed XSD or DTD file and converts it into a collection of Java source code and Hibernate mapping files that allows XML files based on that definition file to be read into a relational database. XSD-to-sses that provide functions needed by many XMLPipeDB database applications. Specifically, the library includes reusable classes for: importing XML files into Java objects, saving these XML-derived Java objects to a relational database, querying the relational database using either HQL (Hibernate Query Language) or SQL, and configuring a client application to communicate with a relational database. Finally, GenMAPP Builder is an application for creating the GenMAPP Gene Database files.

GenMAPP Builder has been tested for use with the open source PostgreSQL relational database, but can be used with any other relational database management system for which a JDBC driver is available. JDBC-to-ODBC connectivity is used to transfer data from this relational database to a Microsoft Access MDB file, which is the format expected by the GenMAPP application.

XMLPipeDB was developed by graduate students as part of a team-taught course in bioinformatics that was then extended into a second workshop course on open source software development. The primary objectives of this course are to gain real world experience with open source software development, learning key open source

development concepts, gaining proficiency with tools that are frequently used in the open source community (many of which are used in effective software development in general), and making a concrete contribution to an open source software project. XMLPipeDB is available under the GNU Library or Lesser General Public License (LGPL) at <http://sourceforge.net/projects/xmlpipedb>.

3.12. BioRuby Shell and RubyTools

Toshiaki Katayama, Human Genome Center, Institute of Medical Science, University of Tokyo
Pjotr Prins, Wageningen University

Saturday August 5th, 1:30pm

As one of the Open Bio* projects hosted by Open Bio Foundation, we have been developing the BioRuby, a Ruby library for bioinformatics. In this February, we have released the BioRuby version 1.0 coming with various new features, bug fixes, unit tests and documentations. In this talk, I will describe what we have achieved with the 1.0 release and our future plans.

The Open Bio* libraries have been successful as toolkits to develop customized bioinformatics pipelines, however, it was still difficult for the biologist to utilize the libraries as their daily tool. There can be two main reasons, (1) learning a programming language is a burden for their spare time and (2) there are **too** many ways to do it! Thus, we included the BioRuby shell, a newly developed CUI (command line user interface) for the BioRuby library. BioRuby shell integrates and abstracts various ways of entry retrieval, flatfile processing, and accessing of web services, even with enabling users to utilize all functionality of Ruby and BioRuby without writing any script file. In the shell, objects and the history are conserved across the sessions, and a script to reproduce the procedure can be automatically generated.

Other enhancements in BioRuby 1.0 include unit tests and documentations. Hohman started to add unit tests for some essential classes in BioRuby and Nakao has completed >1000 tests to make our library stable. We also added a English guideline to contribute our projects, and thanks to it, several developers joined to our project. Aerts and Raaum have been worked on the documentation format specification (RDoc format) and adding detailed API documentations. Goto translated Japanese tutorial and Prins improved the English version. Wennblom launched <http://bioruby-doc.org/> site to develop further BioRuby documentations.

With the world wide Ruby user base growing (in part thanks to Ruby on Rails) we are putting an effort in to increase the adoption of Ruby in bio-informatics by both improving the documentation of the BioRuby project and dedicating a significant section to Bio-informatics on the SciRuby project: a portal for all things scientific and Ruby. The latter an initiative of the Pragmatic Bookshelf team - which is planned to lead to a book on Ruby in science (see <http://sciruby.codeforpeople.com/>).

In this talk we present two open source initiatives namely TaxonSearch - a BLAST post- processor for the genomic mining of taxonomic information and Ruby Queue (rq), a free and straightforward clustering job management tool. TaxonSearch stores full taxonomic information of all matches of a large scale BLAST exercise and allows

for complex regular expression based queries by end users. We used it, for example, to find putative lateral gene transfer candidates from bacteria to nematodes (BLASTing EST databases against the NCBI non-redundant database). The program features a split search and query procedure because the BLAST search can take a lot of (cluster) time. Once the database has been built, queries can happen on a simple user workstation with no other dependencies beside the Ruby interpreter.

Ruby Queue (rq) can be used to drastically reduce the overhead and complexity of distributing work to a collection of commodity workstations. Ruby Queue does not depend on other clustering tools and can be run on a number of Linux machines mounting a shared network file system (NFS).

Both TaxonSearch and Ruby Queue tools are also relevant to non-Ruby users. As a conclusion we will try to clarify why we think Ruby as a language is to play a major role in bio-informatics. TaxonSearch will be published for BOSC 2006 through both BioRuby and SciRuby projects under the GPL.

BioRuby code is Ruby licensed (GPL, IIRC) Ruby Queue (rq) is Ruby licensed (GPL, IIRC) The bio-informatics sections in SciRuby are published under a creative commons license. Pjotr Prins (see <http://thebird.nl/>) is a researcher at the Department of Nematology, Wageningen University and a visiting research fellow at the Department of Bio-informatics at the University of Groningen.

Pjotr initiated the OSS Cfruby, xParrot and GenEst projects and contributes to BioRuby and SciRuby.

Ara Howard (see <http://sciruby.codeforpeople.com/>) is a professional research assistant at the National Geophysical Data Centre (NGDC).

3.13. Informatics for Evolutionary Science and Synthesis

Hilmar Lapp, Ph.D. National Evolutionary Synthesis Center (NESCent), Durham, NC

Saturday August 5th, 1:30pm

The National Evolutionary Synthesis Center (NESCent) was founded in November 2004 with funding from the U.S. National Science Foundation with the goal of addressing "grand challenge" questions in evolutionary biology. NESCent pursues this goal by sponsoring synthetic science that is highly collaborative, involving teams of researchers from institutions around the globe, and also highly interdisciplinary, reaching across disciplines such as genetics, developmental biology, systematics, ecology, geography, and paleontology. Much of this research is also informatics-intensive, utilizing and combining datasets, annotation, and analysis methods from multiple domains.

NESCent is well positioned to play a leading role in improving the infrastructure for synthetic science in evolutionary biology. Areas of critical need include data and metadata sharing standards and technologies, open libraries to support standard data exchange formats, software interoperability and usability, and training a critical mass of open source software developers within the discipline.

We describe three collaborative informatics projects at NESCent that are helping to enable synthetic science in different ways. First, we are extending GMOD tools and data models to accommodate a key evolutionary datatype, population variation. Second, we are establishing a searchable meta-data registry for evolutionary data to promote data reusability. Third, we are helping systematists and developmental geneticists connect knowledge about zebrafish mutants, on the one hand, and natural phenotypic diversity among related fish in the Order Cypriniformes, on the other, through semantic mediation.

Finally, we encourage participation and seek input from the informatics community at large. We are sponsoring 'hackathons' and software-oriented working groups for the development of open software libraries in evolutionary biology. We are also soliciting whitepapers to identify areas of focus for the future (see <http://www.nescent.org/>). NESCent is committed to open-source development and to partnering with existing open-source projects wherever possible.

JBB 2006: The Joint BioLINK-Bio-Ontologies Meeting

A Joint Meeting of
The ISMB Special Interest Group on Bio-Ontologies and the
BioLINK Special Interest Group on Text Data Mining

In association with ISMB 2006, Fortaleza, Brazil

August 5, 2006

Program Chairs:

BioOntologies

Phillip Lord, School of Computing Science, Newcastle University, UK
Robert Stevens, School of Computer Science, University of Manchester, UK
Robin McEntire, GlaxoSmithKline, USA
Jim Butler, GlaxoSmithKline, USA

BioLINK

Hagit Shatkay, School of Computing, Queen's University, Canada
Lynette Hirschman, MITRE, USA
Alfonso Valencia, Spanish National Cancer Research Centre (CNIO), Spain
Christian Blaschke, Bioalma, Spain

Program Committee:

Sophia Ananiadou, University of Manchester
Judith Blake, The Jackson Laboratories
Olivier Bodenreider, NLM, NIH
Kevin B. Cohen, University of Colorado School of Medicine
Nigel Collier, National Institute of Informatics, Japan
Carol Friedman, Columbia University
Udo Hahn, Jena University
Midori Harris, European Bioinformatics Institute
William Hayes, Biogen
Marti Hearst, University of California at Berkeley
Eivind Hovig, University of Oslo
Larry Hunter, University of Colorado School of Medicine
Lars Jensen, EMBL Heidelberg
Cliff Joslyn, Los Alamos National Labs
Michael Krauthammer, Yale University
Marc Light, University of Iowa
Joel Martin, National Research Council, Canada
Jong Park, KAIST, South Korea
Helen Parkinson, European Bioinformatics Institute
Luis Rocha, University of Indiana
Andrey Rzhetsky, Columbia University
Lorrie Tanabe, NCBI, NLM, NIH
Jun-ichi Tsujii, University of Tokyo
Karin Verspoor, Los Alamos National Labs
Marc Weeber, KnewCo Inc.
John Wilbur, NCBI, NLM, NIH

Special thanks to:

- Andrew Gibson and Simon Harper for technical expertise and advice.
- Steven Leard, the meeting planner, for accommodating our logistical needs.

Keynote Speakers

Suzanna Lewis, NCBO-Berkeley

The Joy of Ontology

In 1998, FlyBase announced that we intended to turn our personal collection of controlled vocabularies and techniques into an ontology. Ontology building would no longer remain a private passion for FlyBase. FlyBase was a celebrated model organism database that sensed that we were not alone in the need for a no-nonsense, practical resource in research. So, mustering what assets we (plus SGD, and MGI) had, we self-published *The Gene Ontology: A Compilation of Reliable Terms with a Casual Textual Definitions*. Out of these unlikely circumstances was born the most authoritative ontology in biology, the ontology your lab mate and mother probably learned to annotate from. To date it has annotated more than 200 thousand gene products. We, at the new National Center for Biomedical Ontology are updating the field with a series of All About ontology tutorials. With notes and techniques scattered throughout, along with helpful illustrations, we will continue our tradition of offering ontologies that take a reasonable amount of time to prepare but result in flexible, accurate, researcher-friendly ontologies that are beautiful enough to provide to anyone.

Udo Hahn, Jena University

Language & Information Engineering (JULIE) Lab

What Do We Share? Conceptualizations of Biological Knowledge from the Viewpoints of Biology and Natural Language Processing

In this talk, two breeds of conceptual resources (ontologies) for biological knowledge will be compared. Those developed by biologists for the purpose of data curation, and those developed by NLP researchers for the purpose of information extraction and text mining. The gaps and misconceptions we encounter will give rise to proposals how we might improve interoperability among the products from both camps on a large(r) scale.

Agenda

START	END	TITLE	AUTHORS
9:00	9:10	Introduction	
9:10	9:30	<i>Can Literature Analysis Reveal Similarities among Cellular Processes?</i>	M. Chagoyen, P. Carmona-Saez, C. Gil, J.M. Carazo, A. Pascual-Montano
9:30	9:50	<i>The Design of a Wiki-based Curation System for the Ontology of Functions</i>	R. Hoehndorf, K. Prüfer, J. Kelso
9:50	10:10	<i>Distributed Representations of Bio-Ontologies for Semantic Web Services</i>	C.A. Joslyn, D.D.G. Gessler, S.E. Schmidt, K.M. Verspoor
10:10	10:30	<i>SpindleViz: A Three Dimensional, Order Theoretical Visualization Environment for the Gene Ontology</i>	C.A. Joslyn, S.M. Mniszewski, S.A. Smith, P.M. Weber

10:30-11:00 Coffee Break

11:00	12:00	Invited Speaker	Suzanna Lewis
-------	-------	-----------------	---------------

12:00-13:00 Lunch

13:00	13:20	<i>Exploring the Construction and Applications of a Protein Description Corpus</i>	M. Krallinger, R. Malik, A. Valencia
13:20	13:40	<i>Improving Biomedical Corpus Annotation Guidelines</i>	Z. Lu, M. Badal, P.V. Ogren, K.B. Cohen, L. Hunter
13:40	14:00	<i>leXML: Towards a Framework for Interoperability of Text Processing Modules to Improve Annotation of Semantic Types in Biomedical Text</i>	D. Rebholz-Schuhmann, H. Kirsch, G. Nenadic
14:00	14:20	<i>Automatically Adapting an NLP Core Engine to the Biology Domain</i>	E. Buyko, J. Wermter, M. Poprat, U. Hahn
14:20	14:30	10 Minutes Break	
14:30	15:30	Invited Speaker	Udo Hahn

15:30-16:00 Coffee Break

16:00	16:20	<i>Improving Biomedical Text Categorisation with NLP</i>	M. Matthews
16:20	16:40	<i>Use of Text Mining for Protein Structure Prediction and Functional Annotation in Lack of Sequence Homology</i>	A. Rechtsteiner, J. Luinstra, C.E..M. Strauss
16:40	16:50	Short Poster Advertisements	
16:50	17:50	Discussion	
17:50	18:30	Poster Session	

Short Papers

Included in this Volume

<i>Automatically Adapting an NLP Core Engine to the Biology Domain</i> E. Buyko, J. Wermter, M. Poprat, U. Hahn	65
<i>Can Literature Analysis Reveal Similarities among Cellular Processes?</i> M. Chagoyen, P. Carmona-Saez, C. Gil, J.M. Carazo, A. Pascual-Montano	69
<i>The Design of a Wiki-based Curation System for the Ontology of Functions</i> R. Hoehndorf, K. Prüfer, J. Kelso	73
<i>Distributed Representations of Bio-Ontologies for Semantic Web Services</i> C.A. Joslyn, D.D.G. Gessler, S.E. Schmidt, K.M. Verspoor	77
<i>SpindleViz: A Three Dimensional, Order Theoretical Visualization Environment for the Gene Ontology</i> C.A. Joslyn, S.M. Mniszewski, S.A. Smith, P.M. Weber	81
<i>Exploring the Construction of Resources for Detecting Protein Descriptions from the Literature</i> M. Krallinger, R. Malik, A. Valencia	85
<i>Improving Biomedical Corpus Annotation Guidelines</i> Z. Lu, M. Badal, P.V. Ogren, K.B. Cohen, L. Hunter	89
<i>Improving Biomedical Text Categorisation with NLP</i> M. Matthews	93
<i>IeXML: Towards an Annotation Framework for Biomedical Semantic Types Enabling Interoperability of Text Processing Modules</i> D. Rebholz-Schuhmann, H. Kirsch, G. Nenadic	97
<i>Use of Text Mining for Protein Structure Prediction and Functional Annotation in Lack of Sequence Homology</i> A. Rechtsteiner, J. Luinstra, C.E..M. Strauss	101

Poster Abstracts

DigraBase: A Graph-theoretic Framework for Semantic Integration of Biological Data

D. Jacobson, dan@nbn.ac.za

National Bioinformatics Network, South Africa

A method for integrating large volumes of biological data for efficient retrieval and searching is investigated. We present a model for the semantic representation of data in a distributed environment.

DigraBase implements a graph-theoretic model for capturing dyadic relations between schematically distinct databases. An abstract typing system is implemented for representation of concepts through use of typed logic. Methods for adapting external data to the framework schemata are described, thereby allowing capture of conceptual information independently of extrinsic data sources. The prototypical set-theoretic model is explained as well as the envisaged framework extensions for addition of an automated reasoning engine.

DigraBase is a generic data integration engine which enables the semantic description of biological objects for the purpose of representing complex and dynamic biological systems. This method of data integration will allow better modelling of biological systems.

Announcing the Second BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) Challenge Evaluation, 2006-2007

A. Morgan, alexmo@stanford.edu

MITRE Corporation

The BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) challenge evaluation is a community-wide effort for evaluating text mining and information extraction systems applied to the biological domain. BioCreAtIvE arose out the needs of working biologists, biological curators and bioinformaticians to access the wealth of information in the literature, and to link this information to biological databases and ontologies. BioCreAtIvE focuses on the comparison of methods and community assessment of scientific progress, rather than on the purely competitive aspects. The first BioCreAtIvE challenge evaluation in 2003-2004 attracted broad interest within the bioinformatics and biomedical text mining community, with participation from 27 groups from 10 countries. BioCreAtIvE is organized through collaborations between text mining groups, biological database curators and bioinformatics researchers.

BioCreAtIvE II will be held during October of 2006, with the workshop to be held in Spring 2007. It will consist of three tracks. The first will focus on finding the mentions of genes and proteins in sentences drawn from MEDLINE abstracts and is the same as Task 1A (Tanabe, Xie et al. 2005) from BioCreAtIvE I. The second track will involve producing a list of the EntrezGene identifiers for all the human genes/proteins mentioned in a collection of MEDLINE abstracts and is similar to BioCreAtIvE I Task 1B (Hirschman, Colosimo et al. 2005). The third track of BioCreAtIvE II is new and will involve identifying protein-protein interactions from full text papers, including extraction of excerpts from those papers that describe experimentally derived interactions, for curation into one of two interaction databases: IntAct (Hermjakob, Montecchi-Palazzi et al. 2004) and MINT (Zanzoni, Montecchi-Palazzi et al. 2002). Detailed descriptions of the tasks and a preliminary schedule can be read on the website: <http://biocreative.sourceforge.net>.

The approximate time line for BioCreAtIvE is as follows:

June 2006 Release of training data; Oct 2006 Release and evaluation of test data; Dec 2006 Results returned to participants; Feb 2007 Workshop papers submitted; Mar 2007 Workshop; Jun 2007 Journal articles submitted.

FUNC: Detecting Significant Groups of Annotations in an Ontology

K. Pruefer, pruefer@eva.mpg.de

Max-Planck-Institute for Evolutionary Anthropology, Leipzig, Germany

Genome-wide expression and association studies typically yield large sets of gene candidates which require further analysis and interpretation. By taking advantage of the annotation of genes to various ontologies it is possible to identify relationships between the gene sets identified by experimental methods, and to use this knowledge in the interpretation of results. In recent years a number of methods that use the gene ontology to find over-represented functional categories of genes in large scale experiments have been introduced.

We present here FUNC, a comprehensive and flexible tool which reproduces and expands on the functionality provided by existing approaches. Four different statistical tests have been implemented in FUNC, enabling the user to analyze different types of annotation related data. In addition to correction for multiple testing by comparison to randomized gene data, FUNC provides a global test statistic. The global test statistic determines whether the complete distribution of functional annotations differs from a random distribution, and in this way allows determining an overall significance for the data set before selection of a FDR. Additionally, FUNC is not restricted to use with a particular ontology, and is able to accept any ontology which is represented as a tree or DAG as input. The tree or DAG structure itself is also used for a graph based refinement which extracts the categories with the most precise description of the analysed data from the result set of significant groups. FUNC is implemented as a command line tool in C++, is open source, and can be compiled for various platforms. A web-interface for analyses using GO and eVoc is available at <http://func.eva.mpg.de>.

Integrating Image Data into Biomedical Text Categorization

H. Shatkay, N. Chen, D. Blostein, shatkay@cs.queensu.ca

School of Computing, Queen's University, Kingston, Ontario, Canada

Categorization of biomedical articles is a central task for supporting various curation efforts. It can also form the basis for effective biomedical text mining. Automatic text classification in the biomedical domain is thus an active research area. Contests organized by the KDD Cup (2002) and the TREC Genomics track (since 2003) defined several annotation tasks that involved document classification, and provided training and test data sets. So far, these efforts focused on analyzing only the text content of documents.

However, as was noted in the KDD'02 text mining contest – where figure-captions proved to be an invaluable feature for identifying documents of interest – images often provide curators with critical information. We examine the possibility of using information derived directly from image data, and of integrating it with text-based classification, for biomedical document categorization.

We present a method for obtaining features from images and for using them – both alone and in combination with text – to perform the triage task introduced in the TREC Genomics track 2004. The task was to determine which documents are relevant to a given annotation task performed by the Mouse Genome Database curators. We show preliminary results, demonstrating that the method has a strong potential to enhance and complement traditional text-based categorization methods.

Status of the Functional Genomics Investigation Ontology (FuGO)

P. L. Whetzel (on behalf of the FuGo working group), fugo-devel@lists.sourceforge.net

University of Pennsylvania

The application of high-throughput functional genomics technologies such as microarrays and mass spectrometry to a wide variety of biological conditions has facilitated the investigation of new kinds of biological questions and has also resulted in the generation of massive amounts of data and metadata. The Functional Genomics Investigation Ontology (FuGO) is being developed as a collaborative, international effort to build an ontology that will provide a common resource of terms to annotate functional genomics investigations as well as assays that generate large amounts of data and metadata (Whetzel et al., The Development of FuGO – An Ontology for Functional Genomics Investigations, OMICS (in press)). Those involved in the FuGO project represent biological and technological domains and currently include representatives from the following communities: Metabolomics, Proteomics,

Transcriptomics, Crop Sciences, Environmental Genomics, Flow Cytometry, Genome Sequence, Immunogenomics, Nutrigenomics, Polymorphism, Toxicogenomics. Although these are diverse communities, there is a shared need for terms to describe universal features of investigations such as the design, protocols, assays, material, and units of measurement. The scope of FuGO will cover both terms that are universal to all functional genomics investigations and those that are domain specific. In addition, FuGO will reference out to existing mature ontologies for additional domain content, e.g. Foundational Model of Anatomy, to avoid the need to duplicate these resources, and will do so in such a way as to foster their ease of use in annotation.

The initial ontology building process includes the collection of use cases from communities, the identification of terms of interest for annotation. From these use cases, and the binning of terms into top-level containers such as Object and Process. In the process of building FuGO, a style guide is being developed to provide recommendations for naming conventions for terms, formats for definitions and policies for the development of the ontology such as those for the addition of synonyms and numeric identifiers for terms. In addition to these style recommendations, FuGO is being developed as an OBO Foundry Application ontology and will meet OBO requirements (<http://obofoundry.org>).

During our first FuGO workshop in February, a draft version of FuGO was developed using Protégé/OWL. This initial version of the ontology focuses on the need for universal terms from the involved communities and includes an extension for technology specific terms for the Proteomics community. Thus far, the ontology building process has focused on the development of *is_a* relations, while noting where other relations, such as *part_of*, are needed. Future development of FuGO will proceed as an iterative process in which a section of the ontology will be reviewed by all interested communities allowing terms, both universal and community specific, to be added to the ontology. In addition, *non-is_a* relations will also be added to join the branches of the ontology. In conclusion, FuGO is being developed to provide a common resource for annotation of functional genomics investigations and in this way, the ontology will serve as the ‘semantic glue’ to provide a common understanding of the annotated data from across disparate data sources.

A Question Answering System in Biomedical Domain

Y. Yamamoto, yayamamo@hgc.jp

University of Tokyo

In the biomedical domain, as in many others, it is critical for researchers to obtain their most wanted knowledge from vast amount of information efficiently. Information Retrieval (IR) and Information Extraction (IE) have been developed in recent years in this domain, but IR often returns too many relevant results with relatively large number of irrelevant ones. IE is too specific for each task such as extracting protein-protein interaction or relationships between proteins and diseases. In this situation, there is an urgent need of a system to answer a variety of biomedical questions. We developed a question-answering (QA) system called Dr. Kurt that accepts a sentence as an information need of a user and returns answers as terms with the supporting evidences (i.e., links to literature). QA systems have been developed in general domains such as answering names of politicians or companies for more than 30 years, but no attempt has been made in this domain. The efforts in the general domains revealed that linguistic resources and tools prepared for a target domain such as thesauri / ontologies and named entity recognizers were important to get a good result. Recent vigorous studies in the biomedical domain have made such resources and tools in this domain available. Accordingly, we utilized them and examined the feasibility of applying a QA approach developed in the general domains to the biomedical domain.

Our system consists of 1) question analyses, 2) query generation and expansion followed by literature retrieval, and 3) extraction, scoring, and evaluation of candidate sentences for answers. The data source from which our system extracts to answer is MEDLINE, and we used UMLS, Disease Ontology, and GENE to categorize questions or identify domain specific terms and expand queries. A full parser called Enju that has been tuned for MEDLINE abstracts is used for question analyses and scoring candidate sentences for answers. To solve the issue of decreasing literature retrieval performance due to the query expansion, we developed a novel method of scoring term weights. In addition, to improve retrieval performance, we took an approach of expanding domain specific verbs. The literature retrieval engine we used is zettair, an open source program, and we modified the ranking method and the query processing component.

Our system answered 57 biomedical questions made by a biologist who did not participate in the system development. As a result, F-measures were 27% for 30 questions in categories that our system can accept currently and 13% for all the questions. The obtained result was comparable to those achieved in the general QA domains to date. Consequently, the feasibility of applying approaches developed in the general domains to the biomedical domain has been shown. Considering that specific IE tasks have been dominant, we believe that our attempt to develop a general or versatile IE system in this domain is significant.

Automatically Adapting an NLP Core Engine to the Biology Domain

Ekaterina Buyko^{*1}, Joachim Wermter¹, Michael Poprat¹ and Udo Hahn¹

¹Jena University Language and Information Engineering (JULIE) Lab,
Friedrich-Schiller Universität Jena, Fürstengraben 30, D-07743 Jena, Germany

Email: {buyko|wermter|poprat|hahn}@coling-uni-jena.de;

*Corresponding author

Abstract

Background: Rather than specifying rules, constraints and lexicons for NLP systems manually, we advocate a procedure for automatically acquiring linguistic knowledge using machine learning (ML) methods. In order to demonstrate how feasible this approach is, we automatically adapt OPENNLP, an open source ML-based NLP tool suite, to the sublanguage domain of biology.

Results: In the first evaluation ever of a ML-based ensemble of core NLP components in the biology domain, which are all based on the maximum entropy method, we demonstrate that the performance of OPENNLP's sentence splitter, tokenizer, part-of-speech tagger, chunker and parser on bio texts matches up with state-of-the-art performance figures from the newspaper domain.

Conclusions: Core BioNLP systems can automatically be derived from non-bio domain settings by (re-)training rather than by manual rule re-writing and lexical re-specification.

Background

The prevailing approach to build large-scale BioNLP systems relies heavily on human efforts and expertise, i.e., manual rule, constraint and lexicon specification and maintenance (e.g., [1,2]). It is often deplored that

such an approach is time-costly, error-prone and heavily iterative (lots of test/modify cycles). Within the NLP community proper, this approach is almost completely abandoned and, as an alternative, procedures for automatically acquiring linguistic knowledge, i.e., machine learning (ML) techniques, dominate the field (cf., e.g., various CoNLL activities¹). On the flip side of this success story, (semi-)supervised ML methods at least, require annotated corpora to be available as a training resource. Fortunately, in the molecular biology domain two richly annotated language resources (GENIA² and PENNBIOIE³) already exist. In the following, we stipulate that the exchange of domain (e.g., from newspapers to biological articles), merely boils down to re-training NLP tools on annotated corpora rather than laborious manual re-writing and changing of rule-based systems and lexicons. For this purpose, we chose OPENNLP TOOLS⁴ as an open-source tool suite which contains a variety of JAVA-based NLP components. Our focus is here on five “core” NLP components, *viz.* sentence detection, tokenization, POS tagging, chunking and parsing. OPENNLP is a homogeneous package based on a single machine learning approach, *viz.* maximum entropy (ME) [3,4]. The rationale behind ME for any collection of facts is to choose a model which is consistent with all the facts but otherwise as uniform as possible. Each OPENNLP tool requires an ME model that contains statistics about the component’s default features combining diverse contextual information such as words around end-of-sentence boundaries for the sentence splitter or word/tag combinations in five-word/tag-window for the chunker. The components are partly based on publications like the chunking model described by Sha & Pereira [5] as well as the part-of-speech tagger and the parser described by Ratnaparkhi [6]. For training and testing of all OPENNLP components we considered, the above-mentioned two major biomedically annotated corpora, *viz.* GENIA and PENNBIOIE, were employed which currently contain the most elaborate annotations relevant for syntactic analysis in the bio field. We performed ten-fold cross validation for all OPENNLP tools on both corpora.

Results

The *sentence splitter* achieved an accuracy of approximately 99.0% on GENIA and 97.4% on PENNBIOIE (see also Table 1).⁵ False positives [7] which constitute the majority of errors (67% in GENIA) usually do not occur in biological terms but rather occur in abbreviations of names in literature citations. In PENNBIOIE false negatives are mainly due to section headings ending with “:” such as “*OBJECTIVE:*”.

¹<http://ilps.science.uva.nl/~erikt/signll/conll>

²<http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/>

³<http://bioie ldc.upenn.edu>

⁴<http://opennlp.sourceforge.net>

⁵By comparison, the accuracy of the sentence splitter on newspaper data in [6] is 98.9%.

	Sentence Splitter (Accuracy)	Tokenizer (Accuracy)	PoS Tagger (Accuracy)	Chunker (F-Score)	Parser (F-Score)
GENIA	99,0%	99,6%	98,9%	93,6%	87,0%
PENNBIOIE	97,4%	99,0%	98,9%	89,5%	85,2%

Table 1: Evaluation Results of OPENNLP Components

The *tokenizer* achieved an accuracy of 99.6% on GENIA and 99.0% on PENNBIOIE (Table 1). On both corpora the tokenization of commas and parentheses also occurring in a lot of chemical names does not seem to be problematic. The lower accuracy of the tokenizer with regard to “-” and “/” in PENNBIOIE, however, reveals the fact that in PENNBIOIE (though not in GENIA) expressions containing these symbols are meticulously token-annotated and thus raise the bar for the tokenizer.

The *part-of-speech tagger* achieved an accuracy of 98.9% on GENIA and 98.9% on PENNBIOIE (Table 1).⁶ For these runs, we also used the tagger option to include an automatically generated tag-dictionary from the training data enumerating a fixed set of possible tags for each word to delimit the decision space. Without the dictionaries, the tagging accuracy drops to 98.2% on GENIA and 98.3% on PENNBIOIE. With the exception of “*binding*” the tagging mistakes in the top 10 are not biomedical terms. The erroneous taggings seem to replicate the tagging mistakes discussed already in [6].⁷

The *chunker* achieved an overall F-score of 93.6% on GENIA and 89.5% on PENNBIOIE (Table 1). At a first glance, it seems that the chunker performs markedly better when trained on GENIA, in particular, concerning the important recognition rate for NPs (92.3% on GENIA *vs.* 85.1% on PENNBIOIE). However, these results must be treated with caution because of some inadequacies of the CHUNKLINK script⁸ in converting treebank annotations into IOB-notation, especially when PENNBIOIE is concerned. The overall F-score of 93.6% on GENIA is a state-of-the-art figure (in comparison with performance figures from CoNLL 2000) and even better than when trained on newspaper data.⁹ In particular, the good recognition rates for NPs, PPs (96.9% in GENIA) and VPs (95.9% GENIA) are essential for deeper linguistic analysis. We evaluated the *parser* along standard parameters such as *Bracketing recall* and *Bracketing precision* [8,9] on all sentences and on those with a length less than 40 words (about 90% of the sentences in GENIA and 88,6% in PENNBIOIE).¹⁰ The parser F-score on GENIA 87.0% is about two percentage points

⁶The ME-based POS tagger in [6] achieves 96.6% accuracy on newspaper data.

⁷For example, the word “*that*” is confused being used as a relative pronoun (WDT), a subordinating conjunction (IN) or a determiner (DT); similarly, the word “*both*” by its use as a determiner (DT) and a conjunction (CC).

⁸<http://ilk.kub.nl/~sabine/chunklink/>

⁹The OPENNLP chunker obtained an overall F-score of 92.4 on the CoNLL 2000 data (own evaluation).

¹⁰These proportions are similar to those in the PENN TREEBANK [9].

better than on PENNBIOIE 85.2% (Table 1). It should also be noted that the number of parsing failures is quite low. For GENIA, only one out of 4370 sentences could not be parsed (0.02%), and for PENNBIOIE, this number amounts to eight out of 6296 (0.1%).¹¹ Particularly interesting from the perspective of deeper linguistic analysis of biological literature is the parser's capacity to predict semantic function tags of constituents. For these, the parser's F-scores only decrease about one point on each corpus.

Conclusion

There are two major results of this study. We have the first evaluation of an ensemble of NLP components all of which are based on the same ML paradigm (maximum entropy) within the biology domain. Second, our results from the biology domain match up to state-of-the-art performance figures from the newspaper domain. As OPENNLP TOOLS stand for an ML-based approach to acquire linguistic regularities (which is quite unusual in the biology domain, up until now) the state-of-the-art performance figures we determined provide ample evidence for challenging common wisdom. We might no longer create and maintain rule systems and lexicons manually but simply (re-)train NLP components on properly annotated data. Our current efforts are directed at exploring an ML paradigm which helps minimize the amount of annotated data required for (still) effective training of NLP tools.

Acknowledgments. This research was funded by the EC's 6th Framework Programme (4th call) and conducted within the BOOTStrep consortium under grant FP6-028099.

References

1. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A: **Genies: A natural language processing system for the extraction of molecular pathways from journal articles.** *Bioinformatics* 2001, **17**(Suppl. 1):74–82.
2. Gaizauskas R, Demetriou G, Artymiuk P, Willett P: **Protein structures and information extraction from biological texts: The Pasta system.** *Bioinformatics* 2003, **19**:135–143.
3. Berger A, Della Pietra S, Della Pietra V: **A maximum entropy approach to natural language processing.** *Computational Linguistics* 1996, **22**:39–71.
4. Rosenfeld R: **A maximum entropy approach to adaptive statistical language modeling.** *Computer, Speech and Language* 1996, **10**:187–228.
5. Sha F, Pereira F: **Shallow Parsing with Conditional Random Fields.** In *Proceedings of the HLT-NAACL 2003* 2003.
6. Ratnaparkhi A: **Maximum Entropy Models for Natural Language Ambiguity Resolution.** *PhD thesis*, University of Pennsylvania 1998.
7. Palmer DD, Hearst MA: **Adaptive Multilingual Sentence Boundary Disambiguation.** *Computational Linguistics* 1997, **23**(2):241–267.
8. Collins M: **Three generative, lexicalised models for statistical parsing.** In *Proceedings of the ACL'97/EACL'97* 1997:16–23.
9. Clegg AB, Shepherd AJ: **Evaluating and integrating Treebank parsers on a biomedical corpus.** In *Proceedings of the ACL 2005 Workshop on Software* 2005.

¹¹A flat chunk phrase structure is returned in such a case.

Can literature analysis reveal similarities among cellular processes?

Monica Chagoyen^{1,2§}, Pedro Carmona-Saez¹, Concha Gil^{3,4}, Jose M. Carazo¹, Alberto Pascual-Montano²

¹Biocomputing Unit. Centro Nacional de Biotecnología – CSIC, Madrid, Spain. ²Dpto. Arquitectura de Computadores y Automática. Universidad Complutense de Madrid, Madrid, Spain. ³Dpto. Microbiología II. Facultad de Farmacia. Universidad Complutense de Madrid, Madrid, Spain. ⁴Unidad de Proteómica UCM - Parque Científico de Madrid, Madrid, Spain

§Corresponding author; Email address: monica.chagoyen@cnb.uam.es

Abstract

Background: Methods to narrow the gap between current definitions of cellular processes in functional schemes and the complex structure of molecular networks are required to understand biological systems. To this aim we explore the use of the scientific literature to compute a similarity metric among cellular processes in the context of an organism.

Results: We have established a similarity score between cellular processes in *Saccharomyces cerevisiae*. This score was obtained using latent semantic indexing of documents relevant to GO biological process annotations. We compare the results obtained by this literature-based metric with two ontology-based measures as well as with gene product co-annotation relationships.

Conclusions: The biomedical literature is an important source of information to obtain potential relationships between cellular processes. The associations discovered by literature analysis are a valuable complement to those encoded in existing functional schemes, and to those that arise by genome co-occurrence annotation.

Background

Recent research in molecular biology emphasizes the importance of understanding the interlinked nature of biological networks within the cell. In this sense, the focus of attention is being moved from the assignment of functions to individual proteins, to the global analysis of cellular processes. Several current developments in the bioinformatics area are aiming at this goal, analyzing different sets of experimental data and functional information.

Currently there are a number of functional schemes that provide a rich description of processes relevant for the biology of the cell. A good example of this is the Biological Process ontology within the Gene Ontology (GO) project [1]. In addition, GO term relationships (in its three aspects) have been previously studied by the analysis of experimental data, e.g. gene expression data [2], as well as genome annotation [3] and linguistic content [4].

In the context of text analysis, most of the research has been devoted to the development of new methods to compare functional information at the gene and protein level. Among them, a number of methods use different document similarity measurements to establish potential gene relationships and to perform functional classification [5-9]. However, little work has been done on the analysis and comparison of the biological processes themselves. In this work we use the scientific literature to establish relationships among cellular processes as described by the GO. To this end, we define a similarity score between biological processes using Latent Semantic Analysis [10] of relevant documents. To verify that our proposal is valid, we create a pair-wise similarity matrix of the GO biological processes annotated for the *Saccharomyces cerevisiae* genome. We compare our results with those obtained using two previously reported ontology-based measurements [11, 12], and a metric based on gene product co-annotation [3].

Methods

A literature set was obtained for each GO biological process term (GOp) annotated for *S. cerevisiae*, as provided by SGD (www.yeastgenome.org). This set comprises all the references included as evidence for a particular GOp, independently of the genes and the hierarchical relationships. For each GOp, a meta-document was created by concatenating the titles and abstracts of the corresponding references. Subsequently, a vector space representation of GOp-terms (**A**) is obtained similarly as in [5]. Terms were filtered out if they did not appear in at least 2 of the GOp documents.

Once the biological processes are represented in the vector space as a sparse matrix (**A**), we applied a SVD factorization ($\mathbf{A}=\mathbf{USV}^t$) to find a low-rank approximation to **A**. The number of factors selected ($k=200$) was chosen using the scree test [13]. A similarity metric is then obtained for all pairs of biological processes using the cosine correlation between each pair of rows in **VS**.

Similarities among a subset of biological process pairs (those containing at least 5 references) were compared to those obtained by other metrics. Two ontology-based metrics were computed as: semantic similarity established by means of the information content-based metric defined by Lin, 1998 [14] and proposed by Lord *et al.*, 2003 [11]; and Czekanowski-Dice similarity, used in the GOTOolBox as described by Martin *et al.*, 2004 [12]. Even though both measurements were originally proposed to relate genes/proteins by their functional information, in this work we applied the same metric for the comparison of GO biological process terms.

In addition, another measure of similarity based on gene product annotation was also used for comparison purposes. This measure was computed as a modified version of [3], where each GO process is described by a vector of gene products corresponding to the associations provided by the SGD. Similarity is computed as the cosine of the angle between process vectors.

Results

We postulate that similarities among cellular process of a given organism (*S. cerevisiae* in our experiments) can be calculated as similarities among the set of bibliographic references related to them. To ensure the relevance of the documents to analyze, we selected those articles provided as evidence in the GO annotation provided by the *Saccharomyces* Genome Database (SGD).

In the absence of a gold standard to validate similarities among biological processes, we compared our results with two sources of information: ontological relationships and gene product co-annotation. To this aim, only references directly associated to a given GO term were considered relevant for that particular process. This ensures that conserved hierarchical relationships are genuinely discovered from document similarity.

	Pearson		Spearman's rho		Kendall's tau		Uncentered	
	All	Subsume	All	Subsume	All	Subsume	All	Subsume
Slit-Slin	0.3516	0.4456	0.2143	0.5660	0.1483	0.3979	0.6321	0.8114
Slit- Czdice	0.2582	0.4401	0.1715	0.5031	0.1161	0.3579	0.7008	0.8103
Slit- Sann	0.5336	0.5794	0.2310	0.6158	0.1659	0.4365	0.5820	0.8461
Sann-Slin	0.4519	0.8461	0.1058	0.9174	0.0765	0.7602	0.5276	0.9625
Sann-Scd	0.2889	0.5860	0.1149	0.5995	0.0822	0.4267	0.3875	0.9194
Slin-Scd	0.5793	0.6631	0.6137	0.5991	0.4478	0.4257	0.7577	0.9595

Table 1: Correlation among the four similarity metrics, calculated for all process pairs (All), and for process pairs with inclusion relationship (Subsume). Correlation coefficients are calculated as: Pearson, Spearman, Kendall and uncentered dot product. Slit (literature-based similarity); Slin (Lin similarity); Scd (Czekanowski-Dice similarity); Sann (Annotation based similarity).

Table 1 contains the correlation coefficients calculated among the four similarity metrics. Analysis of scatter plots revealed that there is no linear relationship among the literature metric with any of the metrics used for comparison (therefore Pearson's correlation coefficient should be carefully interpreted). According to our results, the similarity obtained by document comparison is slightly more related to gene product co-association than to metrics based on ontological relationships. Nevertheless, the correlation with ontology-based metrics increases significantly when only inclusion relationships are analysed (i.e. those process pairs in which one process contains all the genes associated with the other). This is an indicator that our literature analysis conserves a large number of subsumption relationships encoded in GO, although it also produces complementary results.

In order to discover the nature of these complementary relationships we analyzed those biological process pairs for which we obtained most contradictory values. For the sake of space only the three top illustrative examples are shown in table 2.

Slit	Scd	Slin	CG	Biological Process A	Biological Process B
0.94	0.23	0.03	7	phosphoinositide dephosphorylation (7)	inositol lipid-mediated signalling (13)
0.94	0.37	0.04	0	high affinity iron ion transport (5)	iron ion homeostasis (29)
0.06	1	0.82	0	Rho protein signal transduction (17)	Ras protein signal transduction (17)

Table 2: Sample illustration of the biological process pairs analyzed which reveal the complementary nature of the similarities obtained from the literature analysis and those obtained from ontology-based metrics. Slit (literature-based similarity); Scd (Czekanowski-Dice similarity); Slin (Lin similarity); CG (number of genes co-annotated). Total number of genes annotated with each biological process term is shown in brackets.

The first two correspond to biological process pairs that have been found to be highly similar using our method, while showing low similarity scores according to both ontology-based methods (meaning that they are not highly related in the GO). The similarity found between 'phosphoinositide dephosphorylation' and 'inositol lipid-mediated signalling' is justified by the fact that all *S. cerevisiae* gene products annotated as being involved in the first (7 in total), are also annotated with the second. Therefore, this similarity could also be found by analysis of GO term co-annotations. On the other hand, no gene is already described to be involved, as far as SGD GO annotation, in the second biological process pair 'high affinity iron ion transport' and 'iron ion homeostasis'. Nevertheless GO defines metal ion homeostasis processes as the regulation of the levels, transport, and metabolism of metal ions within a cell or between a cell and its external environment. Therefore, a certain level of relationship between homeostasis and transport mechanisms is expected by definition.

The last pair, 'Rho protein signal transduction' and 'Ras protein signal transduction' is highly similar according to both ontology-based metrics. The reason is that both are direct descendants of 'small GTPase mediated signal transduction' category. Hence, the similarity between processes can be explained by the fact that both signalling cascades are mediated by proteins belonging to the same family. However, our literature analysis provided a very low similarity score to these two signalling events. On the other hand, the most similar process to 'Rho protein signal transduction' according to the literature is 'maintenance of cell polarity (sensu Fungi)', which reveals a clear relationship among the two processes in the context of *S. cerevisiae*.

Discussion and conclusions

The biomedical literature is a valuable source of information from which to obtain potential relationships between cellular processes. Processes described by a representative set of references can be compared and related, even if they were encoded in different functional schemes. The relationships computed from document similarity show a good agreement with those computed from semantic similarity in the case of closely related processes in the GO. In addition, we were able to find a number of similarities that are not revealed by metrics computed from the GO structure or by genome co-annotation.

Therefore, the relationships discovered by literature analysis are a helpful complement to those encoded in existing functional schemes and those that arise by gene product co-annotation. Our results also indicate that a full exploitation of the complementary nature of currently available similarity metrics among biological processes might provide new biological insights and therefore constitutes an interesting line for further research.

Acknowledgements

This work has been partially funded by the Spanish grants CICYT BFU2004-00217/BMC, GEN2003-20235-c05-05, TIN2005-5619, PR27/05-13964-BSCH and a collaborative grant between the Spanish CSIC and the Canadian NRC (CSIC-050402040003). PCS is recipient of a grant from CAM. APM acknowledges the support of the Spanish Ramón y Cajal program.

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.
2. Yu T, Sun W, Yuan S, Li KC: **Study of coordinative gene expression at the biological process level.** *Bioinformatics* 2005, **21**:3651-3657.
3. Bodenreider O, Aubry M, Burgun A: **Non-lexical approaches to identifying associative relations in the gene ontology.** *Pac Symp Biocomput* 2005:91-102.
4. Ogren PV, Cohen KB, Acquaaah-Mensah GK, Eberlein J, Hunter L: **The compositional structure of Gene Ontology terms.** *Pac Symp Biocomput* 2004:214-225.
5. Chagoyen M, Carmona-Saez P, Shatkay H, Carazo JM, Pascual-Montano A: **Discovering semantic features in the literature: a foundation for building functional associations.** *BMC Bioinformatics* 2006, **7**:41.
6. Chaussabel D, Sher A: **Mining microarray expression data by literature profiling.** *Genome Biol* 2002, **3**:RESEARCH0055.
7. Glenisson P, Antal P, Mathys J, Moreau Y, De Moor B: **Evaluation of the vector space representation in text-based gene clustering.** *Pac Symp Biocomput* 2003:391-402.
8. Homayouni R, Heinrich K, Wei L, Berry MW: **Gene clustering by latent semantic indexing of MEDLINE abstracts.** *Bioinformatics* 2005, **21**:104-115.
9. Shatkay H, Edwards S, Wilbur WJ, Boguski M: **Genes, themes and microarrays: using information retrieval for large-scale gene analysis.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:317-328.
10. Deerwester S, Dumais S, Landauer T, Furnas G, Beck L: **Improving Information-Retrieval with Latent Semantic Indexing.** *P Asis Annu Meet* 1988, **25**:36-40.
11. Lord PW, Stevens RD, Brass A, Goble CA: **Semantic similarity measures as tools for exploring the gene ontology.** *Pac Symp Biocomput* 2003:601-612.
12. Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B: **GOToolBox: functional analysis of gene datasets based on Gene Ontology.** *Genome Biol* 2004, **5**:R101.
13. Cattell RB: **Scree Test for Number of Factors.** *Multivar Behav Res* 1966, **1**:245-276.
14. Lin D: **An Information-Theoretic Definition of Similarity.** In *Fifteenth International Conference on Machine Learning; July 24-27; Madison, Wisconsin, USA*. Morgan Kaufmann Publishers Inc.; 1998: 296-304.

The design of a wiki-based curation system for the Ontology of Functions

Robert Hoehndorf^{1,2}, Kay Prüfer², Michael Backhaus², Johann Visagie² and Janet Kelso^{*2}

¹Research Group Ontologies in Medicine (Onto-Med), Institute of Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Härtelstraße 16–18, 04107 Leipzig, Germany

²Department of Evolutionary Genetics, Max-Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany

Email: Robert Hoehndorf - hoehndorf@eva.mpg.de; Kay Prüfer - pruefer@eva.mpg.de; Janet Kelso* - kelso@eva.mpg.de;

*Corresponding author

Abstract

Recently, studies [1] argued that statistical and linguistic methods can be applied to extract information from biomedical ontologies, and represent the identified relations in the top-level ontology Ontology of Functions (OF) [2]. However, human intervention is required in order to clear noise from the generated data. A simple platform for collaborative work is needed. We propose here the use of a semantic wiki to represent relations between terms. We provide a relationship model for this semantic wiki, and add a core ontology as top-level type system to this wiki. We then discuss a design for the implementation of a platform for the curation of the OF, thereby enabling the community to curate the results of automatic extraction methods, and to add and correct ontology and annotation information. The aim of this collaborative effort is to provide a means to extend and correct the numerous ontologies which are used to describe biological functions in the OF.

1 Background

In recent years biologists have accumulated more and more knowledge about the functions of cells, cell components and gene products. Many databases exist that summarize this information, often based on the hard work of curators. Their dedicated work and expertise helps in ordering and formalizing the available knowledge about functions in

biomedicine. Ontologies play an increasingly important role in the process of linking these different biomedical databases together, not only by providing a common vocabulary, but by structuring the knowledge in a way that permits large scale analysis of the data.

One of the main design issues with many biomedical ontologies, like those provided by OBO (see <http://obo.sf.net>), is the simplicity of their structure. In order to formalize more knowledge than possible with this structure, we propose in [2] to add a top-level ontology to the OBO ontologies, in order to formalize more knowledge than currently possible.

The Ontology of Functions (OF) is based on the assumption that functions can be specified using *requirements* and a *goal*. An entity, e.g. a gene product, plays the *role* [3] specified in the function – it *has the function*. A biological process is, in turn, the *realization* of the described function. This formalization provides all necessary means for describing biological functions in more detail than is possible by using the current OBO ontologies. Such a formalization, however, heavily relies on the vocabulary provided by these ontologies in order to express this description.

In order to make OF useful for the community it must provide knowledge of real biological functions. Large scale analyses as well as individual data mining require the modeled knowledge to be of high quality. However, given the currently available amount of data concerning functions and processes, it seems impossible to achieve this goal in a reasonable time through manual curation performed by a few people.

Fortunately, there have been several publications such as [1] regarding the application of techniques to extract links between different ontologies using linguistic and statistical methods. These methods can complete formalized knowledge, and can also serve as a basis for populating OF. However, the result will contain errors and will be restricted to links between the terms used by existing ontologies. These approaches cannot replace the work of curators, but merely provide a basis from which an experienced biologist can begin to add more information to the OF.

We present a wiki-based solution which involves the user in the process of reviewing the results of automatic extraction methods, and which enables the user to add to or edit the ontology. The success of community-based ontology curation has already been shown in [4]. The use of a wiki provides users with an easy-to-use interface and makes true community editing possible.

2 Method

Wikipedia defines a wiki as “a type of website that allows anyone visiting the site to add, to remove, or otherwise to edit all content, very quickly and easily, sometimes without the need for registration.” Today, thousands of wikis are in use, and they have become a common way to enable collaborative work on a subject, while achieving an

exceptionally high quality [5].

Tools for the development and curation of ontologies in a collaborative environment exist (i.e. multi-user protege). However, wikis give more freedom in respect to the natural language descriptions of terms and are therefore better suited for a dynamic, open community. The main problem, however, is the representation of the structure of ontologies in the wiki. Fortunately, possible solutions are already being developed, predominantly as wikis that use Semantic Web technologies [6]. These *semantic wikis* use a formal model for representing the content of individual pages. They treat wiki pages as concepts, and hyperlinks as relations between these concepts. Most semantic wikis use the Semantic Web formalisms OWL and RDF as underlying representation languages [7].

We find OWL and RDF unsuitable for our application. The reasons are, besides their bloated syntax, that we need to model n -ary relations, which is made difficult by OWL and RDF (using reification), and leads to ontological errors: reified RDF-statements should be considered wiki-pages, while genuine n -ary relations should not. We propose here another relationship model, based on [3, 8]. Relations have a name, a number of argument roles given by a role name, a type restriction for each argument role, and a minimality condition. For example, the *hasFunction* relation is given as $\langle hasFunction | (function, T_1), (object, T_2), (context, T_3) \rangle$ with the condition that the function and object roles must be filled¹ T_i are types. For example, fillers of the *function* role will be restricted to the type *Biological function*. We need some form of type system as a top-level reference for this type construction. A biomedical core ontology defines the most general types of the biomedical domain, such as *Biological process* or *Biological Object*, and relations between them. There are a few core ontologies already available². We use GFO-Bio (<http://onto.eva.mpg.de/gfo-bio.html>), a core ontology based on the General Formal Ontology [9] (GFO). Wiki pages can be explicitly classified according to the type system provided by the core ontology. Furthermore, the use of wiki pages (concepts) as fillers for typed argument slots can be used to classify these wiki pages further and automatically, thus providing more structure and better error detection.

3 Results

We have begun the implementation of a prototype on which to test our proposal for OF. None of the semantic wikis of which we are aware supports n -ary relations or the integration of a core ontology. Therefore, we choose to adapt the Semantic MediaWiki [7] to our needs by (i) extending the syntax for semantic relations as well as the database to support n -ary relations, and (ii) including the types and relations of a biomedical core-ontology in the wiki, which provides a way to use typed relations and to set the type of wiki pages explicitly. The second extension will be used to perform type-checking (automated reasoning) whenever a relation is used.

¹Compare *hasFunction(transport O₂, red blood cell, circulating system)* and *hasFunction(accumulate O₂, red blood cell, respiratory system)*.

²For a good example see <http://www.cs.man.ac.uk/~rector/ontologies/simple-top-bio/>.

4 Conclusion

Our goal is to implement a curation framework for the Ontology of Functions (OF). We want to use linguistic and statistical methods of ontology or information extraction in order to generate a proto-ontology. This proto-ontology will then be made available in the formalism provided by our framework. Hoping that OF will provide value to the community involved in the curation and maintenance of biomedical ontologies, we encourage the potential user community to take an active role in the development of OF. A prototypical implementation is available via the website <http://onto.eva.mpg.de>.

The developed framework is flexible and can be used by other biomedical ontologies as well, in order to minimize curation effort and improve user involvement. Another possible application is for use as a counterpart to a gene function wiki [10]. This would be beneficial both for the annotation of gene functions and for biomedical ontologies, because it will lead to direct feedback and annotation-driven curation and error-discovery. Additionally, this wiki can be used to integrate several biomedical ontologies, by allowing its users to create links between them.

Our project will be successful, if it helps to realize the power that lies in community-based knowledge acquisition in a highly-specialized domain such as biology.

Although we have concentrated here on the representation of structure in the wiki, an immediate benefit is clearly the natural language descriptions that it provides. Hidden in the extension to other domains in biology and science in general is the power to collaboratively develop a structured, high quality knowledge resource which can provide value for scientists all over the world.

References

1. Bodenreider O, Aubry M, Burgun A: **Non-lexical approaches to identifying associative relations in the gene ontology**. *Pac Symp Biocomput* 2005, :91–102.
2. Burek P, Hoehndorf R, Loebe F, Visagie J, Herre H, Kelso J: **A top-level ontology of functions and its application in the Open Biomedical Ontologies**. In *Proceedings of the ISMB 2006* 2006.
3. Loebe F: **Abstract vs. Social Roles: A Refined Top-level Ontological Analysis**. In *2005 AAAI Fall Symposium Roles, an Interdisciplinary Perspective*, Menlo Park (California): AAAI Press 2005:93–100.
4. Good BM, Tranfield E, Tan P, Shehata M, Singhera G, Gosselink J, Okon E, Wilkinson M: **Fast, cheap and out of control: a zero curation model for ontology development**. In *Proceedings of the PSB 06* 2006.
5. Giles J: **Internet encyclopaedias go head to head**. *Nature* 2005, **438**:900–901.
6. Berners-Lee T, Hendler J, Lassila O: **The Semantic Web**. *Scientific American* 2001.
7. Völkel M, Krötzsch M, Vrandečić D, Haller H, Studer R: **Semantic Wikipedia**. In *Proceedings of the 15th international conference on World Wide Web* 2006.
8. Devlin K: *Logic and Information*. Cambridge University Press 1991.
9. Herre H, Heller B, Burek P, Hoehndorf R, Loebe F, Michalek H: **General Formal Ontology (GFO) – A Foundational Ontology Integrating Objects and Processes**. Onto-Med Report 8, University of Leipzig 2006.
10. Wang K: **Gene-function wiki would let biologists pool worldwide resources**. *Nature* 2006, **439**(7076):534.

Distributed Representations of Bio-Ontologies for Semantic Web Services

CA Joslyn^{*1}, DDG Gessler², SE Schmidt³ and KM Verspoor¹

¹Computer Science Division, Los Alamos National Laboratory, PO Box 1663, Los Alamos, NM, 87545, USA

²National Center for Genome Resources, Santa Fe, NM 87505 USA

³Technische Universität Dresden, Germany

Email: CA Joslyn - joslyn@lanl.gov; DDG Gessler - ddg@ncgr.org; SE Schmidt - midt1@msn.com; KM Verspoor - verspoor@lanl.gov;

*Corresponding author

Abstract

We introduce the Semantic Moby/VPIN distributed refactoring of the Gene Ontology optimized for semantic web architectures. We show through an order-theoretical analysis that this alternative representation preserves the original topology, and through Formal Concept Analysis that additional information about ambiguous inheritance may be revealed by the reconstruction.

1 Introduction

Bio-ontologies such as the Gene Ontology (GO) [5] are increasingly important as researchers seek to standardize vocabulary and integrate results into a common framework. They are also being used as the basis for critical bioinformatics tasks, including function prediction [9] and protein family classification [6]. These ontologies are often large (the GO currently contains on the order of 20,000 concepts), making them unamenable for semantic web architectures that need rapid access to only a few terms across many ontologies. Additionally, the common practice of building ontologies with deep, fixed subsumption assertions (e.g., successive, nested `owl:subClassOf` assertions) means that creating these ontologies is a low-throughput, labor intensive task, yielding third-party extensions vulnerable to fragility and rigidity.

We seek an alternative approach that retains the high value present in extant static ontologies, while refactoring them into distributed representations more appropriate for a semantic web services infrastructure. The advantages include supporting dynamic integration, querying of ontological terms from distinct ontologies, and avoiding massive redundancy in distributed ontological representation. In this paper we first describe a distributed representation of GO within the framework of the Virtual Plant Information Network (VPIN)¹ and the SemanticMoby effort [7]². This mapping produces an “individual/property-centric” rather than a “subsumption-centric” model, allowing inference of subsumption assertions on demand. This can be used to generate or expand ontologies *de novo*.

But such distributed representations also come with risks and costs. In particular, it's crucial that the original ontology be easily reconstructible from the decomposed pieces in a lossless manner. We can use order theory [8] to demonstrate that this distributed structure preserves the original GO in a lossless manner. Moreover, we can use Formal Concept Analysis (FCA) [4] to produce a reconstruction which may reveal additional information not present in the original GO, in particular allowing for the unique resolution of potential ambiguities of inheritance.

¹<http://vpin.ncgr.org>

²<http://www.semanticmoby.org>

2 The Semantic Moby/VPIN GO Representation

Viewed ontologically, the GO has very few properties (or “predicates” in RDF terminology) from which a reasoner could *infer* subsumption amongst classes. Rather, subsumption is *asserted* explicitly with **is-a** relationships, such as “cellular process **is-a** biological process”. In order to produce a knowledge structure more amenable to semantic web applications, we seek a distributed refactoring of the GO which allows all GO subsumption relations to be determined, yet minimizes explicit static subsumption statements.

We begin by replacing the subsumption-centric representation of GO with an individual/property-centric representation. Let P be the set of all GO nodes, and for nodes $A, B, C \in P$, consider that we have C **is-a** B and B **is-a** A , and that A is the root of the hierarchy. We could represent this ontologically as the subsumption statements “ C is a sub-class of B ” and “ B is a sub-class of A ”, or mathematically as $C \subseteq B \subseteq A$. In the spirit of REpresentational State Transfer architecture [3], we replace the representation of A, B, C as classes with their representation as instantiated individuals, members of a single class P , resources which reside at URLs. We then introduce a generic property **superProperty** whose domain and range are $P \cup \{\text{null}\}$, where **null** may optionally be used to signify no individual. The semantics of **superProperty** are that its object (right-hand side) is an individual which is a member of a super-class of the class of the subject (left-hand side). We thus assert **C superProperty B**, **B superProperty A**, **A superProperty null**, where the triple is read as subject, predicate, object.

We extend the same convention exactly one level down the hierarchy by introducing the property **subProperty**, and define it dually. We only include the immediate children, since listing all successors for the root would reproduce the entire ontology in the root’s definition. It is often desirable to know an individual’s root class directly (e.g., Biological Process), instead of following a long chain of **superProperty** statements. We therefore finally introduce a property **rootProperty** for all individuals that points to the root of the DAG, e.g. **A rootProperty A**, **B rootProperty A**, etc.

Note that we specifically wish to avoid static subsumption statements as in the use of `owl:subClassOf`, or by naming our property something like **hasParent**. Rather, we wish to encode the necessary information to derive subsumption dynamically. Also, the statement **C superProperty A** could be inferred from **C superProperty B** and the definition of B , but making all **superProperty** statements explicit in the definition of C allows one to build complete definitions within a single file. This has significant advantages in semantic web services architectures. Finally, solely reading the definition of C , one cannot determine if $B \subseteq A$ or $A \subseteq B$. This is an important encapsulation, because the only explicit statements about C that should appear in its definition are those where C is the subject.

Our final definitions are therefore:

A: A rootProperty A	B: B rootProperty A	C: C rootProperty A
A subProperty B	B superProperty A	C superProperty B
	B subProperty C	C superProperty A

The Semantic Moby/VPIN (SMV) refactoring of the GO is available at <http://ontologies.ncgr.org>, an implementation of the abstract Open Biomedical Ontology (OBO) standard. Thus we define the abstract classes for OBO at <http://ontologies.ncgr.org/OpenBiomedicalOntologies> with the extension that includes the properties **superProperty**, **rootProperty**, and **subProperty**. We then map these into specific terms for the Gene Ontology at <http://ontologies.ncgr.org/GeneOntology>. Finally, we map each GO concept into an individual, whose definition files are available at

http://ontologies.ncgr.org/GeneOntology/<ontology>/<GO_id>,

where **<ontology>** is one of **BiologicalProcess**, **CellularComponent**, or **MolecularFunction**, and **<GO_id>** is the GO ID of the term, e.g., **GO_0000001**. In this proof-of-concept implementation, only subsumptive **is-a** relations are included, but **has-part** is also transitive, and thus amenable to the same analysis.

3 Order Theoretical Representation and Reconstruction

The SMV distributed GO implementation has a natural mathematical representation in order theory, the theory of ordered sets and lattices [8]. In this section, we outline this representation, illustrate some of its features, and show how it demonstrates that the original GO can be reconstructed exactly from its SMV factors. Moreover, in the event that there is a portion of the GO with a branching structure complex enough to not allow identification of a least common subsumer, the methodology of FCA [4] will disentangle and disambiguate those connections and allow such identification.

3.1 Factor Reconstruction

Fig. 1 shows a collection of models of a hypothetical portion of the GO, all of which are Directed Acyclic Graphs (DAGs). The left side shows a set of GO nodes P as we typically encounter them in a graphical viewer. Nodes in P are GO categories, connected by arrows indicating **is-a** relations. This is a structure called the Hasse diagram of a partially ordered set (poset) \mathcal{P} on the set of nodes P . Technically, we deal only with bounded, finite posets, and insert an inconsequential virtual bottom (here D) if lacking.

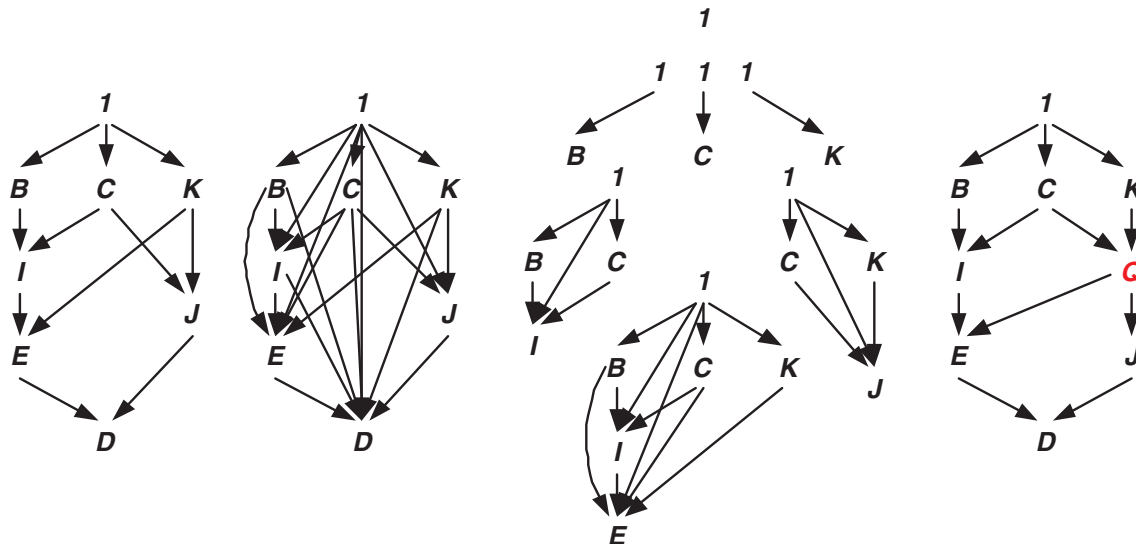


Figure 1: (Left) The Hasse diagram of a model of a GO portion. (Center Left) Transitive closure. (Center Right) Principle filters. (Right) Dedekind-MacNeille completion.

In the center left is the transitive closure of the Hasse diagram, which is a complete relational representation of the ordered set \mathcal{P} . This structure represents the inclusion of all transitive links, effectively what happens when the transitivity of the GO’s “true path rule”³ is followed through to completion, as in the SMV representation. In order theory, a “principle filter” of a node $X \in P$ in a poset \mathcal{P} is a structure $\uparrow X$ consisting of the node X and all of its ancestors, all the way up to the root. In the example, the principle filter of J is the set $\uparrow J = \{C, J, K, 1\}$. The center right of Fig. 1 shows the collection of all the principle filters $\uparrow X, X \in P$ “exploded out”, except $\uparrow D$, which is identical to the center left diagram.

In the SMV representation of the GO, the **superProperty** relation results in the storage of the principle filters $\uparrow X$ for each node $X \in P$ in the database. An important result from order theory is that the original structure of a poset \mathcal{P} is completely recoverable from the principle filters of its atoms (here $\uparrow E$ and $\uparrow J$). Effectively, all that is required is unioning together the filters, and then constructing the transitive reduction [1], resulting in the original Hasse diagram. Thus the SMV representation of the GO is lossless.

³<http://www.geneontology.org/GO.usage.shtml#truePathRule>

	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>I</i>	<i>J</i>	<i>K</i>	1
<i>B</i>	✓							✓
<i>C</i>		✓						✓
<i>D</i>	✓	✓	✓	✓	✓	✓	✓	✓
<i>E</i>	✓	✓		✓	✓		✓	✓
<i>I</i>	✓	✓			✓			✓
<i>J</i>		✓				✓	✓	✓
<i>K</i>							✓	✓
1								✓

Table 1: The formal context of our example.

3.2 Lattice Completion

Given two GO classes *A* and *B*, what do they have in common? Technically, this is the question of calculating their least common subsumer (LCS) [2], a critical and foundational ontology operation (e.g. “what is the common function of these two genes?”). In our example in Fig. 1, the LCS of *I* and *J* is *C*. But what about *E* and *J*? Here there are two possible LCSs, *C* and *K*. In general in bounded posets, not all pairs of nodes need have unique LCSs, but if they do, then the poset is called a lattice.

So while we generally prefer ontologies to be lattices, and not proper posets, if a GO portion happens to be one, then the SMV GO representation provides a solution through the use of FCA [4]. In particular, the **superProperty** properties of the SMV entries for each node $X \in P$ determine a row in a matrix in $P \times P$, called a formal context, where a cell in an *X* row indicates an ancestor of *X*. Our example formal context is shown in Tab. 1. FCA then provides a canonical method to calculate a lattice which precisely represents the context, even if the original ontology was not a lattice. Technically, this is called calculation of the Dedekind-MacNeille completion (Theorem 4 in [4]), and the example is shown on the far right of Fig. 1. Notice the inclusion of *Q* as the new unique parent of *E* and *J*, our two nodes lacking an LCS. *Q* thus acts as a placeholder for whatever is held in common between *C* and *K*, from which it multiply inherits.

4 Conclusion

The SMV representation of large taxonomic ontologies such as the GO is distributed, flexible, and conducive to semantic web architectures, and together with our order theoretical analysis promises significant advances. For example, the lattice-like properties of the GO remains an uninvestigated empirical question: are LCSs always available in the GO? If not, to what extent is it a proper poset and not a lattice, and can we identify the offending portions? The SMV representation allows this calculation to be performed relatively easily, constructing the formal context directly from the database entries, and subtracting the GO from its FCA reconstruction. Moreover, our approach points the way for the use of distributed semantic web implementations and order theoretical technology in other ontology tasks such as induction.

Partial support for this work was provided by NSF BD&I grant 0516487.

References

1. Aho, AV; Garey, MR; and Ullman, JD: (1972) “The Transitive Reduction of a Directed Graph”, *SIAM Journal of Computing*, v. 1:2, pp. 131-137
2. Baader, Franz; Sertkaya, Baris; and Turhan, Anni-Yasmi: (2004) “Computing the Least Common Subsumer w.r.t. a Background Terminology”, in: *Proc. JELIA 2004, Lecture Notes in AI*, v. 3229, pp. 400-412
3. RT Fielding: (2000) *Architectural Styles and the Design of Network-based Software Architectures*, PhD Dissertation, UC Irvine
4. Ganter, Bernhard and Wille, Rudolf: (1999) *Formal Concept Analysis*, Springer-Verlag
5. Gene Ontology Consortium: (2000) “Gene Ontology: Tool For the Unification of Biology”, *Nature Genetics*, v. 25:1, pp. 25-29
6. AG Maguitman, A Rechtsteiner, KM Verspoor, CE Strauss, and LM Rocha: (2006) “Large-Scale Testing Of Bibliome Informatics Using Pfam Protein Families”, *Pacific Symposium on Biocomputing*, 11:76-87.
7. G Schiltz, D Gessler, L Stein: (2004) “Semantic MOBY”, Position Paper for the W3C Workshop on Semantic Web for Life Sciences, W3C, <http://lists.w3.org/Archives/Public/public-swls-ws/2004Sep/att-0036/smoby-w3c-sw-ls.pdf>
8. Schröder, Bernd SW: (2003) *Ordered Sets*, Birkhauser, Boston
9. Verspoor, KM; Cohn, JD; Mniszewski, SM; and Joslyn, CA: (2006) “Categorization Approach to Automated Ontological Function Annotation”, *Protein Science*, 15:6, pp. 1544-1549

SpindleViz: A Three Dimensional, Order Theoretical Visualization Environment for the Gene Ontology

CA Joslyn^{*1}, SM Mniszewski¹, SA Smith² and PM Weber²

¹Computer Science Division, Los Alamos National Laboratory, PO Box 1663, Los Alamos, NM, 87545, USA

²Decision Applications Division, Los Alamos National Laboratory, PO Box 1663, Los Alamos, NM, 87545, USA

Email: CA Joslyn - joslyn@lanl.gov; SM Mniszewski - smm@lanl.gov; SA Smith - sas@lanl.gov; PM Weber - pmw@lanl.gov;

*Corresponding author

Abstract

We introduce SpindleViz, a three dimensional visualization environment for the Gene Ontology (GO) and other taxonomically structured bio-ontologies. SpindleViz exploits the mathematical properties of the partially ordered sets (posets) implied by identified portions of the GO to properly represent the interval-valued vertical levels of GO nodes, and then orient them around a central “spindle” with respect to their centrality in the structure. A novel force-directed layout algorithm allows a natural distribution of the GO nodes in 3-D space, while still respecting the underlying mathematical constraints.

1 Introduction

Currently available tools for visualizing and interacting with portions of large bio-ontologies such as the Gene Ontology (GO) [3] (identified from e.g. gene expression experiments, functional annotation applications, etc.) are inadequate in various ways:

- Classical tools, including early versions of Amigo¹, and DAG-Edit, represent the GO as an indented list, similar to file browsers. While appropriate for tree structures, GO’s Directed Acyclic Graph (DAG) structure results in multiply inheriting nodes being “flattened” by being listed multiple times.
- Other systems also provide an actual visualization of the GO structure, but still rely on a flattening approach to transform the inherent DAG structure of the GO into a tree [1].
- Amigo now provides a graphical viewer, and others (e.g. GeneInfoViz²) display the DAG structure of the GO correctly, but in a manner which doesn’t accurately reflect the underlying mathematical properties such as vertical distance, and doesn’t include such basic tasks as showing all the descendants of a node. Resulting layouts can be difficult to read and maneuver.
- Finally, in our opinion, the size and complexity even of the portions of the GO typically encountered by researchers requires embedding visualizations in three dimensions. While limited 3-D lattice visualizers are available from the mathematics community (see Freese’s work in particular [2]³), we are not aware of any 3-D GO visualizers currently available.

We introduce SpindleViz, a three dimensional visualization environment for the GO. SpindleViz exploits the mathematical properties of the partially ordered sets (posets) implied by GO portions to properly represent the interval-valued vertical levels of GO nodes, and then orients them around a central “spindle” with respect to their centrality in the structure. A force-directed layout algorithm allows a natural distribution of the GO nodes in 3-D space, while still respecting the underlying mathematical constraints.

¹<http://godatabase.org>

²<http://genenet.org/geneinfoviz>

³<http://www.math.hawaii.edu/~ralph/LatDraw>

2 Method

Consider the example diagram of a model portion of the GO DAG shown in the left side of Fig. 1. While this example is very small, it illustrates some of the key structural features of the GO, including multiple inheritance and a “bushy” structure widening towards the bottom, with many leaves.

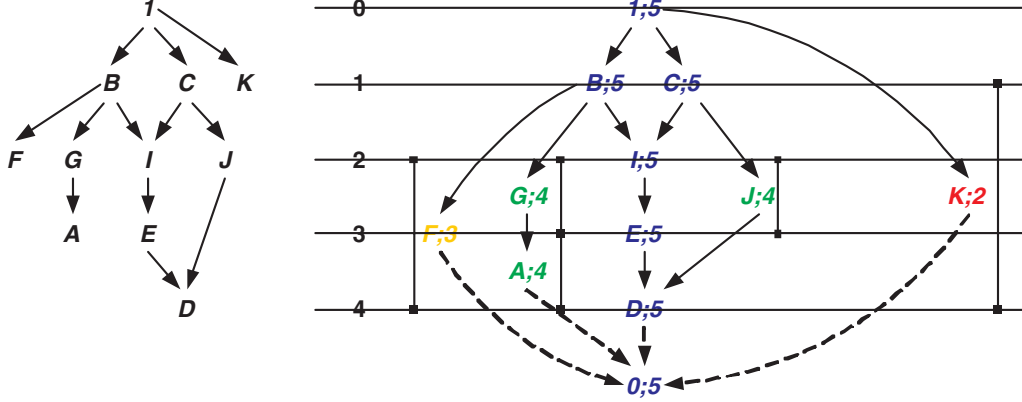


Figure 1: (Left) A model portion of the GO DAG. (Right) The same portion shown with: centrality on each node, including color coding; the interval rank of each node; and laid out vertically by interval rank midpoint and horizontally by centrality.

Following our prior work [6, 8], we represent the GO DAG as a discrete mathematical structure called a partially ordered set (poset) \mathcal{P} on a set of GO nodes P . The right side of Fig. 1 shows the same structure with mathematical quantities related to the poset structure identified, and laid out accordingly. While space precludes a detailed technical description of the poset mathematics here (see [5–8]), we will note:

- We have added a **virtual bottom** node $0 \in P$, useful for completing the mathematical analysis.
- Each node $a \in P$ can be characterized by the set of **complete chains** it sits on connecting the top node $1 \in P$ to the virtual bottom node $0 \in P$ through a . For example, B sits on the three chains $1 \rightarrow B \rightarrow F \rightarrow 0$, $1 \rightarrow B \rightarrow G \rightarrow A \rightarrow 0$, and $1 \rightarrow B \rightarrow I \rightarrow E \rightarrow D \rightarrow 0$.
- Consider depth in the structure as vertical level or rank. The node $C \in P$ is at the “first level”, in that it is one down from the top, which we call “top rank” $r^t(C) = 1$. But while K is also one down from the top, unlike C , it is also a leaf, and thus *also* close to the *bottom*. In fact, K is both one down from the top and one up from the bottom, while C is four up from the bottom. In this way, each node also has a bottom rank, so that $r^b(C) = 1$, $r^b(K) = 4$, and thus an **interval rank** $R(a) = [r^t(a), r^b(a)]$. So $R(K) = [1, 4]$, existing at levels one through four *simultaneously*, as shown in the figure. Similar concepts have been used by Freese [2].
- The **centrality** $s(a)$ of a node $a \in P$ is the length of the largest complete chain it sits on. For example $s(K) = 2$, while $s(C) = 5$. All the blue nodes have maximum centrality 5, and are oriented horizontally at the center along a (possibly complex) central “spindle”. Nodes are also colored from cold to warm with decreasing centrality.

For our 3-D layout algorithm, a set of initial 3-D cylindrical coordinates $\langle \rho, z, \theta \rangle$ are established where z is vertical placement at the interval rank midpoint $r^m(a) = (r^t(a) + r^b(a))/2$; ρ is horizontal placement proceeding outwards from the spindle with decreasing centrality; and θ is the angular position out of the plane around the spindle randomly assigned based on the number of nodes with the same initial z . Then a force-directed layout algorithm allows the nodes to relax in three dimensions: z is constrained between the rank interval midpoint and bottom $[r^m(a), r^b(a)]$, limiting “vertical violations” due to overlapping interval ranks; ρ is constrained to limit “horizontal violations” where a more central node is positioned outside of a less central node; and θ is free to rotate out of the plane around the spindle.

Since vertical position z is a direct function of the interval rank $R(a)$ reflecting the inheritance relations, no vertical violations are tolerated. However, horizontal position ρ is a function of centrality, whose semantics may be somewhat arbitrary based on the particular chain lengths present in a part of the GO; thus limited horizontal violations can be tolerated to improve the layout. Finally, reflecting its status as a mechanism both for the mathematics and for the layout, the virtual bottom is “locked” into place below the top node, and then hidden from the user.

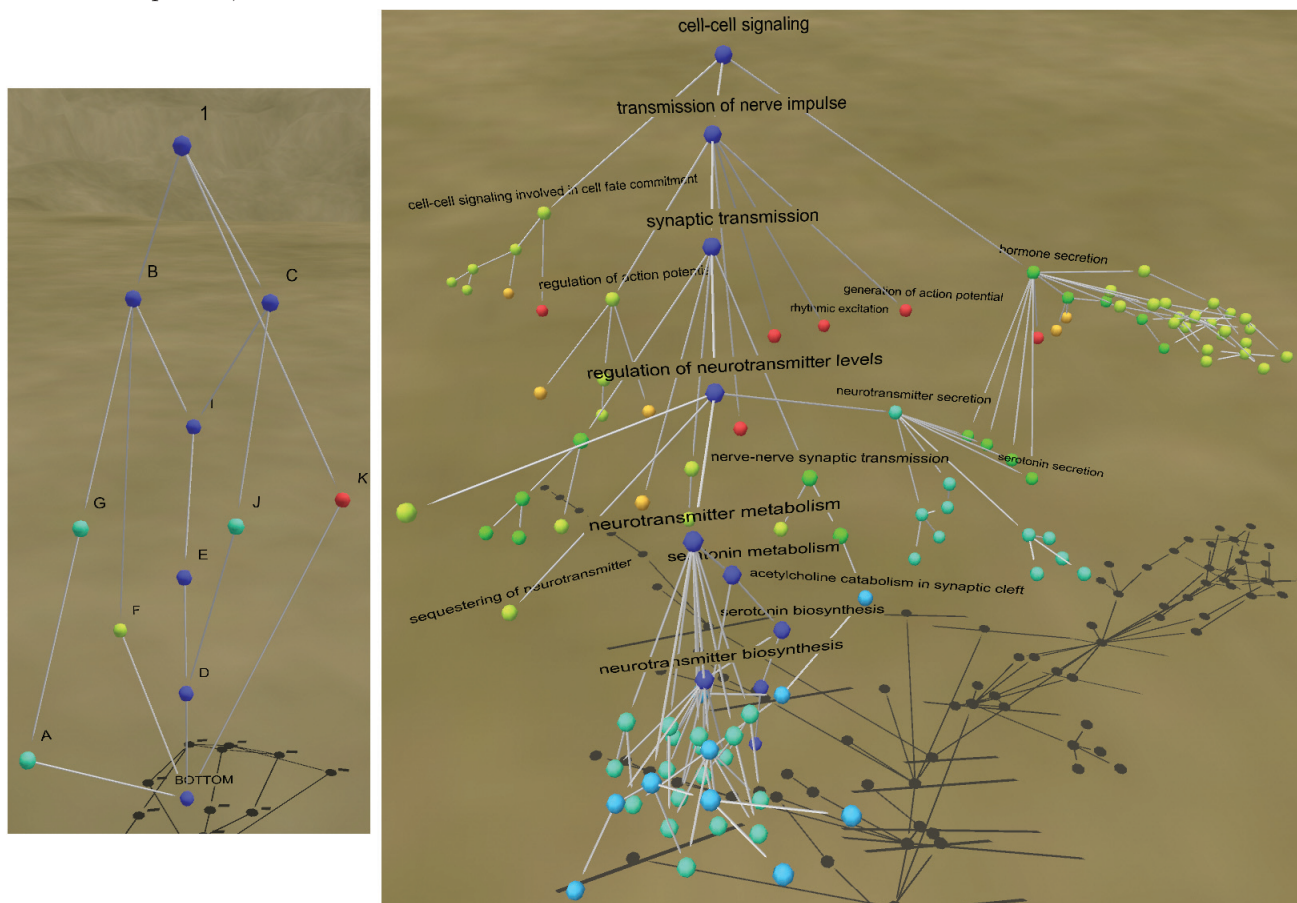


Figure 2: (Left) Example from Fig. 1 in SpindleViz. (Right) Cell-Cell Signalling and its descendants in BP.

3 Implementation and Example

SpindleViz is implemented in Flatland⁴, a 3-D visualization environment developed by the University of New Mexico. While designed to run in a fully immersive environment such as a CAVE or Head Mounted display, Flatland also runs on a workstation or even a modern laptop. In flat, fixed images such as presented in this paper, depth cues like perspective scaling, occlusion, distance attenuation (fog), lighting, shadows, and textures are all that can be relied upon to perceive 3-D effectively.

SpindleViz uses a novel generalized force-directed layout algorithm developed by us for use in Flatland. As in standard force-directed approaches (which are widely used in general bioinformatics applications, e.g. [4]), nodes are assigned mass and charge, and edges length and strength. Nodes experience a force proportional to the spring strength K divided by the difference between the length of the edge and its specified length. We enhance the standard inverse-square Coulomb-like repulsion force $F = A/R^2$, where R is the distance between nodes and A is a scaling factor, by freeing the exponent and introducing an

⁴http://www.hpc.unm.edu/research/scientific_visualization/homunculus_project.htm

attractive force inspired by the Lennard-Jones potential used in molecular modeling, resulting in a two-term expression $F = A/R^\alpha - B/R^\beta$. Continuous adjustment of α, β , and the ratio A/B , allows a wide variety of results to be displayed, and exploration and exposure of different features of graphs.

For poset display, central nodes (blue) are connected by short, strong edges while peripheral nodes (red) are connected by longer, weaker edges. In particular, spring length and the spring constant for each edge is mapped to the square of the minimum of the two centralities of the nodes it connects. We also adjust α and β to encourage both tight packing of highly connected components (clustering) and broad separation of loosely disconnected branches. The left side of Fig. 2 shows the example from Fig. 1 in SpindleViz, while the right side shows the GO node “Cell-Cell Signalling” in the Biological Process branch, GO:0007267, and all its descendants. Both use $\alpha = .5, \beta = 1.0, A = 50$, and $B = 80$, yielding $A/B = .625$. Space considerations will limit us to showing and discussing this single specific real example.

We note a number of features drawn out by the layout, and reflecting the underlying mathematical structure. First, the central spindle in blue descends a relatively deeper portion of the GO, down to “neurotransmitter biosynthesis”. Leaf nodes, which indicate the most specific functions and would tend to accumulate the most protein annotations, are readily visible, for example “serotonin secretion”. However, not all leaves are the same, and in particular red nodes such as “rhythmic excitation” indicate functions which are specific in the sense of accumulating annotations, but are also shallow, indicating a region of the GO which is not as fully developed. Areas of multiple inheritance are also evident, for example “serotonin secretion” drawing from both “neurotransmitter secretion” and “hormone secretion”. Most significant is the clustering effect achieved, with two clear groups under each of those nodes and “neurotransmitter metabolism” (on the spindle), but with “hormone secretion” the largest and deepest.

4 Conclusions and Future Work

SpindleViz is a 3-D capability which is difficult to translate to a flat image in this paper. The ability to rotate, orbit, and zoom, with text becoming more apparent as nodes come closer, is essential to a full appreciation of SpindleViz’s capability to reveal the underlying structure of arbitrary GO portions. Indeed, visualization of such large discrete structures as the GO with tools such as SpindleViz and Flatland is best done within immersive, interactive virtual environments such as CAVEs and RAVEs.

However, we also intend to deploy SpindleViz within a web environment coupled to a simple query system allowing the specification not only of arbitrary GO nodes to visualize, but also expressions such as all the ancestors $\uparrow a$ or descendants $\downarrow a$ of a node $a \in P$, both ancestors and descendants $\Xi(a) = \uparrow a \cup \downarrow a$, the most specific common ancestor (least common subsumer) $a \vee b$ of a pair of nodes, and combinations (unions and intersections) of these. The resulting system will provide a high degree of utility in bioinformatics tasks such as microarray analysis and protein inference.

In this prototype implementation, to aid in debugging and design, color codes for centrality, but this is redundant with the radius ρ . We intend to explore color mappings and other visual cues which show additional information, for example density of protein annotation, inheritance structure, user selections, etc.

References

1. Eric H Baehrecke, Niem Dang, Ketan Babaria and Ben Shneiderman: (2004) “Visualization and analysis of microarray and gene ontology data with treemaps”, *BMC Bioinformatics*, 5:84 doi:10.1186/1471-2105-5-84
2. R Freese: (2004) “Automated Lattice Drawing”, in: *Concept Lattices (ICFCA 04)*, LNAI v. **2961**, pp. 112-127
3. Gene Ontology Consortium: (2000) “Gene Ontology: Tool For the Unification of Biology”, *Nature Genetics*, v. **25**:1, pp. 25-29
4. K Han and Y Byun: (2004) “Three-dimensional visualization of protein interaction networks”, *Computers in Biology and Medicine*, 34 127-139
5. Joslyn, Cliff: (2004) “Poset Ontologies and Concept Lattices as Semantic Hierarchies”, in: *Conceptual Structures at Work* LNAI v. **3127**, ed. Wolff, Pfeiffer and Delugach, pp. 287-302, Springer-Verlag, Berlin
6. CA Joslyn, SM Mniszewski, A Fulmer and G Heaton: (2004) “The Gene Ontology Categorizer”, *Bioinformatics*, v. **20**:s1, pp. 169-177
7. Schröder, Bernd SW: (2003) *Ordered Sets*, Birkhauser, Boston
8. KM Verspoor, JD Cohn, SM Mniszewski, and CA Joslyn: (2006) “A Categorization Approach to Automated Ontological Function Annotation”, *Protein Science*, in press

Exploring the construction of resources for detecting protein descriptions from the literature

Martin Krallinger

Dep. of Struct. Comp. Biology
Spanish National Cancer Centre
Madrid, Spain
mkrallinger@cnio.es

Rainer Malik

Dep. of Computer Sciences
University of Utrecht
Utrecht, the Netherlands
rainer@cs.uu.nl

Alfonso Valencia

Dep. of Struct. Comp. Biology
Spanish National Cancer Centre
Madrid, Spain
valencia@cnio.es

Abstract

We introduce the Protein description sentence (Prodisen) corpus, a useful resource for the automatic identification and construction of text-based protein description records using information extraction and text classification techniques. Basic guidelines and criteria relevant for the construction of a text corpus of functional descriptions of genes and proteins are proposed.

The used steps for the corpus construction and its features are presented. Moreover, some of the potential applications of the Prodisen corpus for biomedical text mining purposes are explored and the obtained results are presented.

The Prodisen corpus construction steps can be regarded as a general guideline for the construction of gene description corpora for text mining and information extraction in the biology domain. It is freely available, easy to revise and extend and complements well existing resources such as the GENIA and MedTag corpora.

Background

The rapid growth of the biomedical literature as deposited in databases such as PubMed together with the demands posed by the biology community for efficient access to gene descriptions increased the interest in using information extraction (IE) and text mining strategies to identify relevant functional descriptions from scientific literature. Attempts to use computational tools to extract a variety of functional information from the literature, ranging from protein interactions [Blaschke and Valencia2001], gene product annotations [Stoica and Hearst2006] to gene-disease association [Perez-Iratxeta et al.2002] have been made in the past.

For the development and evaluation of such systems, high quality training and test data collections are crucial. Despite the number of described applications, only few attempts to construct such data sets have been made so far. In most cases, biomedical text mining systems were evaluated based on small data collections, lacking sufficient descriptions on both, data selection criteria as well as quality of the data in terms of inter-annotator agreement. Although developing suitable biomedical corpora is a time consuming and labor-intensive task which requires the involvement of domain experts some high quality biomedical corpora exist, such as the GENIA corpus [Kim et al.2003] and MedTag corpus [Smith et al.2005]. Moreover Cohen et al [Cohen et al.2005] explored basic aspects which influence the usage of existing biomedical corpora.

An aspect often neglected when using existing biomedical corpora is that they are often based on previous information retrieval (IR) steps using very specific query terms. In case of the GENIA corpus, a previous selection step for PubMed articles with the MeSH terms *human*, *blood cell* and *transcription factor* was done. Previous filtering steps using very specific query terms or selecting only certain journals may result in limiting the type of functional descriptions contained in the derived corpora.

Other available data collections were only tailored for very narrow text mining tasks such as the identification of gene mentions in text [Yeh et al.2005] and are thus not suitable for the discovery of functional descriptions where proteins are referred to using for instance anaphoric expressions.

The BioCreAtIvE contest [Blaschke et al.2005] resulted in a useful text passage corpus for the extraction of human protein annotations based on controlled vocabulary terms (Gene Ontology concepts). Nevertheless it does not consider other relevant descriptions such as protein interactions or gene-disease associations. Other resources used

by text mining applications were derived from data stored in existing biological databases, mainly the Gene Ontology Annotation (GOA) database [Camon et al.2004] and the GeneRif database [Mitchell et al.2003]. Both of these data collections were not constructed having in mind their use by biomedical text mining applications. Their use by information extraction tools show significant limitations. To complement existing resources and to help in the efforts to bridge the gap between IR approaches and Text Mining, we developed the Protein descriptions in sentences (Prodisen) corpus. It contains a set of positive sentences corresponding to gene descriptions and a contrast set of negative sentences with no functional information associated. Prodisen explores the whole PubMed collection without previous document restriction using specific query terms. The Prodisen sentences derived from a large corpus of PubMed abstracts have been manually categorized by domain experts.

The most significant gene description types considered include aspects related to their interactions with other biomolecules, their molecular function, cellular location or associations to disease conditions.

The procedure for the construction of Prodisen is designed to be reproducible and extendable by distributing the corpus together with the used corpus construction script. The Prodisen gene description sentences include anaphoric cases and are not limited to genes or gene products and also include general protein family names. To assess the difficulty of identifying gene descriptions from PubMed the inter-observer agreement was measured. Also basic characteristics such as the relative sentence position of gene descriptions within abstracts were measured.

1 Corpus construction

For the construction of a biomedical corpus the PubMed database of life sciences literature citations was used [Wheeler et al.2003].

Only a fraction of the PubMed entries are related to molecular biology, and many of the abstracts belonging to this domain do not contain relevant information on gene or gene product functions.

For a system which tries to extract useful gene descriptions from PubMed it is crucial to identify the actual sentences describing the association between genes and their functions.

To take into account gene descriptions which do not explicitly state the characterized genes, in case of the Prodisen construction, domain experts (two Ph.D. students in Molecular Biology) were presented sequentially with the sentences together with the corresponding abstract. Then they had to classify each sentence into one of three basic classes: (Y) the sentence is useful as gene description, independent of whether the gene name is men-

tioned or a referential expression is used; (N) the sentence is not a gene description or (D) uncertain ambiguous cases, where the expert was not sure.

Two sample cases belonging to the negative class (i.e. not suitable to derive protein descriptions) are:

"N 15337019 10 Group B-1 (54 patients) underwent lateral prophylactic sternal reinforcement before placement of peristernal wires."

"N 10444591 1 ATP-dependent nucleosome remodeling and core histone acetylation and deacetylation represent mechanisms to alter nucleosome structure."

The uncertain class (D) included sentences where it was not clear from the context if it was referring to a gene description. This class also included cases of wrong sentence splitting.

Corpus	Sent.	Abs.	Words	Y	N	D
R	10,039	1,234	224,890	1,704	7,899	436
E	11,125	1,232	244,549	7,693	2,949	483
All	21,164	2,466	469,439	9,397	10,848	919

Table 1: Prodisen corpus in numbers. R: Prodisen random corpus, E: Prodisen enriched corpus, Sent.: total number of sentences, Abs.: total number of abstracts, Words: total number of words, Y: total number of sentences useful to derive gene descriptions, N: total number of sentences which are not useful as gene descriptions, D: total number of sentences which correspond to uncertain cases.

1.1 Gene description classes

The sentences containing gene description information were additionally classified in those containing information on relevant aspects of a gene, gene product, gene group, protein family or protein domain based on the analysis of the contextual information. Those cases include descriptions where the gene names appear as well as cases where referring expressions were used or it could be inferred that the sentence contains a relevant gene description (based on the context). The classes considered include:

- Descriptions related to molecular functions, biological processes and cellular locations.
- Descriptions related to associations to diseases, symptoms or treatments.
- Descriptions referring to interactions, e.g. protein interactions and dimerization or protein-compound interactions.
- Information related to the gene expression (e.g. in which tissues a given gene is expressed)
- Descriptions of sequence and structural features (including mutations, protein family, isoforms, post-translational modifications, SNP, chromosome mapping) and homology information.

- Other useful gene descriptions such as information related to phenotypes, experimental usage (markers) and enzyme kinetics.

All these basic types of gene descriptions refer to different relevant aspects that characterize genes, proteins and protein families and represent the diversity of annotation information stored in different biological annotation databases.

DB	GOA	GeneRif	UniProt	OMIM	PDB
GOA	29,248	3,972	15,409	9,465	135
GeneRif	3,972	84,380	4,890	6,637	620
UniProt	15,409	4,890	112,476	19,859	5,061
OMIM	9,465	6,637	19,859	88,766	296
PDB	135	620	5,061	296	11,790

Table 2: PubMed article usage overlap between different biological databases. For each database the number of (non-redundant) PubMed articles used was extracted and the overlap between different databases was calculated.

1.2 Prodisen random corpus

To address the detection of gene description sentences from the whole PubMed collection (only a fraction of abstracts contained in PubMed are related to Molecular Biology) and to estimate the total amount of functional description sentences in PubMed, we constructed a set of randomly selected abstracts.

We have classified each of the sentences from those abstracts into one of the three previously described categories. As shown in Table 1, the Prodisen random corpus contains over 10,000 sentences, where around 15 percent have been classified as gene description sentences. Based on the PubMed size we expect that there are around 9,5 million sentences in Pubmed which contain gene description related information.

1.3 Prodisen enriched corpus

As the proportion of positive cases, i.e. sentences corresponding to gene descriptions, is relatively small in the random Prodisen corpus, we propose a strategy to construct an enriched set of abstracts in terms of gene descriptions. This strategy is based on PubMed article citation overlap between different biological annotation databases, each of them with a different focus regarding the gene description type.

To obtain a collection of abstracts that covers all the previously defined relevant gene description types we extracted first for each database the number of (non-redundant) PubMed articles used for their annotations. Then we identified the articles which were used as citations by several different databases. Table 2 illustrates the

binary overlap between the citations extracted for each of the biological databases. The core set of the enriched Prodisen corpus consisted in the articles cited by the following biological databases: Uniprot, OMIM, GeneRif and GOA. Additional articles cited in GOA and PDB, Pfam or IntAct were included. The resulting Prodisen enriched corpus contained over 11 thousand sentences, with over 7 thousand gene description sentences (see table 1).

2 Initial analysis of the Prodisen corpus

The initial inspection of the corpus reveals an association between the relative position of the informative sentence within an abstract and the classification as a gene description. Many of the gene descriptions are contained in the final part of abstracts (Figure 1). This is in line with the general discourse structure of scientific abstracts: (1) very general introduction, (2) more specific aspects, aims or problems, (3) experimental methods and (4) obtained results and conclusions. Previous studies already pointed out the importance of text zone analysis for information extraction applications [Mullen et al.2005].

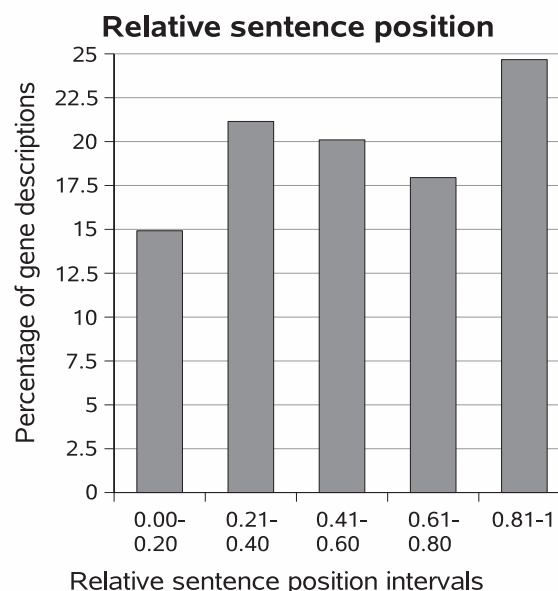


Figure 1: Relative sentence position intervals vs. percentage gene description sentences, derived from the enriched Prodisen corpus. Most of the gene description sentences are contained in the final part of abstracts. The obtained percentages of description sentences for each position are confirmed by the generally used discourse structure of abstracts, where the first sentences are very general statements, followed by more detailed aspects, experimental methods and obtained results.

3 Prodisen availability

The Prodisen corpus (both the random and the enriched set) is available online at:

<http://www.pdg.cnb.uam.es/martink/PRODISEN/>

It is provided in an easy to use format together with the Prodisen construction script, to ensure both, the reconstruction and the possible customized extension of the Prodisen corpus.

4 Discussion and Conclusions

We explored the construction of a corpus of gene descriptions, which is not limited to specific description types. It has thus a more extended use, for instance to construct gene summaries, which commonly are not limited to a single description type. The classification includes a set of practically relevant gene description types, directly related with the information stored in existing biological annotation databases.

The Prodisen corpus which has been constructed from randomly selected PubMed entries is useful to identify functional descriptions from the whole PubMed. The fraction of gene descriptions in this corpus reflects the gap between the amount of gene descriptions contained in biological databases and those stored in the scientific literature.

One of the observations made during the process of building Prodisen is that many of the gene descriptions do not mention explicitly the gene or protein names but use referring expressions (e.g. anaphora) instead.

We also propose a strategy to construct a gene description enriched corpus by selecting those articles which have been used by multiple different databases, each focussing on different description types. Using this approach, the selected abstracts have a higher probability of containing a large number of gene description sentences, and increase the probability of the sentences to be based on experimental results.

Many of the gene descriptions required inference processes based on contextual information as well as on the background knowledge of the domain experts, with potentially more than one interpretation. Considerable reader variability is often encountered also in other domains such as in clinical sciences for agreement in screening of mammography [Gram et al.2005].

We can foresee the potential uses of the Prodisen corpus in the training and testing of information extraction techniques dedicated to the identification of gene/protein descriptions in text. It can be a useful resource for both, bag of words based approaches focussing on word frequencies, as well as for the discovery of description patterns. Additionally the corpus can be used to derive contextual word frequencies for protein mention discovery (disambiguation). Finally, the availability of the Prodisen con-

struction script will facilitate the extension of the corpus with additional data, or with the creation of more specific types of gene descriptions.

References

- C. Blaschke and A. Valencia. 2001. The potential use of SU-ISEKI as a protein interaction discovery tool. *Genome Inform Ser Workshop Genome Inform.*, 12:123–134.
- C. Blaschke, E. Andres Leon, M. Krallinger, and A. Valencia. 2005. Evaluation of BioCreative assessment of task 2. *BMC Bioinformatics.*, 6:S16.
- E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler. 2004. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, 32:262–266.
- K.B. Cohen, L. Fox, P.V. Ogren, and L. Hunter. 2005. Corpus Design for Biomedical Natural Language Processing. *Proc of ACL-ISMB 2005 Workshop*, pages 38–45.
- I.T. Gram, Y. Bremnes, G. Ursin, G. Maskarinec, N. Bjurstam, and E. Lund. 2005. Percentage density, wolfe's and tabar's mammographic patterns: agreement and association with risk factors for breast cancer. *Breast Cancer Res.*, 7(5):854–861.
- J.D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus—semantically annotated corpus for bio-textmining. *Bioinformatics.*, 19:i180–i182.
- J.A. Mitchell, A.R. Aronson, J.G. Mork, L.C. Folk, S.M. Humphrey, and J.M. Ward. 2003. Gene indexing: characterization and analysis of NLM's GeneRIFs. *AMIA Annu Symp Proc.*, pages 460–464.
- T. Mullen, Y. Mizuta, and N. Collier. 2005. A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. *ACM SIGKDD Explorations Newsletter*, pages 52–58.
- C. Perez-Iratxeta, P. Bork, and M.A. Andrade. 2002. Association of genes to genetically inherited diseases using data mining. *Nat Genet.*, 31:316–319.
- L.H. Smith, L. Tanabe, and T. Rindflesch. 2005. MedTag: a collection of biomedical annotations. *Proc of ACL-ISMB*, pages 32–37.
- E. Stoica and M. Hearst. 2006. Predicting Gene Functions from Text Using a Cross-Species Approach. *Pac Symp Biocomput. 2006*.
- D.L. Wheeler, D.M. Church, S. Federhen, A.E. Lash, T.L. Madden, J.U. Pontius, G.D. Schuler, L.M. Schriml, E. Sequeira, T.A. Tatusova, and Wagner.L. 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, 31:28–33.
- A. Yeh, A. Morgan, M. Colosimo, and L. Hirschman. 2005. BioCreAtIvE Task 1A: gene mention finding evaluation. *BMC Bioinformatics.*, 6:S2.

Improving Biomedical Corpus Annotation Guidelines

Zhiyong Lu^{*1}, Michael Bada¹, Philip V. Ogren¹, K. Bretonnel Cohen¹ and Lawrence Hunter¹

¹Center for Computational Pharmacology, School of Medicine, University of Colorado, Aurora, CO, 80045, USA

Email: Zhiyong Lu^{*} - Zhiyong.Lu@gmail.com; Michael Bada - Mike.Bada@uchsc.edu; Philip Ogren - Philip.Ogren@uchsc.edu; K. Bretonnel Cohen - Kevin.Cohen@gmail.com; Lawrence Hunter - Larry.Hunter@uchsc.edu;

^{*}Corresponding author

Abstract

Background: We annotated over a thousand GeneRIFs with respect to different biological entities (e.g., protein) and processes (e.g., protein transport) in order to develop and evaluate text mining methods in molecular biology.

Results: We monitored inter-annotator agreements (IAAs) between two human annotators on a weekly basis and found that IAAs had a significant improvement over the first 10-week annotation period. In this paper, we present in detail our span selection guidelines along with other useful annotation experiences.

Conclusions: Clear annotation guidelines play a critical role in high quality corpus annotation. The guidelines presented in this paper are designed to be general purpose and easy to follow. Complete guidelines are publicly available at <http://compbio.uchsc.edu/grifs/transport/guidelines>

Background

Clear annotation guidelines are important for achieving consistent high quality annotations of textual corpora [1–3]. However, for the biomedical domain, there are few published annotation guidelines [1, 4–6], and these mostly concern only gene/protein names [1, 4, 5]. Few published corpora provide annotation beyond this level, most notably GENIA [7] and BioIE [8].

In support of a project that has annotated over a thousand GeneRIFs (Gene Reference Into Function) with 31 semantic classes in the protein transport domain, we have produced a novel set of guidelines for span selection with four advantages over other published guidelines: (1) They are generally applicable throughout molecular biology. Originally developed for annotating biological entities pertinent to protein transport (e.g., transport mechanisms, locations), their generality is validated by the application of the same set of rules to annotate many other concepts in

gene expression; (2) They can be learned and applied well by annotators; (3) They generally require only simple, straightforward judgements; and (4) Some of these rules can be applied to produce data for new tasks. For example, the rules for the embedded entities and appositives are useful for NLP tool development (see Discussion for details).

Methods

Annotation process

We hired two annotators with advanced degrees in molecular biology. They each had 20 hours of training that included learning Knowtator [9], a general purpose text annotation tool that was developed as a Protégé plug-in [10] and thus takes advantage of its knowledge representation capabilities. After annotation commenced, biweekly meetings were conducted with the annotators to raise concerns and discuss problems, which often led to refinements in the annotation guidelines.

Annotation guidelines for span selection

Every new annotation starts with a word in a GeneRIF that best corresponds to a class in an ontology. We call this special word the *anchor word*. Typically, the anchor word is the base noun in a noun phrase (e.g., the word “receptor” in the phrase “nuclear estrogen receptor” [PMID: 9522357] refers to **protein**). (Words and phrases in *Courier* denote classes of the ontology). And the noun phrases here do not include any prepositional phrases. Less frequently, another part of speech will fill the role of anchor word, for instance: (i) adjectives (e.g., “nuclear” in “nuclear estrogen receptor” refers to **nucleus**); (ii) modifying nouns (e.g., “estrogen” in “nuclear estrogen receptor” refers to **small molecule**); and (iii) verbs (e.g., “translocates” in “IMP1 translocates to the nucleus” [PMID: 12921532] refers to **transport**). After identifying an anchor word, annotators follow the rules below:

1. Rule for modifiers: Include all preceding nouns or adjectives that modify the anchor word, as well as *trailing variant specifiers* (letters or numbers used to distinguish a specific entity from a more general one). For example, in “estrogen receptor alpha” [PMID: 16271083], “receptor” is the anchor word, and it has a preceding modifier “estrogen” and a trailing variant specifier “alpha”. According to this rule, all three words “estrogen receptor alpha” should be selected as one annotation. It is to be noted that we do not consider preceding articles (e.g., a, the), demonstratives (e.g., this, that), pronouns (e.g., its, they), and quantifiers (e.g., one, all) as modifiers; hence, they should not be included.

2. Rule for embedded entities: Annotate embedded entities separately. Because of the previous rule, an annotation can sometimes include one or more other entities. We call those entities *embedded entities*. For example, an annotation of a protein “estrogen receptor alpha” includes another entity “estrogen,” the ligand to which the receptor binds. In this case, we ask annotators to make a separate annotation for the embedded entity “estrogen.” Together with the previous rule, for “nuclear estrogen receptor,” three annotations should be made (underlined below), one for each anchor word. Note that when the anchor word is “estrogen,” the preceding adjective “nuclear” is not included in the text span because it modifies “receptor,” not “estrogen.”
nuclear estrogen receptor (anchor word “receptor”)
 nuclear estrogen receptor (anchor word “estrogen”)
nuclear estrogen receptor (anchor word “nuclear”)

3. Rule for appositives: Make separate annotations for appositives. An appositive is a noun phrase that renames or describes another noun phrase. An appositive often appears immediately after its expansion, as in “estrogen receptor-alpha (ER-alpha)” [PMID: 14691461]. However, “Nur77” in the phrase “TR3/Nur77” [PMID: 14500374] is also appositive because “Nur77” and “TR3” are synonyms for the same entity (Entrez GeneID: 3164). Thus, according to our rule, both “ER-alpha” and “Nur77” should be annotated separately.

4. Rule for punctuation: Normally, punctuation will not be included in span selection. For example, the parentheses in “estrogen receptor-alpha (ER-alpha)” should never be part of an annotation span. However, punctuation that is a part of a word or name should be included in span. For instance, the connecting hyphen in “ER-alpha” should be included in an annotation.

5. Rule for conjunctions: Mark up each constituent of the conjunctions. The conjunction words “and”, “or”, “&” and sometimes “/” and “-” create interesting annotation challenges such as “estrogen receptor alpha and beta heterodimers” [PMID: 15803276]. In this case, a total of four separate annotations should be made (underlined below), which correspond to **molecular complex**, **protein**, **protein**, and **small molecule**, respectively.

estrogen receptor alpha and beta heterodimers
estrogen receptor alpha and beta heterodimers
estrogen receptor alpha and beta heterodimers
estrogen receptor alpha and beta heterodimers

Results and Discussion

IAAs over time

We compared annotations between two human annotators by using inter-annotator agreement (IAA):

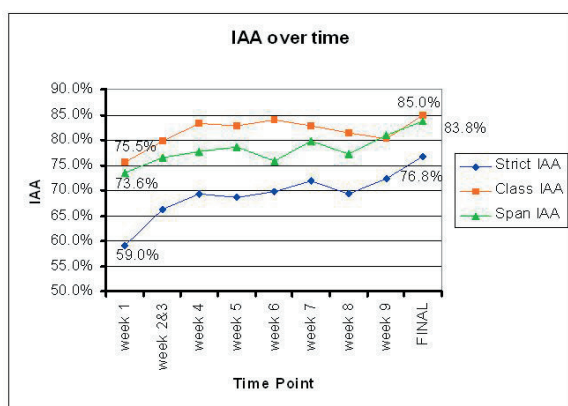
$$IAA = \frac{M}{M + NM}$$

where M is the number of annotations considered a match and NM is the number of non-matches. The sum of M and NM equals the total number of annotations. Three different IAAs were measured by changing the match criteria: *Strict IAA* requires that the annotations have the same class and span. *Type IAA* requires that two annotations have the same class but relaxes the span matching criteria such that the annotations' spans need only overlap (rather than match exactly). *Span IAA* requires that two anno-

tations have only exact matching spans but does not compare their classes.

We computed the three IAAs on a weekly basis (in Figure 1) for the first ten weeks and found that the final IAAs are approximately 10% to 18% higher than the initial IAAs. There are multiple reasons for such increases, one of which we believe is due to the refinements in annotation guidelines. Other possibilities include that annotators became more familiar with the task, and that annotators followed the guidelines more strictly.

Figure 1 - IAA improvement over time



Rules can be well learned and applied

To investigate how well annotators use the rules after a period with relatively small or no rule changes (vs. constant changes during the first ten weeks), we examined the spans of 569 annotations that were recently marked up for 188 GeneRIFs. Table 1 shows the number of usages of each individual rule described in the previous section. As can be seen, approximately 40% (249/569) of the annotations required the annotator to apply one or more span selection rules (e.g., “fibroblast growth factor receptor 3 (FGFR3)” [PMID: 11731410]). We only found 10 cases where the annotator failed to follow the rules.

Table 1: Usage of span selection rules. Rules I – V corresponds to the rules for modifiers, embedded entities, appositive acronyms, punctuation, and conjunctions, respectively.

Span Rules	I	II	III	IV	V	Total
Correct Usages	92	43	14	96	4	249
Errors	2	6	2	0	0	10

Guideline design rationale

The design of span selection rules is based on two main criteria to guarantee the rules are: (1) general and applicable to different entities; and (2) capable of letting annotators produce consistent annotations.

The first criterion makes sure that when there is an ontological change (e.g., the addition of new concepts or even an entirely different ontology), the same rule can be applied. The second criterion makes sure the rules are practically useful to achieve high IAAs. An example is the rule for preceding modifiers, where annotators need only make simple, straightforward decisions regarding whether or not to include preceding modifiers, thereby facilitating consistent annotations.

Comparison to other guidelines

Our guidelines are comparable to those used in other projects but are different in that they: (i) are more explicit and consistent; (ii) require only simple, straightforward judgements; and (iii) can be applied to produce data for many new NLP tasks, such as recognizing nested named entities and identifying abbreviation definitions.

We will use the following examples for comparison of our guidelines to other published guidelines:

1. classII-positive B cell
2. IL-2 receptor
3. P-glyconprotein (P-gp)-related compounds

With regard to the first example, the entire phrase “classII-positive B cell” would be selected according to our rule for modifiers. The same annotation would also be made in the case of GENIA following a policy of “more specific concepts” [11], which is similar to our rule for modifiers because preceding modifiers typically lead to more specific concepts. However, our rule is more explicit. Furthermore, it rarely relies on annotators’ subjective judgement on what “more specific” means in various conditions, which according to [7] “may seem arbitrary”.

In the second example, a single annotation for “IL-2 receptor” but not “IL-2” would be made in GENIA and GENETAG [1] according to similar policies “mentioned substance only” [11]. Following our rule for embedded entities, we would make two separate annotations, “IL-2 receptor” and “IL-2”, because they refer to two distinct entities. This is desirable for two reasons. First, separate annotations are useful for developing and evaluating au-

tomatic methods of recognizing nested named entities [12]. Second, separate annotations are useful for more complex knowledge representations. For example, if we decide to add a slot (attribute) into the class **protein** to formally represent the ligand a protein binds to, then a separate annotation “IL-2” will be required to fill the slot. Finally, if some tasks (e.g., BioCreAtIvE 1A [13]) require only the longer expression “IL-2 receptor”, we could easily remove “IL-2” during post-processing because we could determine “IL-2” is embedded in the longer expression “IL-2 receptor” by examining their spans. However, it is not a straightforward procedure to programmatically extract a nested entity from a single, undivided text span. For example, it is difficult to retrieve embedded entities such as “human lung” (which refers to **organ**) from a longer expression “human lung epithelial cell” (which refers to **cell**).

In the third example, a separate annotation for “P-gp” would not be made in BioIE [6] because it is nested in a longer term (“P-glycoprotein (P-gp)-related compounds”). However, the BioIE guidelines, in general, do require annotators to mark up abbreviations separately (e.g., annotate “GIST” separately in “gastrointestinal stromal tumor (GIST)”). In contrast, our rule for appositives requires annotators to always mark up abbreviations, which resulted in more consistent annotations that can be useful for developing and/or evaluating automatic tools like [14]. Furthermore, since our rule has no exceptions, it helps to reduce the total number of the rules, thus making it less difficult to be learned.

Other helpful experiences

In addition to detailed span selection guidelines, we found it is important to provide concrete examples, especially for classes that are sometimes hard to differentiate from one another (e.g., **small molecule** vs. its parent **molecule**).

Another useful experience is to decompose the complex annotation project into coherent subparts (Martha Palmer, personal communication), each of which can then be focused on individually. We switched to this approach by first asking the annotators to only mark up mentions of **cellular component** and its subclasses in all GeneRIFs, then mentions of **protein** in a second pass, and finally mentions of **protein transport** in a third pass.

Conclusions

We have described several notable features in our span selection guidelines by focusing on noun phrases because their annotation is usually more

difficult and thus requires more attention. Guidelines for others (e.g., verb phrases) can be found at the paper supplementary website. We conclude that our guidelines have advantages over other published guidelines and can be useful for many other similar annotation projects.

Acknowledgements

This work was supported by NIH grant R01-LM00811 (LH). We thank Sue Brozowski, Lynne Fox, and Manuel Miranda. We thank people from BioIE, GENETAG, and iProLink for making their guidelines publicly available.

References

1. Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ: **GENETAG: a tagged corpus for gene/protein named entity recognition**. *BMC Bioinformatics* 2005, **6 Suppl 1**.
2. Colosimo ME, Morgan AA, Yeh AS, Colombe JB, Hirschman L: **Data preparation and interannotator agreement: BioCreAtIvE task 1B**. *BMC Bioinformatics* 2005, **6 Suppl 1**.
3. Blaschke C, Leon EA, Krallinger M, Valencia A: **Evaluation of BioCreAtIvE assessment of task 2**. *BMC Bioinformatics* 2005, **6 Suppl 1**.
4. Inderjeet M, Zhangzhi H, Bae JS, Ken S, Matthew K, Jon Pa: **Protein name tagging guidelines: lessons learned**. *Comparative and Functional Genomics* 2005, **6(1-2)**:72–76.
5. Vlachos A, Gasperin C, Lewin I, Briscoe T: **Bootstrapping the recognition and anaphoric linking of named entities in drosophila articles**. In *Proceedings of Pacific Symposium on Biocomputing* 2006:100–111.
6. **BioIE online annotation guidelines** [http://bioie ldc.upenn.edu/wiki/index.php/Main_Page].
7. Kim JD, Ohta T, Tateisi Y, Tsujii J: **GENIA corpus—semantically annotated corpus for bio-textmining**. *Bioinformatics* 2003, **19 Suppl 1**.
8. Seth K, Bies A, Liberman M, Mandel M, McDonald R, Palmer M, Schein A: **Integrated annotation for biomedical information extraction** 2004.
9. Ogren P: **Knowtator: A Protege plug-in for annotated corpus construction**. In *Proceedings of HLT-NAACL 2006 Demonstrations*, NY, NY 2006.
10. Noy NF, Sintek M, Decker S, Crubezy M, Fergerson RW, Musen MA: **Creating Semantic Web Contents with Protege-2000**. *IEEE Intelligent Systems* 2001, **2(16)**:60–71.
11. Ananiadou S, McNaught J: **Corpus Annotation in Biology**. In *Text Mining for Biology And Biomedicine*, Artech House Publishers 2005:188–211.
12. Gu B: **Recognizing nested named entities in GENIA corpus** [abstract]. In *Proceedings of BioNLP workshop of HLT-NAACL 2006*, Brooklyn, NY 2006.
13. Yeh A, Morgan A, Colosimo M, Hirschman L: **BioCreAtIvE task 1A: gene mention finding evaluation**. *BMC Bioinformatics* 2005, **6 Suppl 1**.
14. Schwartz A, Hearst M: **A Simple Algorithm For Identifying Abbreviation Definitions in BioMedical Text**. In *Proceedings of PSB 2003*, Lihue, Hawaii 2003.

Improving Biomedical Text Categorisation with NLP

Michael Matthews

Informatics Department, University of Edinburgh, Buccleuch Place, Edinburgh, UK

Email: m.matthews@ed.ac.uk;

Abstract

Background: Text categorisation has been used in bioinformatics to help identify documents containing protein-protein interactions. Standard text categorisation methods have used the bag-of-words approach with little input from NLP. While this has proved effective in the past, there is some evidence that the techniques are not adequate in some biological domains. Here we examine how chunking, named-entity recognition and relationship extraction can be combined with traditional text categorisation techniques to improve the classification of documents containing protein-protein interactions.

Conclusions: A system that combines the output of an NLP system with the standard techniques of text categorisation can produce results that exceed the performance of either system on its own. The F_1 of a system that combined features of an NLP system with standard text categorisation features was 68.1 compared with 62.0 using text categorisation alone and 61.9 using relationship extraction alone.

1 Background

1.1 Introduction

Automatic text categorisation has been used in the biomedical domain to help find documents that contain protein-protein interactions. Typically, text categorisation largely ignores the structure and semantics of a document and rather treats a document as a bag-of-words. There is some evidence, however, that standard techniques are not always sufficient. Therefore it is worth considering how NLP can possibly improve performance. In particular, we examine how chunking, named-entity recognition (NER), and relationship extraction (RE) can be used to improve text categorisation.

1.2 Previous Work

In text categorisation, a document is typically represented as a vector of the words in the document with associated weights. The words are frequently stemmed using the porter stemmer and words in a stop word

list are often removed. The weight is usually some function of the number of times the word occurs in the document [1]. A subset of all of the features is often selected using some metric, most commonly information gain [2]. Finally, given a training corpus of documents which have each been marked with their true class, a model can be created which predicts the most likely class of a document given the document representation. There are many different classification techniques, some of the most common being naïve Bayes [3], SVMs [4] and KNN [1]. Many of these techniques have been successfully applied in the bioinformatics domain. [5] use a Bayesian approach while both [6] and [7] use SVMs to classify documents containing protein-protein interactions. However, in the KDD Challenge 2002 [8], Regev et al [9] found evidence that in at least that specific classification task, the identification of papers suitable for FlyBase gene-expression database curation, information extraction techniques were more suitable than classic text categorisation techniques. Information extraction is a subset of NLP which covers a broad range of techniques for processing human language with computers with the general goal of extracting information from text with minimal human interaction. NLP has been used extensively in bioinformatics [10], particularly in NER to identify proteins and other biological entities [11] and in RE to extract protein-protein interactions [12, 13].

2 Data and Methods

The experiments were run on a set of 2025 PubMed abstracts that were all analysed to determine if they contained protein-protein interactions as part of the Text Mining programme (TXM). Abstracts containing protein-protein interactions were considered curatable while documents not containing protein-protein interactions were considered not-curatable. In all, 467 documents were found to be curatable and 1558 not-curatable. The collection was split 64% for training, 16% for heldout testing and 20% for testing with each set having the same proportion of curatable and not-curatable documents. Each document was processed with an NLP pipeline consisting of a tokeniser and chunker based on the LTG toolkit [14], a part-of-speech tagger based on the Curran and Clark tagger (C&C) [15] trained on the MedPost data [16], a NER tagger also based on the C&C tagger trained on documents with proteins annotated, and a maximum entropy RE model [17] trained on documents with protein-protein interactions annotated. Results are given for naïve Bayes ¹ using all features occurring at least 3 times and selecting the top 1500 features based on information gain. Numbers were all converted to the # symbol, punctuation was removed and Greek symbols were converted to their English equivalents. The term frequency is used as the weight for each feature. The prior probability used for the naïve Bayes classifier was modified to optimise the F_1 score.

¹Experiments were also run with SVMs and Maximum Entropy Models, but naïve Bayes performed the best. The reasons for this are being studied and may be the subject of a future paper.

Results are given for 10 fold cross validation using the training and heldout sets, for the heldout set when training on the training set and for the test set when training on the train and heldout sets. The test set was not used until the final evaluation.

3 Results and Discussion

The following experiments were run with results reported in Table 1.

Chunks We compare the results of using bigrams as features with those of using the chunks provided by the chunker as features under the hypothesis that the chunks will provide more meaningful groupings of words and thus higher performance.

NER We compare the performance of using the proteins identified using NER as features with the results of using words matching a list of 500,000 proteins names derived from RefSeq as features under the hypothesis that NER will provide a more reliable indication of proteins than a word list.

RE The RE module predicts protein-protein interactions and assigns a probability indicating the confidence of the interaction. This output can be used on its own to classify documents as containing protein-protein interactions or can be used as an additional feature for the text classification system. We compare the results of using standard text categorisation on its own, RE on its own, and results of combining both sets of features. For the combined results, we also experiment with weighting the features by their F_1 calculated as described in [2].

Features	Cross-Validation			Held Out			Test		
	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1
Chunk	54.4	64.0	58.6	57.0	60.8	58.8	55.0	71.0	62.0
Bigram	53.6	62.2	57.4	55.8	58.1	57.0	55.8	67.7	61.2
NER	54.0	69.4	60.5	54.1	71.6	61.6	57.7	76.3	65.7
Protein List	53.4	65.1	58.4	54.8	62.2	58.2	53.5	73.1	61.8
Text Categorisation Alone	54.4	64.0	58.6	57.0	60.8	58.8	55.0	71.0	62.0
RE Alone	54.8	71.9	62.1	59.0	66.2	62.4	55.6	69.9	61.9
RE Combined Simple	56.4	73.4	63.6	54.0	63.5	58.4	57.5	74.2	64.8
RE Combined F_1 Weighting	59.6	79.7	68.0	62.1	79.7	69.8	59.1	80.6	68.2

Table 1: Experimental Results

4 Conclusions

The chunker appears to provide a slight advantage over simple bigrams and NER provides an improvement over using a gazetteer. Not surprisingly, RE provides the greatest improvement. Adding features derived

from the output of the RE module to a text classification system and weighting the features in proportion to their individual F_1 scores resulted in an improvement of 10% over using either RE or text categorisation alone. Thus a system that combines the output of an NLP system with standard techniques of text categorisation can produce results that exceed the performance of either system on its own.

5 Acknowledgements

The work reported here was supported by the ITI Life Sciences Text Mining programme (www.itilifesciences.com).

References

1. Sebastiani F: **Machine learning in automated text categorization**. *ACM Computing Surveys* 2002, **34**:1–47.
2. Forman G: **An extensive empirical study of feature selection metrics for text classification**. *J. Mach. Learn. Res.* 2003, **3**:1289–1305.
3. McCallum A, Nigam K: **A comparison of event models for Naive Bayes text classification**. In *AAAI-98 Workshop on Learning for Text Categorization* 1998.
4. Joachims T: **Text categorization with support vector machines: learning with many relevant features**. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*. Edited by Nédellec C, Rouveirol C, Chemnitz, DE: Springer Verlag, Heidelberg, DE 1998:137–142.
5. Marcotte E, Xenarios I, Eisenberg D: **Mining literature for protein-protein interactions**. *Bioinformatics* 2001, **17**:259–363.
6. Polavarapu N, Navathe SB, Ramnarayanan R, ul Haque A, Sahay S, Liu Y: **Investigation into Biomedical Literature Classification Using Support Vector Machines**. In *CSB* 2005:366–374.
7. Donaldson I, Martin J, de Bruijn B, Wolting C, Lay V, Tuekam B, Zhang S, Baskin B, Bader GD, Michalickova K, Pawson T, Hogue CWV: **PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine**. *BMC Bioinformatics* 2003, **4**:11.
8. Yeh A, Hirschman L, Morgan A: **Background and overview for KDD Cup 2002 task 1: information extraction from biomedical articles**. *SIGKDD Explor. Newsl.* 2002, **4**(2):87–89.
9. Regev Y, Finkelstein-Landau M, Feldman R, Gorodetsky M, Zheng X, Levy S, Charlab R, Lawrence C, Lippert RA, Zhang Q, Shatkay H: **Rule-based extraction of experimental evidence in the biomedical domain: the KDD Cup 2002 (task 1)**. *SIGKDD Explor. Newsl.* 2002, **4**(2):90–92.
10. Blaschke C, Hirschman L, Valencia A: **Information extraction in molecular biology**. *Briefings in Bioinformatics* 2002, **3**(2):154–165.
11. Finkel J, Dingare S, Manning CD, Nissim M, Alex B, Grover C: **Exploring the boundaries: gene and protein identification in biomedical text**. *BMC Bioinformatics* 2005, **6**:S5.
12. Blaschke C, Valencia A: **The Frame-Based Module of the SUISEKI Information Extraction System**. *IEEE Intelligent Systems* 2002, **17**(2):14–20.
13. Hao Y, Zhu X, Huang M, Li M: **Discovering patterns to extract protein-protein interactions from the literature: part II**. *Bioinformatics* 2005, **21**(15):3294–3300.
14. Grover C, Tobin R: **Rule-Based Chunking and Reusability**. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy 2006.
15. Curran JR, Clark S: **Language Independent NER using a Maximum Entropy Tagger**. In *Proceedings of CoNLL-2003*. Edited by Daelemans W, Osborne M, Edmonton, Canada 2003:164–167.
16. Smith L, Rindfleisch T, Wilbur WJ: **MedPost: a part-of-speech tagger for bioMedical text**. *Bioinformatics* 2004, **20**(14).
17. Nielsen LA: **Extracting protein-protein interactions using simple contextual features**. In *BioNLP'06 Linking natural language processing and biology: towards deeper biological literature analysis*, Brooklyn, USA 2006.

IeXML: towards an annotation framework for biomedical semantic types enabling interoperability of text processing modules

Dietrich Rebholz-Schuhmann¹, Harald Kirsch¹, Goran Nenadic²

¹ European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

² School of Informatics, University of Manchester, Manchester, UK
rebholz@ebi.ac.uk, kirsch@ebi.ac.uk, g.nenadic@manchester.ac.uk

Abstract

Identification of biological, medical, chemical and other terms in the biomedical literature is the key to successful information extraction. Annotation of these semantic types is subject to ongoing research and new solutions are constantly being proposed. Unfortunately, the corresponding text processing modules usually fail to be easily incorporated into a different text mining infrastructure, because more often than not they follow special format requirements leading to substantial integration overheads. Furthermore, they are rarely ready to incorporate annotations of other modules (e.g. POS taggers or other named entity recognisers) to enable additional complex annotations (e.g. event identification). We have carefully analysed the requirements for modules in an information extraction pipeline to be interoperable. We propose a generalised framework that defines an annotation and data exchange format for a set of components (namely tokeniser, POS tagger, NER, etc.) that facilitate annotation of semantic types. The suggested annotation types include tokens, terms, chunks and sentences, and semantic categories are attributed to terms. In our annotation schema, the annotations follow a hierarchical order, whereas the order of text processing modules is not predefined. We define a basic but extendable set of tags and attributes that the modules need to provide in order to facilitate interoperability for annotation of semantic types. In the case of ambiguities, alternative annotations are gathered and kept in a list. We also propose an XML-based implementation (called IeXML) of the framework, which serves as a general exchange format between text mining modules. Exchange and composition of such modules will increase collaborative progress of research work done in the bio-text mining community.

1. Background

Information extraction (IE) and text mining (TM) from the scientific literature are complex tasks that require a number of processing resources (e.g. POS taggers, gazetteers, named entity recognisers, parsers, etc.). Identification of named entities and terms (such as names of genes, proteins, gene products, organisms, drugs, chemical compounds, etc.), in particular, is the key factor for accessing and integrating the information stored in the literature [10]. Although techniques for term identification are becoming more widely-used in biomedical and bioinformatics research [6, 8, 9], there are very few community-wide efforts to facilitate their interoperability [1, 3]. Interoperability is here regarded as the ability to combine modules and exchange data, meta-data and other resources to maximise their re-use [5, 7]. This is in contrast to widespread attempts to standardise exchange and semantic descriptions of non-textual biomedical data (e.g. through SBML [15]), ontological descriptions (e.g. through the OBO format [12]) and specifications of bioinformatics services (e.g. through the Taverna/myGrid framework [16]).

Still, at present, many researchers in bio-text mining openly contribute and disseminate data (e.g. annotated corpora such as Genia¹ or BioCreAtive²) or modules (such as POS taggers, parsers and specific named entity recognisers). Unfortunately, adaptation of available resources to a convenient format usually requires substantial efforts, since they have been designed to meet a particular aim and tend not to comply with any common data format, which leads to overheads of wrapping modules and adapting input and output formats. This has significantly reduced the ability of many users as well as researchers to reuse available modules.

Remote procedure calls, Web Services, common APIs or platforms are usual ways to achieve service-oriented interoperability of software modules. An example for an open source JAVA-based platform is IBM's UIMA (Unstructured Information Management Architecture³) that has recently been introduced to the bio-text mining community for creating and composing text processing modules and integrating their results. The UIMA platform primarily targets *solution development and integration* rather than interoperability of individual text processing components. It addresses the challenge of "standardizing and improving the process of development, deploying and integrating the operation of multiple text-analysis methods" [11], within the given platform. For example, UIMA provides various ready-to-use JAVA-based methods for reading, consuming or creating annotations, but their semantics is defined by and may not be shared between different applications.

Another approach to interoperability of bio-text mining processing modules is to rely on common data exchange formats that allow for all components to communicate with other components if they comply with the defined format(s). Several groups and projects have suggested specific processing formats (e.g. [4] or SciXML⁴), but there is not yet a common initiative to provide a community-wide solution that would improve interoperability and co-operation within the community. Previous efforts to standardise generic annotation of general-language corpora (e.g. Text Encoding Initiative (TEI)⁵, Corpus Encoding System (CES)⁶) have not been widely accepted in domain-specific text mining applications, as they have been mainly focused on resources that are manually developed.

³ <http://www.research.ibm.com/UIMA/>

⁴ <http://www-tsujii.is.s.u-tokyo.ac.jp/jw-tmnlp/simone.pdf>

⁵ <http://www.tei-c.org/>

⁶ <http://www.cs.vassar.edu/CES/>

¹ <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/>

² <http://biocreative.sourceforge.net/>

In this paper we propose a task-oriented data exchange framework that facilitates interoperability for annotation of semantic types, and provides a basis for other more complex annotations (e.g. event annotations). It includes the basic principles for marking up biomedical text to enable interoperability through a 'common exchange format', where various levels of mark-up are gathered within the document. We have identified the major tasks (modules) and defined a basic set of mark-up tags and attributes that enables interoperability. These are briefly discussed below (Section 2). We also propose an inline XML-based implementation (called IeXML) of the framework and illustrate it via a processing pipeline example (Section 3).

2. Towards a common annotation framework: requirements and principles

A common data exchange format should be general enough to allow all software modules to be interoperable, and – on the other hand – it has to be specific enough to gather all information provided by all modules contributing annotations to the text. Also, a common annotation schema has to account for the fact that selected modules rely on input from other text mining components (e.g. a shallow parser might expect POS information). These dependencies have to be kept in mind when building a framework for the annotation of natural language text.

It is the main aim of the framework presented here to facilitate the replacement of text processing modules that perform the same task and comply with the common exchange format. For a general text processing pipeline, this format has to serve all components in a pipeline, with minimal pre-assumptions on their dependences and ordering of their application. For example, tokenisation and separation of text into sentences are the basic text processing steps, and typically do not rely on any previous pre-processing, whereas chunk or dependency parsing as well as named-entity recognition (NER) could be based on POS annotation, but need not. Consequently, we can assume that tokenisation and sentence splitting are typically performed first, followed by POS tagging and NER in this or the reverse order, and that any further processing is completed afterwards. Further, we do not exclude that the annotations are modified at a later stage, as long as the common data format is preserved.

2.1. Basic principles for annotations provided by TM/IE modules

We have identified a set of tasks and corresponding modules that are typically used to support annotation of semantic types. These include: sentence finder, tokeniser, POS tagger and term tagger. The modularisation used here may be more fine-grained than encountered in some TM/IE systems today. This supports a precise specification of tasks as well as input/output behaviours, and should not hinder merging modules (e.g. tokenisation and POS tagging) as long as they match the requirements.

Our framework is based on a basic (minimal) set of conceptual tags and attributes that each component should provide. However, this should not preclude any module from producing other tags or attributes, nor should the tags represent any rigid meaning. As an example, we propose

how a token should be annotated to allow for the reuse of tokenisers, but we do not attempt to define what a token is, so not to restrict the choice in available tokenisers. Furthermore, in addition to identifying tokens, a tokeniser may supply additional attributes, not included in the basic set of attributes.

Any IE/TM module adds or changes tags in an input document to produce a result document by generating at least a set of basic elements for that type (as defined below). We distinguish between the following basic types of annotation elements:

- **tokens** (w elements) are minimal individual constituents of text; apart from words, numbers etc., these include punctuation, references, links, inline formulas, dates, measurements, etc. They do not contain any of the other three elements, and have c attributes that specify their category (e.g. POS information).

- **terms** (e elements) comprise one or more tokens denoting concepts, objects or entities, with possibly ambiguous semantics. This includes names, terminology and nomenclature strings or sequence mutations. Attributes are used to attach semantics (*sem* attribute) and POS information (c attribute). They can contain token and term elements only.

- **chunks** (c elements) denote syntax units, such as noun, verb or prepositional phrases. They can contain token, term and chunk elements, and their type is assigned via a t attribute.

- **sentences** (s elements) refer to sentence elements. They can contain any of the above types.

Here, the annotation types follow a hierarchical order, whereas the order of text processing modules (tasks) is not predefined. Given the above requirements and tag types, a brief summary of the descriptions of the identified tasks is as follows:

- 1) A **sentence finder** provides s elements to an input document.

- 2) A **tokeniser** is responsible for providing w elements to an input document.

- 3) **NER and term taggers** annotate documents with e elements and specify a mandatory *sem* attribute that points to semantic types (see also below).

- 4) A **POS-tagger** generates c attributes for any w and e elements found in a document.

Each component can amend the annotations existing in documents as long as the resulting annotation is in line with the relevant principles (e.g. a POS tagger may either merge multiple tokens and assign a joint POS tag (c attribute), or split a token into multiple units and assign POS tags to each of them). Also, a component may gather various alternatives using an *alt* (alternatives) element, which lists all alternatives as an *ali* (alternative item) element each (see also Section 3).

Furthermore, we demand that the original document should be completely recoverable from any result document. This enables client applications to display IE/TM annotations as semantic enrichment of documents.

In the following subsection we give more details about the requirements for term and POS taggers.

2.2. An annotation task example: term tagging

Although there are notable differences between terms and named entities [10], they are typically used uniformly in further processing steps and are referred here as terms.

Therefore, we suggest a common element (*e*) to denote terms, named entities and any other domain-specific sequences (such as mutation notations) that need a semantic attribute. Similarly, we refer to ATR (automatic term recognition) and NER modules as term taggers.

If we apply the above mentioned principles to a software component that is classified as a term tagger, then the software component has to adhere to the following requirements:

(1) Domain-specific terms are marked up as *e* elements with a mandatory attribute *sem* to specify the semantics of the term. (It is our intention to allow a yet unspecified set of possible values for the *sem* attribute.) The term tagger may assign POS-information as a *c* attribute to a term.

(2) If the term tagger is applied to a tokenised document (text containing *w* elements), each term must contain at least one token. If it is applied to an untokenised document, each term must contain tokenisable text that would result in at least one token if tokenised.

(3) The input document may already contain *e* elements. Furthermore, terms can be nested.

2.3. An annotation task example: POS tagging

In the case of a POS-tagger, the input to a POS-tagger is a tokenised document, and POS information is encoded with the attribute *c* assigned to each *w* element. This attribute is also added to every term (*e*) element. The POS-tagger may add POS information to tokens that are inside terms. Also, the POS-tagger may change POS information in tokens or terms that have already POS information assigned. The attribute value may be an empty string to signal that the POS tagger does not know which POS applies.

In general, this framework allows for modification of the order of applied text processing modules; only in the case of the POS-tagger we require that it receives a tokenised input document. This does not exclude modules that merge a tokenizer with a POS-tagger at the same time. Furthermore, our framework leaves enough flexibility to allow, for example, the introduction and specification of *c* attributes before the POS tagger is applied, or having tokens with or without POS information inside of terms, or transformation of multiple tokens into a single token ("multi-word tokens").

3. IeXML: towards an implementation of the framework

A variety of document formats are currently in use that have to be considered as possible inputs for text mining systems. Some of them do not follow an open standard initiative (e.g. commercial formats), while others are well established in both scientific and commercial domains (e.g. XML and Html). Although XML is well suited for interoperability of electronic data and although scientific literature is increasingly available in XML data formats, there is not yet a standard exchange format defined that harmonises available document formats and that considers semantic annotation and enrichment of text that is produced automatically by real-world applications [13].

We have developed a straightforward XML-based implementation (called IeXML) of the above annotation schema for biomedical text mining (see Figure 1 for an example). In our proposal, we have used the inline annotation approach where XML elements are integrated

into the text, as opposed to stand-off annotations where references to text components are kept separated from the text itself (either within or outside the document).

We have decided to start with simple inline annotations for various reasons. Firstly, according to a recent survey in the biomedical field [2], 50% of developers rely on inline annotation in their text mining solutions, showing that there are no preferences for any of the two annotation approaches. It is well known that inline annotation suits different text encodings (e.g. ASCII, UTF-8, Unicode, etc.), whereas stand-off annotations better overcome XML's limitations to describe ambiguous and overlapping elements [4, 13]. However, a vast majority of results of basic processing steps (such as sentence boundaries, tokenisation, POS) as well as annotation of semantic types can be represented with a limited level of ambiguities. Also, most pragmatic solutions in real-world applications resolve and scarcely represent (or use) complex ambiguities if they are not resolved at an early processing stage already. Furthermore, standard software tools are widely available to handle and visualise inline annotation (e.g. XML Document Object Model, XML DOM). Such tools are currently not or rarely available for stand-off annotation, since offset referencing is not yet standardised (there are some attempts to use token-based offsetting as opposed to character-based [4]). Lastly, stand-off annotations might raise substantial conflicts in a schema if document formats defined by the publishers of biomedical journals (e.g. Medline abstracts, Pubmed Central and Biomed Central documents) already contain stand-off annotations (e.g. MeSH terms, journal information, authors, date of publication, etc).

```

<s>
<e c="n" sem="uniprot:P50144,uniprot:P50145">
  <w c="n">Cholecystokinin</w>
</e>
<w c="cnj">and</w>
<e c="n" sem="uniprot:002686,uniprot:P01350">
  <w c="n">gastrin</w>
</e>s
<w c="v:P">differed</w>
<w c="prep">in</w>
<w c="n">stimulating</w>
<e c="n" sem="uniprot:002686,uniprot:P01350">
  <w c="n">gastrin</w>
</e>
<e c="n" sem="GO:0046903" onto="BP">
  <w c="n">secretion</w>
</e>
<w c="prep">in</w>
<e c="n" sem="species:9986">
  <w c="n">rabbit</w>
</e>
<w c="adj">gastric</w>
<w c="n:p">glands</w>
<w c="pun">.</w>
</s>
. . .

```

Figure 1: An excerpt tagged using the IeXML schema, after applying a sentence splitter, tokeniser, POS tagger and various term taggers

We have applied the IeXML schema to the problem of the annotation and disambiguation of semantic types using a cascaded approach [13]. A set of tools from different sites (EBI and the University of Manchester) is being

ported to comply with IeXML in order to assess benefits of combining various modules for annotation of semantic types. The existing modules (including separate term taggers for various semantic types (proteins, GO terms, drugs, species etc.) that are embedded in EBIMed [14]) are being ported to comply with IeXML. Our methodology is based on the idea of separating the process into clearly defined tasks applied one after another in a pipeline (workflow), where each module operates continuously on an input *stream* and performs its function on stretches or windows of text that are usually much smaller than the whole input. In this case, inline annotations are much more suitable than the stand-off approach. Communication of information between the modules is strictly downstream and all meta-information is contained in the data stream itself in the form of inline XML mark-up. Figure 1 presents an excerpt of a tagged document.

In the case of ambiguity, we suggest using a generic *alt* element. The *alt* element should span the minimal number of elements necessary to cover all alternatives. The modules may provide weight attributes for individual alternatives. More precisely, alternatives are represented inline as follows:

```
<alt>
  <ali>first alternative</ali>
  <ali>second alternative</ali>
  <ali>...</ali>
</alt>
```

For example, alternative term structuring (nesting) can be represented as in the following example:

```
<e c="n" sem="...">leukaemic
<alt>
  <ali>T <e c="n" sem="...">cell line</e></ali>
  <ali><e c="n" sem="...">T cell</e>
line</ali>
</alt>
</e>
```

4. Conclusions

Despite the huge efforts in developing and providing tools for the community, there are no common processing formats for the biomedical domain that could lead to wider availability and interoperability. In this paper we have suggested a framework to improve combining modules and results of annotation of semantic types, and to maximise their re-use. The suggested approach is task-oriented, as it involves the identification of basic modules that perform well-defined tasks in text processing and that can act as a part of a basic text processing workflow. For each such module, basic but extendable tag sets and basic attributes that the modules need to provide are defined.

We are currently developing an inline XML implementation of the framework. Note that an equivalent implementation can be done with the stand-off approach. Since some phenomena (e.g. sentences, tokens) can be represented more naturally using inline markup, while others are better modelled through stand-off annotations, we expect a combination of the two to be best suited for the annotation of biomedical texts.

The principles presented here allow natural extension to other module types such as parsers, anaphora resolution modules, etc. We believe that this approach could be used

as a basic step for a community-wide discussion on developing common exchange formats for biomedical text processing, which will improve interoperability, tool availability and reusability in the domain. Of course, after providing interoperability on the tag level, the next challenge for the community is to define the minimal common values for tags in order to enable semantic interoperability (i.e. to specify the values for the *sem*, *c*, *t* and other attributes).

Acknowledgements

This work has been partially supported by the Network of Excellence "Semantic Interoperability and Data Mining in Biomedicine" (NoE 507505), and BBSRC grant "Mining Term Associations from Literature to Support Knowledge Discovery in Biology" (BB/C007360/1).

References

1. Carroll J., Evans R, Klein E: **Supporting text mining for e-Science: the challenges for Grid-enabled natural language processing.** *UK e-Science Programme All Hands Meeting 2005*
2. Corpora Survey, http://compbio.uchsc.edu/corpora/Survey_Summary.htm
3. Grover C, Halpin H, Klein E, Leidner JL, Potter S, Riedel S, Scrutchin S, Tobin R: **A Framework for Text Mining Services.** *UK e-Science Programme All Hands Meeting 2004*
4. Grover C, Matthews M, Tobin R: **Tools to Address the Interdependence between Tokenisation and Standoff Annotation,** *Workshop on Multidimensional markup with XML (XMLNLP), EACL 2006.*
5. Interoperability Focus, www.ukoln.ac.uk/interop-focus/about/leaflet.html
6. Jensen LJ, Saric J, Bork P: **Literature mining for the biologist: from information retrieval to biological discovery.** *Nat Rev Genet.* 2006 Feb;7(2):119-29
7. Kirsch H, Gaudan S, Rebholz-Schuhmann D: **Distributed modules for text annotation and IE applied to the biomedical domain.** *International Journal of Medical Informatics.* (doi:10.1016/j.ijmedinf.2005.06.011), 2005
8. Krallinger M, Alonso-Allende Erhardt R, Valencia A: **Text-mining approaches in molecular biology and biomedicine.** *Drug Discovery Today* 10, 439-445 (2005).
9. Krallinger M, Valencia A: **Text mining and information retrieval services for Molecular Biology.** *Genome Biology*, 6 (7), 224 (2005).
10. Krauthammer M, Nenadic G: **Term identification in the biomedical literature.** *Journal Biomedical Informatics*, 37(6):512-26. 2004.
11. Mack R et al.: **Text analytics for life science using the Unstructured Information Management Architecture.** *IBM Systems Journal, Vol 43, No 3, 2004*
12. Open Biological Ontologies, <http://obo.sourceforge.net/>
13. Rebholz-Schuhmann D, Kirsch H, Gaudan S, Arregui M, Nenadic G: **Annotation and Disambiguation of Semantic Types in Biomedical Text: a Cascaded Approach to Named Entity Recognition.** *Workshop on Multidimensional markup with XML (XMLNLP), EACL 2006.*
14. Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, Stoehr P: **EBIMed – Text crunching to gather facts for proteins from Medline.** *Bioinformatics* (accepted)
15. Systems Biology Markup Language (SBML), <http://sbml.org/index.psp>
16. Taverna project, <http://taverna.sourceforge.net/>

Use of Text Mining for Protein Structure Prediction and Functional Annotation in Lack of Sequence Homology

Andreas Rechtsteiner^{*1}, Jeremy Luinstra², Luis M Rocha³, Charlie E M Strauss²

¹Center of Genomics and Bioinformatics, Indiana University, Bloomington, IN 47401

²Bioscience Division, Los Alamos National Lab, Los Alamos, NM 87545

³School of Informatics, Indiana University, Bloomington, IN 47401

Email: Rechtsteiner, A.* - andreas@cgb.indiana.edu; ; Rocha, LM - rocha@indiana.edu; Strauss, CEM - cems@lanl.gov;

*Corresponding author

Abstract

Background: Linking of information from different data sources, specifically literature, becomes increasingly important to annotate the growing number of new genome sequences. For the large percentage of genes with no known sequence homologs, new, possibly integrative, methods need to be developed. Ab-initio structure prediction and comparison is a method some of us pursued previously for functional annotation of sequences with no known homologs [1]. Here we use a large set of sequences of known structure to evaluate a new method that uses keyword information from literature to improve our previously used ab-initio structure prediction method.

Results: We report two results: first, the literature and keyword similarity measure we employ here performs well in identifying functional and/or structural relationships even if there is little or no sequence homology between the compared proteins, the difficult, but frequent, so-called “twilight zone” case in annotation and structure prediction. Second, our novel method that uses literature to assist SCOP super-family prediction [2] significantly improves on our original ab-initio structure prediction algorithm.

Conclusions: We show that the literature keywords and similarity measure used here are of great value for the increasingly important field of functional annotation of new sequences with no or little sequence homology.

Background

Each newly sequenced genome is principally annotated by comparison of its sequences to previously annotated genomes. Typically 40 to 60% of a new genome can be reliably annotated in this fashion. However, this method is most successful for the genes we often care least about, placing a premium on methods that can annotate unusual or highly diverged sequences. In this twilight recognition realm, ab-initio structure prediction based annotation has proven valuable [1]. By prediction of a protein’s approximate structure we can compare it’s structure to proteins of known function. Because this approach is less specific than sequence based annotation, it is useful to confirm ab-initio structure prediction based annotations by other means. Here we present a novel method that uses text mining to improve and screen genome-scale structure

predictions in an automated fashion assisting further human curation efforts.

We generate a body of text for test sequences from sequence-based comparison to the non-redundant UniProt sequence database by selecting the literature associated with the top BLAST hits and extracting MeSH keywords [3]. Literature is obtained similarly for the member sequences of all structural SCOP super-families [2]. The literature for each super-family is combined. We then rank the super-families by decreasing cosine keyword vector similarity for each test sequence (details of this cosine measure can be found in [5,9]). The keyword based and ab-initio structure based rankings are further combined in a single ranking with the rank-product method [6]. We find that the annotation (i.e. SCOP super-family) rankings based on this approach dramatically improve the annotation accuracy over structure based annotation alone. This is demonstrated on a large benchmark set of sequences with known structures that is carefully screened for homolog removal to simulate highly diverged sequences.

MeSH terms were shown to be useful for extracting functional information about genes or proteins previously. For example, Masys et al. [7] showed that MeSH terms associated (through publications) with two clusters of co-expressed genes were informative about the medical conditions of the gene expression samples. MacCallum et al. [8] extracted keywords for proteins from the SwissProt/UniProt database and used the cosine similarity to improve remote homolog detection over using sequence similarity alone. They evaluated their method on a set of 100 known remote homologs. Only SwissProt keywords from the exact match of the remote homolog candidates in the protein database was considered, however.

The work presented here extends work we reported previously [4,5,9]. There we showed the power of the keyword similarity method to infer functional relationships for close sequence homologs, i.e. to predict protein families that are based on sequence homology. Here we go beyond that and show that sequence similarity is not required for the keyword similarity method to detect functional (and/or structural) relationships among proteins, its potential usefulness is therefore wider than supported by our previous results.

Data and Methods

400 non-redundant test sequences with known structure were selected at random from 320 (randomly selected) SCOP super-families from all 4 main SCOP classes. We predicted 1000 model structures for each sequence using repeated runs of the Rosetta algorithm [10]. For each set of 1000 models, these were clustered for structural similarity into typically 20 clusters. The model structures of the cluster centers were compared with MAMMOTH [11] to a non-redundant set of known SCOP structures (i.e. Astral40) which resulted in a ranking of SCOP super-families based on decreasing structure similarity as reported by MAMMOTH. Details on the prediction algorithms and methods can be found in [10,11].

We used BLAST to find the top matches for the 400 test sequences in the UniProt [12] protein database. We proceeded similarly for the non-redundant (Astral40) member sequences of all 1280 SCOP super-families (ver. 1.63), except that we removed SCOP sequences if they had a BLAST e-value smaller than 5 to one of the test sequences and both were members of the same SCOP super-family. The goal of this filter was to test for the difficult twilight cases expanded on above, where we do not have close sequence homology between a sequence we want to annotate and the member sequences of the correct super-family. Literature references for the BLAST hits are obtained from UniProt. MeSH keywords¹ for these references were extracted from PubMed/MEDLINE [3]. After frequency based keyword filtering and weighting, the cosine similarity measure between the keyword vectors of the test sequences and each of the pooled SCOP super-family vectors were calculated and used to rank the super-families by decreasing similarity. Further details can be found in [4,5,9].

To combine the independent structure prediction and keyword similarity based super-family rankings we used the non-parametric rank-product method [6]. For a given test sequence i and super-family j we have the structure based ranking $s_{i,j}$ and the keyword based ranking $k_{i,j}$. We obtain a rank product score for sequences i and super-families j by calculating $rp_{i,j} = s_{i,j} * k_{i,j}$ for all i and j . We then rank for each test

¹Other sets of keywords are evaluate currently.

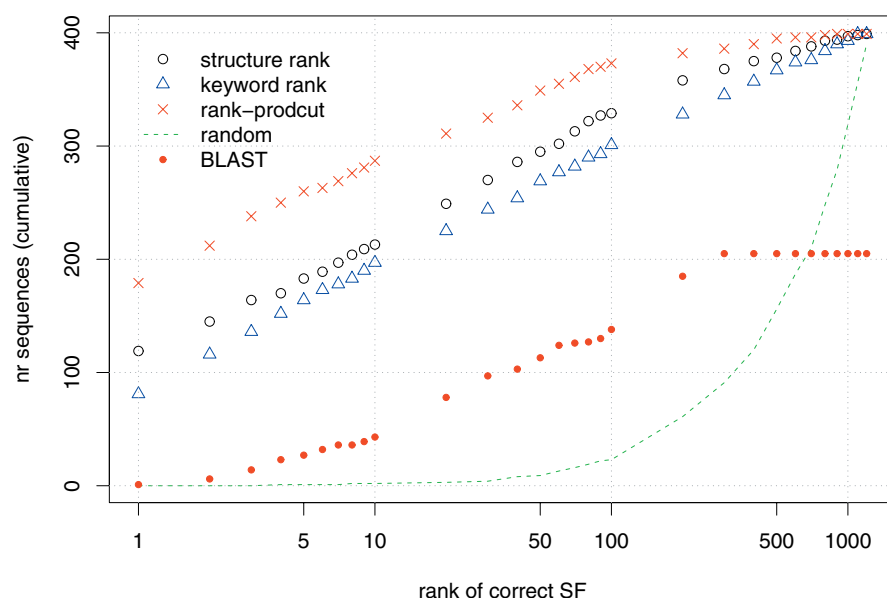


Figure 1: Ranking of correct SCOP super-families for 400 test sequences by the following methods: (i) BLAST e-value, (ii) keyword similarity, (iii) ab-initio structure prediction method, (iv) by combining the keyword and structure based rankings with the rank-product method, and (v), for comparison, picking and ranking super-families randomly. Our novel combined method performs significantly better than either the original structure prediction or keyword based prediction methods alone. The keyword method performs well even though the literature comes from sequences with little (BLAST) detectable sequence homology.

sequence i the SCOP super-families j by increasing $rp_{i,j}$ and obtain a new ranking of super-families based on the structure and the keyword ranking.

Results and Discussion

Figure 1 shows for our above explained 3 super-family prediction methods the number of sequences (y-axis) that had their correct super-family ranked in the position indicated by the x-axis or higher (i.e. better)². The structure based ranking performs slightly better than the keyword similarity based method by itself. The combined rank-product method improves prediction considerably over both methods individually. For example, a typical set of structure predictions provided to a human curator are the top 10 predictions made by Rosetta and Mammoth. Among the top 10 predictions of the structure based method the correct super-families for 210 of the 400 sequences can be found. For the combined method, the correct super-families for 285 sequences can be found, an improvement of 35%.

Also shown are results if super-families are ranked based on sequence similarity, by increasing BLAST e-value. Sequence homologs to the test sequences had been removed thoroughly, as BLAST was not able to rank the correct super-family at the top for a single test sequence³. This result shows very clearly the

²Note that the number of sequences on the y-axis is cumulative, e.g. the number of sequences shown for rank 5 includes all sequences that had their super-family ranked at position 5 or higher.

³The leveling off of BLAST rankings after about 200 sequences is due to the e-value cutoff of 10,000 (the database size was

strength of the keyword similarity method, as it is able to correctly indicate functional (and/or structural) similarity between proteins with no detectable sequence homology.

Conclusions

We show that literature keyword similarity measures can infer functional and structural relationships among proteins even if there is no, or very little, sequence homology among the respective protein sequences, an ability searched for by the community to make predictions in the difficult twilight zone. We were able to show further that our combined method of predicting structural super-families with ab-initio structure prediction performs significantly better than either method individually. Current and future work focuses on different keyword sets, like abstract keywords, and obtaining confidence measures for the keyword based predictions.

In conclusion, our results are encouraging to further pursue text mining for Bioinformatics for the challenging tasks at hand and to search for ways to link different methods and sources of information.

Acknowledgements

Judith Cohn set up the internal Protein Function Inference Group (PFIG) database at LANL which was very valuable to the tasks at hand. She helped further in obtaining data from the database and provided many useful comments. Lars Malmström assisted greatly in obtaining the ab-initio structure predictions for the test sequences. The CGB at Indiana University as well as Los Alamos National Lab, specifically the PFIG grant, funded the work by the authors.

References

1. Bonneau R, Strauss C, Rohl C, Chivian D, Bradley P, Malmstrom L, Robertson T: **De novo prediction of three-dimensional structures for major protein families.** *Journal of Molecular Biology* 2002, **322**:65–78.
2. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J. Mol. Biol.* 1995, **247**:536–540.
3. National Library of Medicine: **PubMed.** <http://www.ncbi.nlm.nih.gov/entrez/> 2006.
4. Rechtsteiner A, Rocha L, Strauss C: **Clustering of Protein Families in literature keyword space.** In *Currents in Computational Molecular Biology (RECOMB 2005)*, Boston, MA 2005.
5. Maguitman AG, Rechtsteiner A, Verspoor K, Strauss CE, Rocha LM: **Large-Scale Testing of Bibliome Informatics Using Pfam Protein Families.** In *Pacific Symposium on Biocomputing, Volume 11* 2006:76–87.
6. Breitling R, Armengaud P, Amtmann A, Herzyk P: **Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.** *FEBS Lett.* 2004, **573**:83–92.
7. Masys D, Welsh J, Lynn Fink J, Gribskov M, Klacansky I, Corbeil J: **Use of keyword hierarchies to interpret gene expression patterns.** *Bioinformatics* 2001, **17**(4):319–26.
8. MacCallum R, Kelley L, Sternberg M: **SAWTED: structure assignment with text description-enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons.** *Bioinformatics* 2000, **16**(2):125–129.
9. Rechtsteiner A: **Multivariate Analysis Of Gene Expression Data And Functional Information: Automated Methods For Functional Genomics.** *PhD thesis*, Portland State University 2005.
10. Rohl C, Strauss C, Misura K, Baker D: **Protein structure prediction using Rosetta.** *Numerical Computer Methods* 2004, **383**:66+.
11. Ortiz A, Strauss C, Olmea O: **MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison.** *Protein Science* 2002, **11**:2606–2621.
12. SIB/EBI: **UniProt/Swiss-Prot.** <http://www.ebi.ac.uk/swissprot/> 2004.

10,000 sequences)