

UNIVERSITY OF
CAMBRIDGE

Retrieving Hierarchical Text Structure from Typeset Scientific Articles

Bill Hollingsworth

Ian Lewin (presenter)

Dan Tidhar

University of Cambridge Computer Laboratory

Email: ian.lewin@cl.cam.ac.uk

E-Science Context

- FlySlip: *Integrating Literature, Experiments and Curation in Drosophila Genomics Research*
 - www.cl.cam.ac.uk/users/av308/Project_Index
 - BBSRC grant no: 16291
- CitRaz: *Rhetorical Citation Maps and Domain Independent Argumentative Zoning*
 - EPSRC grant no: GR/S27832/01

Overview

- Hierarchical text structure versus Presentational structure
- Value of hierarchical text structure to NLP
- PTX: PDF to XML processing framework
- An evaluation
- Where next?



***Drosophila* Tbx6-related gene, *Dorsocross*, mediates high levels of Dpp and Scw signal required for the development of amnioserosa and wing disc primordium**

Takashi Hamaguchi,^a Shigeharu Yabe,^b Hideho Uchiyama,^b and Ryutaro Murakami^{a,*}

^aDepartment of Physics, Biology, and Informatics, Yamaguchi University, Yamaguchi 753-8512, Japan

^bGraduate School of Integrated Science, Yokohama City University, 22-2 Seto, Kanazawa, Yokohama 236-0027, Japan

Received for publication 19 June 2003, revised 29 September 2003, accepted 29 September 2003

Abstract

Regional differentiation along the dorsoventral (DV) axis of the *Drosophila* embryo primarily depends on a graded BMP signaling activity generated by Decapentaplegic (Dpp) and Screw (Scw). We have identified triplicated Dpp and Scw target genes *Dorsocross1*, 2 and 3 (*Doc1*, 2, 3) that have a conserved T-box domain related to the vertebrate Tbx6 subfamily and act redundantly to induce dorsal structures. *Doc* genes are expressed in the dorsal region in the early blastoderm. After gastrulation, newly expressed *Doc* appears in a segmental pattern in the ectoderm. This expression correlates spatially with the second phase of Dpp expression in the ectoderm. *Doc* expression in the early blastoderm is abolished in either *dpp* or *scw* mutant embryos, whereas the ectodermal segmented expression depends only on Dpp. Inactivation of *Doc* genes with RNAi dramatically affected the development of amnioserosa and wing disc primordia, both of which depend on high levels of BMP signaling, although leg disc primordium, which depends on low levels of BMP, remained intact. *Doc1* RNAi expressed in *Xenopus* embryos induced ventral mesoderm, suppressed activin-induced events and induced *Xvent* genes, which are analogous to the effects of native *Tbx6* and its upstream regulator, BMP-4. These results suggest that the Tbx6 subfamily act in the BMP signaling pathway required for embryonic patterning in both animals.
© 2003 Elsevier Inc. All rights reserved.

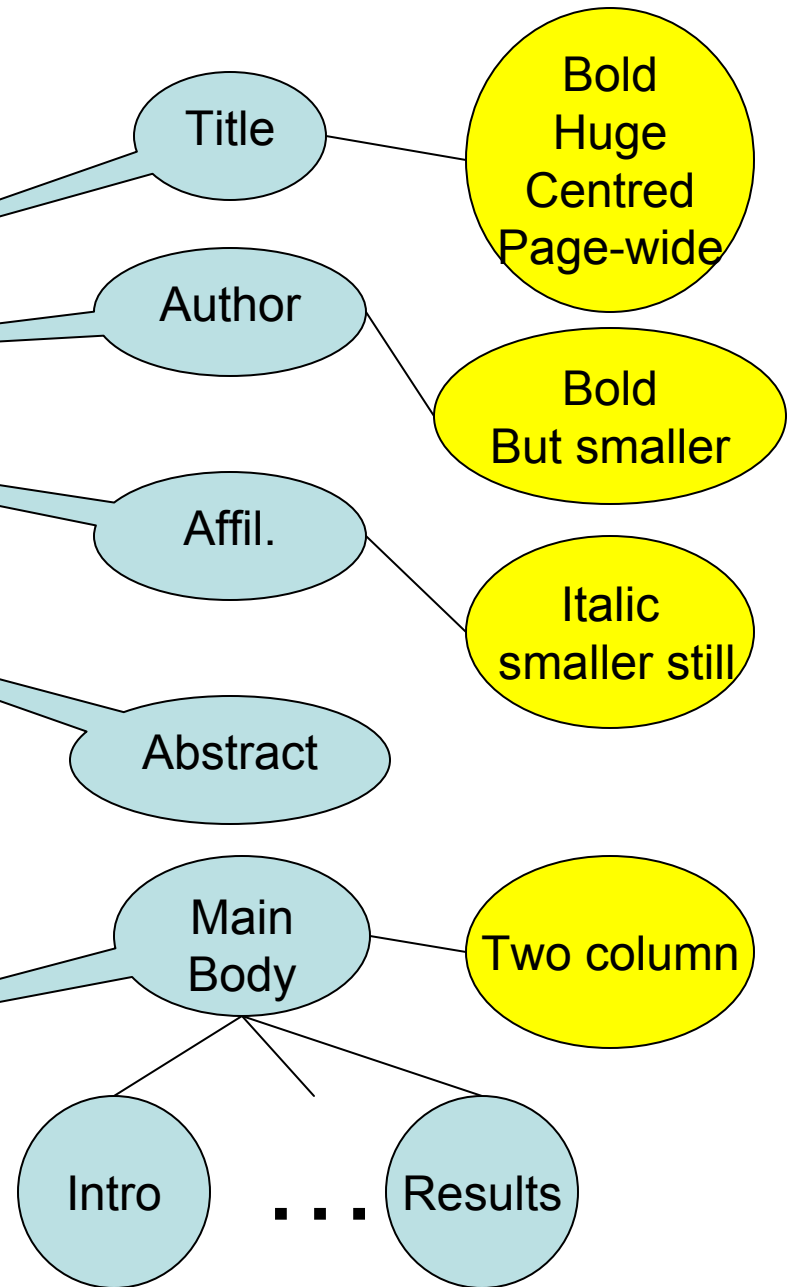
Keywords: Tbx6; *Dorsocross*; BMP; Dpp; Scw; pMad; Wing disc primordium; Amnioserosa; *Drosophila*; *Xenopus*

Introduction

T-box genes, which encode transcription factors characterized with a DNA-binding motif, are highly conserved across the two major animal groups, protostomes and deuterostomes, and even in *Hydra* (Bollag et al., 1994; Herrmann et al., 1990; Kispert et al., 1994; Pflugfelder et al., 1992b; Technau and Bode, 1999; Yasuo and Satoh, 1998). T-box genes have been further classified into several subfamilies, as outlined below (Papaioannou and Silver, 1998; Smith, 1999; Wattler et al., 1998). The *Brachyury* gene, a member of the T subfamily that has been identified in mouse and other deuterostomes, is a founding member of this gene family and plays an essential role in the development of the axial mesoderm (Herrmann et al., 1990). An ortholog of the *Brachyury* gene, *brachyenteron* or *byn* (also

known as *Trg* and *aproctous*), was identified in *Drosophila* and found to play an essential role in specifying the ectodermal hindgut (Kispert et al., 1994; Murakami et al., 1995; Singer et al., 1996). Three other subfamilies, Tbx1, Tbx2 and H15, are also conserved in both vertebrates and *Drosophila* (Bollag et al., 1994; Brook and Cohen, 1996; Griffin et al., 2000; Pflugfelder et al., 1992b; Porsch et al., 1998). While the members of the Tbx6 subfamily play important roles in specifying various mesodermal and endodermal tissues in vertebrates (Chapman and Papaioannou, 1998; Horb and Thomsen, 1997; Trug et al., 1997; Kimelman and Griffin, 1998; Lustig et al., 1996; Mitani et al., 1999; Stennard et al., 1996; Uchiyama et al., 2001; Zhang and King, 1996), until now, no corresponding genes had been identified in *Drosophila*. We have identified and analyzed *Dorsocross1* (*Doc1*, formerly named *Doc*, accession No. AB035412), a *Drosophila* T-box gene that is related to the vertebrate Tbx6 subfamily. *Doc1* and its homologues, *Doc2* and *Doc3*, have also been described in

* Corresponding author. Fax: +81-83-933-5696.
E-mail address: ryu@yamaguchi-u.ac.jp (R. Murakami).

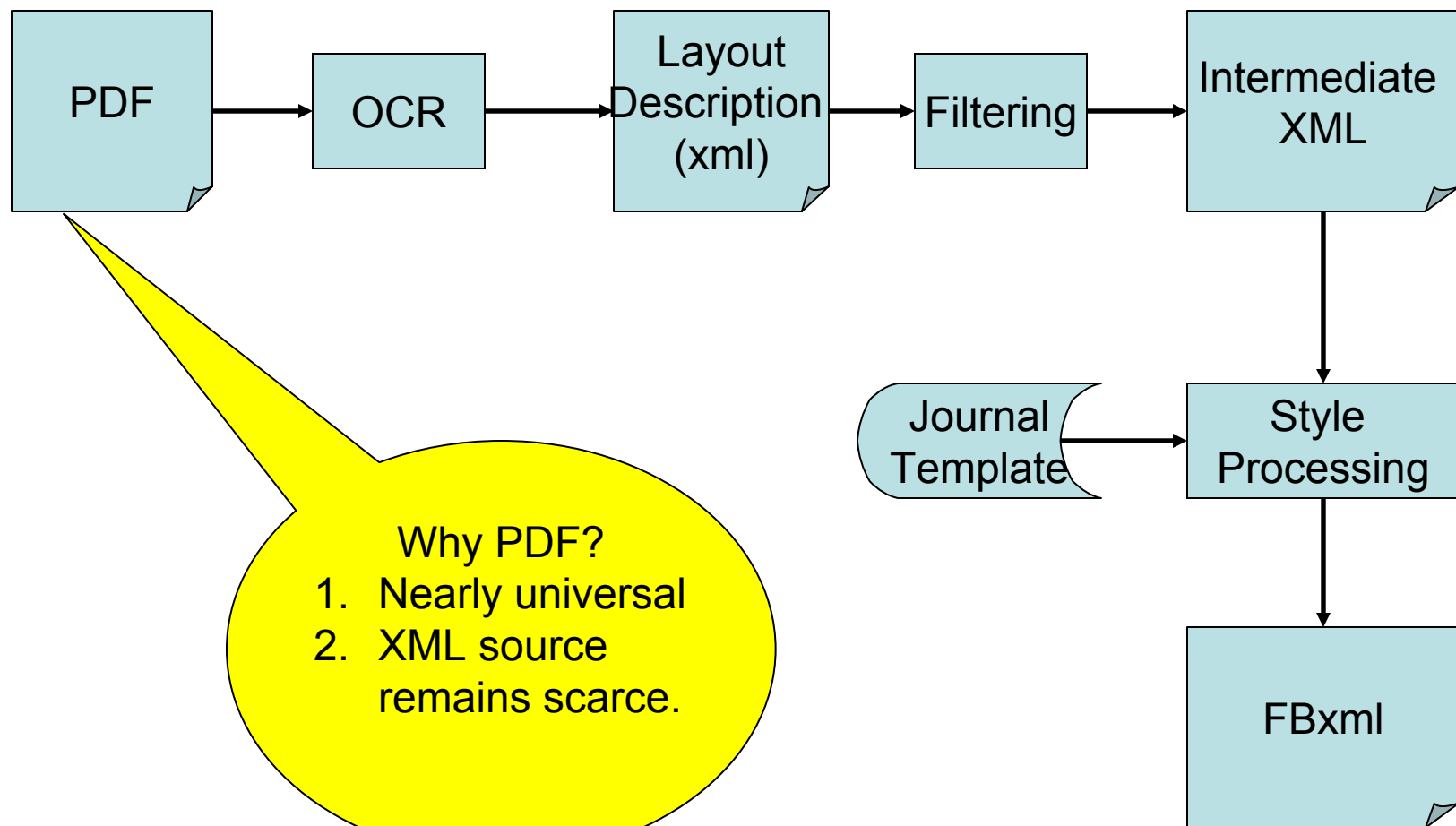


Hierarchical Structure for NLP

- Information Retrieval, Extraction, Classification
 - Build patterns over parsed structures over words
 - Build chains of co-reference between named entities
 - Discern which structures (abstract, results, methods) patterns are matched in
 - Label parts of structure (e.g. in *Discussion*, this is Background information, this is others' results, this is new results)

- Information re-presentation. Present an article back to the user but with “added value”
 - Gene names highlighted
 - Navigation by author’s sectioning
 - Navigation by repeated references
- Linkage of article to FlyBase database contents

PTX : processing framework



- Why PDF?
1. Nearly universal
 2. XML source remains scarce.

Why OCR?

- Standard pdf “text extraction” tends to mishandle or omit important information
- Textual characters that are encoded graphically
- Styles (bold, italics, positional info) that are important for recovering *structure*.
 - *Single line paragraphs in italics are (sub) headings*
 - *Centred italics after authors are affiliations and not the beginnings of abstracts*
 - *Small font, small size paragraphs at foots of pages not following a full stop (or other sentence delimiter) are likely footnotes*
- Styles important for content. In genomics, italicized words often indicate genes, superscript suffixes indicate alleles

Layout XML

- Documents
- Pages
- Regions
- Paragraphs
- Lines
- Words
- Characters
- Stylistic annotations on these elements
 - Font size
 - Font type
 - Indentation
 - Alignment
 - coordinates

...

```
<paragraph para-type="text" align="justified" left-indent="0" right-indent="72" >  
<ln baseline="2511" ff="Times New Roman" fs="1000">  
<wd l="1243" t="2381" r="1579" b="2520" char-attr="italic">  
<ch l="1243" t="2381" r="1397" b="2515">D</ch>  
<ch l="1402" t="2424" r="1493" b="2520">o</ch>  
<ch l="1498" t="2424" r="1579" b="2520">c</ch>  
</wd>  
<wd l="1661" t="2419" r="2098" b="2558">  
<ch l="1661" t="2419" r="1752" b="2558">g</ch>  
<ch l="1757" t="2419" r="1838" b="2520">e</ch>  
<ch l="1843" t="2419" r="1944" b="2515">n</ch>  
<ch l="1944" t="2419" r="2026" b="2520">e</ch>  
<ch l="2035" t="2419" r="2098" b="2520">s</ch>  
</wd>  
<wd l="2179" t="2419" r="2414" b="2520">  
<ch l="2179" t="2419" r="2266" b="2520">a</ch>  
<ch l="2261" t="2419" r="2333" b="2515">r</ch>  
<ch l="2333" t="2419" r="2414" b="2520">e</ch>  
</wd>
```

...

Journal Template Processing

- Generic processing: Top down, left to right traversal of the layout description
- Journal specific: attach program code that executes when certain tags (new zone, zone finishes, new page, new paragraph...) are encountered

Current....

- Programming framework that enables structure discovery by the writing of good rules
 - “a change of zone with font-style change when processing the abstract indicates that the abstract has finished”
 - “a change of zone where top-of-zone is top of page and ‘high enough’ and zone encompasses a one line paragraph indicates this is a page header”

Near future

- We're moving to a more structured approach with a generic *scientific document* model (expressed as a finite state machine) and a predefined set of document features whose detection permits state transition
 - Simpler to generate new journal templates
 - Possible to learn templates from examples
 - Possible to learn common mistakes and correct

Evaluation

- Acid test will be: do the FlyBase curators find the outputs *usable*?
 - There will be mistakes; some could be crippling; others could be irrelevant
- Meantime, we can measure some coarse statistics
 - F-Measure on structure tags
 - Character recognition & style accuracy
 - Greek characters
 - Superscripts/subscripts

Evaluation

<TITLE>Proteoglycan UDP-Galactose:/3-Xylose
/31,4-Galactosyltransferase I Is Essential for
Viability in <i>Drosophila melanogaster*</i>
</TITLE>

<ABSTRACT>Heparan and chondroitin sulfates play
essential roles in growth factor signaling during
development and share a common linkage
tetrasaccharide structure,
GlcA β 1,3Gal β 1,3Gal β 1,4Xyl β 1-*0*-Ser. In the
present study, we identified the <i>Drosophila</i>
proteoglycan| UDP-galactose: β -xylose
 β 1,4-galactosyltransferase I (d β 4GalTI), and
determined its substrate specificity. The enzyme
transferred a Gal to the - β -xylose (Xyl) residue,
confirming it to be the <i>Drosophila</i> ortholog
of human proteoglycan UDP-galactose: β -xylose
 β 1,4-galactosyltransferase I.

Does “abstract” occur
in the same context
window (upto 4 words
either side) in a
Gold Standard
annotation?

Does β occur in the
same context window
in an HTML version of
this published paper?

Evaluation results

- Structure:
 - Journal A (dev. set of 8 papers)
 - 18 papers from same year
 - Precision = 97.5 Recall = 96.2
 - 10 papers from previous year
 - Precision = 93.0 Recall = 94.2
 - Journal B
 - Precision = 73.9 Recall = 90.4
- Styles
 - Superscript: Precision = 100% Recall = 97.6%
 - Greek characters: Precision = 73.1%

Evaluation

- Structure Recognition – generally good
 - Many mistakes in tabular data which OCR misrecognizes as tiny textual paragraphs (looks correctable)
 - Some mistakes where citation strings in *another* English font is mis-recognized as Greek font (may be less easy to correct!)
- Superscript/subscript recognition excellent
- Greek character recognition generally good (bar citation string problem)

Future Work

- Develop generic “scientific document” model; ideally train journal specific variations via machine learning
- Find out what really matters to FlyBase database curators
- Incorporate feedback into curation cycle for iterative NLP improvement
- Incorporate images and tables