ISSN 1346-5597

NII National Institute of Informatics

NII Technical Report

Annotation of Biomedical Texts for Zone Analysis

Yoko MIZUTA, Tony MULLEN and Nigel COLLIER

NII-2004-007E
Oct. 2004

URL: http://www.nii.ac.jp/ Tokyo 101-8430, JAPAN +81-3-4212-2000 FAX: +81-3-3556-1916

Annotation of Biomedical Texts for Zone Analysis

Yoko MIZUTA, Tony MULLEN and Nigel COLLIER
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku,
Tokyo, Japan, 101-8430
{ymizuta, mullen, collier}@nii.ac.jp

Abstract

This document provides the framework of the annotation scheme of biomedical texts for zone analysis (ZA, Mizuta and Collier 2004a, 2004b) in the hope that this annotation scheme will be useful for enabling improved access to information contained in biomedical texts. It is intended to serve as a set of guidelines for building annotated corpora for ZA and is neutral to the mark-up language/tool to be used. We briefly introduce our set of zones and describe the procedure for the annotation task. In the Appendices, we provide practical knowledge about the annotation task in terms of solutions to controversial cases, and a sample of annotated full text accompanied by notes. We also provide the information about our dataset of 20 full text articles from four major online journals in the biological domain (i.e. EMBO, PNAS, NAR, and JCB).[1]

1 Introduction

In (Mizuta and Collier, 2004a, 2004b), we proposed annotating texts into rhetorical zones with a view to enabling improved access to information contained in biomedical texts, taking Teufel et. al's work (Teufel, Carletta, and Moens, 1999; Teufel and Moens, 1999, 2000, 2002) on text summarization in the domain computer science as a starting point. In contrast with discourse-oriented rhetorical analysis of texts such as Rhetorical Structure Theory (Mann and Thompson, 1987; Marcu et. al, 2002), which focus on the logical relations between sentences in a hierarchical structure, we focus on the *global* rhetorical status of the elements of texts and propose a set of zone classes with respect to the problem-solving process, intellectual attribution, and scientific argumentation. An annotated text is a sequence of zones with a shallow nesting.

In what follows, we provide the framework of our annotation scheme for zone analysis (ZA) as guidelines for an annotation task.

2 Zones

2.1 Zone classes

A total of ten zone classes are defined as belonging to three groups. These are the classes which can be used in an annotation. The OWN class is exceptional in that its subclasses, rather than OWN itself, are used in annotation.

Group 1: This group concerns major elements with respect to the problem-solving process, intellectual attribution, and scientific argumentation. The zone classes and their coverage are as follows.

BKG (Background): given information (reference to previous work; general assumptions)

PBM (Problem-setting): a problem or an open issue which the authors identify or introduce, and which motivates the authors' work presented in the paper. Typically, it's the goal of the present research/paper and the goal of a specific experiment performed.

OWN: various aspects of the authors' own work:

- **MTH** (Method): statements about experimental procedure and materials used;
- **RSL** (Result): experimental results as observed;
- **INS** (Insight): the authors' interpretation of the results in terms of a biological process or the role of a biological entity behind the observed results, i.e. insights and findings. It also applies to the authors' insights from previous work.

[1] The dataset is subject to copyright permission: Copyright permission for these articles are currently pending discussion with the publishers.

- **IMP** (Implication): various kinds of implications of the authors' work described in the present paper, typically those of the authors' experimental results (e.g. assessment, applications, limitations, future work). It also covers the authors' conjectures and hypotheses to be examined.
- **ELS** (Else): any other kind of information within OWN (e.g. the naming of a gene discovered by the author).

Group 2: This group deals with comparative or contrasting relations between items which fit into Group 1 classes.[2] Specifically, similarities or differences are described between results, insights, etc. presented in the work at hand and between items pertaining to the work at hand and those pertaining to previous work.

CNN (Connection): correlation, consistency
DFF (Difference): contrast (and more generally, comparison), inconsistency

Group 3: This group concerns statements about the paper/work at hand. It consists of one zone class.

OTL (Outline)[3]: a characterization or summary of the paper; excerpts from the paper; statements about the section organization (e.g. "Section 2 provides ……"). [4]

2.2 Distinction between zone classes

Some zone classes are not easy to distinguish from each other on the surface or on a purely intuitive basis. Here are some clarifications.

2.2.1 INS vs. IMP

INS concerns the authors' idea and/or finding obtained from their own experimental results. Importantly, statements qualifying for INS should be made with certainty and from an objective point of view: Otherwise, they fit into IMP (as 'weaker' insights or conjectures).

The following sentence qualifies for INS, because the hypothesis now receives a positive evaluation.

(1) Our results supports the hypothesis that ~.

In contrast, statements made in a modal context (introduced by *could, may, might*, etc. and adverbials such as *probably, perhaps*, etc.) fit into IMP.

2.2.2 PBM vs. IMP

Statements of a problem or an open issue may well qualify for either PBM or IMP: The distinction depends on the status of the problem. If the problem motivates the authors' present work, then the statement qualifies for PBM. If it motivates the authors' (or somebody else's) future work, then the statement qualifies for IMP.

2.2.3 CNN and DFF

As mentioned in Section 1, CNN and DFF are only considered between one element in some class under OWN and another element in the same class under OWN or an element in BKG. In other words, relations between elements in different classes under OWN are out of concern. For example, correlation between the authors' results (i.e. between elements in RSL) qualifies for CNN, but a cause-effect relation between an experiment and its result (i.e. between elements in MTH and RSL) does not. Neither is correlation between a result and the authors' insight obtained from it (i.e. between RSL and INS).

For example, sentence (1) relating the authors' results (c.f. RSL) to a hypothesis (c.f. IMP) does not license a CNN zone. (The sentence qualifies for an INS zone only.)

3 Annotation

3.1 Scope of annotation

Both abstracts and full texts are the scope of annotation.

For full texts, only the four main sections are to be considered, which are 'INTRODUCTION', 'RESULTS', 'DISCUSSIONS', and 'METHODS and MATERIALS'. Currently, only the main text is considered for annotation. Figures and Tables as well as their legends are out of the scope.

The section and abstract headings are annotated as 'SECT'. The main text is annotated using the ten zone classes defined in Section 1. If some unit lacks a corresponding zone class, it will be left unannotated.

[2] In many cases, however, the items at issue are *not* annotated as zones in their own right. Because, for example, they are mentioned only in a noun phrase or in a citation.

[3] TXT (Textual) proposed in an earlier work (Mizuta et. al 2004a) has been incorporated into OTL.

[4] OTL does not apply in the case of abstracts or summaries.

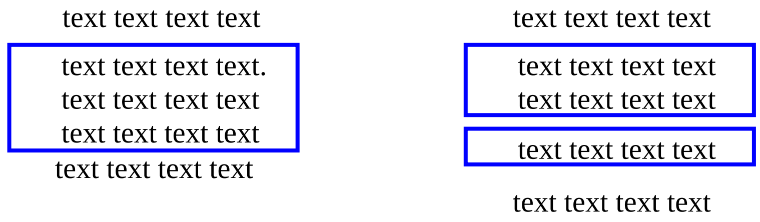
3.2 Considerations of the context

The content and the rhetorical status of an element of a text is dependent on the context in which it appears. Therefore, it is necessary to look at a wider span of text in order to figure out the correct zone class for the element under consideration. The recommended way is to first scan a paragraph for an overview of its content and then to annotate its elements.

3.3 Unit of annotation

3.3.1 Overview

Generally speaking, annotators may proceed sentence by sentence. If a sentence fits semantically into a single zone class, it qualifies as this zone. If adjacent sentences fit into the same class, they may be annotated either together (Fig. 1a) or separately (Fig. 1b). The choice is left to the annotator's convenience: in some cases, s/he may want to annotate these sentences little by little (e.g. in some groups reflecting the discourse structure), and in other cases, s/he may want to annotate the whole sequence to save time. Both versions of annotation equally mean that each component sentence comes with the context of that zone class.



a. A combined annotation **b.** Separate annotations

Figure 1. Options for Annotation

There are some exceptional cases where larger areas of text qualify as zones in a non-compositional way, that is, cases where a sequence of sentences *as a whole* constitutes a single zone. This is typical of an OTL zone as we usually observe it at the end of an INTRODUCTION section. To illustrate with an example:

(2) We report here the results of experiments that directly address both of these questions. In brief, we have asked …… To address the first question, we utilized affinity isolation of Scp160p-associated mRNPs, followed by microarray and quantitative RT?PCR analyses of the mRNAs released from these complexes. We found ……Together, these results not only confirm that Scp160p associates with specific mRNAs in yeast, but also that these interactions are biologically meaningful.

Here a sequence of sentences, but not a component sentence by itself, constitutes an outline of the paper. The annotation of such a sequence should follow along the lines of a combined annotation, as illustrated in Fig.1a above. That is, the whole sequence should receive a single annotation.

In other cases, a single sentence may involve more than one class. There are two cases.

If the whole sentence simultaneously fits into two zone classes, then it results in combined zones (Section 3.5). Annotate the sentence as nested zones in whichever order.

In contrast, if the sentence sequentially fits into multiple zone classes (i.e. if the sentence consists of multiple constituents fitting into different zone classes), then it results in a sequence of different zones. Note however that only a certain types of constituents qualify for an independent zone and therefore get separate annotation.[5] Given this, the annotator should check the type of constituents to see if it deserves of an independent zone (Sections 3.3.2 and 3.3.3). For example, the following sentence in the form of 'Although *S1*, *S2*' (, where *S1* and *S2* are clauses) gets annotated as a sequence of MTH and PBM zones, as follows, because by definition (Table 1) subordinate clauses such as this deserve of an independent zone.

(3) **[** Although both we and others have hypothesized previously that Scp160p associates with mRNAs in vivo (2?5), **]$_{\textbf{BKG}}$** **[** whether those mRNAs are random or specific, and whether these associations are biologically significant, has remained unclear. **]$_{\textbf{PBM}}$**

However, the annotation below is incorrect, because the subject NP is not the appropriate constituent type to be annotated in its own.

[5] For useful discussions on the unit of annotation in different frameworks, see Carlson et. al (2001) and Prasad et. al (2003).

(4) **[**Parallel reverse transcription reactions using total RNA isolated from whole cell soluble lysates of both strains**]** $_{\textbf{MTH}}$ **[**resulted in indistinguishable strong smears.**]** $_{\textbf{RSL}}$: microscopic annotation

Instead, the whole sentence should be annotated as a RSL zone.

3.3.2 Types of constituents deserving of an independent zone

The following provides a complete list of constituent types which may be annotated as belonging to an independent zone. Other types of constituents do *not* qualify for an independent zone. The content of such constituents will be ignored: only a larger constituent which includes the constituent should get annotated.

Table 1. Constituent types which license an independent zone[6]

- A sequence of sentences
- A sentence
- Coordinate clauses
  e.g. 'A but B' ⇒ [A] [but B]

(1) **[** The molecular details of the role of AdoMet in cleavage are not yet clear, **]** $_{\textbf{PBM}}$ **[but** it has been suggested that AdoMet binding causes conformational changes in the restriction enzyme which are essential for cleavage (16).**]** $_{\textbf{BKG}}$

Note 1. The second zone includes the conjunct *and*.
Note 2. Other coordinate conjuncts: *and*, *whereas*, *while*(used for a contrast)

- Subordinate clauses
  e.g. 'A when B' ⇒ [A] [when B], 'When B, A' ⇒ [When B,] [A]

(2) **[** When an assay containing 4.2 pmol of R.*Eco*P15I and supercoiled pUC19 DNA containing 4.6 pmol of *Eco*P15I sites was carried out, **]** $_{\textbf{MTH}}$ **[** the kinetics of cleavage showed that most of the DNA was cleaved within 20 min (Fig. 2B, lane 1). **]** $_{\textbf{RSL}}$

Note. Other subordinate conjuncts: *because, since, although, when* etc.

- Nonrestrictive relative clauses
  e.g. [….., [which ~,] …]

(3) **[** Based on the results described here, **[which indicate** that the efficiency of restriction enzymes increases with decreasing affinity for cleaved DNA,**]** $_{\textbf{INS}}$ **[** we propose a functional evolutionary hierarchy for R-M systems illustrated schematically in Figure 5. **]** $_{\textbf{IMP}}$

Note. The inserted clause constitutes an embedded zone. Restrictive relative clauses (i.e. those provided without a comma) do not qualify for an independent zone

- Present or past participle version of nonrestrictive relative clauses
  e.g. [...,] [indicating that ~.]
  cf. […,] [which indicates that ~.]

(4) **[** Initially, the amount of DNA cleaved is directly proportional to the enzyme concentrations (Fig. 1, phase A),**]** $_{\textbf{RSL}}$ **[indicating that** the R.*Eco*P15I catalysed reaction is stoichiometric with respect to enzyme concentration and that it performs a single round of catalysis *in vitro*.**]** $_{\textbf{INS}}$

- *to*-infinitives expressing the goal/ purpose or the result of what's stated in the remainder of the sentence

(5) **[To** address the question of a potential involvement of the flagellum in the trypanosome cell cycle,**]** $_{\textbf{PBM}}$ **[**we decided to perturb flagellum formation~.**]** $_{\textbf{MTH}}$

Note. The *to*-infinitive expresses the purpose of the experiment performed (and therefore qualifies for a PBM zone), whereas the remainder describes the experiment (and therefore qualifies for a MTH zone).

3.3.3 Examples of constituents NOT deserving of an independent zone

Below are some examples of 'smaller' constituents which should *not* get annotated in their own right. 'Microscopic view' illustrates the incorrect annotation. (See the next subsection for nested zones.)

[6] Sentence numbers in Table 1 are separated from those in the main text.

(5) The correct annotation: a single RSL zone

[A progressive reduction in both (TbDHC1b) and (TbIFT88) RNAi mutant growth rates was noticed **during the course of induction of RNAi** (Figure 1F), **]**$_{\textbf{RSL}}$

Microscopic view (incorrect annotation): with an embedded MTH zone

[A progressive reduction in both (TbDHC1b) and (TbIFT88) RNAi mutant growth rates was noticed **] [during the course of induction of RNAi]**$_{\textbf{MTH}}$ (Figure 1F), **]** $_{\textbf{RSL}}$

(6) The correct annotation: a single RSL zone

[This cellular organization was severely perturbed **in non-flagellated cells** .... **]**$_{\textbf{RSLe}}$

Microscopic view (incorrect annotation): with an embedded MTH zone

[This cellular organization was severely perturbed **[in non-flagellated cells]**$_{\textbf{MTH}}$ ... **]**$_{\textbf{RSL}}$

(7) The correct annotation: a single RSL zone

[Parallel reverse transcription reactions using total RNA isolated from whole cell soluble lysates of both strains resulted in indistinguishable strong smears.**]** $_{\textbf{RSL}}$

Microscopic view (incorrect annotation): a sequence of MTH and RSL zones

[Parallel reverse transcription reactions using total RNA isolated from whole cell soluble lysates of both strains**]** $_{\textbf{MTH}}$ **[**resulted in indistinguishable strong smears.**]** $_{\textbf{RSL}}$

3.3.4 Summary

A single sentence or a sequence of sentences may receive multiple annotations, in parallel or in a sequence, if it semantically fits into more than one zone class (and if the constituents at issue are of the type deserving of an independent zone, in the case of a sequential annotation).

3.4 Nesting

Elements within a zone may form a zone in its own right. The passage below illustrates such a case.

(8) We report here the results of experiments that directly address both of these questions……**We found not only that yeast mRNA sequences are present in these samples, but also that the sequences present are specific, not random.**……. Together, these results not only confirm that Scp160p associates with specific mRNAs in yeast, but also that these interactions are biologically meaningful.

The passage as a whole forms an OTL zone, and a boldfaced sentence within it qualifies for a RSL zone by itself. This results in nested zones as illustrated in Figure 2 below.

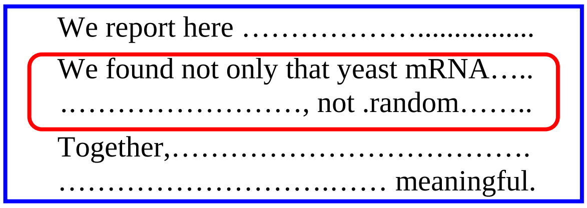
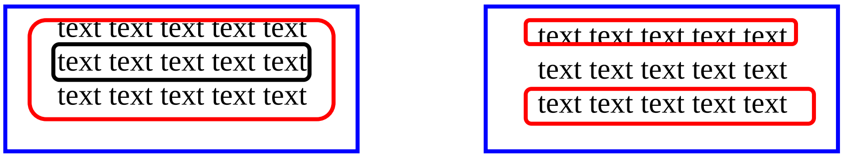

We report here ……………….............
We found not only that yeast mRNA…..
……………………, not .random……..
Together,………………………………
…………………….…… meaningful.

Figure 2. Nested Zones (two-level)

For practical reasons, nesting is limited to two levels within each group provided in Section 1. So, semantically most important zones should be identified, if candidate zones involve three-level nesting (Fig.3a). So long as nesting is at two-levels, there may be any number of embedded zones (Fig.3b).



a. Ill-formed (three-level) **b.** Well-formed (two-level)

Figure 3. Ill-formed and Well-formed Nesting

3.5 Combined zones

The same constituent may fit into more than one zone classes simultaneously. For example, the sentence below qualifies for a RSL (Result) zone, because it provides an experimental result, and also for a CNN (Connection) zone, because it compares the result with other results:

(9) Similar DNA links were also observed in the complexes with … (ref.), which show structural similarities with….

Such a case results in 'combined zones'. For practical reasons, we treat combined zones as a special case of nested zones, that is, nested zones which have the identical scope and which are not sensitive to ordering (i.e. outer/inner). Thus, combined zones are also limited to two levels.

text text text text
text text text text

a. Annotation 1 **b.** Annotation 2

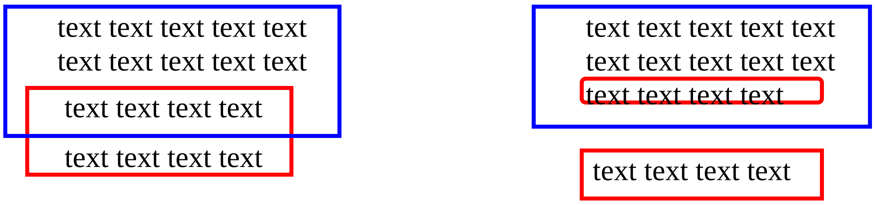
Figure 4. Combined zones as nested zones

3.6 Other constraints

3.6.1 Zone boundaries

Zones at a level smaller than a sentence should be annotated separately. Thus, if such a smaller zone is preceded (or followed) by a sentence-level zone of the same class, the two zones should be annotated separately, even if they fit into the same zone class. A single annotation of such whole sequence (e.g. a sentence plus a *to*-phase in the following sentence) is ill-formed.

In the case of embedding, the inner zone should close first. That is, annotation should not go across a zone boundary. Partial overlap should be resolved by dividing the bridging zone into smaller ones, as follows.



a. Ill-formed **b.** Well-formed

Figure 5. Zone boundaries

3.6.2 Scope of CNN and DFF zones

CNN and DFF semantically have two arguments (i.e. elements which are related to each other).

When a minimal unit of annotation contains both arguments, the scope of CNN/DFF is that unit:

(10) **[Consistent with** the low level of c-Myc in Mnt$^{-/-}$ MEFs at passage 4 (Figure 4), it was difficult to detect c-Myc and Max bound to Cdk4 proA in these cells (Figure 5B).]$_{\text{CNN}}$

However, the arguments may be provided in a separate annotatable unit as follows:

(11) (element 1). In contrast, (element 2).
(element 1), whereas (element 2).
Whereas (element 2), (element 1).

Whereas it is natural to annotate the whole sequence as DFF, what we proposed is to annotate only one part, including the discourse connective etc. attached to it (e.g. *in contrast*, *whereas*):

(12) (element 1). [In contrast, (element 2).]$_{\text{DFF}}$
[(element 1),] [whereas (element 2)]$_{\text{DFF}}$
[Whereas (element 2),]$_{\text{DFF}}$ [(element 1)]

The main reason for this is that the other argument (i.e. 'element 1') is not always explicitly provided in the text and if it is at all, there may well be also some other constituents between the two arguments. Given this, we give higher priority to annotating in a consistent manner.

4 Discussions

We have come up with certain controversial cases in the course of our annotation task. Appendix 1 provides solutions to typical cases. We intend to improve our annotation guidelines through feedback from other annotators who have worked with the help of this document.

References

Carlson, L., Marcu, D. and Okurowski, M. E. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. *SIGDAL 2001.*

Mann, W.C. and Thompson, S.A. 1987. Rhetorical structure theory: toward a functional theory of text organization. *Text* 8(3), 243-281.

Marcu, D. and Echihabi, A. 2002. An unsupervised approach to recognizing discourse relations. *ACL2002*.

Mizuta, Y. and Collier, N. 2004a. Annotation scheme for a rhetorical analysis of biology articles. In *Proceedings of the Fourth International Conference on Language and Evaluation (LREC2004).* pp.1737-1740.

Mizuta, Y. and Collier, N. 2004b. Zone identification in biology articles as a basis for information extraction. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA) at the COLING2004 International Conference*. pp.29-35.

Teufel, S., Carletta, J. and Moens, M. 1999. An annotation scheme for discourse-level argumentation in research articles. *EACL '99.*

Teufel, S., and Moens, M. 1999. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In Mani, I. and Maybury, M.T (eds.) (1999) *Advances in automatic text summarization*. Cambridge, MA: MIT Press.

Teufel, S. and Moens, M. 2000. What's yours and what's mine: Determining Intellectual Attribution in Scientific Text. *SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Teufel, S. and Moens, M. 2002. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics* 28(4), 409--445.

Prasad, R., Miltsakaki, E., Joshi, A. and Webber, B. 2004. Annotation and data mining of the Penn Discourse TreeBank. In *Proceedings of the ACL Workshop on Discourse Annotation*.

Appendix 1: Solutions to Controversial Cases (Q & A)

Notational remarks

A pair of square brackets indicate the scope of a zone, and the label accompanying the closing bracket indicates the zone class for the scope.

e.g.1. a PBM zone

> **[** we examined the relationship between c-Myc levels and proliferation in Mnt$^{-/-}$ MEFs. $\mathbf{]_{PBM}}$

e.g.2 a PBM zone embedded in an OTL zone (i.e. combined zones in this case)

> **[[** Here we examined systematically how single amino acid substitutions between the pep-anticodons affect the RF activity. $\mathbf{]_{PBM}\,]_{OTL}}$

Boldfaces used in the examples below are for emphasis. Italics used in the Q & A are for either emphasis or quotation.

1. Sentences with *examine*

(1) **[** we **examined** the relationship between c-Myc levels and proliferation in Mnt$^{-/-}$ MEFs. $\mathbf{]_{PBM}}$ **[** For **these experiments**, growth curves were carried out with passage 5 cells while c-Myc levels were monitored. $\mathbf{]_{MTH}}$

Q. Why is the first sentence annotated as PBM rather than as MTH?

A. Because 'the relationship between c-Myc levels and proliferation in Mnt$^{-/-}$ MEFs' indicates the goal of the experiment, rather than the procedure. Indeed, it also entails experiments performed, which are, interestingly, the reference of *these experiments* (boldfaced). However, the annotation is made in favor of the main information provided.

(2) **[[** Here we **examined** systematically how single amino acid substitutions between the pep-anticodons affect the RF activity. $\mathbf{]_{PBM}\,]_{OTL}}$

As in example (1) above, the sentence states an underlying question, although it entails experiments performed, and thus should be annotated as PBM.

2. The expression *These data showed that ~*

(3) **[** we measured distances from the anterior tip of the cell to the nucleus, from the nucleus to the kinetoplast and from the kinetoplast to the posterior end in the same cells, as above (Figure 3D). $\mathbf{]_{MTH}}$ **[These data showed that** the reduction in size was mostly due to a reduction of the anterior part of the cell, i.e. the zone along which the flagellum is attached. **[** In contrast, the distance between the kinetoplast and the posterior end of the cell was not modified. $\mathbf{]_{DFF}\,]_{RSL}}$ **[**This correlation suggests that flagellum elongation could control cell size. $\mathbf{]_{INS}}$

Q. Why does the sentence 'These data showed that ~ ' fit into RSL rather than into INS?

I observe that such phrases usually signal INS. That is, when I see such an expression, I expect that the authors provide an interpretation of the data (i.e. insights obtained).

I agree that the next sentence starting with 'In contrast' fits into RSL. But the sentence in question i.e. 'the reduction in size…', seems to give some analysis as indicated by *due to*.

A. Please have a closer look at the *content* of the clause following 'These data showed that'.

The information provided by the clause is an observation rather than a biological interpretation of what's observed. Here, 'biological interpretation' means information about some biological process or a property of a biological element. Biological interpretations fit into INS, whereas observations providing data/results fit into RSL. So, the clause at issue, and therefore the whole sentence, fits into RSL.

Cf. See the last sentence annotated as INS -- 'flagellum elongation could control cell size' expresses an insight into a property of a biological element, which is obtained from the correlation observed.)

3. *to*-infinitives expressing an object / a result

(4) **[** To address the second question, $\mathbf{]_{PBM}}$ **[** we used quantitative RT?PCR analyses of the RNAs from cell lysates as well as from defined sucrose gradient fractions representing both wild-type and scp160-null cells $\mathbf{]_{MTH}}$ **[to demonstrate** a significant shift from the membrane fraction to the soluble

fraction for one Scp160p-associated mRNA, and a subtle yet significant shift in the polyribosome association profiles of at least two Scp160p-associated mRNAs relative to a non-target control. **]**$_{\mathbf{RSL}}$ **[** Together, these results not only confirm that Scp160p associates with specific mRNAs in yeast, but also that these interactions are biologically meaningful. **]**$_{\mathbf{INS}}$

Q. Should the *to*-infinitive (in lines 3-6) be annotated in its own right, even without a comma before it?

A. Yes (, although it would complicate the automatic annotation process).

Also, notice that it is annotated as RSL (result), rather than PBM (the goal of the experiment).

4. Annotation unit regarding CNN and DFF

(5) **[[Consistent with** the low level of c-Myc in Mnt$^{-/-}$ MEFs at passage 4 (Figure 4), it was difficult to detect c-Myc and Max bound to Cdk4 proA in these cells (Figure 5B). **]**$_{\mathbf{CNN}}$ **]**$_{\mathbf{RSL}}$ **[[**However, both c-Myc and Max were bound to Cdk4 proA in immortal (passage 35) Mnt$^{-/-}$ MEFs (Figure 5B). **]**$_{\mathbf{DFF}}$**]**$_{\mathbf{RSL}}$ **[** Together with the increased Cdk4 mRNA levels found in passage 4 and immortal Mnt$^{-/-}$ MEFs, **these results argue that** the Cdk4 gene is a direct target of Mnt and that it is derepressed in the absence of Mnt. **]**$_{\mathbf{INS}}$

Annotated as a sequence of combined zones (CNN, RSL), another combined zones (DFF, RSL), and an INS zone.

Q1. Semantically, CNN and DFF both take two arguments (i.e. elements to be compared). But in the example above, the scope of these zones are not consistent. This is confusing.

The CNN zone above is annotated to include the two arguments (i.e. 'the low level of c-Myc …(Figure 4)' and the main clause). But the DFF zone is annotated to include only one argument; the other argument is in the first sentence.

A. You are right. But please check with the rules on the scope of annotation mentioned in the main section.

Q2. There are two consecutive RSL zones at the beginning. Would it be o.k. to annotate both sentences as a single RSL zone (with a CNN and a DFF zone embedded in it)?

A. Yes, it's up to you. It is just for practical reasons that I annotated those sentences as smaller RSL zones. The annotation was made in a rather 'primitive' manner by inserting a starting and an ending label (e.g. brackets and the subscript). It is easy for an annotator to forget to type an ending label of an outer zone (RSL in this case). The risk gets larger if the outer zone is larger, in which case the annotator has to remember for a longer time to close the zone.

5. Nested annotation and the treatment of discourse connectives

(6) **[** For example, **[** although we have demonstrated clear association of specific mRNA sequences with Scp160p-containing complexes,**]**$_{\mathbf{INS}}$ we do not yet know whether these interactions are direct or indirect. **]**$_{\mathbf{IMP}}$

Q. In the annotation above, the *although*-clause annotated as INS is embedded in the IMP zone. But it does not seem to fit into IMP. So, the following annotation makes more sense:

> For example, **[** although we have demonstrated clear association of specific mRNA sequences with Scp160p-containing complexes,**]**$_{\mathbf{INS}}$ **[** we do not yet know whether these interactions are direct or indirect. **]**$_{\mathbf{IMP}}$

Also, *for example* plays no role in zone identification, so it is out of the scope of any zone.

A. In the current version of the guidelines, discourse connectives such as *for example* and *although* are included in the constituent to be annotated, even though they make no contribution to the specification of the zone class. So, the IMP zone should include *for example* at the beginning. This gives rise to a larger IMP zone covering the whole sentence. As a result, the INS zone is embedded in that IMP zone, even though it does not fit into IMP in its own right.

Note. As a solution to this issue, we are interested in having a tool which makes it possible to define the scope of a zone excluding a specific part in it. In the example above, with such a tool, the scope of the IMP zone would be 'the whole sentence excluding the *although*-clause', whereas the *although*-clause itself is annotated as INS.

6. Global and local perspectives of annotation involving nested annotation

(7) **[** The contribution of the flagellum and its associated structures (FAZ and FC) to cell morphogenesis and the cell cycle **can be summarized in the following working model**. **[** First, basal body duplicates (Sherwin and Gull, 1989a) and a new FAZ is assembled, prior to flagellum exit from the flagellar pocket (Kohl et. al, 1999). **]** $_{\mathbf{BKG}}$ These steps are independent from the formation of the new flagellum as they still take place in (TbDHC1b) and (TbIFT88) RNAi mutant cells with an old flagellum but without a new one. Next, the flagellum elongates and somehow drives FAZ elongation. From that point in the cell cycle, FAZ elongation is controlled by flagellum growth as production of a flagellum that does not reach wild-type length also leads to incomplete FAZ. As the new FAZ elongates, it could participate in basal body segregation. In the absence of a new flagellum, the new FAZ is much shorter and basal body segregation is less efficient. **[** During this whole process, the elongation of the cytoskeleton continues at the posterior end (Sherwin and Gull, 1989b). **]**$_{\mathbf{BKG}}$ Once flagellum growth is terminated, FAZ elongation also finishes, the FC is disassembled and the cell initiates cleavage at the anterior end of the FAZ. **]** $_{\mathbf{IMP}}$

Q. Isn’t it contradictory that some **BKG** zones are embedded in **IMP**, given that BKG concerns previous work whereas IMP concerns the authors’ work?

A. No, it is not, for the following reasons.

In this example, the whole paragraph describes an explanatory model which the authors propose, and therefore constitutes an IMP zone. The proposed model consists of a set of biological processes. These include those mentioned in previous work, and therefore the sentences fit into BKG.. Notice that the authors’ proposal is the *whole sequence of* these steps, not (necessarily) each component step. (Critically, the BKG zones, as well as the other zones, wouldn’t fit into IMP by themselves.) In other words, the IMP zone is annotated for a larger unit, whereas the BKG zones are annotated locally. So, the nested annotation (BKG zones within an IMP zone) does not mean that those components equally fit into BKG and IMP. Therefore, the embedding of BKG zones in an IMP zone makes sense.

Appendix 2: Sample Annotation

This Appendix aims to provide practical knowledge of how to annotate a text in the biology domain. It illustrates a concrete example of an annotated full-text article together with notes supporting that annotation. This document thus demonstrates how the guidelines apply in each step of an annotation task.

Labels used

In the annotation below, zones are illustrated by a pair of an opening and a closing label specifying their class. It takes a form analogous to the Xml language. For example, a BKG zone is illustrated as:

<BKG> text </BKG>, where 'text' stands for a span of text forming a BKG zone.

The same format applies to other zone classes:

<PBM> text </PBM>, <MTH> text </MTH>, <RSL> text </RSL> ,
<INS> text </INS>, <IMP> text </IMP>, <ELS> text </ELS>,
<CNN> text </CNN>, <DFF> text </DFF>
<OTL> text </OTL>

Besides zones, the titles of the sections under investigation (e.g. INTRODUCTION) are annotated as <SECT>: <SECT> Section Title </SECT>

Embedded zones are indicated in smaller labels. (e.g. at the end of INTRODUCTION section)

Annotating the text

The sample article is: NAR 2003, Vol. 31, No. 7 1830-1837

In the following, ignore sentence-internal question marks ('?') such as the one in line 4 in the INTRODUCTION section), or any other kind of incomprehensible symbols, you see in the text. These are the byproduct of conversion of graphical symbols into the text format.

<SECT> INTRODUCTION </SECT>

<BKG> The gene SCP160 encodes a 160 kDa protein (Scp160p) originally postulated to function in the maintenance of ploidy in yeast (1).[7] More recently, however, a variety of experimental approaches have all demonstrated that Scp160p associates with polyribosomes as a component of large cytoplasmic complexes, believed to be mRNPs (2?5).[8] In addition to Scp160p, these complexes also contain the polyA binding protein Pab1p, and Bfr1p (4). As would be expected of mRNPs, these complexes are resistant to EDTA, but sensitive to both RNase and high salt (3,4).[9] Together, these data support the hypothesis that Scp160p functions at some level of cytoplasmic mRNA metabolism, and that the scp160 null phenotype, which includes abnormal cell size and shape, increased DNA content, and missegregation of genetic markers through meiosis, may reflect the indirect result of aberrant target gene regulation, rather than a direct loss of Scp160p function from many different biological processes.[10] **</BKG>**[11]
<BKG> Subcellular fractionation studies have demonstrated that Scp160p partitions between the soluble and membrane-bound compartments (2,4,5). Similarly, fluorescence microscopy studies using both anti-Scp160p antibodies and GFP-tagged Scp160p, have demonstrated that while some diffuse signal is evident in the cytosol, a significant enrichment of signal is seen around the nuclear envelope (1,4,5), which is the site of the endoplasmic reticulum in yeast. Finally, localization of Scp160p to the endoplasmic reticulum has been demonstrated to be both RNA-dependent (4), and microtubule-dependent (5).[12] Together, these data support the conclusion that Scp160p associates with both soluble and rough endoplasmic reticulum-bound polyribosomes in vivo.[13] **</BKG>**

[7] A generic statement accompanied by a reference.

[8] a finding by recent work

[9] recent work

[10] (author's) insight obtained from previous work Note. It's annotated as BKG, rather than INS.

[11] It's not necessary to end the BKG zone here, since another BKG zone follows. In this sample, however, annotation is made in smaller units conforming, to some extent, to the paragraph and discourse structure.

[12] findings by previous work

[13] (author's) confirmation of a hypothesis in light of previous work

<BKG> Although little is currently known about the structure of the Scp160p protein, sequence alignment studies have revealed the presence of 14 tandem copies of the hnRNP K homology (KH) domain (2,6), a highly conserved motif found in many RNA-binding proteins (7).[14] Indeed, Scp160p demonstrates significant amino acid sequence homology to a large and extended family of multiple KH-domain proteins, collectively known as vigilins (3,8?12).[15] **</BKG>** **<BKG>** Although all vigilin proteins studied to date have been reported to bind nucleic acid, **</BKG>** **<PBM>** in most cases both the type of nucleic acid bound, and the functional significance of the interaction, remain unclear. [16] **</PBM>** **<BKG>** One notable exception is Xenopus vigilin, which was demonstrated recently not only to bind specifically to a defined sequence in the 3' untranslated region of the vitellogenin message, but also to inhibit cleavage of that sequence by the mRNA endonuclease polysomal ribonuclease 1 (13). In vitro studies have previously demonstrated that Scp160p can bind directly to ribohomopolymers, as well as to yeast ribosomal RNA, but not to tRNA (2). **</BKG>** **<BKG>** Although both we and others have hypothesized previously that Scp160p associates with mRNAs in vivo (2?5), **</BKG>**[17] **<PBM>** whether those mRNAs are random or specific, and whether these associations are biologically significant, has remained unclear.[18] **</PBM>**

<OTL> We report here the results of experiments that directly address both of these questions.[19] **<PBM>** In brief, we have asked (i) Do Scp160p-associated mRNPs contain random or specific subsets of yeast messages, and, if specific, what are they? and (ii) Is there any detectable impact of scp160 loss on its target messages?[20] **</PBM>** **<PBM>** To address the first question,[21] **</PBM>** **<MTH>** we utilized affinity isolation of Scp160p-associated mRNPs, followed by microarray and quantitative RT?PCR analyses of the mRNAs released from these complexes.[22] **</MTH>** **<RSL>** We found not only that yeast mRNA sequences are present in these samples, but also that the sequences present are specific, not random.[23] **</RSL>** **<PBM>** To address the second question,[24] **</PBM>** **<MTH>** we used quantitative RT?PCR analyses of the RNAs from cell lysates as well as from defined sucrose gradient fractions representing both wild-type and scp160-null cells[25] **</MTH>** **<RSL>** to demonstrate a significant shift from the membrane fraction to the soluble fraction for one Scp160p-associated mRNA, and a subtle yet significant shift in the polyribosome association profiles of at least two Scp160p-associated mRNAs relative to a non-target control.[26] **</RSL>** **<INS>** Together, these results not only confirm that Scp160p associates with specific mRNAs in yeast, but also that these interactions are biologically meaningful.[27] **</INS>** **</OTL>**

<SECT> MATERIALS AND METHODS <SECT>

Yeast strains and culture conditions

<MTH> The yeast strains used in this study included JJ52 (MAT gal7 102 ura3-52 trp1-289 ade1 lys1 leu2-3,112, a generous gift from Drs Mark Parthun and Judith Jaehning, University of Colorado Health Sciences Center) and JFy1511, which was derived from JJ52 by substitution of an N-terminally FLAG-tagged allele of SCP160 in place of the wild-type allele (3). All studies comparing wild-type versus scp160-null cells were performed using diploid strains of W303-derived cells homozygous for a genomic scp160 deletion, that

[14] a finding by recent work Note. The problem mentioned by the *although*-clause is a minor one and does not motivate the author's present work. So, it does not deserve a PBM zone.

[15] supporting evidence for the finding just mentioned

[16] finding by previous work (BKG) and an unsolved problem (PBM)

[17] Zone labels are provided *after* a punctuation mark such as a period and a comma. (a stylistic issue subject to change.)

[18] previous work (BKG) and an unsolved problem

[19] The OTL zone continues to the end of the section, with some secondary (embedded) zones This first sentence mentions what the present paper is all about (Note. *Here* refers to the whole paper), whereas the following sentences summarize the paper.

[20] the main research questions (answered in the paper)

[21] goal (focus) of the experiment

[22] an experimental procedure

[23] an observation of the results

[24] the goal (focus) of the experiment

[25] an experimental procedure

[26] an observation of the results

[27] insights into the function of a biological entity and a biological process

either did (JFy4100), or did not, carry a plasmid borne copy of wild-type SCP160 (JF3116, URA3), respectively. Due to concerns over potential and progressive aneuploidy in the scp160-null strains, these strains were always generated fresh from JFy4100 just prior to use by plasmid curing on medium containing 5-fluororotic acid (5FOA) (14). **</MTH>**[28]

<SECT> RESULTS <SECT>

Polyadenylated RNA is present in Scp160p-containing complexes

<PBM> To address directly the question of whether Scp160p-containing complexes include mRNA,[29] **</PBM> <MTH>** we exploited the presence of a FLAG epitope tag engineered onto the N-terminus of Scp160p (see Materials and Methods) to affinity isolate these complexes, essentially as described previously (3).[30] **</MTH> <BKG>** FLAG-Scp160p has been demonstrated previously to function in vivo indistinguishably from the untagged native protein (3).[31] **</BKG> <MTH>** As a control for specificity, cells expressing native, untagged Scp160p also were subjected to the affinity isolation procedure. Total RNA was then released from both isolates and subjected to reverse transcription using an oligo-dT primer in the presence of [-32P]dCTP.[32] **</MTH> <RSL>** As illustrated in Figure 1, a strong smear, centered at approximately 1500 bases in size, was observed in the lane representing FLAG-Scp160p, but not in the control lane, although a larger fraction of the control reaction sample was loaded. Parallel reverse transcription reactions using total RNA isolated from whole cell soluble lysates of both strains resulted in indistinguishable strong smears (data not shown).[33] **</RSL>**

Identification and confirmation of specific mRNA sequences associated with Scp160p

<PBM> To address the question of sequence specificity,[34] **</PBM> <MTH>** RNA samples derived from FLAG-Scp160p-containing complexes versus total RNA from the same cell lysates were used as templates to generate probes for hybridization to Affymetrix YG-S98 yeast gene chips (see Materials and Methods). As a control, corresponding pools of RNA derived from cells expressing native, rather than FLAG-tagged Scp160p, also were prepared.[35] **</MTH> <RSL>** The results, determined by comparing the hybridization results of each Scp160p complex-derived sample against its corresponding total RNA control (see Materials and Methods), were striking.[36] First, although many strong hybridization spots were detected in both test and control samples, the patterns were different, indicating that[37] the Scp160p complex-derived samples did not simply contain a random subset of total cellular mRNAs. Furthermore, those sequences most abundant in the mock-isolated samples were completely distinct from those most abundant in the FLAG-Scp160p complex-derived samples (data not shown), demonstrating[38] specificity of the isolation procedure. In sum, of the >6000 putative yeast gene sequences interrogated on the microarrays in duplicate experiments, only 1% (69 sequences) appeared >2.5-fold enriched in the FLAG-Scp160p complex-derived samples in both experiments (Table 1).[39] **</RSL> <PBM>** To test a subset of these candidates with an independent technology,[40] **</PBM> <MTH>** we performed quantitative RT?PCR using a Roche LightCycler with

[28] Note that the underlying subject of these passive sentences is *we*, i.e. the author(s) (or relevant people).

[29] the question to answer (i.e. focus of the experiment)

[30] experimental procedure

[31] previous work

[32] experimental procedure

[33] To be precise, the subject NP of the last sentence describes an experimental procedure. But it is too small a constituent to be annotated as MTH on its own.

[34] a question to answer (i.e. focus of the experiment)

[35] an experimental procedure

[36] a characterization of the result

Note. The inserted phrase expressing the experimental procedure is *not* annotated as MTH, because a participle phrase like this (i.e. *determined* ~) cannot form a zone in its own right.

[37] What follows *indicating that* is *an interpretation of* the result but crucially, it does not provide an insight into a biological process. So, it is *not* annotated as INS.

[38] What follows *demonstrating* is an interpretation of the result but crucially, it does not provide an insight into a biological process. So, it is *not* annotated as INS.

[39] The RSL zone ends here.

[40] the goal of the experiment

primers designed to amplify small fragments from the 3' ends of each of 12 candidate enriched messages, eight of which are listed in Table 1 (YGR023W, YOR338W, YOL155C, YDR247W, YBR150C, YHR086W, YDL160C, YOL059W), and four of which (YCL029C, YKL203C, YBL109W, YGR110W) are not listed because they appeared enriched in only one of the two microarray experiments performed. Templates analyzed by quantitative RT?PCR were mRNAs derived from three or more independent Scp160p complex isolation procedures, each compared against its corresponding total mRNA control.[41] </MTH> <RSL>Those five sequences that were confirmed as enriched by quantitative RT?PCR are presented in Table 2.[42] </RSL> <PBM> To ensure specificity of these values,[43] </PBM> <MTH> mock enrichment procedures also were performed using cells expressing native rather than FLAG-tagged Scp160p, and apparent fold enrichment of each candidate message in those emock f samples was calculated and subtracted as background from the values presented in Table 2.[44]</MTH> <RSL> As indicated by asterisks in Table 1, four of the eight messages tested from that group did not confirm as enriched by >2.5-fold when measured by quantitative RT?PCR. An additional three of the four esingle microarray candidates f tested also did not confirm as enriched by quantitative RT?PCR.[45]</RSL>

Impact of scp160-loss on DHH1 and YOR338W

<PBM> To address the question of biological significance of Scp160p association with its target messages,[46] **</PBM>** **<MTH>** we first checked both message abundance and message distribution between the soluble and membrane-associated pools for two target sequences, DHH1 and YOR338W, comparing wild-type versus scp160-null yeast.[47] **</MTH>** **<RSL>** As illustrated in Table 3, although no significant change was seen for DHH1, YOR338W demonstrated a significant increase in abundance in scp160-null cells. Furthermore, the distribution of that signal was shifted[48] significantly away from the membrane pellets, and toward the soluble fraction.[49] **</RSL>**

<MTH> Next, we utilized sucrose-gradient fractionation **</MTH>** **<PBM>** to explore the subcellular distribution of DHH1 and YOR338W in both wild-type and scp160-null yeast.[50] **</PBM>** **<RSL>** In brief, both strains were grown to early log phase (OD 1), lysed as described previously (3,4), and the soluble portions subjected to sucrose gradient fractionation, as described previously (3,4).[51] **</RSL>** **<MTH>** Total RNA was isolated from each fraction, and subjected to quantitative RT?PCR using a Roche LightCycler with the appropriate primers (Fig. 2; see Materials and Methods). For each fraction, the target sequence signal detected was normalized to the corresponding signal from a non-target control sequence (enolase, ENO2), so that the data points presented represent ratios.[52]**</MTH>** **<RSL>** As illustrated (Fig. 2, bottom two panels), cells devoid of Scp160p (striped bars) demonstrated a marked enrichment of both DHH1 and YOR338W in the lighter gradient fractions (representing mRNPs), as compared with their wild-type counterparts (solid bars). **</RSL>**

<MTH> Finally, we performed parallel sucrose gradient fractionation experiments on samples derived from the membrane-associated compartments of both wild-type versus scp160-null cells.[53] **</MTH>** **<RSL>** No reproducible differences in the distribution of DHH1 or YOR338W signals were observed in these experiments (data not shown).[54] **</RSL>**

[41] experimental procedure

[42] The sentence points to the results. (It also plays a role of a legend of the table mentioned.)

[43] the goal of the experiment

[44] experimental procedure

[45] results

[46] the goal of the experiment

[47] experimental procedure

[48] Of course, this passive form does NOT have *we* as its underlying subject. Cf. passives in MTH zones

[49] results

[50] the goal of the experiment

[51] results (and related previous work)

[52] experimental procedure

[53] experimental procedure

[54] results

<SECT> DISCUSSION <SECT>

<INS> The results reported here demonstrate two main points.[55] First, Scp160p associates with specific rather than random mRNAs in yeast. Second, loss of Scp160p results in a detectable change in the abundance and membrane association of at least one of its target messages (YOR338W), and in the soluble polyribosome association profiles of at least two of its target messages (DHH1 and YOR338W), relative to a non-target control (ENO2).[56] **</INS> <IMP>** Each of these findings represents an important step forward in our effort to understand the biological function of Scp160p.[57] **</IMP> <IMP>** The first point, **<INS>** that Scp160p associates with only 1% of yeast mRNAs, **</INS>** [58] is important because it rules out the possibility that Scp160p is a general translation factor in yeast. This point is made even stronger considering that only 5 of the 12 candidate targets tested confirmed by quantitative RT?PCR, so that close to half of the other potential target messages currently indicated by microarray analysis alone might also fail to confirm. The actual percentage of messages in yeast that associate with Scp160p may therefore be <1%. Furthermore, considering the disparity between the microarray data obtained and quantitative RT?PCR results, it is reasonable to assume that genuine target messages may also have been missed by the microarray experiments.[59] **</IMP>**

<IMP> Perhaps more important, although the set of Scp160p-associated messages we have presented may not be comprehensive, our data provide a ready list of potential targets for further study -- targets that will likely offer additional insights into the mechanism and impact of Scp160p function in vivo.[60] For example, among the enriched Scp160p-associated messages we have identified and confirmed are DHH1, a putative RNA helicase with close homologs in mammals, including human; BIK1, a putative microtubule binding protein required for microtubule function in mitosis and mating; and NAM8, an RNA-binding protein required for the meiosis-specific splicing of MER2 and MER3.[61] Others (YOR338W and YOL155C) remain hypothetical open reading frames (ORFs); through studies of their interplay with Scp160p, we may also gain insight into their functions, which are currently unknown. Clearly many more interesting potential targets also remain to be studied.[62] **</IMP>**

<IMP> The second point, **<INS>** that loss of Scp160p results in a change in abundance and membrane association for at least one target message (YOR338W), as well as a subtle but significant change in the soluble polyribosome association profile of at least two target messages (DHH1 and YOR338W) relative to a non-target control (ENO2), **</INS>**[63] is equally important, because it demonstrates that the interaction of these messages with Scp160p is biologically meaningful.[64] These data are consistent with the conclusion that Scp160p-loss results in a shift of at least some of its target messages from membrane associated polyribosomes to free mRNPs in the soluble pool. At minimum, these data strengthen the argument that Scp160p functions in some aspect of cytoplasmic mRNA metabolism, perhaps including translation.[65] **</IMP>**

<IMP>Whether the observed shift in polyribosome association reflects altered translational efficiency, stability, or some other property of the target messages, remains to be determined. Furthermore, future studies will be required to determine whether other target messages (e.g. BIK1, NAM8, YOL155C, and others as yet unconfirmed) will demonstrate similar or distinct responses to the loss of Scp160p. **</IMP>**

<IMP> Although the results presented here represent a significant step forward, much work remains to be completed if we are to understand the biological role(s) of Scp160p in yeast, and of its counterparts in other species. **</IMP> <IMP>** For example, **<INS>** although we have demonstrated clear association of specific

[55] a pointer to insights Note. This sentence is included in the INS zone.

[56] findings (regarding a biological process)

[57] implication of the findings (their significance from a wider perspective)

[58] The inserted nonrestrictive relative clause is annotated as INS. This zone is embedded in an IMP zone.

[59] an implication of (or an inference from) the first finding

[60] relation to future work

[61] example of the list just mentioned

[62] potential targets which remain to be studied

[63] The inserted nonrestrictive relative clause is annotated as INS. This zone is embedded in an IMP zone.

[64] the significance of the second finding mentioned above (i.e. assessment)

[65] relate the findings (‘data’) to further argument/conclusion (i.e. support the argument)

mRNA sequences with Scp160p-containing complexes,[66] **</INS>** we do not yet know whether these interactions are direct or indirect.[67] **</IMP>** **<IMP>** Preliminary in vitro RNA-binding studies between recombinant Scp160p and labeled transcript suggest that direct binding can occur with target sequences, although the specificity of that binding is unclear.[68] Furthermore, what features these confirmed target messages exhibit, and perhaps share in common, that enable each to associate with Scp160p, remain to be defined. It is also possible, if not probable, considering the large size and significant number of non-Scp160p proteins apparent in Scp160p-containing mRNP complexes (3,4), that some determinants of specificity may derive from other components of these complexes, not only from Scp160p and transcript. What these other components are, and how they may contribute to the specificity of message association, remains to be defined. **</IMP>** **<IMP>** In addition,[69] **<MTH>** we have, to date, selected target messages for study based only upon their degree of apparent enrichment in Scp160p-containing complexes (e.g. 2.5-fold); **</MTH>** however, these may not be the most biologically important messages impacted by Scp160p. **</IMP>** **<IMP>** Alternative approaches will be required to define pools of messages that not only associate with Scp160p, but that are also specifically impacted in any given way by Scp160p-loss.[70] **</IMP>**

<IMP>[71] **<MTH>** Finally, in the microarray studies reported here, by lysing cells in the presence of EDTA, which disrupts polyribosomes, we have intentionally mixed the soluble and membrane-associated Scp160p pools prior to Scp160p complex isolation. This strategy of isolation was designed to give a ewhole cell f representation of Scp160p, and to minimize the number of microarrays required to perform the experiments.[72] **</MTH>** **<DFF>** However, it is entirely possible that the membrane-associated and soluble populations of Scp160p may interact with different subsets of mRNA.[73] **</DFF>** Future experiments will focus on exploring separately the mRNA and protein components of soluble versus membrane-associated Scp160p-containing complexes, in order to compare and contrast these two populations.[74] **</IMP>**

[66] The inserted *although*-clause is annotated.

[67] remaining problems / future work

[68] related to future work

[69] This IMP zone continues to the end of the sentence.

[70] remaining problems (future work) in relation to alternative approaches to the present work

[71] This paragraph is committed to mentioning implications of the present methodology.

[72] Review of the methodology of the present work (in the context of implication)

[73] provides a counterargument to the methodology just mentioned (DFF) as well as mentions a need for sophistication of it (IMP)

[74] Goal and focus of future experiments

Appendix 3: Dataset

A dataset of twenty articles annotated for zones are available from the authors upon request subject to pending copyright agreement from the publishers.

1. Sources of sample articles

We first downloaded sample articles randomly selected from four major online journals (i.e. EMBO, NAR, PNAS, and JCB) in the Microsoft Word format. We then hand annotated them with regard to the abstract and the main sections of each article in the fashion illustrated in Appendix 2.[75] Then, after saving the annotated files in the text file format, the version in the XML format were created.[76] The sources of sample articles are:

EMBO (European Molecular Biology Organization):	5 articles
PNAS (Proceeding of National Academy of Science):	5 articles
NAR (Nucleic Acid Research):	6 articles
JCB (Journal of Cell Biology):	4 articles
Total:	20 articles

The publication information about these articles (as it appears in the online version) is as follows.

The EMBO Journal (2003) 22, 4584-4596
The EMBO Journal (2003) 22, 5336–5346
The EMBO Journal (2003) 22, 5358–5369
The EMBO Journal (2003) 22, 5370–5381
The EMBO Journal (2004) 23**,** 2059–2070
Proc Natl Acad Sci U S A. 2002 February 19; 99 (4): 1807–1812
Proc Natl Acad Sci U S A. 2002 February 19; 99 (4): 1905–1909
Proc Natl Acad Sci U S A. 2002 February 19; 99 (4): 1921–1925
Proc Natl Acad Sci U S A. 2002 February 19; 99 (4): 1819–1824
Proc Natl Acad Sci U S A. 2002 February 19; 99 (4): 1842–1846
Nucleic Acids Research, 2003, Vol. 31, No. 7, 1830-1837
Nucleic Acids Research, 2003, Vol. 31, No. 7, 1869-1876
Nucleic Acids Research, 2003, Vol. 31, No. 7, 1888-1896
Nucleic Acids Research, 2003, Vol. 31, No. 7, 1974-1983
Nucleic Acids Research, 2003, Vol. 31, No. 7, e36
Nucleic Acids Research, 2003, Vol. 31, No. 8, 2077-2086
The Journal of Cell Biology, Volume 157, Number 4, May 13, 2002 565-570
The Journal of Cell Biology, Volume 157, Number 4, May 13, 2002 591-602
The Journal of Cell Biology, Volume 157, Number 4, May 13, 2002 631-643
The Journal of Cell Biology, Volume 157, Number 4, May 13, 2002 679-691

2. Readme file for the data sets

Below is what we reproduce the readme file for the dataset.

This dataset contains twenty articles which have been hand annotated for rhetorical zones.

Zones are identified by the element *zone*. Zone elements have *scope* attributes which are presently limited to the values *sentence* or *constituent*. In the case of zones with scope *sentence* the open element is placed at the beginning of the sentence, and the close element is placed at the end of the sentence. In the case of zones with scope *constituent* the opening and close elements may occur within the sentence. Any scope greater than a single sentence is not explicitly encoded in this dataset. Thus, a two sentence long zone will be composed of two sentences, each of which are tagged with opening and closing sentence-scope zone elements. The constituent scope is more specific as regards where the zone is within the sentence, but we treat the information as redundant. If a sentence contains a constituent of a particular zone class, the sentence-scope attribute value will also contain the name of this class.

[75] A single annotator (Yoko Mizuta) is responsible for the annotation of all the files. There are quite a few controversial cases as mentioned in Appendix 1, but we hope that the they are annotated in a consist manner.

[76] Tony Mullen has worked on this.

Multiple zones which share the same sentence or constituent scope may be represented with a single zone element as list values. The zones represented are listed for the L1class, L2class, and L3class attributes, corresponding to the three groups mentioned in Section 2. For example, if a sentence belongs to the classes "MTH", "BKG" (in Group 1) and "CNN" (in Group 2), then the L1class value of the zone will be "MTH,BKG" and the L2class value of the zone will be "CNN". On the sentence level, sentence is considered as belonging to a class if any element within the sentence belongs to that class. For this reason, by looking at the sentence scope zone information only, it is not possible to tell whether the zones overlap or whether they apply to distinct constituents within the sentence: This can be determined from the constituent-scope zone information.

Document type: zone-dataset
Elements annotated in this file:

zone-dataset
article: requires filename
section: level (section|subsection)
zone: annotated zone. Scope:(sentence|constituent),
L1class (set of Group 1 classes),
L2class (set of Group 2 classes),
L3class (set of Group 3 classes)
wordindex: word position
word: token]
citation: citation
genseq: gene sequence
number: string of integers