Ways to Increase Clarity and Reduce Reader Effort

BACTER Writing Workshop, June 18, 2009
Science is Hard to Read

“There is no form of prose more difficult to understand and more tedious to read than the average scientific paper.” —Francis Crick
What We’ll be Covering

Word Choice
• Acronyms
• Plain language
• Wordiness

Sentence Structure
• Subject-verb agreement
• Maintaining linkage
• Noun clusters
• Passive voice
References

- Enhancing the mission of academic surgery by promoting writing skills. PA Derish et al. (2007) *Journal of Surgical Research*
- The science of scientific writing. G Gopen and J Swan (1990) *American Scientist*
- Compliance (Communicate please with less abbreviations, noun clusters and exclusiveness) MJ Tobin (2002) *American Journal of Respiratory and Critical Care Medicine*
Why Make the Extra Effort?

• “PLoS ONE staff do not copyedit the text of accepted manuscripts; it is therefore important for the work, as presented, to be intelligible.”

• “In the interests of speed, manuscripts are not extensively copyedited and authors are requested to check their texts carefully before submitting them so that proofs will require only correction of typographical errors.” (Bioinformatics)

• “Manuscripts will not be extensively copy-edited...” (J. Comput. Biol.)
The Goal

To dazzle readers with your science rather than distract them with your writing.
What We’ll be Covering

**Word Choice**
- Acronyms
- Plain language
- Wordiness

**Sentence Structure**
- Subject-verb agreement
- Maintaining linkage
- Noun clusters
- Passive voice

Reviewing each other’s writing
Word Choice: Use Acronyms Sparingly

• Some acronyms (abbreviations) are so familiar, they’re like words to us: ATP, DNA, BLAST

• But discipline-specific abbreviations require us to memorize new terms as we read: DNR, SVM, CBM, CPD
Guidelines for Using Acronyms

- Always define them
- Keep them out of the abstract
- Use an acronym only when an unwieldy word or phrase occurs 10 or more times
- Limit yourself to 2 to 3 discipline-specific acronyms per manuscript
- Substitute a word instead; e.g., Instead of “OR” for olfactory receptors, write “receptors” after the first use

“Some authors apparently believe that they must impress the reader (and the editor) with their erudition and mastery of multisyllabic words in order for their work to be given the appreciation it deserves. This is a mistaken notion. With scientific writing, as with most other forms of communication, the simplest and most direct statement of the intended message is always best.”

--The top 10 reasons why manuscripts are not accepted for publication (2004) *Respiratory Care*
Use Plain Language

Whenever possible, substitute simple words for more complex or unfamiliar ones.

- “Use” instead of “utilize” or “utilization”
- “Include” instead of “comprise” (Note that “comprised of” is incorrect; should be “composed of”)
- “Happen” instead of “transpire”
- “Explain” instead of “elucidate”
- “Most” instead of “a majority of”
- “Cause” instead of “causal factor”
Use Plain Language

Original: “A good proximal understanding of the normal behavioral repertoire of astrocytes is a prerequisite for understanding the large number of pathophysiological conditions that may arise from dysfunctional situations as well as for assessing the potential of the astrocyte as a therapeutic target. However, our current concepts of even long known aspects of this repertoire are quite obscure.”
A better understanding of the normal behavior of astrocytes is needed before we can deduce their role in disease or assess their potential as therapeutic targets. However, our knowledge of even their most basic behavior is still incomplete.

- good proximal (?) → better
- behavioral repertoire → behavior
- prerequisite → needed
- large number of pathophysiological conditions that may arise from dysfunctional situations → disease
- current concepts → knowledge
- long known aspects → most basic
- quite obscure → incomplete
Use Plain Language

Original: “Transposable elements (TEs) are ubiquitous components of most sequenced genomes, but their function, if any, is poorly understood. Comprising ~50% of the human genome, the majority of TEs belong to the short interspersed element (SINE) (>10%), long interspersed element (LINE) (>20%), and endogenous retroviral/long terminal repeat (LTR) (~10%) families. The SINEs encode no open reading frame (ORF) and have utilized LINE-encoded proteins to amplify to >10^6 copies in the human and mouse genomes. On the other hand, only a limited number of LINEs and LTR elements are full-length; many of which are rendered non-functional due to point mutations and deletions. Therefore, the majority of TEs no longer pose a significant burden as insertional mutagens, although many retain the regulatory signals necessary for transcription.”
Revised: Transposable elements (TEs) are present everywhere in most sequenced genomes, but their function, if any, is poorly understood. Composing ~50% of the human genome, most TEs belong to the short interspersed element (SINE) (>10%), long interspersed element (LINE) (>20%), and endogenous retroviral/long terminal repeat (LTR) (~10%) families. The SINEs encode no open reading frame (ORF) and have used LINE-encoded proteins to expand to >10^6 copies in the human and mouse genomes. On the other hand, only a few LINEs and LTR elements are full-length, and many of these no longer function due to point mutations and deletions. Therefore, most TEs no longer act as insertional mutagens, although many retain the regulatory signals necessary for transcription.
A Related Issue: Wordiness

Original: A good proximal understanding of the normal behavioral repertoire of astrocytes is a prerequisite for understanding the large number of pathophysiological conditions that may arise from dysfunctional situations as well as for assessing the potential of the astrocyte as a therapeutic target. However, our current concepts of even long known aspects of this repertoire are quite obscure.

(57 words)
A Related Issue: Wordiness

Revised: A better understanding of the normal behavior of astrocytes is needed before we can deduce their role in disease or assess their potential as therapeutic targets. However, our knowledge of even their most basic behavior is still incomplete.

(38 words)
Wordiness

Wordiness often results from using nouns instead of verbs.

- “is a consequence of” instead of “results from”
- “Our findings lead us to the conclusion” instead of “From our findings, we conclude”
- “the determination of function” instead of “determining function”
- “enlargement of the size” instead of “enlarging the size”
Wordiness also arises from using unnecessary words.

- “more unique” instead of “unique”
- “pooled together” instead of “pooled”
- “in order to” instead of “to”
- “based on the fact that” instead of “because”
- “what the explanation is” instead of “why”
- “it was observed in the course of the experiment” instead of “we observed”
Wordiness

Original: “This discussion is my concept of the way that a scientific paper should or should not be written. It is not intended to criticize, but to improve precision and enhance communication. Not everyone will agree with me, but those who follow these suggestions likely will have fewer rejections and do less revising. In this era of “publish or perish,” it is imperative to produce manuscripts based on quality research, but even the best research may be lost to science if the scientific community is unable to understand the text in which it is presented.” (94 words)
Wordiness

Revised: This discussion is my concept of how a scientific paper should or should not be written. It is not intended to criticize, but to improve precision and enhance communication. Not everyone will agree with me, but those who follow these suggestions likely will have fewer rejections and do less revising. In this era of “publish or perish,” everyone must produce manuscripts based on quality research, but even the best research may be lost to science if it’s not communicated well. (77 words)
Wordiness

Original: “Many people when asked what is the most critical or far-reaching part of a scientific paper would answer: the results or conclusions drawn from the research conducted. I beg to differ. The single most critical item in any scientific paper is—the title. Hundreds, if not thousands, of people will read a title when searching for information on a particular topic in the various bibliographic services on the internet or in libraries...Thus, for the intended audience to be reached, clarity of intended meaning and proper order of words used in a title should be of paramount concern to authors.”

(99 words)
Revised: Many people when asked what is the most critical or far-reaching part of a scientific paper would answer: the results or conclusions drawn from the research conducted. I beg to differ. The single most critical item in any scientific paper is—the title. Hundreds, if not thousands, of people will read a title when searching for information on a particular topic in the various bibliographic services on the internet or in libraries...Thus, for the intended audience to be reached, clarity of intended meaning and proper order of words used in a title should be of paramount concern to authors. (76 words)
Revised 2: Many people when asked what is the most critical part of a scientific paper answer: the conclusions drawn from the research. I disagree. The most critical item in any scientific paper is—the title. Hundreds, if not thousands, of people will read a title when searching for information on a topic...Thus, **whether you reach your intended audience depends greatly on the clarity of your title.**

(65 words)
What We’ll be Covering

Word Choice
• Plain language
• Wordiness
• Acronyms

Sentence Structure
• Subject-verb agreement
• Maintaining linkage
• Noun clusters
• Passive voice
Original: Although the method yielded a relatively high prediction performance (more than 80% accuracy) in cross-validation and usefulness in the comprehensive prediction of target proteins for a given chemical compound with tens of thousands of prediction targets, it suffered from the problem of predicting many false positives when comprehensive predictions were conducted.
Revised: Although The method yielded a relatively high prediction performance (more than 80% accuracy) in cross-validation, and was useful in the comprehensive prediction of target proteins for a given chemical compound with tens of thousands of prediction targets; however, it suffered from the problem of predicting many false positives when comprehensive predictions were conducted.
Subject-Verb Separation

Original: "**Master switches**, a central set of regulatory genes whose states (on/off or activated/deactivated) determine specific developmental fate or cell-fate specification, **play** a pivotal role for whole developmental processes.
Revised: Pivotal in whole developmental processes are “master switches,” a central set of regulatory genes whose states (on/off or activated/deactivated) determine specific developmental fate or cell-fate specification.
Subject-Verb Separation

Original: **Addition** of an Escherichia coli 50S subunit containing a Cy5-labeled L11 N-terminal domain within the GTPase-associated center to an E. coli 30S initiation complex containing Cy3-labeled initiation factor 2 complexed with GTP **leads** to rapid development of a FRET signal during formation of the 70S initiation complex.
Revised: Rapid development of a FRET signal occurs during formation of the 70S initiation complex upon adding an Escherichia coli 50S subunit containing a Cy5-labeled L11 N-terminal domain within the GTPase-associated center to an E. coli 30S initiation complex containing Cy3-labeled initiation factor 2 complexed with GTP.
Original: Ruminants form the cotyledonary placenta at the feto-maternal interface. Two specific types of trophoblast cells, trophoblast giant binucleate cells (BNCs) and trophoblast mononucleate cells (TMCs), play a crucial role in ruminant placentation. The properties of BNC-specific genes... have been investigated, and TMC-expressed interferon-tau is the molecule for maternal recognition of pregnancy. ??? BNC and TMC individually produce numerous proteins of unknown function. ??? It is important to identify the genes that are specifically expressed in each cell type in order to systematically decipher the function of the trophoblast cells.
Maintaining Linkage

Revised 1: Ruminants form the cotyledonary placenta at the feto-maternal interface. Two specific types of trophoblast cells, trophoblast giant binucleate cells (BNCs) and trophoblast mononucleate cells (TMCs), play a crucial role in ruminant placentation. The properties of BNC-specific genes have been investigated, and TMC-expressed interferon-tau is the molecule for maternal recognition of pregnancy. However, (?) BNC and TMC also individually produce numerous proteins of unknown function. ??? It is important to identify the genes that are specifically expressed in each cell type in order to systematically decipher the function of the trophoblast cells.
Maintaining Linkage

Revised 2: Ruminants form the cotyledonary placenta at the feto-maternal interface. Two specific types of trophoblast cells, trophoblast giant binucleate cells (BNCs) and trophoblast mononucleate cells (TMCs), play a crucial role in ruminant placentation. The properties of BNC-specific genes... have been investigated, and TMC-expressed interferon-tau is the molecule for maternal recognition of pregnancy. However, (?) BNC and TMC also individually produce numerous proteins of unknown function. Thus, as a first step toward (?) systematically deciphering the function of these trophoblast cells, we sought to identify the genes that are specifically expressed in each cell type.
A challenge of using genomics databases is the automated retrieval of relevant data. **For example**, naive approaches to retrieve gene expression studies about ‘brain’ will fail to find datasets that only mention ‘cerebrum’ in their descriptions, because free text-based retrieval algorithms generally cannot make the inference that ‘cerebrum’ is part of ‘brain’. **In addition**, using free text for information retrieval can produce false positives due to ambiguity, and false negatives due to synonyms. **For these reasons**, it is valuable to use formal ontologies to describe genomics studies... **However**, tagging studies with terms from ontologies is currently done by human curators.
And what happens when connecting words are removed...

A challenge of using genomics databases is the automated retrieval of relevant data. Naive approaches to retrieve gene expression studies about ‘brain’ will fail to find datasets that only mention ‘cerebrum’ in their descriptions, because free text-based retrieval algorithms generally cannot make the inference that ‘cerebrum’ is part of ‘brain’. Using free text for information retrieval can produce false positives due to ambiguity, and false negatives due to synonyms. It is valuable to use formal ontologies to describe genomics studies... Tagging studies with terms from ontologies is currently done by human curators.
Connecting Words

Transition **ideas** and the connecting **words** that serve them:

**Result:** therefore, thus, hence

**Example:** for example, specifically, an illustration

**Contrast:** but, yet, however, on the other hand

**Time:** meanwhile, after, later

**Sequence:** first, then, finally

--from the Writing Center Sourcebook
Sentence Structure: The Topic Position

• The start of the sentence is called the **topic position**.
• “The information in the topic position prepares the reader for the upcoming material by connecting it backward to the previous discussion.” —Gopen and Swan, 1990
• I.e., the topic position should contain “old” information that links backward. The latter part of the sentence should include new information to be emphasized.
• When the topic position consistently contains new information, the reader gets confused.
The Topic Position

Original: Ruminants form the cotyledonary placenta at the feto-maternal interface. Two specific types of trophoblast cells, trophoblast giant binucleate cells (BNCs) and trophoblast mononucleate cells (TMCs), play a crucial role in ruminant placentation.
Revised: Ruminants form the cotyledonary placenta at the feto-maternal interface. In ruminant placentation, two specific types of trophoblast cells play a crucial role: trophoblast giant binucleate cells (BNCs) and trophoblast mononucleate cells (TMCs).
Original: “DNA breaks arise spontaneously or in response to genotoxic events. Cells respond to double-stranded break (DSB) formation to prevent chromosomal abnormalities. The conserved Mre11-Rad50-Xrs2 (MRX) complex (MRN complex in mammals) is implicated in the DSB response. It binds and holds together the broken extremities, thus preventing chromosome fragmentation...”
Revised: DNA breaks arise spontaneously or in response to genotoxic events. Cells respond to double-stranded break (DSB) formation to prevent chromosomal abnormalities. Implicated in the DSB response is the conserved Mre11-Rad50-Xrs2 (MRX) complex (MRN complex in mammals). It binds and holds together the broken extremities, thus preventing chromosome fragmentation...
What’s broken here?

“The early bi-potential mammalian gonad requires the expression of a Y-linked gene, *Sry*, during a brief window of time to ensure proper testis development. WT1 and its direct target gene *Sf1* function during sex determination as well as in the specified testes and ovaries. We have previously shown that the transcription co-factor CITED2 interacts with WT1 to stimulate the expression of *Sf1* in the adrenogonadal primordium to ensure adrenal development. We now show through genetic interactions and expression analyses that *Cited2* acts in the gonad with *Wt1* and *Sf1* to increase the expression of *Sry* levels to attain a critical threshold to efficiently initiate testis development.”
Maintaining Linkage

Original: “The early bi-potential mammalian gonad requires the expression of a Y-linked gene, *Sry*, during a brief window of time to ensure proper testis development. WT1 and its direct target gene *Sf1* function during sex determination as well as in the specified testes and ovaries. We have previously shown that the transcription co-factor CITED2 interacts with WT1 to stimulate the expression of *Sf1* in the adrenogonadal primordium to ensure adrenal development. We now show through genetic interactions and expression analyses that *Cited2* acts in the gonad with *Wt1* and *Sf1* to increase the expression of *Sry* levels to attain a critical threshold to efficiently initiate testis development.”
Revised: **The (definition)** WT1 and its direct target gene *Sf1* function during sex determination as well as in the specified testes and ovaries. We have previously shown that **WT1 interacts with** the transcription co-factor CITED2 to stimulate the expression of *Sf1* in the adrenogonadal primordium to ensure adrenal development. We now show through genetic interactions and expression analyses that *Cited2* acts in the gonad with *Wt1* and *Sf1* to increase the expression of **the Y-linked gene Sry**, to attain a critical threshold to efficiently initiate testis development **in the early bi-potential mammalian gonad.**
Sentence Structure: Noun Clusters

Mouse cell ✓
Mouse liver cell ✓
Noun Clusters

Mouse cell ✓
Mouse liver cell ✓

Normal mouse mammary gland development ✗
Cultured sheep pulmonary artery endothelial cells ✗
Robust spindle microtubule plus-end attachment ✗
Noun Clusters

Robust spindle microtubule plus-end attachment

- Robust plus-end attachment in spindle microtubules ???
- Plus-end attachment in robust spindle microtubules ???
- Plus-end attachment in microtubules from robust spindles ??
Sentence Structure: Passive and Active Voice

Active voice: The subject of the sentence performs the action expressed by the verb.

- "We conducted the experiment..."
- "Researchers believe that scaffolds have profound effects..."
- "Other studies have found reduced levels of pro-inflammatory cytokines..."
- "Normal mammary gland development requires Wnt co-receptor Lrp6..."
Passive and Active Voice

Passive voice: The subject is acted upon.

- “The experiment was conducted...”
- “Scaffolds are believed to have profound effects...”
- “Reduced levels of pro-inflammatory cytokines have been found...”
- “Wnt co-receptor Lrp6 is required for normal mammary gland development...”
Passive Voice: When to Use It

To take the emphasis off the actor and place it on the thing acted upon:

• “The cells were harvested” versus “We harvested the cells”

• “Stationary-phase promastigotes were washed three times in phosphate-buffered saline...” versus “Phosphate-buffered saline washed the stationary-phase promastigotes...”

• “Lrp5 is required for ductal stem cell activity...” versus “Stem cell activity requires Lrp5...”
Passive Voice: When to Use It

When the subject doing the action isn’t important to the meaning of the sentence:

- “Between the extremes of network models and atomistic simulation a spectrum of models has been developed that…”

- “It is estimated that 33.2 million people were infected with HIV-1 at the end of 2007…”
Passive Voice: When NOT to Use It?

When switching back and forth between discussing your results and the work of others.

Who did what here?

(1) In the present study, we identified another specific feature of SOLD1. (2) Secreted SOLD1 protein was detected under the basement membrane, but only trophoblasts expressed the SOLD1 gene. (3) There is some evidence that trophoblast cells have bilateral secretion ability... (4) In contrast, the basolateral secretion of matrix metalloproteinase-2 and -9 (MMP2 and MMP9) has been confirmed in human syncytiotrophoblasts. (5) The secretion of leptin was confirmed at both the apical and basolateral surfaces of the human trophoblast cell line BeWo.
Use active voice unless there is a good reason not to.

Original: “In this paper, we describe two strategies, namely two-layer SVM and reasonable negative data design, which are used for the purpose of reducing the number of false positives and improve the applicability of our method for comprehensive prediction. In two-layer SVM, outputs produced by the first-layer SVM model are utilized as inputs to the second-layer SVM. In order to design negative data which produce fewer false positives, we iteratively constructed SVM models or classification boundaries and selected negative sample candidates according to pre-determined rules. By using these two strategies, the number of predicted candidates was reduced to around 100 (Table 1) in experiments in which the potential ligands for some druggable proteins [(UniProt ID P10275 (androgen receptor), P11229 (muscarinic acetylcholine receptor M1) and P35367 (histamine H1 receptor)] are predicted on the basis of more than 100,000 compounds in the PubChem Compound database.”
Revised: In this paper, we describe two strategies, namely two-layer SVM and reasonable negative data design, which aim to reduce the number of false positives and improve the applicability of our method for comprehensive prediction. In two-layer SVM, the outputs produced by the first-layer SVM model serve as inputs to the second-layer SVM. In order to design negative data which produce fewer false positives, we iteratively constructed SVM models or classification boundaries and selected negative sample candidates according to pre-determined rules. By using these two strategies, we reduced the number of predicted candidates to around 100 (Table 1) in experiments that predicted the potential ligands for some druggable proteins [UniProt ID P10275 (androgen receptor), P11229 (muscarinic acetylcholine receptor M1) and P35367 (histamine H1 receptor)] on the basis of more than 100,000 compounds in the PubChem Compound database.
Final Thought: Make Revising a Habit

• Issues such as word choice and sentence structure are best addressed during the revision process
• So, make sure to do some revising

“A well-known scientist and journal editor was asked, ‘Do you revise your work?’ He answered: ‘If I’m lucky, only about 10 times.’” (from How to Write and Publish a Scientific Paper)