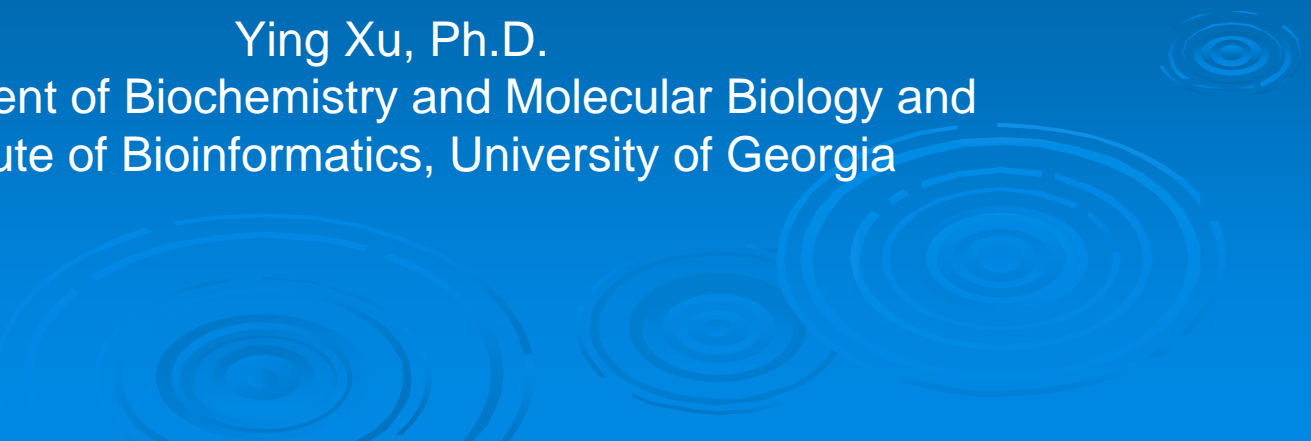


# Learn to Write a Scientific Paper like a Professional (as if I knew)

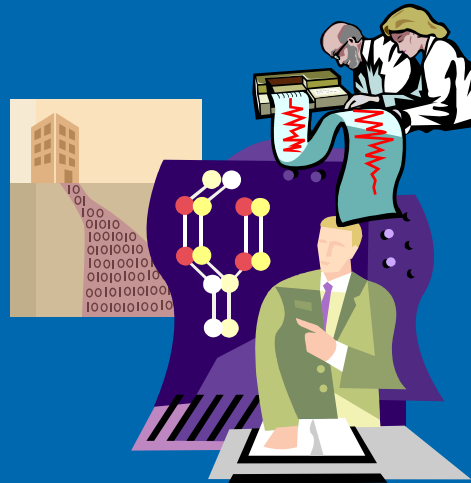
Ying Xu, Ph.D.  
Department of Biochemistry and Molecular Biology and  
Institute of Bioinformatics, University of Georgia



# Writing a Scientific Paper



research



data analysis



computational modeling



thinking



ready to write?

# Goal of This Talk

- Help young scientists to better prepare and write scientific papers by
- Writing a paper is like telling/writing a story; it takes a lot of practice to do it well but there are some rules we can follow
  - sharing some of the rules that I learned through my own struggles in getting papers written and published over the years



# Phase 1: Outlining Your Message

- Step #1: before you start writing, sit back and think hard about
  - What key results I want to write about
  - Why someone might be interested in these results
  - Who am I writing the paper for (who is your reader)



# The Goal of Your Writing

- Define what you intend to accomplish through the writing (**what to present**), e.g.,
  - present your new scientific discovery
  - discuss a new hypothesis with partial evidence
  - demonstrate your new algorithm is better than anything out there
- Know who your readers are (**how to present**), e.g.,
  - biologists, computer scientists, general scientific community, or general public
- Understand the expectation of your reader (**put yourself in reader's position when you write**)



# Outline Your Key Results

- List the key scientific results, each in 1-2 sentences, in logical order, e.g.,
  - modeled the tertiary structure of protein X
  - assessment result of the structural quality of X is good
  - proposed the interaction mechanism between X and Y based on the predicted structure
- Note: make sure that your results have a common theme, and indicate the challenging nature in getting the results
- This forms the basis of your Results section

# Outline How You Got the Results

- Step #2: outline how you have got your results by listing the key methods used, each in 1-2 sentences, e.g.,
  - developed a new protein-protein docking method that can fully utilize limited inter-residue distances from NMR
  - applied the Modeller program
  - applied three threading-based structure prediction programs, X, Y and Z
- Note --
  - make sure your high-level description is convincing; and your detailed description allows your reader to reproduce your results;
- This forms the basis of your Methods section

# Outline Why Your Work is Important

- Step #3: Imagine that you are trying to convince your picky friend why your work is important!
  - Explain the problem that you are trying to solve; and why the problem is important in 2-3 sentences
  - Explain what other people have done in tackling this problem in a few sentences
  - Explain what the key remaining issues are, and how your results are related to the ultimate solution of the problem in 2-3 sentences
  - Outline key contributions of your work in terms of advancing the state of the art on this particular issue in 1-2 sentences
  - Explain the reasons for your advancement
- This forms the basis of your Introduction section



# Outline Implications of Your Work

- Step #4: Sit back (again) and think about the bigger picture issues of your work
  - Assess the potential impact of your new research results in 2-3 sentences
  - Consider possible applications in 2-3 sentences
  - Discuss about limitations and possible solutions to overcome these limitations
- This forms the basis of your Discussion section



# Phase 2: Expand Outlines

- In phase 2, expand the outlines into sequence of short paragraphs
  - Results section
  - Methods section
  - Introduction section
  - Discussion section



# Example: Results section

- Expand on “modeled the tertiary structure of protein X”
  - *Protein X has three domains, A, B, C, determined based on ProDom prediction and multiple sequence alignments. A has a close homolog A' in PDB, with sequence XX%; hence we modeled the structure of A based on A's using Modeller. B does not have a close homolog in PDB but has good threading scores against structure B's by three different threading programs. C is a membrane protein with two helices, and we have modeled its structure using an ab initio approach. Side-chains are then added using Molleder. We then docked the three structures into one complex structure using a new docking program that is capable of utilizing limited NMR distant restraints.*

# Example: Methods section

- Expand on “developed a new protein-protein docking method”
  - *We have extended an existing protein-protein docking algorithm X, allowing it to fully utilize experimental inter-residue distance information as docking constraints. The key idea of the constrained docking program is to add a new energy term, which penalizes inter-domain residue-residue distances that are inconsistent with the provided NMR data, using a well-shaped penalty function. The scaling factor of the new energy term is empirically determined based on a training data of known complex structures, using simulated NMR distance data.*



# Example: Introduction section

- Expand on “the problem that you are trying to solve, and its importance”
  - *In this paper, we present a new and efficient computational method for rigorously solving the protein threading problem using provided residue-residue interactions as distance constraints. Protein threading is a general strategy for template-based protein tertiary structure prediction, which is applicable to 70% of the soluble proteins. However no rigorous and computationally efficient algorithms are currently available for solving this problem, which has hindered the full utilization of this general structure prediction strategy.*

# Example: Introduction section

- **Expand on “the reasons for your scientific advancement”**
  - *The fundamental reason that we are able to develop a both rigorous and efficient computer algorithm for solving the threading problem is the realization that pair-wise interactions in a protein threading problem are highly sparse, which gives rise to an optimization problem on a highly tree-like structure when the problem is formulated as a graph-theoretic problem. Specifically, the protein threading problem can be formulated as a combinatorial optimization problem on a graph, whose computational complexity is an exponential function of only the tree-width of the underlying graph. Interestingly the tree-width for all protein structures in the PDB database is no higher than 5.*



# Example: Discussion section

- Expand on “potential impact of new research results”
  - *Using this new threading algorithm, we can possibly assess the effectiveness of the existing threading energy functions in a systematic manner, to identify energy terms that are truly useful in identifying the correct structure-sequence alignments. This issue has been mostly avoided by researchers since there have not been an effective technique that can distinguish between failed threading examples caused by bad energy terms and by the inability of the existing threading algorithm in finding the global optimal solutions of the current energy terms.*

# Phase 3: From Sketches to Text

converting your ideas to a complete story

- Expand each brief paragraph containing  $N$  points into  $N$  paragraphs so each paragraph has one point
- Define notations for concepts that are repeatedly used across the text, in the beginning
- Avoid repetitive material by referring to earlier description
- Check for inconsistencies throughout the text
- Insert references where needed



# From Sketches to Text

converting your ideas to a complete story

- Use “lead sentence” or “transition sentences” to avoid “sudden thematic changes” between two paragraphs, e.g.,
  - **paragraph N on the existing methods for protein threading:** *A developed the first algorithm in 1990; B improved A’s algorithm by allowing consideration of the residue interaction energy; C improved the effectiveness of the interaction energy by B; ....*
  - **paragraph (N+1) on your contribution:** *We have developed a novel algorithm that can both rigorously and efficiently solve the protein threading problem that considers residue interactions.*
  - **add lead sentences:** *While substantial efforts have been put into development of threading methods, one key issue remains unsolved. That is none of the existing algorithms can rigorously solve the threading problem when residue interaction energy is considered, which has clearly under-utilized the power of the current interaction energy.*

# From Sketches to Text

converting your ideas to a complete story

- Ask yourself: **have I presented the story in a logical, clear and convincing manner?**
  - If not, fix all the issues you identified and repeat the above five steps of phase 3
  - Keep doing the above until you are happy



# Phase 4: Polish Your Story

to make it easy and fun to read

## ➤ Polishing at the sentence level

- Is this sentence really needed?
- Is this sentence in the right place?
- Is the sentence too convoluted, too long? Does it read awkward?
- Is the meaning of the sentence clear and accurate?
- Is this the best way to describe what you wanted to say?
- Avoid using the same word more than once in the same sentence
- .....



# Polish Your Story

to make it easy and fun to read

## ➤ Example:

- *There are two major microarray technologies, cDNA arrays and oligo-nucleotide arrays. They have been extensively used to study the gene expression patterns in biological samples, after they were introduced around 1996. It is essential to use statistical algorithms to explore through the huge volume of microarray data for biological hypothesis.*

## ➤ Comment:

- Sentence #1 contains outdated information
- Sentence #2 is not particularly informative
- Sentence #3 reads awkward and its meaning is not very clear

# Polish Your Story

to make it easy and fun to read

- **Revised:** *Microarray gene expression chips provide a powerful tool for studying transcription and regulation in an organism. Using this technique, scientists have been generating gene-expression data for various organisms under designed conditions, aiming to collect comprehensive enough datasets to allow elucidation of global/local transcription regulation networks. Among the analysis techniques of the gene-expression data is the biclustering method, .....*

# Polish Your Story

to make it easy and fun to read

## ➤ Polishing at the paragraph level

- Does the paragraph contribute the main theme of the paper?
- Does the paragraph have a cohesive theme with a point?
- Does the paragraph cover everything that needs to be said but no more?
- Can you possibly shorten the paragraph?
- Does the paragraph flow well logically, and is easy to follow? Is this the best structure to make the point intended to make?
- Is the presentation clear and to the point?
- .....



# Polish Your Story

to make it easy and fun to read

- *Example: To summarize, the above information sources are surely useful and give many successes in different processes, but they are all limited if being used separately in single species because of the discrepancy between the complexity of organisms and severe simplifications required. For example, functional modules are usually evolutionary among different species. Many operons may have divided into small segments and orthologs or paralogs may acquire new functions in different species. To overcome the drawbacks, it is promise to use many species to construct a functional evolutionary network. The hope is the real information may be retained in most species and can be enhanced by a multi-species network. Based on this idea, we use operon information and homology information to construct a reference graph, and find the function related gene pair in target genome by systematic relationship discovering.*



# Polish Your Story

to make it easy and fun to read

## ➤ This paragraph has the following key points

- Homology and operon information are both useful to prediction of [pathways];
- Using these two sources of information separately has limited usefulness;
- Combining them through application of additional genomes could possibly fully utilize the available information

## ➤ .. however its presentation has various problems

- Overall, lack of clarity and good flow;
- Most of the sentences did not really get to the point (quite vague);
- Some statements are made out of context, hence confusing

# Polish Your Story

to make it easy and fun to read

- **Revised:** *For a target genome, we define a distance between any pair of genes in the genome to measure the level of their functional relatedness in terms of a set of reference genomes. Specifically, two genes are functionally related if they are homologous, sharing a common operon directly or through their homologs in a reference genome, or deemed to be functionally related through combinations of the first three criteria. For any pair of functionally related genes in the target genome, their distance is defined as the minimum number of applications of this recursive definition. Our algorithm identifies genes possibly involved in a target pathway based on their distances to genes already in the pathway.*

# Polish Your Story

to make it easy and fun to read

## ➤ Polishing the whole text

- Is the material organized in a logic manner?
- Do all the pieces hung well together?
- Are there any redundancy or logic gaps?
- Do all the points made follow a common theme?
- Is the paper too dense, too dry and too boring to read?
- Is the paper presented at the right level for the targeted reader?
- Is it possible cut any part of the paper without affecting the integrity of the paper?
- .....



# Polish Your Story

to make it easy and fun to read

## ➤ Is the paper too dense, too dry to read?

- Avoid using too many mathematical formula or technical discussions
- Include some examples in between to dilute the dense material (to give the reader a breather)
- A good practice in dealing with heavy mathematical (technical) part of a paper is to outline the general idea of the mathematics in layperson's language first; and follow the mathematics with examples



# Polish Your Story

to make it easy and fun to read

- **Do all the pieces hung well together?** -- think if you have told your reader
  - all the needed background information;
  - all the exciting results;
  - all the methods used to achieve the results;
  - all the potential implications of your results
- **.. and have done it in such a way that you can summarize your main points to a smart outsider in a few minutes!**
  - Remember the process of polishing is to make the presentation crystal clear, easy to follow so your reader can capture your summarized points!



# Phase 5: Zoom Out

## Think the Big Picture Again

- Does the Introduction set the stage properly for the Results?
  - Make sure that your reader knows what problem you are trying to solve; why it is important to solve it; how your solution is better than the previous ones; and the key ideas that lead to a better solution
- Do the Results look clear and convincing?
  - Make sure not to let your brilliant ideas get buried in technical details; and to provide enough background information for the (targeted) reader to follow the Results section



# Zoom Out

## Think the Big Picture Again

- Does the Methods section provide all the information needed to understand how the Results are derived?
  - Make sure to consider who your reader is!
  - The presentation gives enough details for the reader to reproduce the results
- Does the Discussion section cover all the major points about the Results?
- Check the paper against the background of the anticipated reader – too much or too little details? missing any key background information?



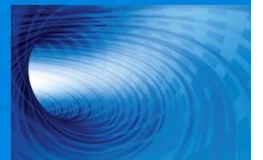
# Phase 6: Complete the Writing

- Include proper references, making sure that you cite all the key relevant papers
- Include all the necessary figures/tables to help you to make your points more clear and more convincing
- Include all the people/agencies that need to be acknowledged
- **Most importantly write Abstract, Conclusion and Title!**

# Writing a Summary of the Story:

## The Abstract

- Include a brief problem definition, the challenging nature and/or the importance of the problem
- State the most interesting results of your study in a succinct and logic manner
- Summarize the key contribution of your work



# Designing the Title

- The title of your paper should clearly and accurately tell your reader what the paper is about
- The title should be eye-catching, making the reader be interested in reading the paper



# Things for the Reader to Remember:

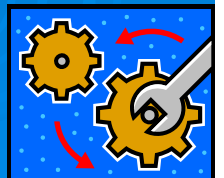
## Concluding remark

- List the key things that you want to your reader to remember about your work
- Present them in a way that your reader will be looking forward to your follow-up studies



# Fixing the Language Bugs

- Use spell-checker to make sure all the words are spelled correctly
- For words whose meanings you are not sure about, try [google language tool](#)
- Fix all grammatical errors as highlighted by Microsoft Word



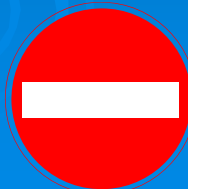
# Do's

- Use simple and short sentences whenever you can
- Use lead and/or transition sentences between paragraphs
- Define short notations for repeatedly used concepts, and use them
- Be consistent with your terminologies throughout the paper
- Remove anything that can be removed without losing any information
- Use Figures/tables to help make your point



# Don'ts

- Do not use convoluted sentences
- Do not use a concept before defining it
- Do not repeat yourself in the paper
- Avoid formalism whenever you can
- Do not let your brilliant ideas buried in the technical details
- Do not include any material that does not contribute to the central theme of the paper
- Do not include material that is inconsistent with the level of knowledge of your targeted reader



# Take-Home Message

- Think hard and organize your thoughts about the messages that you want to deliver
- Outline them using the framework of “Introduction – Results – Methods – Discussion”
- Repeatedly expand and refine your outline to a polished text
  - Making sure the text is clear, easy to follow, convincing, fun to read and should leave the reader the impression that the work is important and understand why it is important
- Add Abstract, Concluding remarks and Title at the end



*Have Fun Writing Your Next Paper!*

*THANK YOU!*

