How to read a scientific paper

Mihai Pop
Computer Science
Center for Bioinformatics and Computational Biology

Reasons to read a paper

- You were told to
- Describes current research
- Allows you to replicate/extend the results
- Provides you with useful data
- Gives you “pre-digested” thoughts
- To decide whether to publish it

- Teaches you how to write.

Not all papers are equal

- Some you read to fully understand everything – may spend a lot of time reading the paper
- Some you read to get the "gist" – may just read abstract/intro and bounce a bit through results

Reading “mechanics”

- Remove distractions (Red Sox or paper - pick one)
- Take notes & save notes for future reference
- Jump around through the text, don't just read it like a Harry Potter book

Types of papers

- Theoretical
  - prove theorems
  - describe new algorithms
- Implementation
  - describe new software tools
- Experimental
  - describe results of experiments
- Survey/Review
  - review current results in a field of research
- NOTE: not mutually-exclusive, most papers are a mix

Primary vs. secondary sources

- Primary
  - actual description of the work/results reported
- Secondary
  - describe work/results of others
  - e.g. background section in most papers
  - survey papers
  - encyclopedias (e.g. Wikipedia)

- Try to read the primary references (though secondary references are quite useful too)!
- e.g. Mozart and babies

Paper organization

- Title & author list
- Abstract
- Introduction
- Materials and Methods
- Results
- Discussion/Conclusion
- Open problems

- Depending on the journal/conference/type of work these can vary in content/order

Theory paper

- Introduction
- Preliminaries
- Specific topic 1
- Specific topic 2, ...
- Future work
- Conclusions

P/L paper (from Mike Hicks' talk)

- Abstract (4 sentences)
- **Introduction** (1 page)
- The problem (1 page)
- My idea (2 pages)
- The details (5 pages)
- Related work (1-2 pages)
- Conclusions and further work (0.5 pages)

Venue

- First things first: Where was the paper published?
- If the work is similar to what you do, this should give you ideas about which journals/conferences you should target with your own work
- Over time, you'll learn to evaluate journal/conference quality based on the quality of papers you read.

Title and authorlist

- Title
  - what is this paper about?
- Author list
  - who did the work? where are they from?
  - try to remember the names: these people may become collaborators, colleagues, or bosses sometime in the future.
  - also useful when planning a postdoc or future job
- Author list conventions
  - alphabetical (traditional CS)
  - ranked: first author did most work, last author (senior author) led the study (usually the PI)

Abstract

- Brief outline of the results presented in the paper
- Read it carefully
  - Can you understand what the paper is about?
  - Do the conclusions make sense?
  - Can you come up with a solution to the problem addressed by the paper?
  - How comfortable will you be reading this paper?

- Note: from any paper you should at least read the title, author list, and abstract
- Mike's opinion – often just jump into the intro

Introduction

- Introduces the problem(s) addressed in the paper and prior art
- Questions to ask:
  - now that the problem is stated in more detail than in the abstract, can you think of a solution (or conclusion)?
  - is enough/any prior art listed? If not, why? Is the author hiding anything?
  - can you see why this paper is an advance over what was done in the past?
- Introduction will also give you pointers to other papers you might want to read

Materials and Methods

- The “meat” of the paper - how the work was performed.
- Play the guessing game: for every problem or theorem stated, try to think of a solution before reading any further.
- Is sufficient information provided for you to understand how the paper “works”? What's missing? Is the paper correct?
- Note: in conferences papers are often “extended abstracts” - many details are missing. Try to fill them in.

Results

- Verbose conclusions of the paper
- Often this section also contains “materials and methods”-type content
- Questions to ask:
  - what conclusions can you draw from the data presented? (ask before the paper “brainwashes” you)
  - does the experiment/data support the conclusions described in the paper?
  - are there alternative conclusions that the authors did not consider?
  - how would you set up the experiment?
- Make sure figures do not lie

Conclusions

- The authors' summary of the contributions provided by the paper.
- Often, also philosophical discussions on the problem, or field of research
- Questions to ask:
  - do you agree with the authors' conclusions?
  - what are your own conclusions?
  - do the authors' conclusions derive logically from the material presented in the paper?

Open problems

- Many “traditional” CS papers end in an open problems section - questions the authors have asked themselves but cannot easily answer.
- This section is very important
  - provides you with problems you might want to work on
  - tests your understanding of the paper - many open problems are questions you should have asked yourself while reading the paper.

- E.g. paper describes an $O(n^{1\text{-epsilon}} \log \log n)$ algorithm - natural question: is this a lower bound as well?

Two papers

- *Initial sequencing and analysis of the human genome.* International Human Genome Sequencing Consortium, Nature 409, 860 (2001).
- http://www.nature.com/nature/journal/v409/n6822/full/409860a0.html

- *Microbial Genes in the Human Genome: Lateral Transfer or Gene Loss?* Steven L. Salzberg, Owen White, Jeremy Peterson, Jonathan A. Eisen. Science 292:1903-1906 (2001)
- http://www.sciencemag.org/feature/data/scope/keystone1/

Paper 1

- Conclusion: at least 223 genes were transfered from bacteria to humans
- (note: this event is extremely unlikely - one should be skeptical)
- Method:
  - find all genes similar between humans and bacteria yet not found in any other “complex” organism
- Logical link:
  - if an ancestor of both humans and bacteria had any of these genes, it's unlikely they would have been lost in all “complex” organisms but preserved in both human and bacteria.

Paper 2

- Conclusion: Not so fast, batman....
- Hypothesis:
  - there are many reasons why one might not find the genes in other “complex” organisms
  - e.g. we haven't sampled enough of them yet
- Method:
  - same as in the previous paper
- Results:
  - many of the “transfered” genes disappeared once more “complex” organisms were found
- New Conclusion: first paper was likely wrong

Other resources

General resources

- General writing resources (at Purdue)
  http://owl.english.purdue.edu/

On reading

- How to read a paper by S. Keshav.
  http://www.sigcomm.org/ccr/drupal/files/p83-keshavA.pdf
- Reading scientific papers (at Purdue)
  http://www.lib.purdue.edu/phys/inst/scipaper.html

Reference organizers

- Connotea – reference organizer
  http://www.connotea.org/
- Zotero – firefox extension reference manager
  http://www.zotero.org/
- Comparison of reference manager software tools available
  http://en.wikipedia.org/wiki/Comparison_of_reference_management_software