Efficient representation of uncertainty in multiple sequence alignments using directed acyclic graphs

Adrienn Szabó

Eötvös University, Budapest (ELTE)
and
DMS Group
MTA SZTAKI

July 2, 2015

Table Of Contents

About me

About me

Education

- MSc: Software engineer, Budapest University of Technology and Economics
- PhD: Data mining techniques on biological data (supervisors: *András Benczúr, István Miklós*), Eötvös University, Budapest (finishing in 2015)

About me

Research interests

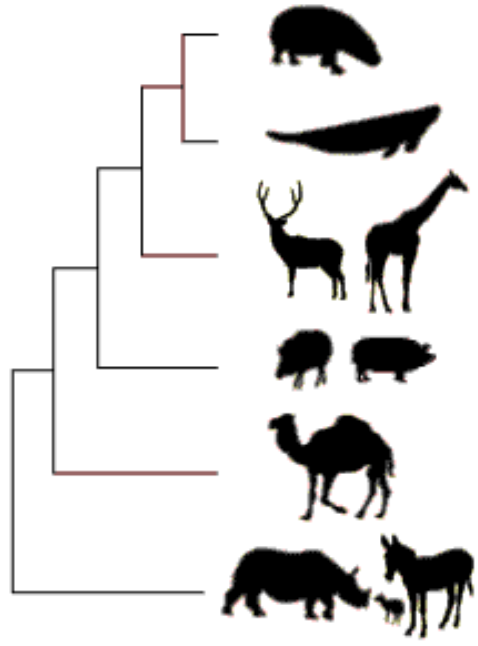
- Bioinformatics, especially multiple sequence alignment, and problems with a lot of data
- Data mining, machine learning, text mining, especially on biological datasets

Work

- Developer and research assistant at Data Mining and Search Group (head: András Benczúr), MTA SZTAKI (2007 -)
- Software engineer intern at Google Zürich (2009)

MSA – Introduction

- Multiple sequence alignment (MSA): alignment of three or more biological sequences
- Needed for phylogenetic analysis, function prediction of proteins, etc.



```
                 :   :      **  *                 : .   :    :: *. *:*   :
1mnmC      TKPYRGHR-FTKENVRILESWFAKNIENPYLDTKGLENLMKNTSLSRIQIKNWVSNRR---RKEKTITIAPEL
1au7A      RKRKR-RTTISIAAKDALERHFG---EHSKPSSQEIMRMAEELNLEKEVVRVWFCNRRQREKRVKT-SLNQSL
1akhA      KSPKG-KSSISPQARAFLEEVFR---RKQSLNSKEKEEVAKKCGITPLQVRVWFINKRMRSK----------
1fjlA      KQRRS-RTTFSASQLDELERAFE---RTQYPDIYTREELAQRTNLTEARIQVWFQNRRARLRKQ----HTSVS
1ftt       MRRKR-RVLFSQAQVYELERRFK---QQKYLSAPEREHLASMIHLTPTQVKIWFQNHRYKMKRQAK-DKAAQQ
1ftz       MDSKRTRQTYTRYQTLELEKEFH---FNRYITRRRRIDIANALSLSERQIKIWFQNRRMKSKKDRTLDSSPEH
```

Basics – pairwise sequence alignment

- The standard edit distance based formulation of sequence alignment leads to $\mathcal{O}(L^2)$
- Dynamic programming: *Smith-Waterman* and *Needleman-Wunsch* algorithms

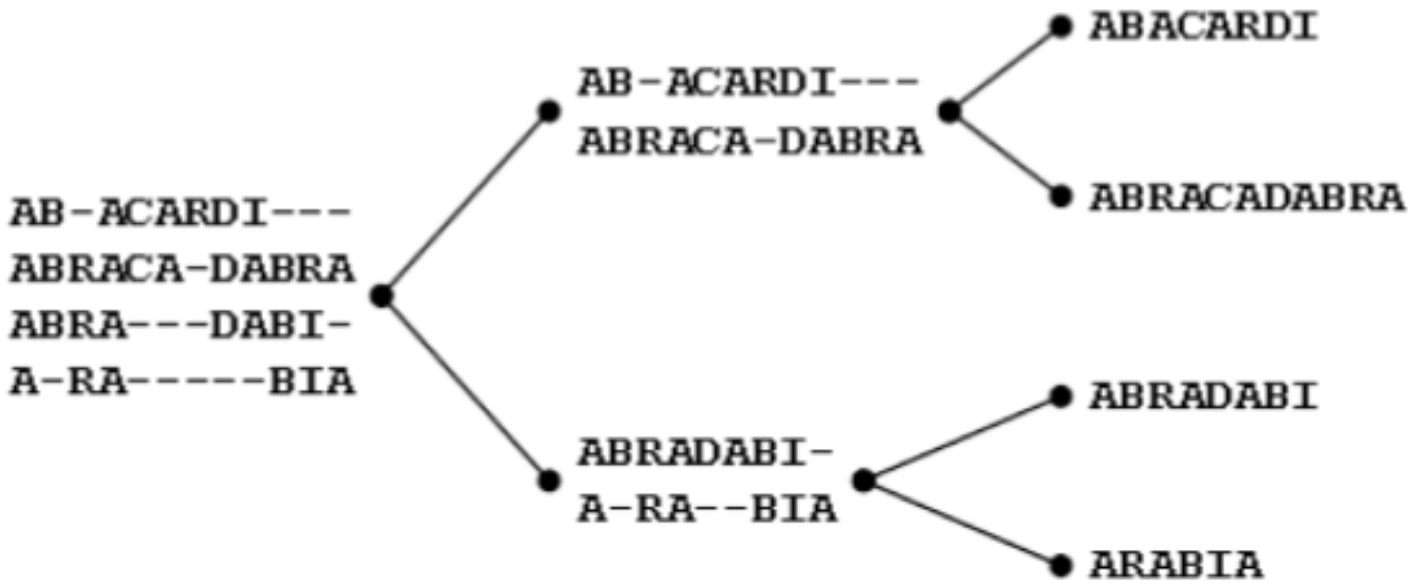
```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :*****.:* ********..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .****************    :  *.: :
```

Problems with MSA

- Simple DP solutions: each additional sequence multiplies the time and memory required
- Finding the optimal alignment is NP-complete
- Corner-cutting methods shrink the search space, but are still exponential in memory and running time
- Heuristics are applied: progressive alignment

Progressive alignment

- Using heuristics: running a pairwise alignment algorithm, many times
- A guide tree defines which pairwise alignments will be done in order (one at each inner node, from leaves to root)
- Polynomial running time :)



Uncertainty of alignments

Because of the heuristics used, errors may be introduced:

- the guide tree migth not be accurate
- a gap inserted near the leaves can not be removed later
- a mis-aligments can not be fixed at upper levels of the guide tree

Dependance on parameters

Even if we do not use any heuristics, parameters of the alignment algorithm might significantly affect the final result:

- similarity matrix (score matrix)
- gap opening penalty
- gap extension penalty

A tiny example

PAM40 matrix, gop = 10, gep = 0.5

```
sponge  SPONGEBOBSQUAR--EPANT--S
barbie  ---------B---ARBIEPARTIES
```

Blosum62 matrix, gop = 10, gep = 0.5

```
sponge  SPONGEBOBSQUAREPANTS--
barbie  ---------BARBI-EPARTIES
```

Blosum62 matrix, gop = 1, gep = 1

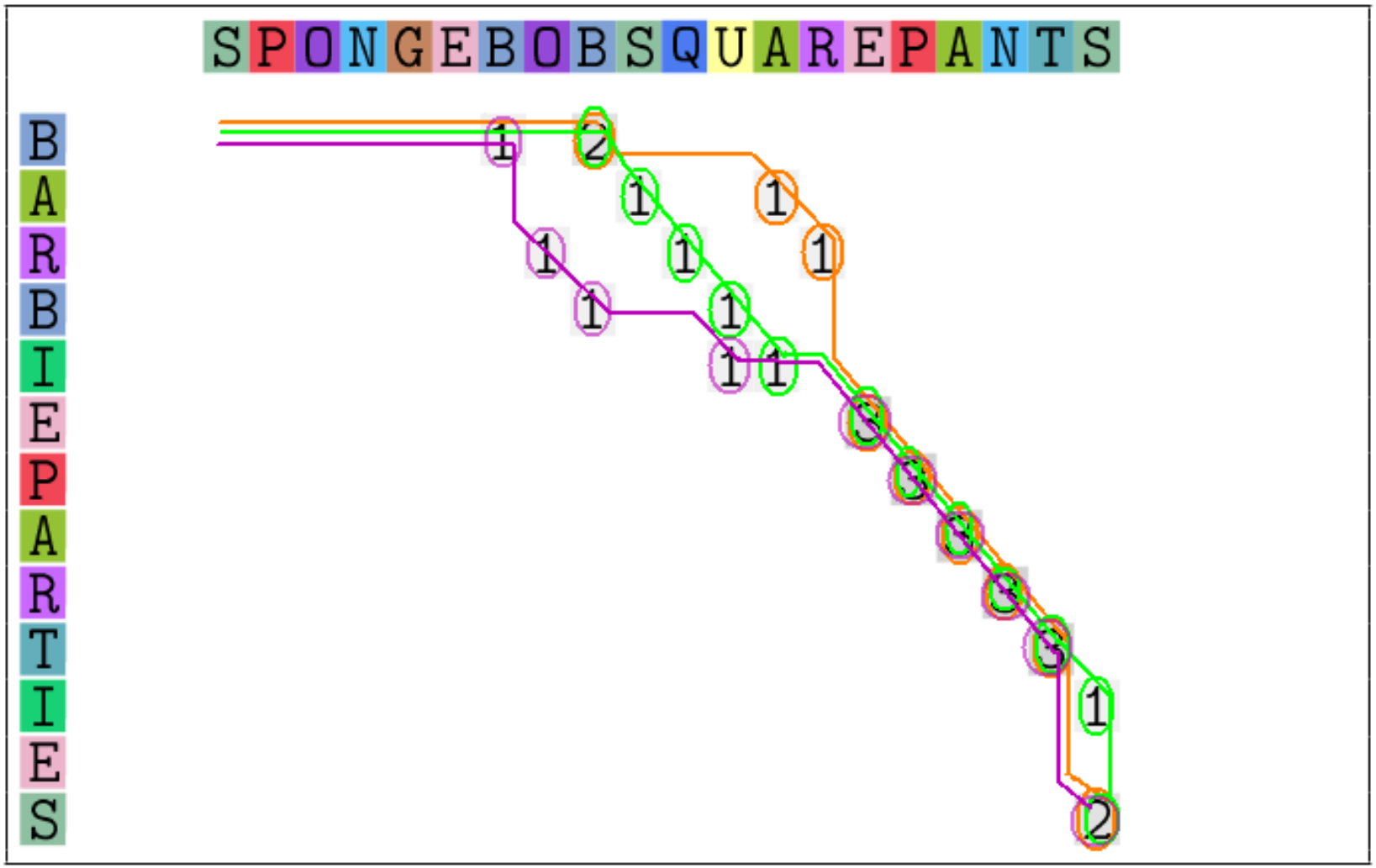
```
sponge  SPONGEB-OBSQUAREPANT--S
barbie  ------BARB--I--EPARTIES
```

How to handle uncertainty?

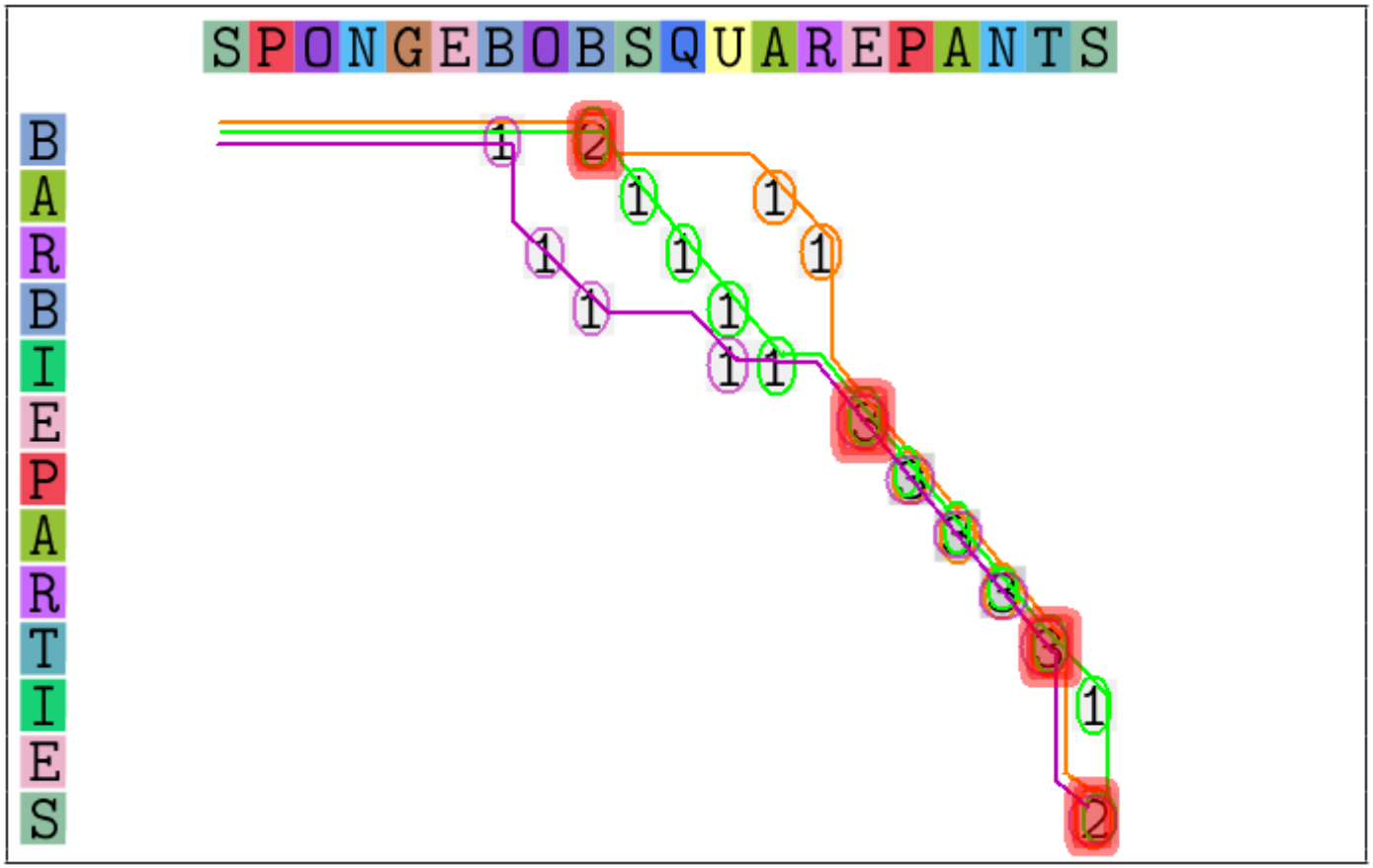
Imagine having thousands of alignments of the same sequence set.

- How do we choose 'the best' one?
- How do we know which parts of an alignment are reliable?
- How could we summarize the many alignment paths efficiently, and use the information from all of them in subsequent analysis?

The tiny example

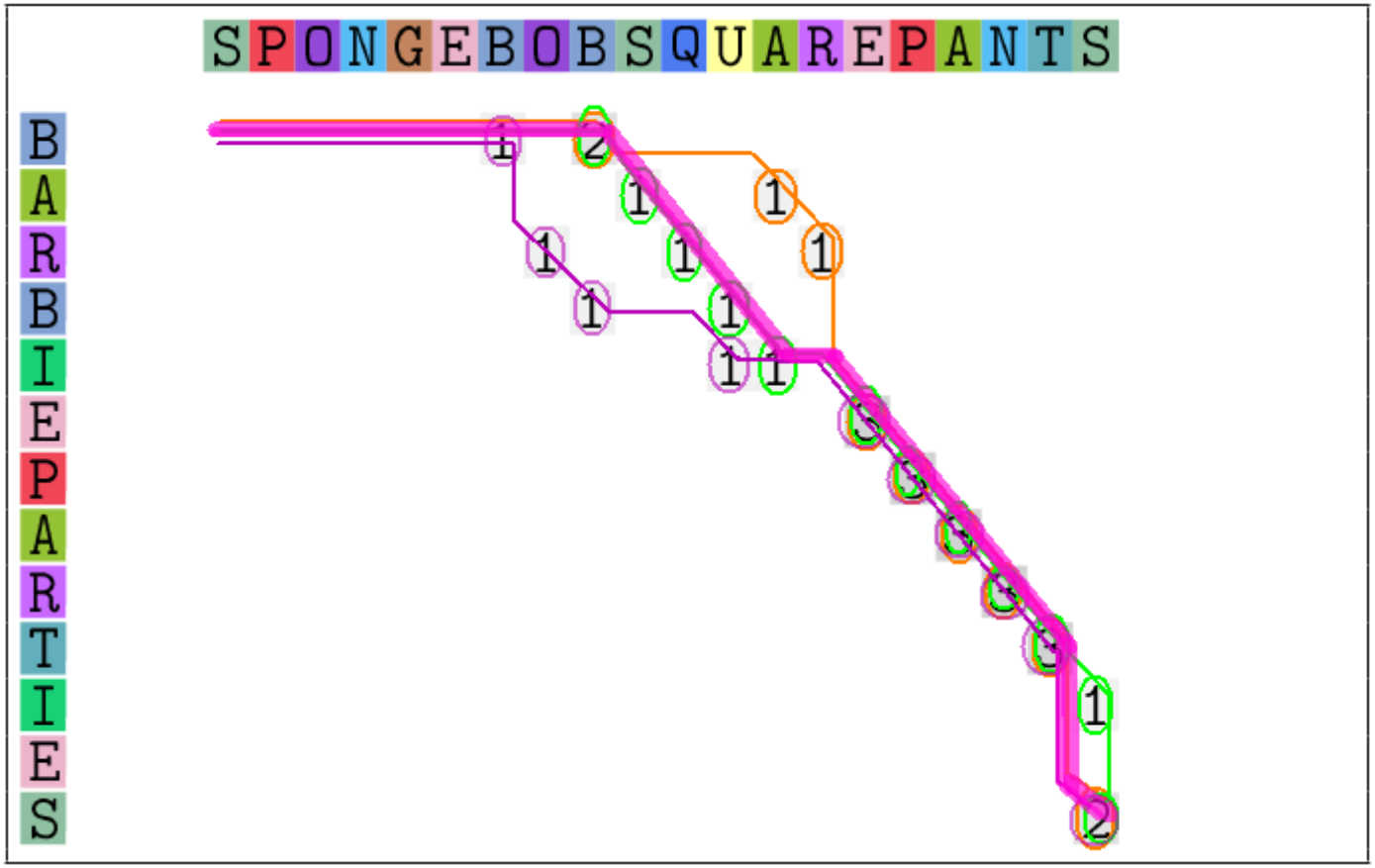


Alignment paths



The input alignment paths can be joined together to form a network (a DAG)...

The tiny example



. . . and new paths can be created.

And now what

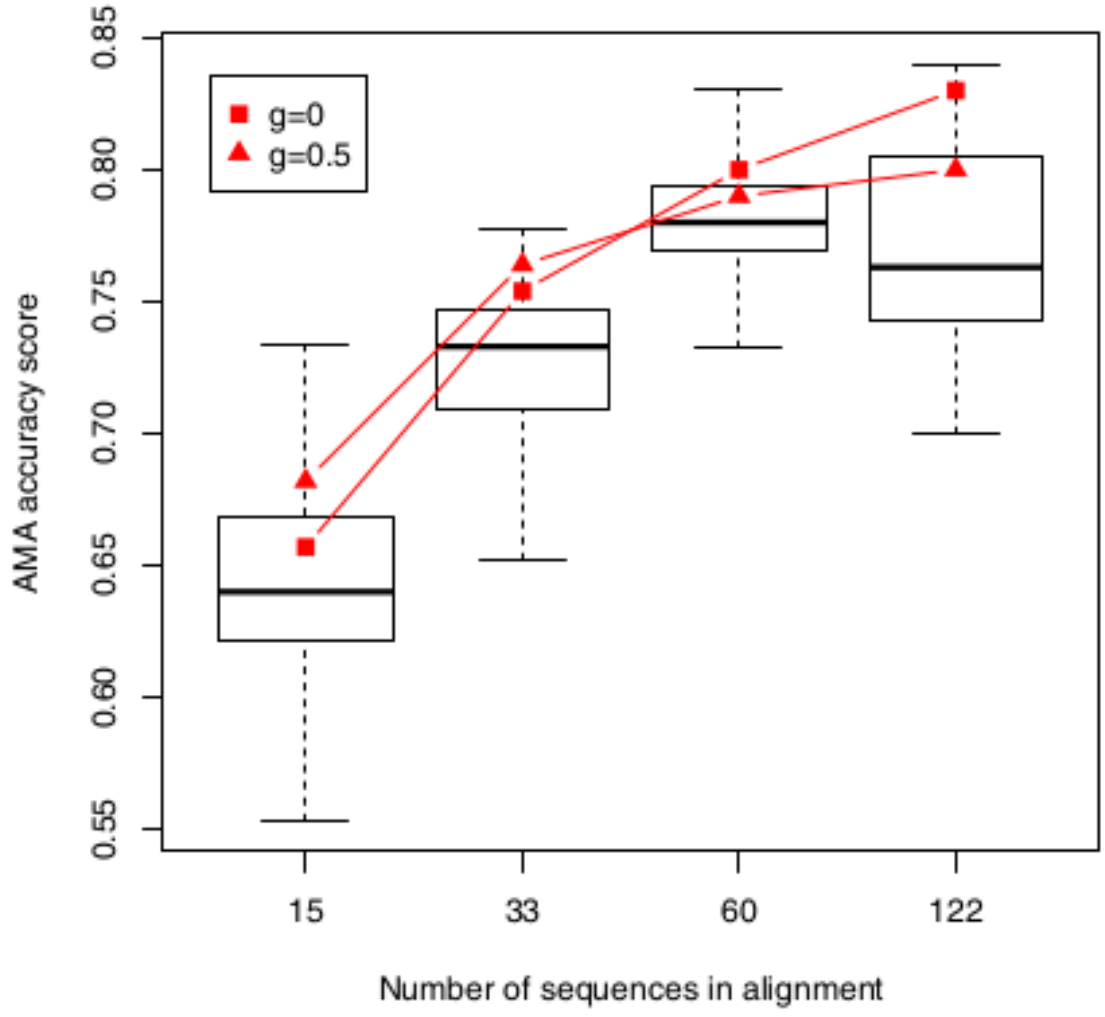
- We can generate orders of magnitudes more new alignment paths via joining the input paths at their common alignment columns
- Then we can take a sample from the paths according to their probability
- Finally, we can derive meaningful statistical estimates of alignment reliability, conserved regions, etc., as well as a most probable summary alignment

Measurements

What we did:

- Generate many (500–5000) alignments for a sequence set (by adding random noise to a similarity matrix)
- Build up the alignment network
- Take a sample from the available paths (according to a statistical model, with an MCMC procedure)
- Create a summary alignment, which is the „best” according to the selected model

Measurements



References

- J. L. Herman, Á. Novák, R. Lyngsø, A. Szabó, I. Miklós and J. Hein:
  Efficient representation of uncertainty in multiple sequence alignments using directed acyclic graphs, BMC Bioinformatics, 2015
  `http://www.biomedcentral.com/1471-2105/16/108`
- J. L. Herman, A. Szabó, I. Miklós and J. Hein:
  Approximate statistical alignment by iterative sampling of substitution matrices, arXiv, 2015
  `http://arxiv.org/abs/1501.04986`

Questions?

Follow me (**adorster**) on twitter!

Reproducible Research in Bioinformatics and Data Mining

Adrienn Szabó

DMS Group, MTA SZTAKI

October 2, 2014

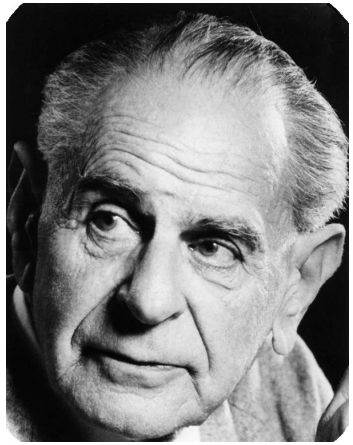
What is (not) Reproducible Research?

Courtesy of S. Harris, ScienceCartoonsPlus.com.

What is (not) Reproducible Research?

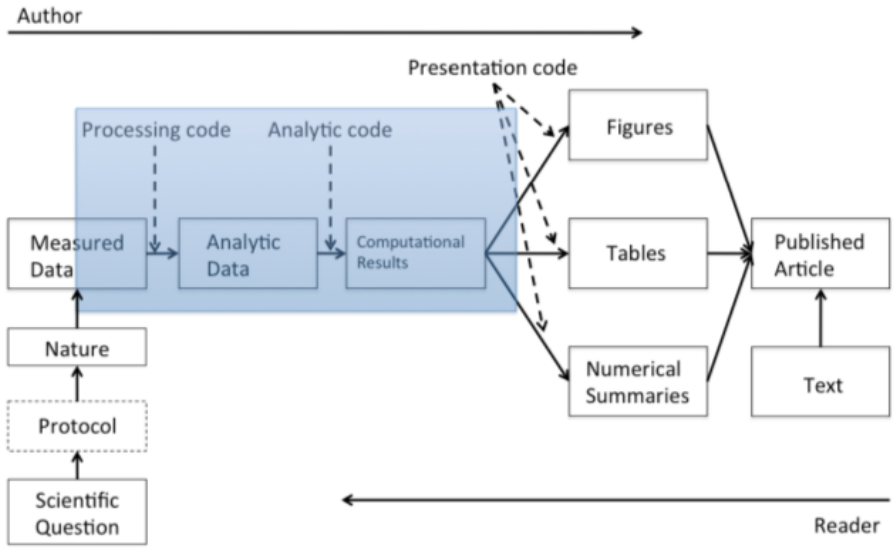
If you observe (or measure, simulate) something but it's not repeatable or reproducible, then it's NO science.

> "***. . . non-reproducible single occurrences are of no significance to science.***"
>
> — Karl Popper

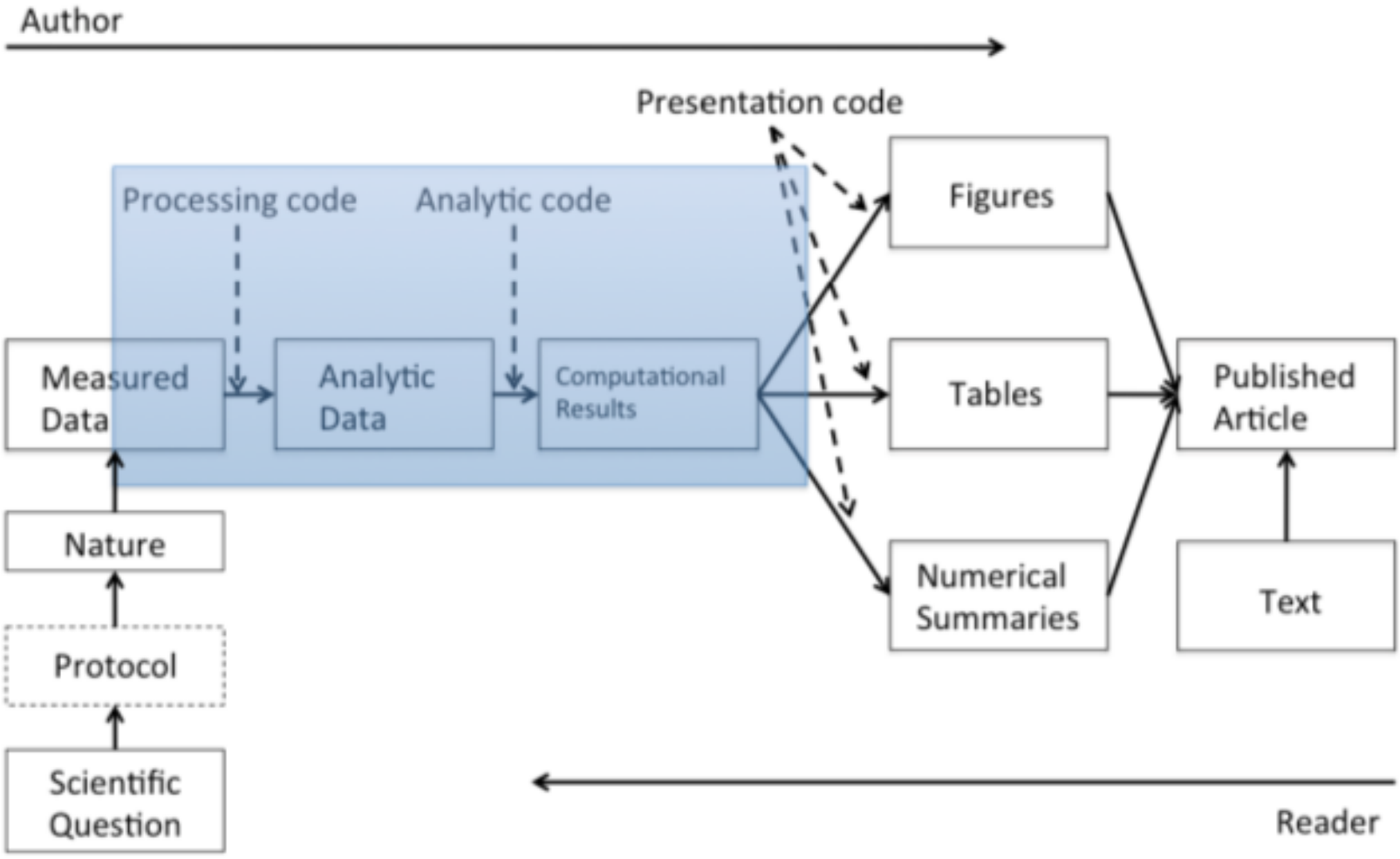


What is Reproducible Research?

Reproducible research is the idea that data analyses, and more generally, scientific claims, are published with their data and software code so that others may verify the findings and build upon them.



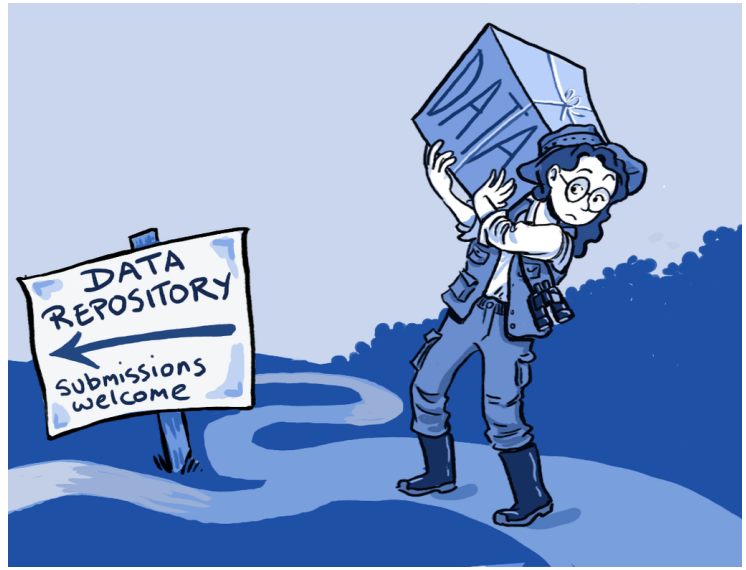
What is Reproducible Research?



What is Reproducible Research?

Related ideas / movements:

- open access
- open source
- open data
- literate programming



A.K.A : Open Science

"*It's a tragedy we had to add the word open to science.*"

How did we end up here?

- "Science is in a crisis of (non) reproducibility."
- "I often found it difficult to replicate previous scientific results."
- "I was frustrated at my inability to identify the precise organisms, probes, antibodies and other scientific materials that underpinned genotype-phenotype assertions in the literature."
- "The lack of specificity in the literature was initially shocking to me"

Source: `peerj.com/about/author-interviews/`

What could the reasons be?

- publication pressure, a feeling that there’s no time to "do it right"
- it is a fairly new phenomenon in science that experiments are run mainly / solely on computers: lack of accepted standards / routines for workflows
- some datasets are not free, or too big: not easy to handle without an expensive infrastructure
- many reserch papers are lacking details on purpose to make sure that a follow-up paper can NOT be done by someone else

Why do we need Reproducible Research?

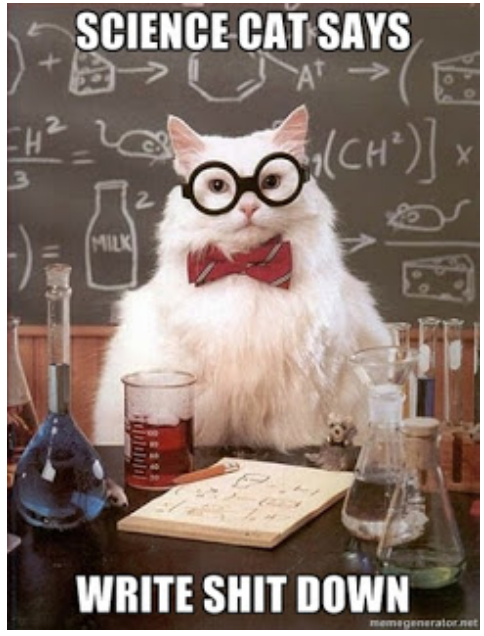
- to reduce the chances of embarrassing errors and faulty results
- to avoid multiplied efforts to reach the same results
- to save time (on the long run)
- to enable others to build upon it
- to increase public trust in science

What has been done?

- Reproducibility manifesto
  `lorenabarba.com/gallery/reproducibility-pi-manifesto/`
- Coursera course on reproducible research
  `www.coursera.org/course/repdata`
- Publications about the issue (see later)
- More and more journals require publication of datasets and codes along with a paper

What can WE do?

- at least write down everything you did (keep "lab notes")
- track & test & document your code
- publish in open access journals
- talk about the problem with other researchers
- take the "Reproducible Research" course on coursera :)



What can WE do? - Manifesto 1

The Reproducibility PI Manifesto

1. I will teach my graduate students about reproducibility.
2. All our research code (and writing) is under version control.
3. We will always carry out verification and validation (V&V reports are posted to figshare)
4. For main results in a paper, we will share data, plotting script & figure under CC-BY

The pledge - Manifesto 2

4. We will upload the preprint to arXiv at the time of submission of a paper.
5. We will release code at the time of submission of a paper.
6. We will add a "Reproducibility" declaration at the end of each paper.
7. I will keep an up-to-date web presence.

arXiv.org

What are the obstacles / challenges?

Factors against reproducibility and open science:

- Laziness: it takes effort to make all data/results/code available
- Lack of convenient tools
- Lack of incentives
- Some are afraid of opening up their "lab notebooks" before everything is published, because someone might steal their ideas

Summary

What is not reproducible is not science

Related publications & sources I

- www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003285
- www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0067111
- www.jove.com/blog/2012/05/03/studies-show-only-10-of-published-science-articles-are-reproducible-what-is-happening
- www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble
- phys.org/news/2013-09-science-crisis.html
- twitter.com/openscience/status/446942010554191872
- peerj.com/about/author-interviews/
- politicalsciencereplication.wordpress.com/2014/02/25/replication-workshop-what-frustrated-students-and-why-they-still-liked-the-course/
- www.wired.com/2014/07/incentivizing-peer-review-the-last-obstacle-for-open-access-science/

Related publications & sources II

- yihui.name/en/2012/06/enjoyable-reproducible-research/
- yihui.name/slides/2012-knitr-RStudio.html#3.2
- biomickwatson.wordpress.com/2014/07/16/how-not-to-make-your-papers-replicable/
- kbroman.org/Tools4RR/assets/lectures/10_bigjobs_withnotes.pdf
- ivory.idyll.org/blog/ladder-of-academic-software-notsuck.html
- www.nature.com/nature/focus/reproducibility/
- ropensci.org/blog/2014/06/09/reproducibility/

Some more collected on the Twiki page:
info.ilab.sztaki.hu/twiki/bin/view/Main/ReproducibleResearch