A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble

PLoS Computational Biology 2009

Motivation

You are at the start of a new project and want to

- pre-process data
- run data through several models, algorithms (code still to be written by you)
- write regular progress reports

Excel, csv, .R, preliminary results, plots, final results, reports.

All files in one folder may not be the best approach.

Motivation

You are at the start of a new project and want to

- pre-process data
- run data through several models, algorithms (code still to be written by you)
- write regular progress reports

Excel, csv, .R, preliminary results, plots, final results, reports.

All files in one folder may not be the best approach.

$\rightsquigarrow$ Author describes "one good strategy" (personal experience) for carrying out computational experiments

- organizing files and directories, and documenting progress.

Core guiding principles

Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why.

Core guiding principles

Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why.

Someone?

Core guiding principles

Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why.

Someone?

- someone who read your published article and wants to try to reproduce your work,
- a collaborator,
- a future student working in your lab,
- your research advisor,
- most commonly, however, that "someone" is you.

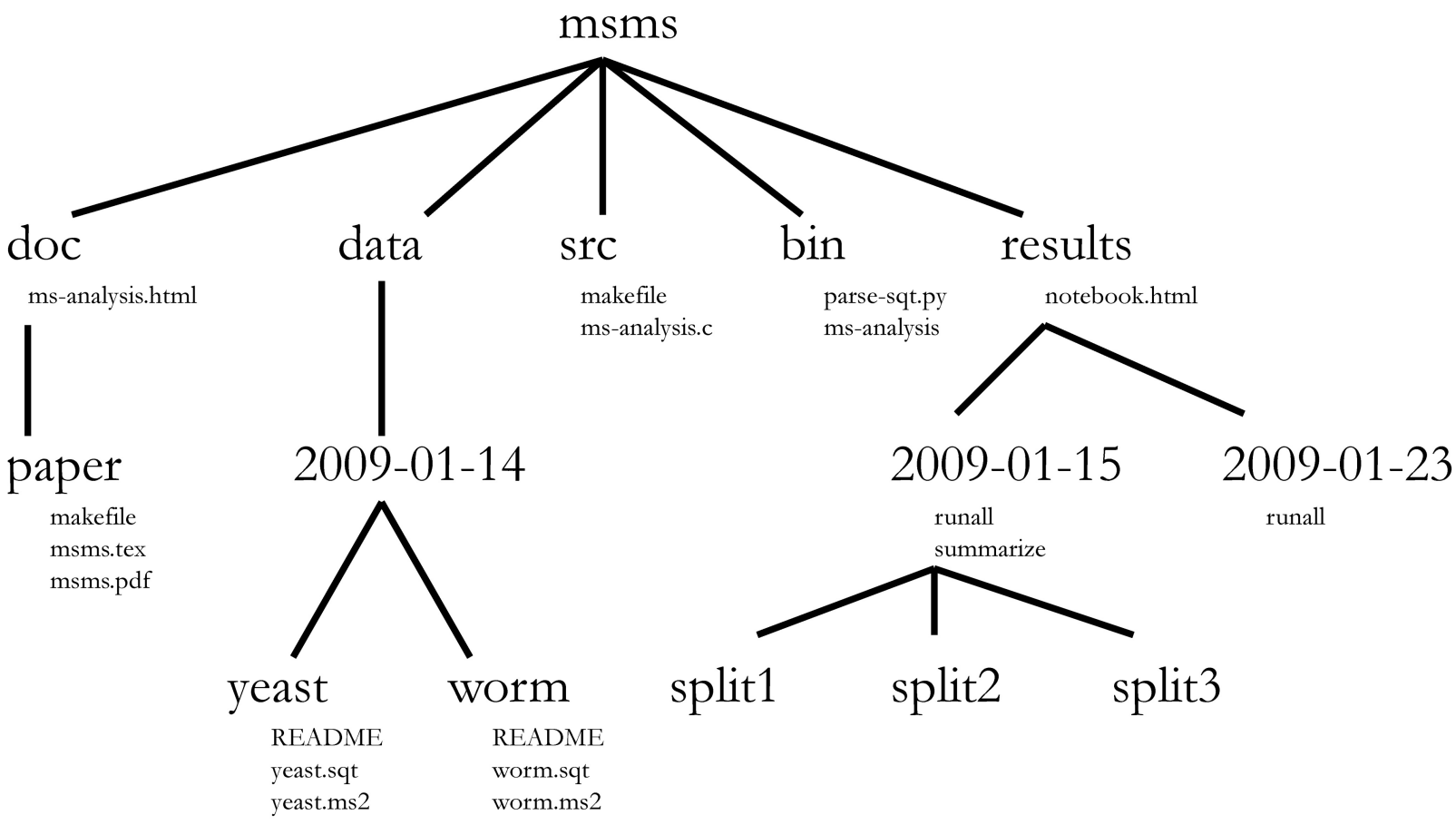
Second principle

Everything you do you will probably have to do over again.

- flaws in algorithm,
- new data,
- broader parametrization,
- reviewers wants modifications.

File and directory organization

Common root directory, keep it chronological.



Lab notebook

Keep in *results* directory

- record observation, your interpretation and conclusion, questions, future ideas.
- esp. when experiment fails or doesn't give expected result. why this is a fail may not be obvious to *that someone*.
- add notes from conversations, emails, meetings with advisor.
- If you want, put notebook online for project team to read.

Single experiment

readme file for each

- have a file runall to make everything automatic, best if this also creates summary. *e.g. run R script first, save plots, put them in LaTeX or Word document.*
- avoid editing intermediate files by hand
- use relative pathnames, not absolute
- if script has long run time use things like
  `if (output does not exists) perform operation; otherwise next step`
- or use function summarize that is called in last line of runall. summarize should then also work with partial results.
- outputs should be temporary files to avoid taking partial results for final/full results.