

Ten (Okay, Twelve) Bioinformatics Commandments



Modified by Dietlind Gerloff
for use in BME110/BIOL181
Winter 2008

© Wiley Publishing, 2007. All Rights Reserved.



1. The Web Is Not a Secure Place

- ✓ Sending your data to a Web server is like making it public
- ✓ If you work for a pharmaceutical company, this might be a bad idea
- ✓ Likewise, don't submit patients' private data over the Web (or anything confidential about yourself)

2. Keep Track of Database and Program Versions

- ✓ Reproducibility is important in science but problematic in applied bioinformatics due to the dynamic nature of the Web
- ✓ The only way to reproduce a bioinformatics experiment is to use the same tools and resources... at the same time
- ✓ Remember (and write down)
 - Which program was used (and which version)
 - Which database was used (and which version)
 - When you accessed the program/server (if it is a dynamic Web)

3. Keep Track of Your Sequences

- ✓ Always write down your sequences' accession numbers and/or database IDs
- ✓ Accession numbers are typically *not* the same in different databases, though the most common large databases will cross-reference the IDs of your sequence in others
- ✓ The best policy is to write down the ID your sequence has in one of the major databases and consult journal instructions!

4. Remember Your Program Parameters

- ✓ If you change default parameters, write down your changes!
- ✓ Keeping a “lab book” (electronically) is essential
- ✓ This will be the only way to reproduce your experiment.
- ✓ If you did not change the default parameters you can report having used those - now you can see why the date on which you used a web server could be very relevant!

5. Save the Right Results

- ✓ Save the flashy graphs for publication & double-checking
- ✓ Save the boring ASCII (text) files for further work!
- ✓ ASCII files include
 - XXXX.aln file for MSAs
 - XXX.dnd or XXX.ph files for trees

6. *Know Thine E-Values*

- ✓ Alignments/similarity are evaluated with E-values
- ✓ People talk about percentages of sequence identity by the coffee machine but they have to **write down** E-values in publications!
- ✓ The lower the E-value, the better the alignment

7. *Know Thine Alignments & Trust Your Expert Judgment*

- ✓ Evaluating alignments (or any bioinformatics result) is difficult
- ✓ Common sense and a bit of experience are often helpful
- ✓ For example it is easy to develop a sense of what is good and bad in a multiple-sequence alignment:
 - Good ⇔ Nice, ungapped blocks
 - Bad ⇔ Messy, gapped blocks

8. Check Borderline Results with Different Programs (Termed “Consistency-checking”, or a “Consensus Approach”)

- ✓ Three short-sighted witnesses are at least as informative than a single eagle-eye witness
- ✓ Since no prediction program/resource is 100% accurate, three **different** programs giving the same borderline result is typically better than one “good” result



9. Be Especially Careful with Unpublished Methods

- ✓ The Internet is **great** because you can find everything on it
- ✓ The Internet is **a pain** because you can find everything on it
- ✓ Good methods are published in peer-reviewed journals **before** being put online
- ✓ Btw just because a method is published does not mean it's accurate, either...

10. Data and Wine

- ✓ Databases are not like good wine . . .
 - They don't improve with age.

- ✓ Databases are more like vegetables, salads, or fruit juice . . .
 - The fresher, the better!
 - As with programs, consult the publication or other documentation to gauge how complete/useful it will be for your purpose

11. Everything Is Not For Free on the Internet

- ✓ Everything is **mostly** free for academics

- ✓ If you are a company, you may have to pay a royalty, even if the information looks free . . .

- ✓ It's all about trust and sharing, this keeps our science and resources open and accessible. 😊

12. *Biting the Right Bullet*

- ✓ You **can** do everything on the Web, but as you are working with large data sets, you will find it
 - Less reliable (servers may be down)
 - More time-intensive (one submission at a time...)
 - Less flexible

- ✓ If you find yourself doing Bioinformatics nonstop for a few months . . .
 - You are way beyond the *For Dummies* level!
 - Time to learn a bit of Perl and Linux!