Guideline for Common Computational Skills (GCCS2019)

- 熟练掌握科研相关的技能是顺利开展科研工作之必须，本文描述 Gao Group 研究生同学在读期间需掌握的计算方面应用技能(skills)
- 相关技能被归组为 track，每个 track 又包括 Basic 和 Advanced 两个级别。
  - 除特别说明外，2014 年之后入学的同学应
    - 在 Qualify 之前掌握全部 track Basic level 的技能
    - 并在毕业前掌握 Track 1、2 中至少一个的 Advanced level 技能
- 掌握程度划分为 了解→熟悉→精通 三个层次
  - 了解：知其然，可在 senior colleagues 的指导下，在科研中应用相关技术
  - 熟悉：知其所以然，可独立在科研中应用相关技术
  - 精通：知其所以“不然”，可根据实际科研需求，自主对相关技术进行拓展和进一步开发

Track 1: Statistical Modeling

- Basic：
  - 结合 R/Bioconductor，熟悉基础统计学[1]及常用统计数据处理技巧[2]；结合 Python 软件包，熟悉常用机器学习模型的应用[3]。
    - 了解 R 和 Python 在命令行下的基本调试方法 (R: debug; Python: ipdb 或 iPython 的 run -d)
    - 了解 Python 常用软件包：sklearn, NumPy, SciPy, pandas
    - 了解基础作图软件包：ggplot/Matplotlib 及其有用的扩展包 ggfortify/GGally

[1] 《统计建模与 R 软件》1~7 章

[2] 如 apply 系列 函数实际应用，data.table、bigmemory 快速读取，reshape2 的变形，plyr 的分组处理

[3] 如 SVM, random forest, decision tree, ridge regression, lasso, elastic net, logistic, glm, knn, kmeans, boosting, bagging, ensemble learning

- 了解基本统计检验方法：分布参数检验 (t 检验，F 检验等)，方差分析，非参数检验，Pearson 相关性检验等
- 了解常用的应用多元分析方法：Clustering/Classification/PCA[4]/SVD
- 熟悉 t-SNE 及类似降维可视化方法的基本思想和应用

- Advanced：
  - 精通常用软件包：
    - 数据处理：NumPy, SciPy, pandas
    - 深度学习：Tensorflow/Keras/Torch/PyTorch
    - 经典统计/机器学习：sklearn
    - 了解 Octave/Matlab/SciLab
  - 熟悉多元统计分析模型、方法与理论，如
    - 回归分析：多元 Logistic 回归分析、非线性回归分析、非参数回归
    - 主成分分析、关联性规则、因子分析、Survey Data Analysis
  - 熟悉主流统计学习模型，以及其在生物学中的应用及潜在应用
    - 以 SVM[5]为代表的 kernel learning
    - 以 CNN/RNN/AutoEncoder/VAE 等为代表的 deep learning
      - 强化学习与 deep learning 的结合
      - 以 GAN 为代表的 generative model
    - 了解以 HMM, Conditional Random Field 为代表的随机过程模型，以 Bayesian network 为代表的 Probabilistic Graphical model
      - 并从 Monocle[6]或类似软件入手，了解在缺乏时间尺度数据的条件下，伪时间序列的基本构建思想和后续分析方法

Track 2: Big Data

- Basic：
  - 熟悉 Linux shell 重要命令和软件，包括

[4] 《统计建模与 R 软件》8~9 章

[5] libSVM(http://www.csie.ntu.edu.tw/~cjlin/libsvm/)

[6] http://cole-trapnell-lab.github.io/monocle-release/

- bash, cat, cut, tr, grep, awk, sed, tar
- screen 基本操作
- ssh, wget, scp, rsync, lftp
- time, date, top, ps, df, du
- 熟悉 emacs/vim 命令行下 (没有鼠标) 的使用。
- 集群作业调度系统 Slurm 的使用

- 熟悉常用代码管理平台，包括：
  - 版本控制软件 Git 及相应 Github 平台的使用
  - 工作流程管理和复现框架 conda + snakemake
  - 至少一种自动报告格式 (能够随输入数据不同而生成内容不同、但格式一样的分析报告文档): Jupyter Notebook, R Markdown
  - 跨平台软件管理系统 conda 及其 channel bioconda
- 了解常用的脚本语言及其生物学扩展库，并熟悉其中一种：Python(BioPython)、Perl(BioPerl)
  - 了解如何编写网络爬虫(至少要有 HTTP Request 和基于 Javascript 的 Ajax 两种)批量获取网站信息
- 熟悉 SQL-based DBMS: MySQL 或 SQLite

- Advanced：
  - 熟悉 Linux 系统的维护，可以独立在 x86 个人电脑或服务器上从头安装 Centos 或者 Ubuntu，了解计算机、集群、云平台的原理和结构，能利用原理解决简单的计算机故障
  - 熟悉以 MongoDB, neo4j, HBase 为代表的 NoSQL DBMS[7]
  - 了解一种编译型语言：Java/C++/C，要求能够在命令行下对编译后的二进制可执行程序设断点动态调试
  - 熟悉一种 Web server-side scripting language： PHP/JSP
  - 熟悉常用的网页前端语言：HTML/CSS/Javascript
  - 熟悉常用的 Web server: Apache, Tomcat 或者 node.js

[7] http://bioinformatics.oxfordjournals.org/content/29/24/3107.full, also read http://www.infoq.com/news/2014/01/genomics-big-data-revolution

Track 3: Writing

- Basic:
  - 熟悉 markdown 的使用
  - 熟悉文献管理软件：Mendeley/Endnote/Zotero
  - 借助北大毕业论文/qualify/中期报告模板，熟悉至少一种基本的论文编写环境：Word / TeX，要求能够实现以下功能：
    - 图/表引用随文档变化而自动更新
    - 文献引用管理系统，文献的正确表示和排版 (Word: Mendeley/EndNote/Zotero; TeX: BibTex)
    - 数学公式的正确插入和排版 (Word: 数学对象; TeX: $$, equation, align, matrix 等重要环境)
    - 修订和审阅 (Word: 修订和审阅; TeX: changes 包)
    - 使用样式来格式化文本内容，而不是手动修改文本格式 (Word: 样式; TeX: section 等环境的自定义格式)