

1. 个人信息

1.1. 基本信息

陈 军

- 男, 25岁(1995年2月), 本科
- 近2年工作经验
- 目标城市: 广州
- 目标岗位: 生物信息工程师
- 联系方式:
 - 电话(微信): [18502829256](tel:18502829256)
 - 邮箱: 1170101471@qq.com



1.2. 个人简介

本人生物专业出身, 对生命活动的各类基本过程较为清晰明确, 系统学习过Python编程, 熟悉生物信息分析的基本方法、流程、软件, 能进行日常的生信数据处理, 项目分析, 拥有较强的流程研发能力, 能完成生物信息软件的基本测试。本次求职中期待能找到与人体健康相关的生物信息分析岗位。

个人经历:

- 2014.9 - 2018.6, 武汉生物工程学院, 本科, 生物工程
- 2018.1 - 2018.6, 中国科学院武汉植物园, 实习兼毕业论文设计
- 2018.6 - 2018.9, 达内培训机构, 参与Python的系统培训
- 2018.9 - 2019.10, 天津诺禾致源生物信息科技有限公司, 任职生物信息工程师

个人能力:

- 在生物信息方面:
 - 熟悉生物信息分析的基本流程与套路;
 - 能使用常见生物信息分析软件, 如Blast, Blat, BWA, SAMTOOLS, GATK, CIRCOS, jellyfish, GenomeScope, gce, SOAPdenovo2, Orthofinder, rawxml等;
 - 能阅读期刊文献, 从中提取数据分析的步骤、原理、软件与方法。
- 在编程方面:
 - 熟练掌握Linux系统操作, 高阶命令使用以及Shell编程;
 - 精通于Python快速编程, 能熟练使用大多数第三方库并快捷完成数据分析及绘图处理, 如对Numpy、Matplotlib、Pandas、BioPython、RE、多进程多线程、网络编程、Request、BeautifulSoup、Django等众多库都有一定的学习与使用, 了解数据库Mysql的基本操作, 了解网页搭建、机器学习的简单使用;

- 拥有较为完善的编程体系，轻松驾驭千行以上代码，具备快速学习与变现的能力。具有规范的、逻辑性的和易维护性的编码风格；日常的文档环境让我现在能从未接触过的技术文档迅速变现；拥有面向对象编程、模块化设计思维、一定的数据结构设计能力，可实现和优化数据处理速度及使用功耗。

个人志向：

- 本人出身于生物，选择生物并借助现代化信息处理工具，志于利用现代化科技探索人体这个精密仪器，期待未来能帮助我们深刻理解生命最基本的生老病死，将生命与健康精准掌控于我们自己手中。

自我评价：

- 对生活与工作都充满积极与热爱态度，不论对人或事物都认真、耐心、富有责任心。
- 性格沉稳而富有创新与探索精神，在长期不断探索、学习、积累、总结中，摸索出一套能根据需求快速学习的高效方法，喜欢挑战难题，对于未知探索勇于尝试，充满热情与创造力。
- 百折不挠，对于未达到的目的，淡定的不断寻找原因与方法。随着自己学习体系逐步强大，一颗热忱的心，富有的挑战精神与不惧向前的勇气，定能带领我在未来创造不菲价值。

2. 工作经验

2.1. 诺禾致源生物信息科技有限公司

2018.9 - 2019.10 —— 生物信息工程师

工作内容：

- **数据分析：**
 - **常规分析：** 利用现有流程，将数据置于指定路径，进行集群任务投递批量并行计算操作，控制运行队列，监控运行状况，杜绝意外发生，解决部分不规则和奇怪的输入数据造成内存和CPU溢出生成的core文件、计算结果不完整等问题。我先后参与了重测序项目，基因组、转录组项目相关分析；
 - **数据个性化处理：** 使用Python编写程序、利用Linux系统各类命令完成数据批量整理和处理操作。进行T量级的测序数据、比对数据、位点标注数据的分割，格式化，统计，校正等信息处理。调试、验证、找出软件以及流程中计算结果中的BUG。设计并优化大文件的比较与合并方法，设计算法使得CPU和占用内存尽可能低耗高效。
- **研发方面：**
 - **新技术实现：** 阅读英文文献，提取新发表论文中所使用分析思想、软件、方法，在Linux上使用非root账号下载安装搭建测试软件，配置软件之间的依赖关系及运行所需环境变量，进行数据分析的重现。
 - **流程串写：** 使用Python将不同软件按“并行”或“串行”的运行方式串写成完整流程，以便多次批量使用。

- **流程优化：** 流程运行中数据的BUG修复(包括python、perl、R脚本)。修改过滤脚本中的名字切割BUG；修改R脚本中的画图脚本，实现循环画图自动化，从而改进此前的多组数据需要每组单独复制脚本后修改脚本内部的操作来运行。
- **基因组网站搭建：** 在新服务器上，配置和搭建基于mysql/cgi/php等开发的基因组数据库网站；在探索的过程中整理的方法、留下的笔记为部门创造了一个对外的新产品。

- **集群管理：**

- **信息管理：** 数据分析前的拷盘需求信息汇总；云冷存储数据的上传与解冻；
- **集群磁盘空间管理：** 使用Python开发自动定期扫盘流程进行磁盘统计，从约50T的扫盘结果数据文件中，过滤出扫盘结果中文件大小大于10M的结果，使用Pandas进行目录分级聚合统计，计算出每个文件夹和每个集群用户下的文件空间大小，尝试指定对数据涉及的集群用户自动化发送邮件，组织大家删除数据分析过程中产生的冗余数据。

2.2. 中国科学院武汉植物园

2018.1 - 2018.6 —— 研究助理（实习）

实习工作内容：

- **数据信息处理方面：** 借用Python编程技能写小程序，辅助博后进行生物数据处理和分析，进行数据格式化，批量替换，表格文件统计等工作；
- **在实验方面：** 帮助实验室的博后测转基因拟南芥种子中的油脂含量；进行十数樟植物的接种，进行组织培养及相关实验仪器管理；
- **毕设方面：** 为毕业设计阅读文献，整理知识，做实验，编写文献综述，撰写论文；

3. 项目经验

3.1. 生物信息方面

3.1.1. 南农8个梨基因组图谱构建及分析

2018/10-2019/05

项目描述：

我在该南农梨项目中的负责内容：

- **个性化分析脚本编写：**

- 对注释得到的原始gff文件进行重新更名。方便后续分析，并避免后续分析由于基因注释名默认太长出现的基因名被截断的BUG。
- 基于不同物种基因间的比对结果，统计9个梨之间的基因共线关系，找出砀山梨在其余九个梨物种中的近缘基因。
- 处理MCscan的文本输入数据，给绘制共线基因区块的每一行非同号染色体前面的数据前面增加颜色标记，使得后续可视化易于观察对比发现规律。
- **个性化流程搭建：**
 - 基于RAxML的批量构树流程的二次开发；
 - 基于itol网站API的Linux集群版进化树可视化构建流程；
 - ▼ 开发场景说明：

在进行基因家族分析的过程中，涉及的研究性状有8个左右，每一个性状相关的基因家族有1~10个不等，有10个物种需要分析，大约需要生成接近1000个构树文件。

集群上没有直接将tre文本文件转化为svg文件或图片的工具，构树的可视化工作量较为巨大，为了快速进行基因的筛选，我开发了基于itol网站API的Linux集群版进化树可视化构建流程。

自动对参考基因和近缘基因上色；构建的基因进化树，结合蛋白表达结构域，鉴定筛选不同物种之间的近缘基因。
- **画图：**
 - 使用R绘制树图(TreeMap)，展示LTR的各部件含量；
 - 辅助改图，调整绘制的Circos的PDF图片，标注物种信息；修改美化扩张收缩图形结果内部文字。

3.1.2. 植物所3个薰衣草泛基因组图谱构建及分析

2019/07-2019/09

项目描述：

我在该薰衣草项目中负责有以下内容：

- **二代转录组数据对三代全长转录组数据进行纠错**
 - 基于IsoSeq流程的重新研发，进行一致性序列获取、引物去除、poly尾巴修建、聚类、抛光、纠错。
- **组装后的基因组注释**
 - 重复序列注释、基因注释（包括基因结构预测和基因功能预测）、非编码 RNA（ncRNA）注释。
- **有参转录组分析**
 - 可变剪切分析、新转录本预测、SNP和InDel分析、基因表达水平分析、RNA-seq整体质量评估、RNA-seq相关性分析、差异表达分析、差异基因GO富集分析、差异基因KEGG富集分析、蛋白互作网络分析
- ▼ 项目执行过程中遇到的问题及解决方法：
 - 技术问题：
 - 问题描述：“二代测序数据对三代测序数据进行纠错”的原始流程由于无人维护，环境变量混乱，导致使用不正常。

- 解决方法：对现有纠错流程重新研发改进。通过同同事交流并自己查阅资料，从[IsoSeq官网的GitHub](#)中提取出了操作方法，并将其串写成了shell流程。
- 流程问题：
 - 问题描述：由于组装组给的数据中，包含许多小片段序列，导致总序列数过多，在执行功能注释的其中一个步骤时，导致流程自动切分过多，总共切分480份，每份投递6个任务，自动投递数量远远超过投递数量2000的上限，导致总是跑不通。
 - 解决方法：在该步中将序列分为两份，两份先后跑完之后再合并进行下一步。

3.1.3. 樱花二代测序数据的叶绿体与线粒体拼接

2018/09-2018/09

项目描述：

- 本项目中，我负责将二代测序的CleanData数据，使用 NOVOPlasty 进行叶绿体和线粒体的组装拼接。由于有50余组数据需要拼接，使用不同的参考基因组得到的结果也不同，而每一组去单独重复配置文件和路径都显得比较麻烦。因而在执行项目的过程中搭建了一个简易批量执行流程。最后，对于拼接效果不好的CleanData，与参考基因组使用BWA进行比对，再使用Samtools提取和过滤出比对到的reads。

3.1.4. 其他生物信息分析项目

• 1. 基因鉴定分析

项目名称：海南大学1个辣椒基因家族分析技术服务（委托）合同（2019/07-2019/08）

项目描述：本项目中，我负责的内容有：基因家族HSP70和CBF的进化树分析，基因结构分析，motif分析，染色体位置分布、共线性分析，基因家族表达模式分析，基因家族启动子分析

• 2. 群体进化分析

项目名称：中国科学院昆明植物研究所159个植物WGS-seq群体进化分析（2019/08-2019/09）、新疆农业大学51个梨WGS-seq群体进化分析（2019/03-2019/03）

项目描述：本类型项目中，我负责下机数据的质控，比对，CallSNP，SNP过滤，群体进化树分析，主成分分析，遗传结构分析，群体多态性分析，连锁不平衡，选择消除分析，基因功能富集分析

• 3. BSA性状定位分析

项目名称：华农3个拟南芥BSA性状定位分析（2018/10-2018/11）、微生物所1个拟南芥BSA性状定位分析（2018/11-2018/11）、4个花生子代池BSA性状定位分析（2018/11-2018/11）

项目描述：本类型项目中，我负责下机数据的质控、比对、CallSNP、子代SNP和INDEL频率差异分析、基于SNP和InDel标记的目标性状区域定位，完成BSA性状定位分析。

• 4. 变异检测分析

项目名称：陕西师范大学4个拟南芥WGS-seq变异检测分析（2018/10-2018/10）、郑州果树所5个苹果WGS-seq变异检测分析（2018/12-2018/12）

项目描述： 本项目中，我使用公司流程完成测序数据的质控，比对，SNP/INDEL/SV/CNV变异检测，进行数据可视化。

3.2. 编程方面

—— 业余探索

3.2.1. 数学模拟计算“三步称盐”的所有可行路径

2020.01-2020.02

项目描述：

该项目为一个数学问题的模拟计算，需要设计出数据结构对所有结果进行遍历并挑选处理。

问题如下：

使用Python求解题目：有一天平，2克和7克砝码各一个，若想利用天平和砝码来将140克盐分成50克和90克两份，规定只能使用3次天平进行称量，有哪些方法？

实例中，

1. 通过对问题的最基本认识，总结出，可以完成所有操作的最基本方法：平分、平移、合并；分析每一个分步骤建立单独计算的单元函数。
2. 随后设计能存储每一种结果的数据结构，可通过完成计算形成一颗庞大的树；总计 33693 种方法。
3. 使用递归函数取出树结构的每一支，判断最终结果是否是符合我们要求。
最终，筛选出 23 种绝对无重复的方法。

技术难点：数据结构设计、递归算法设计

项目地址：<https://github.com/wan230114/chenyang-python>

3.2.2. PythonNote在线文档网页站点的搭建

2019.11-至今

项目描述：

本项目PythonNote是个人在编写系统的Python知识体系，整理的学习思路和笔记的同时，搭建和渲染的一个在线网页教程。内容包含：Python基础，多进程多线程，Django，RE，Numpy，matplotlib，pandas，sk-learn等笔记。

项目中使用Markdown完成笔记文件记录，使用github上开源的docsify框架将markdown实时渲染为网页，对知识体系进行结构构建；将每一个编程小块知识的赋予清晰实例讲解；将Python基础到数据分析，以及最后机器学习总结完整框架。

由于部分文章在CSDN中的公布，也获得了一位图书编辑的邀请私聊。

项目地址：<https://github.com/wan230114/PythonNote>

网站访问主页：<https://wan230114.github.io/PythonNote>

3.2.3. Pywget网络下载器开发

2019.09-2019.10

项目描述：

本项目是开发一款Python版本的文件下载器，使用Python实现类似于linux中wget命令的基本功能，实现与其他开发程序的功能对接，以及进行代理转发下载。在之前生信工作中，由于国内网络环境下下载fasta, gff, pep等文件过于缓慢，有时由于项目的周期也使用该脚本代理转发快速下载。

Pywget分为服务端和客户端，客户端可以不依赖服务端而进行单独运行；其中开启服务端，在客户端操作即可进行数据转发完成代理下载。

技术难点：

1. 网络编程实现：服务端使用get指定数据长度的数据，客户端接收指定长度的数据，使用TCP协议进行转发实现。
2. 粘包问题解决：为解决粘包的问题，在每一段数据前面进行打包为4字节的长度信息，形成包括长度信息的数据报头。
3. 断点续传问题解决：通过请求网址，可以得到请求的服务器对文件是否支持断点续传，若支持，则将文件最后一段截取掉进行续传，从而完成断点续传实现。

项目地址：<https://github.com/wan230114/pywget>

3.2.4. 在线教育培训课程视频的抓取

2019.10-2019.10

项目描述：

本项目使用Python进行网页数据的抓取，爬虫保存1000+培训教学视频课程，获取下载有时限的培训课程视频以及课件文本内容。

技术实现：使用selenium进行自动化测试登录，同时使用browsermobproxy开启代理服务抓取Network中的视频请求网址，通过自动获取到对应视频链接后，抓取到视频地址进行视频下载。

技术难点：selenium的API测试、browsermobproxy代理服务测试、Chrome的Flash设置调试、Pywget下载模块的编写与对接使用

该项目展示了自己的快速学习与变现能力，自己并未系统学习过爬虫，如果手动完成从调试页面抓取视频网址下载并规律命名，付出的时间与精力代价是巨大的，为了快速变现进行了现学现卖。

由于涉及知识产权，我未将代码公开于网络。仅公布代码运行测试视频：

<https://www.bilibili.com/video/av82164815>

3.2.5. 其他内容

KEGG网站数据下载器

开发场景：同同事的一次交流中，他需要打开KEGG网站，找到相应网页上的cds和pep数据保存为本地文件，这样的重复操作需要进行很多次。我通过之前接触过的一点爬虫知识，设计出了代码，快速完成了客户要求的约690个KEGG网页包含的序列数据的提取和批量保存；

4. 技能特长

4.1. 技能/语言

- Python编程，熟练
- Linux系统操作，熟练
- MsOffice，熟练
- Perl、R、C++编程，良好
- Mysql，良好

4.2. 证书

- 2017/3 [全国计算机等级二级](#)
- 2018/8 [大数据分析师](#)

5. 附加信息

5.1. 主题：大学期间活动简介

大一

担任高数老师助教，为公共群300余人解答日常高数难题；

选修网页设计，期末设计的个性网站以90分优秀结束；

学习到计算机的一些基本原理、结构、系统等知识；

大二

学习期间，凭借专业课较好成绩申请获得校级风华学子奖学金；

第一次参加全国数学建模大赛，首次接触matlab和mathematical编程；

大学社团的创始，其中我负责各类技术工作，如使用Photoshop进行宣传海报、团徽设计，Pr视频剪辑；

大三

进入科研楼参与生物技术实验，如克隆、转化、酶切、接种等操作；

使用会声会影x8为老师做网络教学实验视频；

学习c/c++，搭建深度学习caffe框架和运行环境，并初次接触Linux操作系统。

大四

大四做毕业论文时，加入了名师海归博后赵华燕团队，于18年1月份到了中国科学院武汉植物园学习与实习。

主要成就有：

- 学习阅读外文文献，整理了十数樟植物的研究概况，借助谷歌翻译辅助阅读了近上百篇外文文献，将信息整合做成论文的文献综述部分。毕业论文设计中，论文的文献综述部分获得答辩老师赞赏（指导老师认为可以进一步修改投稿）；
- 使用Python语言技巧编程处理FASTA等数据文件，辅助博士进行分子进化树构建的数据处理，提取想要的信息；
- 帮助博士做简单的实验。做过组培、PCR、质粒提取、转化、酶切连接等。

刚毕业，为了进一步提升自己的信息处理能力、编程技术，于达内培训学习系统的Python课程。

5.2. 主题：编程学习过程

主题描述：

本人爱好科学思考、科技创新，大三时一场人机围棋大战博入眼球，我感受到了一个科技爆发奇点的临近，可是我竟然连编程是什么都无法得知，更觉得我对于理解这场机器智能技术进步的科技大变革是如此的空洞，于是我决定开始自学编程，了解计算机方面的知识，为学习与理解未来打下坚实的基础。从C语言到C++，再到Python，perl，R，这个过程让我学习到，无论是什么计算机语言（C，C++，Java，Python等）最重要的内容是思维与算法，如何做这件事，语言只是它的表达形式。

在学习过程中，我就生活中的一些小问题做了一些程序设计。以下的一个实例，见证了我不同阶段的编程能力。

具体实例是设计程序计算思维难题“称盐”的所有可行路径问题，GitHub地址是：

<https://github.com/wan230114/chenyang-python>

需要解决的问题是：有2克、7克砝码各一个，天平一架，如何利用天平和砝码称三次，将140克盐分成90克、50克各一份？一共有哪些不重复的可行方法？

我将3个版本迭代的历程，写成了一个说明文档，链接如下：

https://github.com/wan230114/chenyan-python/blob/master/README_history.md

通过这3个不同时段版本的编程，更展现我不同时期的水平进步，编程效率愈来愈高，逻辑结构愈来愈清晰。



扫描上方二维码可获取网页版链接，或直接输入以下网址：

<https://wan230114.github.io/BioNote/#!/Interview/me>