

生物信息学实践

彭仁海 刘震 刘玉玲 著

中国农业科学技术出版社



生物信息学实践

彭仁海 刘震 刘玉玲 著

中国农业科学技术出版社

图书在版编目 (CIP) 数据

生物信息学实践 / 彭仁海, 刘震, 刘玉玲著. —北京: 中国农业科学技术出版社, 2017. 4

ISBN 978 - 7 - 5116 - 3016 - 2

I. ①生… II. ①彭…②刘…③刘… III. ①生物信息论 IV. ①Q811. 4

中国版本图书馆 CIP 数据核字 (2017) 第 057372 号

责任编辑 崔改泵
责任校对 贾海霞

出版者 中国农业科学技术出版社
北京市中关村南大街 12 号 邮编: 100081
电 话 (010) 82109194 (编辑室) (010) 82109702 (发行部)
(010) 82109709 (读者服务部)
传 真 (010) 82106650
网 址 <http://www.castp.cn>
经 销 者 各地新华书店
印 刷 者 北京富泰印刷有限责任公司
开 本 787mm × 1 092mm 1/16
印 张 11. 25
字 数 267 千字
版 次 2017 年 4 月第 1 版 2017 年 4 月第 1 次印刷
定 价 30. 00 元

◀ 版权所有 · 翻印必究 ▶

目 录

第一章 生物信息学分析基础工具与平台配置	(1)
第一节 文本编辑器	(1)
一、常用的文本编辑器	(1)
二、UltraEdit	(2)
三、Vi 编辑器	(2)
第二节 Linux 系统基础	(3)
一、软件安装	(3)
二、PATH 路径设置	(5)
三、必备 Linux 命令	(5)
四、Linux 系统的输出重定向与管道	(7)
五、中文版 Linux 改为英文版	(8)
六、开启 FTP 服务	(8)
第三节 生物信息学实验室局域网	(10)
一、生物信息学局域网实例	(10)
二、远程登录	(10)
第四节 Windows 系统下构建本地 BLAST	(12)
一、BLAST 的下载安装	(12)
二、Blast 的使用	(12)
三、实例讲解	(14)
参考文献	(18)
第二章 生物信息数据库的使用与构建	(19)
第一节 NCBI 数据库资源	(19)
NCBI 数据库检索	(21)
第二节 数据存储格式	(22)
一、FASTA 格式	(22)
二、FASTQ 格式	(23)
三、Genbank 格式	(23)
四、EMBL 格式	(25)
五、采用 XML 实现生物数据库的整合	(26)
第三节 著名的生物信息学数据库	(28)
第四节 生物信息学数据库的构建方法	(30)

一、Apache 的安装与启动	(30)
二、MySQL 的安装与配置	(31)
三、PHP 的安装与配置	(33)
四、不能安装的情况	(35)
五、利用 Windows Server 搭建数据库服务器	(35)
参考文献	(38)
第三章 基于蛋白质结构的计算机辅助药物设计	(40)
第一节 蛋白质二级结构	(40)
一、 α 螺旋	(41)
二、 β 片层	(41)
三、 β 转角	(41)
四、蛋白质二级结构预测	(41)
第二节 蛋白质结构数据库及其检索	(45)
一、PDB 数据库检索	(47)
二、蛋白质结构数据的存储格式	(48)
三、蛋白质结构可视化	(50)
第三节 蛋白质结构的预测	(53)
一、国际蛋白质结构预测技术评估大赛 (CASP)	(53)
二、利用 SWISS-MODEL 预测蛋白质的三级结构	(54)
第四节 分子对接工具 Autodock	(55)
一、Autodock 程序的安装	(56)
二、小分子的来源和处理	(57)
三、大分子的处理	(57)
四、两个参数文件 (gpf 和 dpf) 的设置	(59)
五、结果的处理	(62)
第五节 分子模拟原理与工具	(63)
一、分子模拟的主要方法	(63)
二、分子模拟常见工具	(64)
第六节 分子动力学模拟工具 Amber	(65)
一、生成小分子模板	(66)
二、处理蛋白质文件	(68)
三、生成拓扑文件和坐标文件	(69)
四、能量优化	(73)
五、LEAP 使用	(75)
六、MD 过程	(77)
七、VMD 的使用	(80)
八、观看并保存图像的步骤	(81)
九、RMS 计算	(81)

十、结果数据处理	(82)
参考文献	(83)
第四章 转座子的生物信息学分析	(86)
第一节 转座子的分类	(86)
一、分类级别	(86)
二、自主与非自主转座子	(89)
三、转座子的命名	(90)
四、转座子的生物信息学分析	(90)
五、重复序列挖掘工具	(91)
第二节 RepeatMasker 和 RepeatModeler	(92)
一、RepeatMasker 的安装	(93)
二、RepeatModeler 的安装	(95)
三、RepeatMasker 的操作	(98)
四、RepeatMasker 搜索的过程	(99)
五、两个 Perl 程序	(100)
六、联合多个重复序列数据库	(101)
七、RepeatMasker 的其他参数	(102)
八、Out 结果文件	(103)
九、RepeatModeler 的使用	(105)
第三节 LTRharvest	(105)
第四节 序列去冗余	(107)
第五节 Circos 绘图	(108)
一、Circos 的安装	(109)
二、Circos 的颜色	(112)
三、图像分析	(114)
四、核型文件	(116)
五、ideogram 标签	(117)
六、连接	(117)
七、图像输出	(118)
八、直方图	(119)
九、Highlights 图	(120)
十、容易出现的问题	(122)
参考文献	(122)
第五章 生物信息学资源	(126)
第一节 网络资源	(126)
一、在线工具链接 Expasy	(126)
二、常用生物软件分类与下载	(127)

三、生物信息学中文论坛	(128)
第二节 期刊与机构	(129)
一、生物信息学期刊	(129)
二、生物信息学机构	(129)
第三节 在线小工具	(131)
一、开放阅读框查找工具 ORF Finder	(131)
二、绘制 GO 注释结果	(132)
三、蛋白质组成和稳定性分析 ProtParam	(133)
四、启动子区预测工具 Promoter	(133)
五、序列 logo	(133)
六、蛋白质序列综合分析工具 PredictProte	(135)
七、信号肽	(135)
八、比较分析图绘制工具 VENNY	(135)
第四节 生物信息学分析软件	(137)
一、EMBOSS	(137)
二、EMBOSS 运行示例	(137)
三、综合序列分析软件 DNASTAR	(140)
四、分子生物学常用工具简介	(144)
参考文献	(147)
第六章 分子进化	(149)
第一节 分子进化基础	(149)
一、构建进化树的算法	(149)
二、进化树格式	(150)
三、进化树的图形显示	(150)
四、进化软件	(152)
第二节 通过 phylip 构建进化树	(153)
一、准备	(153)
二、通过 Clustal 将 Fasta 格式的序列进行比对并保存为 phy 格式	(154)
三、使用 seqboot 设置重复数量	(156)
四、通过似然法计算进化树	(157)
五、构建一致树	(158)
六、图片制作	(160)
参考文献	(161)
第七章 生物信息学编程基础	(163)
第一节 Perl 语言	(163)
一、CPAN	(163)
二、正则表达式	(164)

三、Bioperl 的安装	(165)
第二节 统计语言	(166)
一、R 语言	(166)
二、其他统计分析工具	(166)
参考文献	(166)
附录一 生物信息学常用词汇表一	(168)
附录二 生物信息学常用词汇表二	(170)

第一章 生物信息学分析基础 工具与平台配置

随着生物信息学的普及运用，大多数实验室都需要通过一些生物信息学的工具，甚至是构建生物信息学分析平台服务生物学实验，并提示进一步的实验方案。根据实验室具体需求，集成各种常用分析工具、资源，将可以高效地完成本实验室成员生物信息学分析的需求，还能够完成一些序列批量处理的任务。许多生物信息学的分析软件和数据库可以通过互联网（Internet）免费获取，为生物信息学平台的建设奠定了基础。

生物信息平台的建设是需要逐步推进和实施的。因而在设计和构建过程中要注意平台的实用性并做好备份工作。实用性的高低直接影响到使用者对系统的评价。系统在设计时，在满足实际应用要求的目标前提下，应该多从使用者的角度出发，采用面向用户的设计理念，提供界面友好、操作方便的用户交互平台和操作环境。实验室生物信息学平台一般都由多个实验室成员共同使用，这就导致不可避免会出现一些问题，为此，首先要做好数据的备份工作，其次要注意记录平台软件的安装和使用方法，在平台出现问题后可以方便地进行恢复。

本章的主要内容包括：几乎所有生物信息学分析中都会用到的工具和方法，如：文本编辑器、Linux 操作系统系统下安装软件、系统基础服务的配置等。

第一节 文本编辑器

文本编辑器是用于编写普通文字的应用软件，以纯文本形式进行储存，一般用来编写程序源代码。而 Word 等文档编辑器是以二进制格式进行存储，主要功能是用来排版。文本编辑器具有的典型功能主要有：查找、替换、剪切、复制、粘贴、行号、自动缩排、撤销和恢复等。Windows 系统自带的记事本虽然也是文本编辑器，但功能较小，编写脚本程序时没有语法提示，也不能打开比较大的文件，如：基因组、蛋白组等。在生物信息分析过程中，绝大多数软件的结果都是纯文本格式，不同软件之间的数据传递也经常通过纯文本格式，此外，编写生物信息学程序脚本也需要用到文本编辑器，因此，文本编辑器是生物信息学中一个非常重要的工具。本节将介绍 Windows 下的 UltraEdit 和 Linux 下的 vi 编辑器。

一、常用的文本编辑器

Notepad 是一个开源免费的文本编辑器，可以对多种编程语言实现语法高亮，代码折

叠, 拖放缩放等。Notepad2 是一个相当优秀的轻量级文本编辑器, 具有很多特色功能, 如代码高亮、编码转换、行号显示、多步 Ctrl + Z 等, 是不可多得的记事本替代工具。而 Notepad2-mod 是 Notepad2 的修改版、更新很及时, 支持代码折叠、NSIS、Inno、AHK 语法高亮等。PSPad 是 Windows 平台上免费的适合程序员使用的编辑器, 它可以保持上一次的编辑状态, 下次打开编辑器的时候可以直接显示原来的文件, 此外还支持通过 FTP 进行远程编辑, 支持多文件比较等。Emacs 编辑器具有内置的宏功能以及强大的键盘命令, 几乎被移植到了每一个平台, 并有多个发行版, 是跨平台、完全免费并且开源。Sublime Text3 编辑器支持但不限于 Perl、Python、R、PHP、C、C++、C#、HTML、Groovy、Haskell、HTML、Java、JavaScript、LaTeX、Lisp、Lua、Markdown、Matlab、OCaml、Ruby、SQL、TCL、Textile 以及 XML 等主流编程语言的语法高亮。

Gedit 是 Linux 下的一个纯文本编辑器, 但也可以把它用来当成一个集成开发环境 (IDE), 它会根据不同的语言高亮显现关键字和标识符。

二、UltraEdit

UltraEdit (<http://www.ultraedit.com/>) 是一套功能强大的文本编辑器, 可以编辑文本、十六进制、ASCII 码, 可同时编辑多个文件, 而且即使开启很大的文件速度也不会慢。可以编辑列; 可将文件另存为多种编码格式从而解决乱码问题, 有 Perl 脚本的语法错误提示。生物信息分析中, 可以通过 Ultraedit 打开基因组文件, 支持超过 4GB 的文件。

在 UltraEdit 使用过程中, 建议修改以下配置以方便使用。

(1) UltraEdit 默认保存一个临时文件以备份修改前的文字, 虽然很安全, 但是一般情况下需要不断地进行删除, 造成了不必要的麻烦, 因此, 可以设置成不备份; 高级—配置—文件处理—备份—不备份 (图 1-1)。

(2) 在鼠标右键中添加 UltraEdit, 这样可以很方便地打开文件。高级—配置—文件关联, 选择集成到资源管理项。

三、Vi 编辑器

Vi 编辑器是 Linux 系统的一个文本编辑器, 可通过终端进行操作。因此是必须掌握的工具之一。但习惯 Windows 系统 txt 编辑器的用户需要一个适应过程。以下是操作过程中比较重要的几点:

(1) vi filename 打开一个文本文件, 刚打开时, 是出于 vi 编辑器的命令行状态, 不能对文本进行修改, 这时候, 按下“i”“a”或“o”可以进入编辑状态, 对文本进行修改。其中:

按 i 从光标当前位置开始输入文件;

按 a 从目前光标所在位置的下一个位置开始输入文字;

按 o 是插入新的一行, 从行首开始输入文字。

(2) 对文本修改完毕后, 需要保存退出, 这时候按 ESC 键, 可以对文本进行不同的处理, 在底部可以输入:

: w filename 以指定的文件名 filename 保存编辑内容

: wq 存盘并退出 vi 编辑器

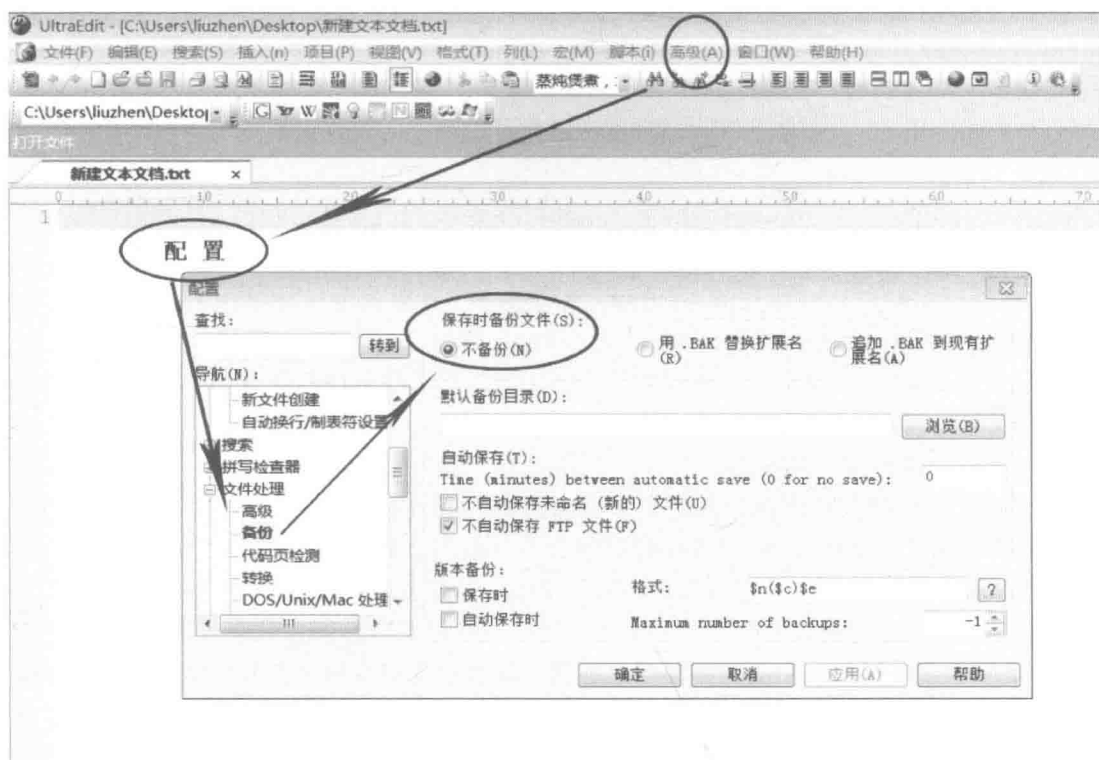


图 1-1 设置 UltraEdit 不产生临时文件

: q! 不存盘强制退出 vi 编辑器

注意这里有“:”半角冒号键

保存之后，可以通过 linux 的 more 命令查看是否修改成功。

第二节 Linux 系统基础

生物信息学分析平台需要使用 Linux 操作系统，虽然 Linux 操作系统的界面没有 Windows 那么友好，软件使用也不很方便，但是很多生物信息学工具都是在 Linux 系统的基础上运行的，虽然也有一些虚拟机等工具可以模拟 Linux 系统，但如果想要深入学习生物信息，则 Linux 是不可避免的。Linux 操作系统没有那么神秘，只要学会在 Linux 系统下安装软件的方法、常用命令和系统设置就可以完成大多数的生物信息学分析工作。

一、软件安装

与 Windows 相比较，在 Linux 系统下安装软件就显得比较麻烦，因此，在下载 Linux 应用软件之后，首先要看软件自带的安装说明然后再安装。这里总结了 Linux 系统下三种常用的软件安装方法供大家参考。

(1) 下载的软件格式类似 `software_name - 1.2.3 - 1. tar. gz`, `software_name` 表示软件名称, 1.2.3 表示版本号, 1 表示修正版本。`.tar.gz`、`tar.Z`、`tar.bz2` 或 `.tgz` 是使用 linux 系统打包工具 `tar` 打包, 再做一次压缩。因此在安装之前, 首先要解压缩, 不同扩展名解压缩命令也不相同, 一般情况下, 运行下面的命令就可以一步完成解压与解包工作:

```
tar-xvzf software_name - 1.2.3 - 1. tar. gz
```

阅读软件附带的 `INSTALL` 或 `README` 等文件, 了解软件安装和使用的基本情况, 这类程序的安装一般需要以下几个步骤:

执行 “`./configure`” 命令为编译做好准备;

执行 “`make`” 命令进行编译;

`Make` 可指定特定 `file` 文件为对象文件。如果没有 “`-f`” 参数, 则系统将默认当前目录下名为 `makefile` 或者名为 `Makefile` 的文件为对象文件。

执行 “`make install`” 完成安装。

到此如果系统没有提示安装错误信息的话, 就表示安装成功了。但是安装的程序却不一定能正常运行, 因为, 安装程序的可执行文件必须在系统的 `PATH` 路径下, 系统才可以找到相应的程序。如: “`/usr/local/bin`” 是一个系统默认的执行目录, 然而, 我们的程序不一定安装在该目录下, 这就需要在设置 `PATH` 变量。

(2) `rpm` 使 Linux 的软件安装工作变得更加简单容易。`rpm` 是 `ReHat Package Manager` (`Red Hat` 包管理器) 的缩写。`rpm` 的安装基本命令为:

```
rpm-ivhsoftware_name.rpm
```

更多参数:

- i 安装软件
- t 测试安装, 不是真的安装
- p 显示安装进度
- f 忽略任何错误
- U 升级安装
- v 检测套件是否正确安装

这些参数可以同时采用。更多的内容可以参考 `RPM` 的命令帮助。

`rpm` 软件的卸载命令为:

```
rpm-e software_name
```

要注意的是，后面使用的是软件名，而不是软件包名。例如，要安装 `software - 1.2.3 - 1.i386.rpm` 这个包时，应执行：

```
rpm-ivh software - 1.2.3 - 1.i386.rpm
```

而当卸载时，则应执行：

```
rpm-e software
```

(3) 软件本身是可执行文件，将文件的目录添加到 `PATH` 变量后，就可以直接运行。

二、PATH 路径设置

Linux 系统环境下，通过命令行运行程序时，系统会在设定的路径范围内查找对应的程序，如果安装的程序没有在指定的路径中，程序就不能运行。

通过 `vi` 编辑器打开账户目录下的 `.bash_profile` 配置文件（图 1-2），修改其中的 `PATH` 变量，多个路径之间通过冒号分开，保存后运行一下命令刷新 `.bash_profile`，新安装的程序就可以运行了。

```
> source .bash_profile
```

需要注意的是刷新 `.bash_profile` 只能在命令行打开的状态下使用，一旦关闭就失效了，如果想永久更新，重启系统即可。

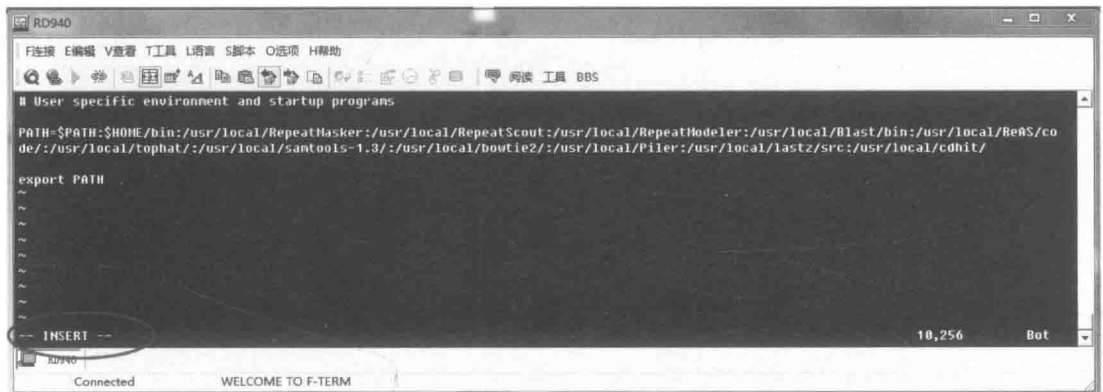


图 1-2 通过 `vi` 编辑器编辑 `home` 目录下的配置文件 `.bash_profile`，可以看到该文件中的 `PATH` 变量，示例中使用的终端是 `Fterm`，从左下角可以看出，`vi` 编辑器处于插入状态

三、必备 Linux 命令

Linux 中的命令的确是非常多，但我们只需要掌握我们最常用的命令就可以了。

(1) cd 命令，用于切换当前目录，它的参数是要切换到的目录的路径，可以是绝对路径，也可以是相对路径。如：

```
cd /usr/local/    切换到/usr/local/目录
cd .              切换到上一层目录
cd ~              转到 home 目录
```

(2) ls 命令，查看文件与目录

- l: 列出长数据串，包含文件的属性与权限数据等
- a: 列出全部的文件，连同隐藏文件（开头为 . 的文件）一起列出来（常用）
- h: 将文件容量以较易读的方式（GB, kB 等）列出来
- R: 连同子目录的内容一起列出（递归列出），等于该目录下的所有文件都会显示出来

注：这些参数也可以组合使用
如：ls-lh

(3) find 命令

功能是查找文件，命令格式为：

```
find [PATH] [option]
```

即：在某一路径下查找某一文件，可以添加与时间有关的参数，与用户或用户组名有关的参数，与文件权限及名称有关的参数等。

(4) cp 命令

该命令用于复制文件，c 它的常用参数如下：

- a: 将文件的特性一起复制
- p: 连同文件的属性一起复制，而非使用默认方式，与 -a 相似，常用于备份
- i: 若目标文件已经存在时，在覆盖时会先询问操作的进行
- r: 递归持续复制，用于目录的复制行为
- u: 目标文件与源文件有差异时才会复制

例如：

```
cp file1 dir/file2 #把文件 file1 复制到 dir 目录下的，文件名改为 file2
```

(5) mv 命令，用于移动文件、目录或更名

```
mv file1 file2 #把文件 file1 重命名为 file2
```

(6) rm 命令，用于删除文件或目录

-f: 就是 force 的意思, 忽略不存在的文件, 不会出现警告消息
 -r: 递归删除, 最常用于目录删除, 它是一个非常危险的参数
 例如:

```
rm-fr dir #强制删除目录 dir 中的所有文件
```

(7) tar 命令, 用于打包、压缩和解压, 它的常用参数包括:

-c: 新建打包文件
 -t: 查看打包文件的内容含有哪些文件名
 -x: 解打包或解压缩的功能, 可以搭配 -C (大写) 指定解压的目录, 注意 -c、-t、-x 不能同时出现在同一条命令中
 -z: 通过 gzip 的支持进行压缩/解压缩
 -v: 在压缩/解压缩过程中, 将正在处理的文件名显示出来
 -f filename: filename 为要处理的文件

(8) cat 命令, 查看文本文件的内容, 后接要查看的文件名。可利用该命名合并文件

```
cat file1 file2 > file
```

将 file1 和 file2 中的内容合并到 file 中

(9) chmod 命令

该命令用于改变文件的权限, 一般的用法如下:

-R: 进行递归的持续更改, 即连同子目录下的所有文件都会更改

同时, chmod 还可以使用 u (user)、g (group)、o (other)、a (all) 和 + (加入)、- (删除)、= (设置) 跟 rwx 搭配来对文件的权限进行更改。

例如:

```
chmod 0755 file #把 file 的文件权限改变为 -rxwr-xr-x
```

```
chmod g+w file #向 file 的文件权限中加入用户组可写权限
```

四、Linux 系统的输出重定向与管道

在 Linux 命令行模式中, 如果命令所需的输入不是来自键盘, 而是来自指定的文件, 这就是输入重定向。同理, 命令的输出也可以不显示在屏幕上, 而是写入到指定文件中, 这就是输出重定向。在生物信息学分析中, 经常需要将程序的过程信息或程序使用说明等信息保存到文件上, 这时候就需要进行输出重定向。Linux 系统下, 可以通过“>”实现输出重定向。如:

```
ls > file.txt #覆盖内容到 file.txt
```

```
ls >> file.txt #追加内容到 file.txt
```

注意：一些程序的提示信息，直接通过“>”不能实现将信息重重定向到文件，这时候可以通过>&实现。

Linux 系统使用管道将多个命令组合起来，就可以形成了非常强大的工具组合，能完成非常复杂的工作。Linux 使用管道符“|”将两个命令隔开，管道符左边命令的输出就会作为管道符右边命令的输入。连续使用管道意味着第一个命令的输出会作为第二个命令的输入，第二个命令的输出又会作为第三个命令的输入，依此类推。

五、中文版 Linux 改为英文版

在通过终端访问 Linux 系统时，如果系统是中文版，则可能会有乱码的问题，因此推荐将 Linux 设置成英文版。首先确定使用的是 Root 账户。

(1) 修改语言配置文件

```
# vi/etc/sysconfig/i18n 使用 vi 编辑器进行编辑
//编辑此文件，将语言设置为英文
LANG = 'en_ US'
```

```
可以将原来使用的语言前加#注释掉
//将语言设置为中文
//LANG = 'zh_ CN' (这个配置项需要注释掉，如需再改为中文，打开此项，注掉
上面一项)
保存配置文件，退出
: wq
```

(2) 重启系统

```
# reboot
```

六、开启 FTP 服务

FTP 是 File Transfer Protocol (文件传输协议) 的英文简称。用于 Internet 上的控制文件的双向传输。基于不同的操作系统有不同的 FTP 应用程序，而所有这些应用程序都遵守同一种协议以传输文件。目前，大多数的基因组数据都是通过这种协议进行传输。在 Linux 系统下，需要进行一些设置开启该功能。步骤如下：

(1) 以 root 登录，输入以下两个命令

```
setsebool ftp_ home_ dir on
service iptables stop
```

(2) 查看，修改防火墙状态


```

/etc/init.d/iptables status
添加开放 21 号端口
/sbin/iptables-I INPUT-p tcp——dport 21-j ACCEPT
保存配置
/etc/rc.d/init.d/iptables save
service iptables stop
chkconfig——level 35 vsftpd on
service vsftpd start

```

(3) 重启防火墙，启用、停用

```
service iptables {start | stop | restart}
```

如果不想每次都输入，可以在/etc/rc.d/rc.local 文件中写上这两个命令，系统开机自动运行。如下：

```

#
# This script will be executed *after* all the other init scripts.
# You can put your own initialization stuff in here if you don't
# want to do the full Sys V style init stuff.
touch /var/lock/subsys/local
setsebool ftp_home_dir on
service iptables stop

```

Linux 系统的 FTP 服务器打开之后，可以使用 Windows 系统下的 FlashFXP 等工具登录到 Linux 的 FTP 服务器，从而实现 Windows 系统和 Linux 系统的文件互传（图 1-3）。



图 1-3 Windows 操作系统和 Linux 操作系统通过 FTP 实现文件互传