生物信息学最佳实践

主编　冉隆科

科学出版社

生物信息学最佳实践

主编　冉隆科

科学出版社
北京

内 容 简 介

本书共六章，第一章介绍生物信息学使用环境搭建，主要介绍 Linux 的发行版、安装、基本配置及远程访问工具；第二章介绍生物信息学分析中主要用到的基本 Linux 命令，并使用生物信息数据进行实例操作；第三章介绍生物信息学的基本序列比对，包括 BLAST 比对、BLAT 比对及 Clustal W 多序列比对等；第四章和第五章介绍目前生物信息学研究领域的热点——高通量数据分析方法，包括基因芯片分析和 RNA-seq 分析；第六章介绍蛋白质结构预测的基本方法。全书内容介绍由浅入深，重视对生物信息学实践能力的培养，通过生物信息学工具和方法来分析具体的生物信息学数据，从而使读者逐步打开生物信息学的大门。本书特别适合刚刚涉入生物信息学研究的初学者，同时也适合对生物信息学感兴趣的研究生参考使用。

图书在版编目(CIP)数据

生物信息学最佳实践 / 冉隆科主编. —北京：科学出版社，2016.3
ISBN 978-7-03-047561-9

Ⅰ.生… Ⅱ. 冉… Ⅲ. 生物信息论 Ⅳ.Q811.4

中国版本图书馆 CIP 数据核字(2016)第 043552 号

责任编辑：威东桂 / 责任校对：彭 涛
责任印制：徐晓晨 / 封面设计：陈 敬

科 学 出 版 社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com
北京中石油彩色印刷有限责任公司 印刷
科学出版社发行 各地新华书店经销
*
2016年3月第 一 版 开本：B5(720×1000)
2016年3月第 一 次印刷 印张：8 3/4
字数：154 000
定价：48.00元
(如有印装质量问题，我社负责调换)

《生物信息学最佳实践》编写人员

主　编　冉隆科

副主编　张帆涛

编　者　（按姓氏汉语拼音排序）

李　广　重庆医科大学

彭　睿　重庆医科大学

冉隆科　重庆医科大学

谭鹏程　重庆医科大学

唐娅琴　重庆医科大学

汪克建　重庆医科大学

张帆涛　江西师范大学

张永红　重庆医科大学

前　言

随着人类基因组测序的完成，大量的高通量数据（包括基因芯片数据和二代测序数据）涌现，越来越多的人类基因组数据得到解读，为临床诊断各种疾病提供了强有力的支撑。由此可以看出生物信息学学科的重要性及发展潜力，但同时这门课程是一门交叉学科，只有具备计算机科学、数学、统计学、生物化学等综合知识才能更好地进行生物信息学数据挖掘。编者在高校从事各个层次学生的生物信息学课程教学，从而更能理解学生对这门课程的渴求和极大兴趣，但同时又碍于无从着手，对实际问题缺乏分析解决能力的现状。

为此，编写本书力争从基本的概念着手，包括 Linux 操作系统的安装及主要命令使用、基因序列比对、基因芯片分析、RNA-seq 分析和蛋白质结构分析等。每章都配备有具体的操作案例和代码，每个步骤讲解详细，由浅入深，使初学生物信息学的读者能很快上手，并逐步掌握各种生物信息学分析软件的使用，最后达到对具体的生物信息学数据进行分析挖掘的目的。本书的一大特色是案例丰富，注重实践能力的培养，并附有大量的分析数据和源代码。

本书的编写得到了重庆医科大学基础医学院管理部门的大力支持，以及各编写人员的大力配合，在此表示衷心的感谢。

本书可供初学生物信息学人员使用，也可供与生物信息学结合紧密的学科人员参考，同时也可作为生物信息学培训课程教学使用。由于编者能力有限和时间仓促，书中不足之处在所难免，敬请读者批评指正。

冉隆科

2015 年 12 月

目　录

第一章　生物信息学使用环境搭建

第一节　Linux 系统简介

Linux 是一个基于 Unix 的操作系统，最初是由芬兰赫尔辛基大学计算机系学生 Linux Torvalds 编写而来，并把它放到网上，使之成为“开源”。因此，该操作系统不属于任何人所有，但可以免费下载和使用。每个人对 Linux 的修改都可能被采纳，这样逐渐被发展成为一个强大的并被全世界的大量用户广泛采用的系统，特别是在寻找替换 Windows 的用户。

Linux 系统之所以会成为目前最受关注的系统之一，主要原因是它具有以下的优势。

一、免 费 获 取

Linux 系统最明显的优点是可以免费获取，而微软 Windows 产品，不仅庞大而且往往还需要收费，并且其系统仅允许安装在单一的一台计算机上。而 Linux 发布版可以同时安装在多台计算机上，而不必付额外的费用。

二、跨平台的硬件支持

Linux 目前支持 X86、Alpha、AMD 和 SPARC 等处理器平台，从大型计算机到服务器、桌面系统、移动平台，甚至包括嵌入式系统在内的各种硬件设备等。多数在计算机上使用的巨大外部设备，Linux 都支持。

三、丰富的软件支持

安装了 Linux 操作系统后，绝大多数的软件，包括常见的办公软件、各种图形图像处理工具、多媒体播放软件及各种网络工具等软件都已安装，用户不必另外安装。对于程序开发者来说，Linux 更是一个很好的操作系统，支持多个软件包，包含 Gcc、Cc、C++、Tcl/Tk、Perl、Fortran77、python 等多种程序语言与开发工具。Linux 系统下的软件与 Windows 系统下的软件相比更倾向于拥有更多的

特性和众多的帮助文档，并且绝大多数的 Linux 软件是开源和免费的。对于用户来说，不仅获取软件是免费的，而且还可以修改软件的源代码使之满足更多的特性要求。

四、多用户多任务

Linux 与 Unix 系统一样，是一个真正的多用户多任务的操作系统。在 Linux 系统下，每个用户对自己的文件设备拥有特殊的权利，从而保证各个用户之间相互独立、互不影响。多任务是现代计算机最主要的一个特点，在 Linux 下，系统调度每一个进程是平等地访问处理器，因此它可以使多个程序同时并独立地运行。

五、可靠的安全性

Linux 系统通过使用安全认证，包括密码保护、控制访问等方式来访问每个特定的文件和加密数据，从而使系统安全可靠。Linux 系统是一个先天具有病毒免疫能力的操作系统，很少受到病毒的攻击。当然对于一个开放式系统来说，在方便用户的同时，很可能存在安全隐患。但利用 Linux 系统自带的防火墙、入侵检测和安全认证等手段，可以及时修补系统漏洞，从而能大大提高 Linux 系统的安全性，使黑客攻击者无机可乘，使用户无需另外购买系统安全防护软件。

六、良好的稳定性

Linux 是基于 Unix 概念发展起来的，因此继承了 Unix 稳定并且高效率的特点。加之 Linux 系统内核源代码是以标准规范的 32 位或 64 位计算机来进行最佳化设计的，从而可确保其系统的稳定性。正因为 Linux 的稳定，才使得 Linux 服务器可以常年于无关机状态下运行。

七、完善的网络功能

Linux 内置了很多丰富的免费网络服务器软件、数据库和网页开发工具，如 Apache、Sendmail、Samba、WuFtp、SSH、MySQL、PHP 和 JSP 等。现在越来越多的企业利用 Linux 的这些强大功能来构建自己全方位的网络服务器。

第二节　Linux 操作系统安装及基本配置

1991 年，芬兰赫尔辛基大学计算机系大学生 Linus Torval(李纳斯·托瓦兹)开发了一个自由的、类似于 Unix 的操作系统，并将其源代码通过 Internet 发布在网上供大家修改。随着电脑黑客、编程人员加入到开发过程中，Linux 逐渐成长和壮大起来。Linux 遵从通用公共许可协议(general public license,GPL)，开发源代码。

目前生物信息学领域对基于 Linux 的计算机和软件依赖性很强，虽然绝大多数的生物信息学程序能在 Mac OS X 和 Windows 操作系统下编译和运行，但对于 Linux 系统来说，一个预编译的二进制程序更容易获得、并且能提供给用户许多的程序文档，因此在 Linux 下安装和使用生物信息学软件更加方便；加之目前的绝大多数软件都是用 C、perl 及 python 开发的，这些语言对 Linux 具有很好的兼容性。

目前，对于绝大多数用户来说，最简单访问 Linux 系统的方法是通过使用 Mac 或 Windows 机器连接访问来实现。这种安排的好处是允许多个用户同时运行一个专门有丰富经验的系统管理员担任维护的 Linux 系统中的软件。当然，对于一个没有什么经验的用户，也可以通过在 PC 级上亲自安装 Linux，或者使用一个 Live CD 来运行 Linux 虚拟机。当然，在实际的生物信息学分析中，用户主要通过远程登录、使用基于文本的终端访问远程 Linux 服务器的方式，来运行统计和生物信息学软件。因此，本书主要是基于远程访问方式来使用 Linux 系统。

一、Linux 发行版介绍

Linux 主要作为 Linux 发行版的一部分而使用。一个典型的 Linux 发布版是由众多的软件集合组成的一个操作系统，该操作系统基于 Linux 内核，通常是一个管理系统包。Linux 用户通过下载 Linux 发布版来安装 Linux 操作系统，该操作系统可以在嵌入式设备、个人计算机及功能强大的超级计算机中使用。一个典型的 Linux 发布版包括 Linux 内核、GUN 工具集合和链接库、软件包、各种文档、图形界面的 Windows 管理器及桌面环境等。并且绝大多数的 Linux 软件是以免费和开源的形式发布，从而允许用户修改该软件。几乎所有的 Linux 发布版都类似于 Unix，但 Android 系统是一个例外，它既不包括基于命令行的接口，也不由典型的 Linux 发布程序组成。

目前 Linux 发行版超过 300 个，最普遍被使用的也有十多个。下面对最常见

使用的几个发行版进行简单介绍。

（1）Ubuntu：Ubuntu 是 Debian 的一款衍生版，也是当今最受欢迎的免费操作系统。Ubuntu 侧重于它在这个市场的应用，在服务器、云计算，甚至一些运行 Ubuntu Linux 的移动设备上很常见。作为 Debian Gnu Linux 的一款衍生版，Ubuntu 的进程、外观和感觉大多数仍然与 Debian 一样，它使用 apt 软件管理工具来管理和安装程序包，是非常适合新手用户使用的一款操作系统。

下载 Ubuntu ISO 映像文件的网址：http://www.ubuntu.com/download。

（2）Fedora：Fedora Linux（第 7 版以前为 Fedora Core）是众多 Linux 发行版之一。它由 Fedora 项目社区开发、Red hat 公司赞助，其目标是创建一套新颖、多功能并开放源代码的操作系统。Fedora 基于 Red Hat Linux。在 Red Hat Linux 发行版终止后，Fedora 就取代了 Red Hat Linux 在个人领域的应用，而 Red Hat 企业版 Linux 则取代 Red Hat Linux 在商业应用的领域。 Fedora 官方支持 x86、x86-64 及 PowerPC 等处理器。Fedora 有庞大的用户论坛和为数不少的软件库，使用 YUM 来管理软件包。下载 Fedora 18 DVD ISO 映像文件的网址：http://mirrors.ustc.edu.cn/fedora/linux/releases/18/Fedora/x86_64/iso/。

（3）OpenSUSE：其是 1992 年由德国的 4 位 Linux 爱好者 Roland Dyroff、Thomas Fehr、Hubert Mantel 及 Burchard Steinbild 共同推出的 SuSE Linux 操作系统下的一个项目（Software und System Entwicklung）。2003 年底，SuSE Linux 被 Novell 公司收购。目前 OpenSUSE 拥有大批满意的用户，并拥有漂亮的桌面环境——KDE 和 GNOME，以及优秀的系统管理工具 YaST。下载 OpenSUSE 13.1DVD ISO 映像文件的网址:http://software.opensuse.org/131/zh_CN。

（4）Debian：Debian 运行起来极其稳定，这使得它非常适用于服务器。Debian 维护三套正式的软件库和一套非免费软件库，这给另外几款发行版，如 Ubuntu 带来了灵感。Debian 这款操作系统派生出了多个 Linux 发行版。它有 37 500 多个软件包。Debian 使用 apt 或 aptitude 来安装和更新软件。Debian 这款操作系统并不适合新手用户，主要适合系统管理员和高级用户使用。Debian 支持如今绝大多数处理器架构。下载 Debian 7.5 ISO 映像文件的网址：http://cdimage.debian.org/debian-cd/7.5.0/kfreebsd-i386/iso-dvd/。

(5)Slackware：Slackware 是 Linux 操作系统中最古老的发行版。它于 1992 年底由 Patrick Volkerding 创建。1993 年 7 月发布了第一个 Linux 发行版的 Slackware。Slackware 特别适合那些喜欢学习和玩弄个性化需求的用户使用。Slackware 的稳定性和简单化是至今为止人们还继续使用它的原因。Slackware 桌面能运行任务 X Windows 管理器和桌面环境，在服务器市场仍然很流行。下载 Slackware 映像文件的网址：http://www.debian.org/distrib/。

(6)CentOS：CentOS 是 community enterprise operating system 的缩写。它提供一个免费的、企业级的、社区支持的计算平台。它和红帽企业级 Linux(RHEL)功能相兼容。2014 年，CentOS 宣布和红帽进行正式合作，但仍和 RHEL 保持独立。因此 CentOS 同样使用 YUM 来管理软件包，并拥有非常稳定的程序包；如果想在桌面端测试一下 Linux 服务器的运作原理，都应该试试这款操作系统。

下载 CentOS 6.5 64 位映像文件的网址：http://isoredirect.centos.org/centos/6/isos/x86_64/。

二、Linux 操作系统的安装

由于 Fedora 18 操作系统对服务器、桌面版都能提供很好的支持。因此本书讲述的所有例子都以 Fedora 18 操作系统进行。常见的 Fedora 18 的安装有两种方法：即单独安装 Fedora18 操作系统和 Windows 系统与 Fedora 18 共存两种方式，下面就 Fedora 18 操作系统的安装步骤进行详细说明。

(1)安装前的准备及系统需求：首先从上述 Fedora 18 提供的映像网址下载 Fedora 18 DVD ISO 映像文件 Fedora-18-x86_64-DVD.iso，然后通过 DVD 刻录软件——nero 等刻录成 DVD 光盘，用这张光盘来安装 Fedora 18 系统。在安装 Fedora 18 系统之前，必须确保计算机至少具有以下配置：硬盘 2GB 以上，内存 4GB。

(2)单独安装 Fedora 18 操作系统：重新启动电脑，按“Del”键，进入电脑“cmos”设置，选择引导盘为光盘，保存退出。光盘开始引导系统，如图 1-1 出现 Fedora 18 启动界面，有三个选项，具体如下所述。

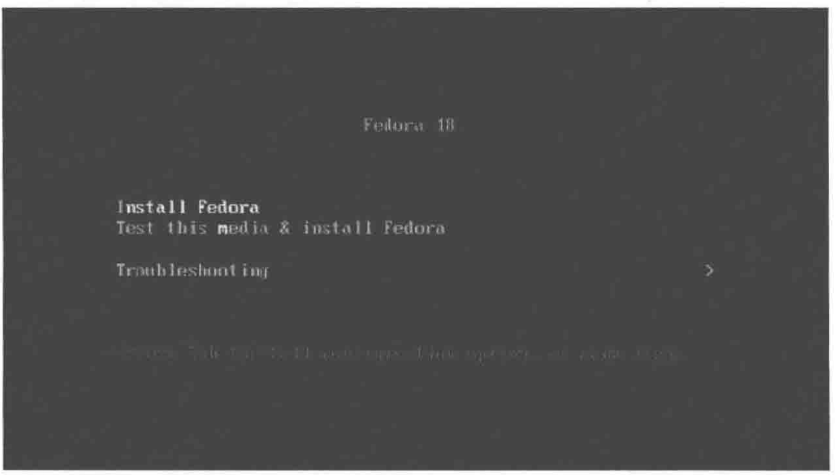

图 1-1　Fedora 18 启动界面

1）“Install Fedora”：选择这个选项，表示采用图形界面来安装 Fedora 18 系统到用户的电脑。

2）“Test this media & install Fedora”：这是默认选项，在安装 Fedora 18 系统之前，检测安装盘的完整性。

3）“Troubleshooting”：选择该选项会出现几个其他的引导选择。

如果确认安装盘没有问题，则选择第一项“Install Fedora”，然后按回车键。

(3)语言选择：如图 1-2 出现“欢迎使用 Fedora 18”界面。移动鼠标选择语言为“中文(中国)”，也可以直接在下面的搜索文本框中输入用户喜欢的语言。如果要将键盘布局设置为与选中的语言一致，请同时选中搜索框下面的复选框。然后按“继续”按钮。

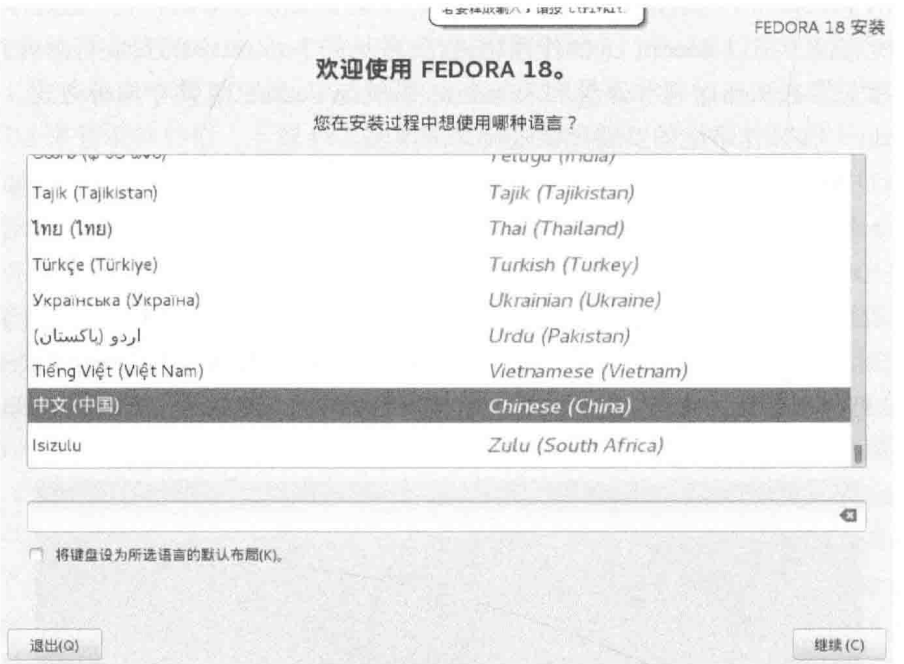

图 1-2“欢迎使用 Fedora 18”界面

(4)安装信息摘要：如图 1-3 所示，该界面有几大选项：“本地化”项有“日期&时间”及“键盘”，“软件”项里面有“安装源”、“网络配置”及“软件选择”，“存储”项里面有“安装目标位置”等内容。操作方法：用鼠标选择要安装的配置项菜单，但完成该项配置后，通过点击该项中的“完成”按钮来完成该项的配置。

点击“日期&时间”项选择区域为“Asia”，城市为“Shanghai”，开启网络时间，然后设置时间和日期。最后点“完成”按钮来结束该项安装配置。

图 1-3 “安装信息摘要”界面

点击“键盘”配置项来配置键盘，如图 1-4 所示，选中“Chinese(Chinese)”，然后点“完成”按钮来完成键盘配置。

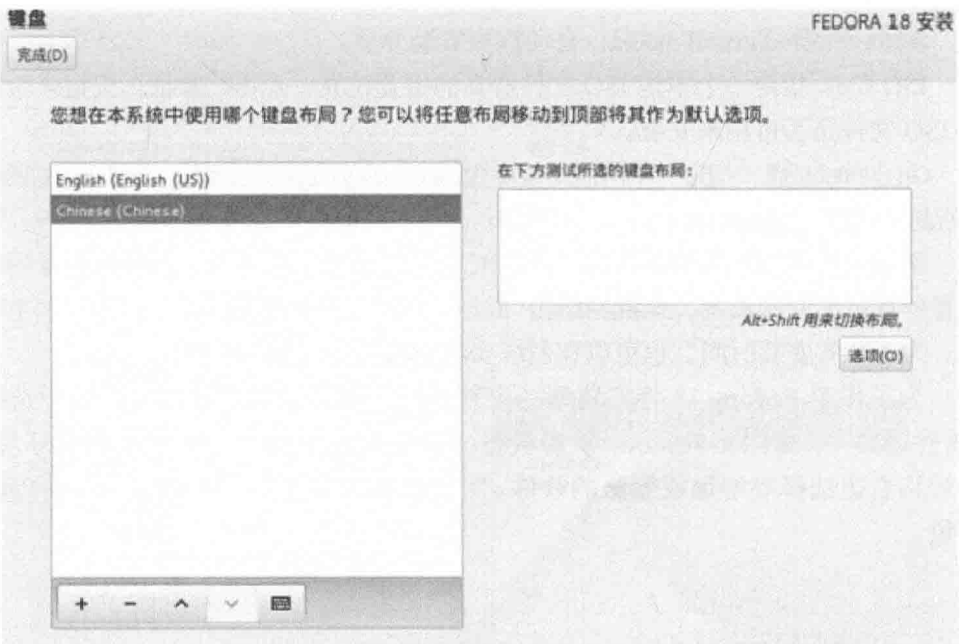

图 1-4　键盘布局选择

从安装信息摘要菜单中选择“安装源”项来定位需要安装的 Fedora 18 的来源文件，如图 1-5 所示，选择可用的本地安装媒体(包括 DVD 或 ISO 文件)或网络安装位置。选择下面的几个选项。

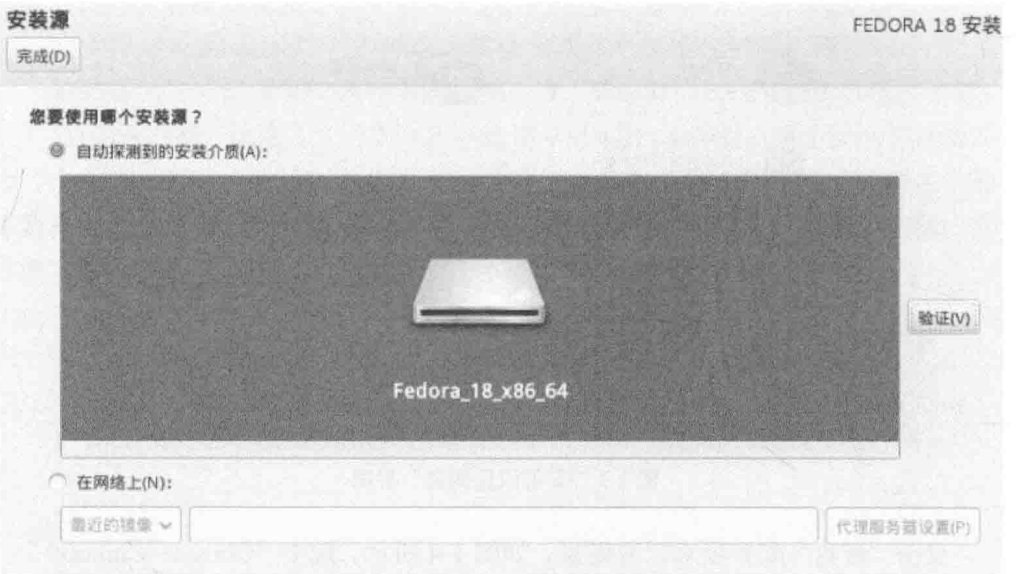

图 1-5　安装源选择

Auto-detected install media：自动探测安装介质。

ISO file：指定一个本地 ISO 文件所在的磁盘位置。点击“验证”按钮来验证该 ISO 文件是否可用来安装。

On the network：指定一个网络安装位置，会出现以下几个选项，根据具体的网络情况进行选择：“Closest mirror：”、“http：//：”、“https：//”、“ftp：//：”、“nfs：”。

从安装信息摘要菜单中选择“网络配置”，如图 1-6 所示，该网络配置用来配置计算机的联网设置，包括网络 IP 地址、网关、子网掩码，以及 DNS 和主机名。为了节省安装时间，也可以在系统安装完成后进行单独配置。

为了指定 Fedora 18 将安装哪一个软件包，则从安装信息摘要菜单中选择“软件选择”，如图 1-7 所示。缺省状态下，Fedora 18 安装 GNOME 桌面环境，然后从右边选择要增加或删除的软件。可以选择开发工具、Fedora Eclipse 等软件包。

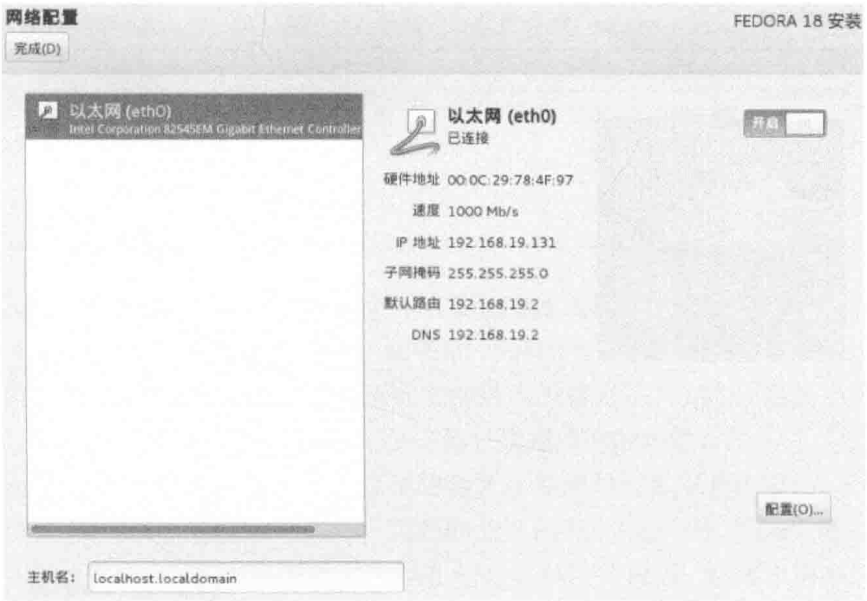

图 1-6　网络配置选择

从安装信息摘要菜单中选择“安装目标位置”，如图 1-8 所示，选中将要安装到的磁盘分区位置。用户可以选择自动创建磁盘分区，或者选择手动分区来创建自定义布局。

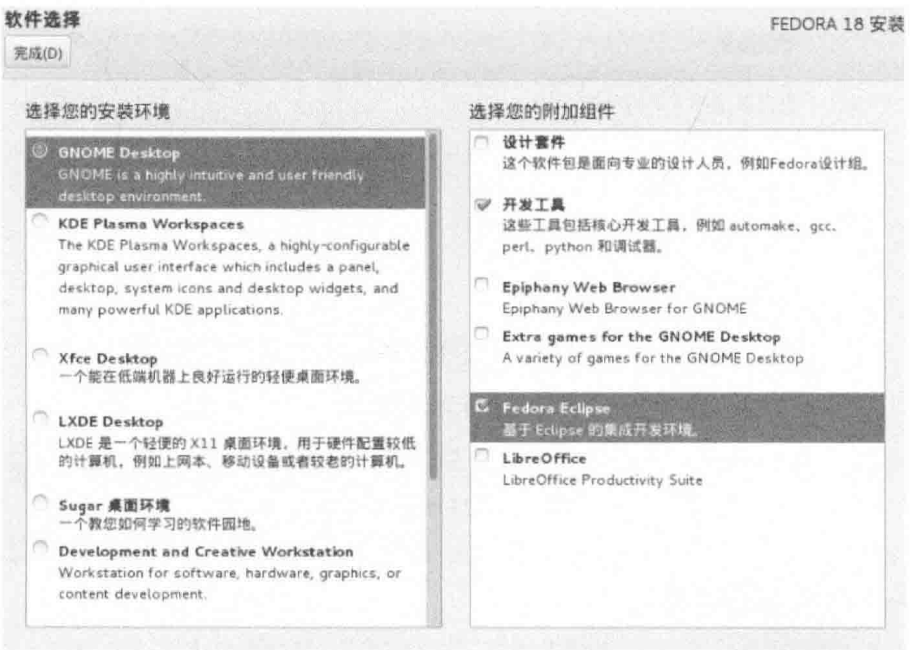

图 1-7　安装软件选择界面