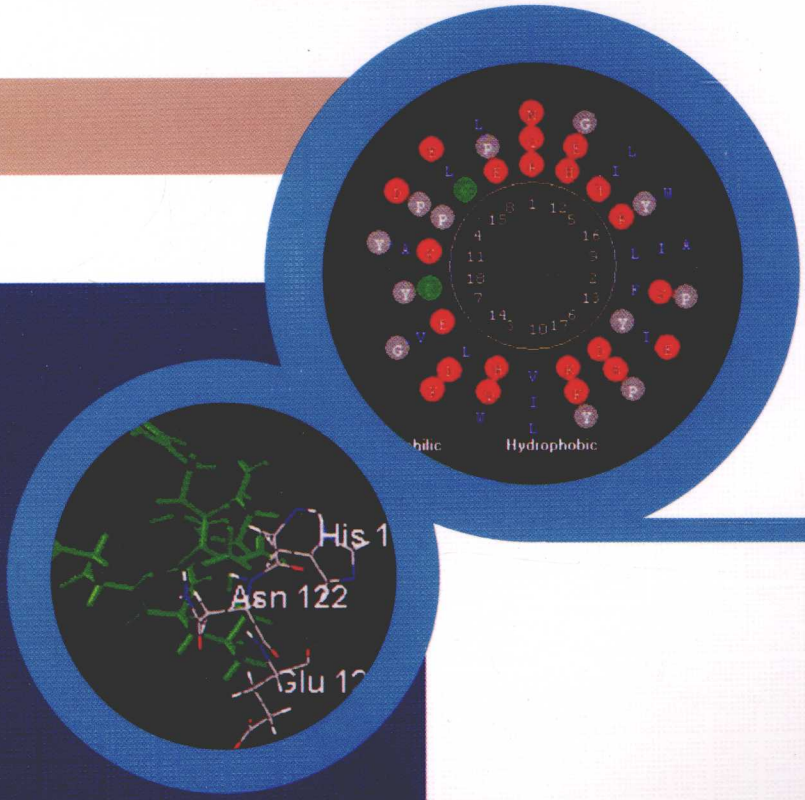
西南交通大学“323实验室工程”系列教材

生物信息学
实践基础教程

王万军 茆灿泉 主编

西南交通大学实验室及设备管理处 主审



西南交通大學出版社
Http://press.swjtu.edu.cn

西南交通大学“323 实验室工程”系列教材

生物信息学实践基础教程

Training Course for Bioinformatics

王万军　茆灿泉　主编

西南交通大学出版社
·成　都·

内容简介

本书从初学者及应用的角度对生物信息学分析所涉及的几个方面进行了介绍，提供了一些重要的常用软件的基本性能、获取方法和简单的使用步骤，力图通过最简单直观的方式使初学者能在最短的时间内了解生物信息学软件的基本功能，而不着眼于对软件功能以及所涉及算法的全面介绍。全书共分为四章，从计算机基础知识入手，以核酸和蛋白质序列分析为重点，由浅入深地介绍涉及核酸与蛋白质序列分析与研究的各个方面。

本书是在西南交通大学实验室及设备管理处“323 基础实验教学平台”建设项目的资助下，由生命科学与工程学院生物系生物信息学专业部分师生集体编著，面向生物信息学软件的初学者，是一本实用性极强的操作指南。本书内容新颖、语言流畅、图文并茂、篇幅适当，操作实例具体而典型，可作为生命科学各专业本科生、研究生实验教材，也可供其他专业师生和科研人员以及广大生物信息学入门和提高的读者参考使用。

图书在版编目（CIP）数据

生物信息学实践基础教程 /王万军，茆灿泉主编. —成都：西南交通大学出版社，2009.6
（西南交通大学“323 实验室工程”系列教材）
ISBN 978-7-5643-0285-6

Ⅰ.生… Ⅱ.①王…②茆… Ⅲ.生物信息论－高等学校－教材 Ⅳ.Q811.4

中国版本图书馆 CIP 数据核字（2009）第 105588 号

西南交通大学“323 实验室工程”系列教材

生物信息学实践基础教程

王万军　茆灿泉　主编

*

责任编辑　张　雪
特邀编辑　牛　君
封面设计　本格设计

西南交通大学出版社出版发行
（成都二环路北一段 111 号　邮政编码：610031 发行部电话：028-87600564）
http: //press.swjtu.edu.cn

四川森林印务有限责任公司印刷

*

成品尺寸：185 mm×260 mm　　印张：15.5
字数：386 千字
2009 年 6 月第 1 版　　2009 年 6 月第 1 次印刷
ISBN 978-7-5643-0285-6
定价：25.00 元

《生物信息学实践基础教程》

编 委 会

主　编　王万军　茆灿泉

编　者（按姓氏笔画排序）：

王万军　刘　岩　江　年　吴　坚

宋晓婕　陈　颖　茆灿泉　郭志云

郭泰林　唐静仪　樊晓霞　魏大木

前　言

以人类基因组计划为代表的生物基因组研究的顺利实施及后基因组和蛋白质组等研究的深入开展，产生了大量的生物大分子数据，这些生物大分子数据具有丰富的内涵及目前我们尚不清楚的生命信息。生物信息学（Bioinformatics）就是通过收集、组织、管理、分析生物大分子数据，得到深层次的生物学知识，加深对生物世界的认识；在生物学、医学研究和应用中，利用生物大分子数据及分析结果，可大大提高研究和开发的科学性及效率，例如，可根据基因功能分析结果检测疾病相关基因，根据蛋白质分析结果进行新药设计等。

生物信息学是目前生命科学研究与发展中的璀璨明珠，可以说，如果没有生物信息学理论技术的发展与工具开发，目前的生命科学研究与实践将仍停留在瞎子摸象般的缓慢进展之中，正是有了生物信息学，生命科学各领域快速渗透、交叉，许多生命现象可以在分子水平上统一起来，同时，一些生命科学中的新东西又不断涌现，因此，生物信息学已成为现代生命科学研究与学习中不可或缺的重要内容。现在，国内外已有众多的生物信息学书籍与资料出版，它们对生物信息学的发展及生命科学的整体推动起到了重要作用；同时，随着生物信息学理论的建立及不断完善，在国内外又产生了众多的基于计算机算法的生物信息学软件工具（程序）及网络资源，它们的开发与应用又对生物信息学的推广普及起到了促进作用。然而，目前大多数书籍都是系统性地介绍生物信息学理论与算法分析以及生物信息学资源，鲜有生物信息学实践操作方面的内容，而这是从事有关研究与学习所急需的，尤其是对初学者而言更是如此；另一方面，在生物信息学软件工具（程序）及网络资源方面，又存在着一个软件（或网络资源）有多个功能、不同软件（或网络资源）可同时具备同一功能，以及不同软件（或网络资源）之间的格式差异，这虽然给使用者提供了广泛的选择余地，但同时也给使用者（尤其是初学者）带来了无所适从之感；还有，由于生物信息学软件工具（或网络资源）不可避免地需要相关的计算机技术及算法知识，使许多人想学习生物信息学但又不敢贸然涉足。我们在长期的教学、科研实践中，对初学者的这种无所适从之感有深刻的体会，我们期望通过一本“step by step”或“follow me”形式的册子使初学者能迅速入门并初步掌握一个软件的使用，从而使他们能很快树立学习信心并逐步深入了解与掌握软件的全部功能，进而使其触类旁通、举一反三，最终达到既能了解生物信息学的内容也能理解生物信息学根本含义的目的。本书不具体涉及生物信息学的理论与算法，只是通过一个个的实际操作，一步步地将初学者带到生物信息学天地中，让他们逐渐体会生物信息学能为自己做些什么以及自己要怎么去做。

全书共分为四章。第一章简单介绍了在生物信息学学习中要用到的几种计算机基础知识，这主要是为那些对计算机还不熟悉的读者准备的，已具有一定计算机技术基础的读者可直接跳过这一章。第二章首先介绍了一种序列格式转换软件，使一种格式的序列数据可被不同软件调用；接着介绍了几种序列分析工具软件，读者可体会使用不同软件实现同一功能；紧接着再介绍引物设计与 DNA 多态性分析；最后介绍了一种凝胶图像分析软件。第三章介绍了多序列比对、系统进化以及蛋白质结构与功能分析方面的几个软件，还详细介绍了蛋白

质结构比较、显示与打印的几个软件。第四章简单介绍了几个生物信息学常用数据库的使用与 GCG 软件包，最后还介绍了一种文本挖掘与聚类分析方法。

本书是在西南交通大学实验室及设备管理处“323 基础实验教学平台”一期建设项目“生物信息学基础实验教学平台”的资助和 2006 年四川省省级精品课程“生物信息学”的推动下，由生命科学与工程学院生物系生物信息学专业部分教师及部分 2007 级生物化学与分子生物学专业研究生集体编著而成，是在各位师生长期对生物信息学软件及网络资源的熟练使用基础之上，参考各个软件的帮助信息并结合个人使用体会以及一些网络信息综合编著而成。其中，魏大木讲师编写了第一章的第一、二、三、四节；王万军教授编写了第二章的第一、二、六节以及第三章的第一、三、五、七节，带领唐静仪同学编写了第二章的第三、四节，带领宋晓婕同学编写了第二章的第七节，带领樊晓霞同学编写了第三章的第二节；茆灿泉教授带领江年、陈颖同学编写了第一章的第五节、第三章的第四、六节以及第四章的第一节；郭泰林讲师编写了第二章的第五节；郭志云讲师编写了第四章的第二节；吴坚副教授带领刘岩同学编写了第四章的第三节。感谢学校实验室与设备管理处给予经费资助；感谢以上各位老师与同学的辛勤工作，使本书能顺利付梓；感谢西南交通大学出版社为本书的出版所做的工作。

生物信息学发展迅速，新的理论、知识、软件及网络资源不断出现，而我们的知识水平有限，再加之时间仓促、取舍不当，书中难免存在遗漏及不妥之处，敬请读者批评指正。

编　者

2009 年 4 月

目　录

第一章　计算机应用基础

第一节　Word 的使用

Word 是功能极强的文字处理和版面编排软件，简单易学、操作界面友好、智能化程度高，可以编辑各种文档（如报告、文章等）以及对各种段落进行设置，在进行文档编辑时，还可以设置字体以及各种格式。Microsoft Word 2003 是 Word 的较新版本，保持了以前版本的优点，同时具有更强大的网络功能和通信功能。

一、Word 的基本操作

(1) 文件的打开、关闭、保存和页面设置；
(2) 对文字和段落格式的设定；
(3) 在文档中插入并制作表格；
(4) 文档编辑、修改。

二、Word 的使用步骤

1. 文件的打开、关闭、保存和页面设置

(1) 新建文档：用鼠标点击桌面 Word 图标或按“开始”→“程序”→“Microsoft Office Word”顺序启动，运行 Word 后选择图 1.1.1 中“文件”菜单中的“新建”命令即可。

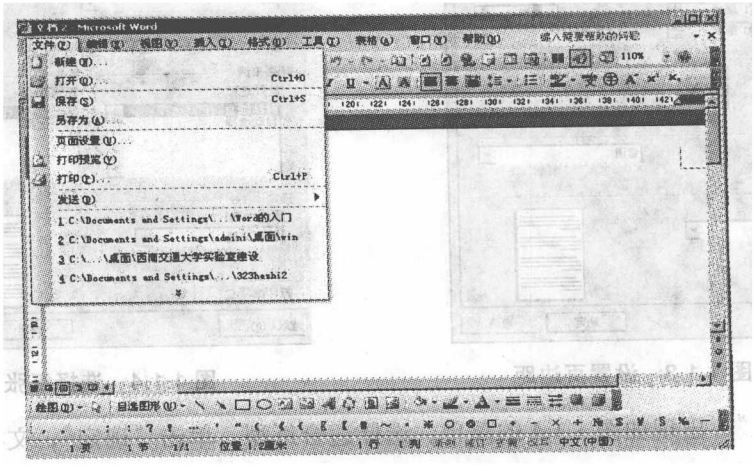


图 1.1.1　Word 中的文件菜单

（2）打开文档：如图 1.1.1 所示，打开已存在的文件则选择“文件”菜单中的“打开”，选择要打开文件的存放路径。同时，单击 ，可以回到目前所处文件夹的上一层文件夹之中。单击 ，可以打开[工具]下拉式框，使用[查找]、[删除]、[打印]等工具。

（3）保存文档：如图 1.1.1 所示，要保存正在编辑的文件，则选择“文件”菜单中的“保存”，选择保存路径和保存格式。在 Word 中使用[保存]指令，如果文件已经保存过一次，那么单击[保存]指令时，文件就直接进行保存，而不出现[另存为]对话框了。当然，也可以把该文件另存为其他文件名或者文件格式，如图 1.1.2 所示。

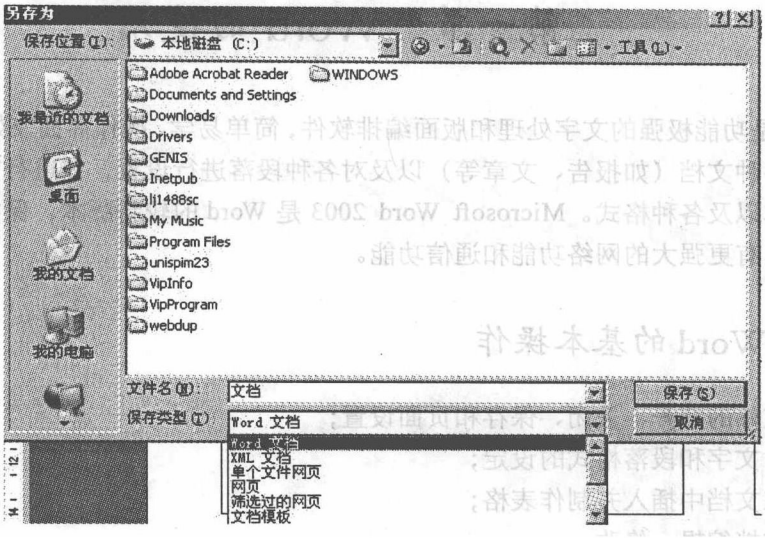


图 1.1.2　保存文档

（4）页面设置：设定文件版面。要对编辑的文件进行页面设置，如图 1.1.1 所示，选择“文件”菜单中的“页面设置”，出现图 1.1.3 所示的菜单。在这个菜单中选择“页边距”，设置合适页边距，单击“确定”，文档将根据所设置的页边距出现。

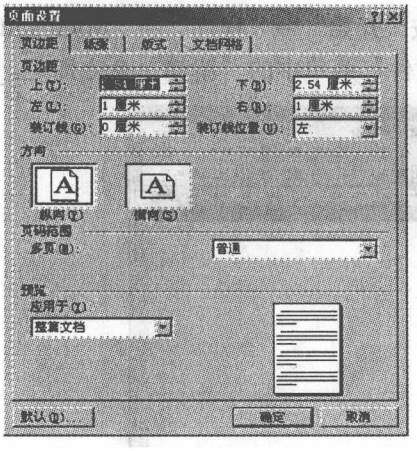


图 1.1.3　设置页边距

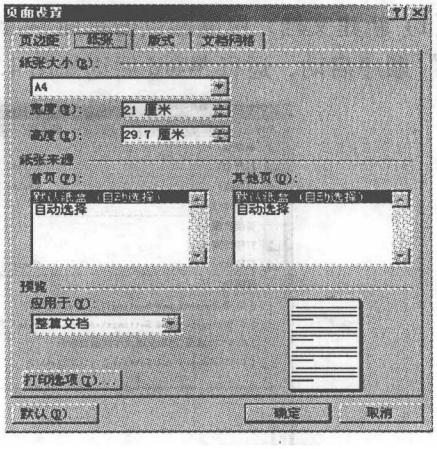


图 1.1.4　选择纸张

选择“纸张”，出现如图 1.1.4 的菜单，选择适合纸型，单击“确定”，文档将根据所设置的纸型出现。

2. 文字和段落格式的设定

(1) 文字字体、格式的设定。

首先，选定需要设置字体的文字，单击“格式”，出现如图 1.1.5 所示的菜单。

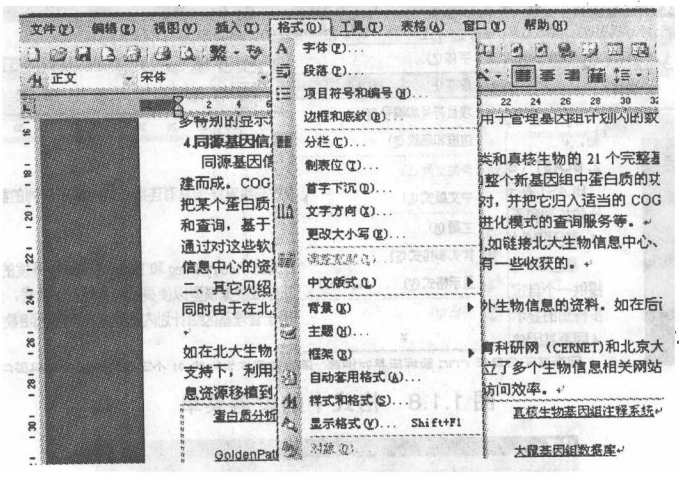


图 1.1.5　格式菜单

然后选择“字体”菜单，出现如图 1.1.6 所示的菜单。在“字体”标签中选择所需字体、字形、字号，单击“确定”。

同时，在图 1.1.6“字体”标签中，可设置文字的效果，如上下标，也是选中需要设置的文字，然后在图 1.1.6 中选择上标或下标，单击“确定”即可。

选择“字符间距”，出现如图 1.1.7 所示菜单，选择间距和位置来设定文字之间的距离，单击“确定”即可。

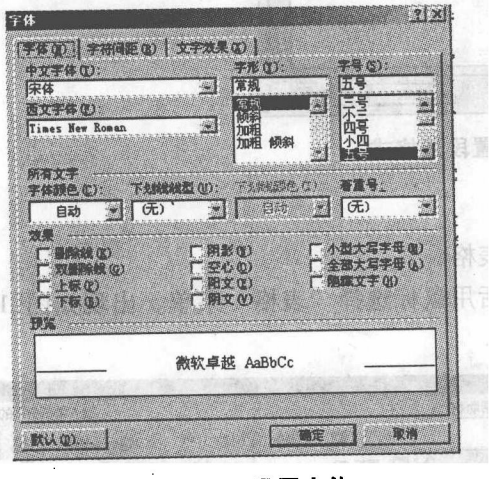


图 1.1.6　设置字体

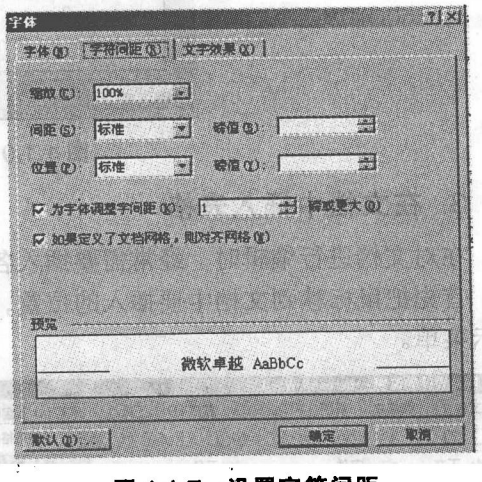


图 1.1.7　设置字符间距

(2) 段落格式的设定。

如图 1.1.8 所示，选择“格式”中的“段落”菜单，出现如图 1.1.9 所示菜单：

单击“缩进和间距”标签，在“对齐方式”中选择一种段落的对齐方式，单击“确定”。在 Word 中，除了可以在“段落”对话框中设定段落对齐方式外，还可以在“格式”工具栏中设

定段落的对齐方式。在 Word 的“格式”工具栏中有“两端对齐”“居中”“右对齐”及“分散对齐”四项工具按钮，使用这四项工具按钮，就可以改变段落的对齐方式。当然，“段落”菜单中也有其他一些如缩进、间距等对段落的设置。

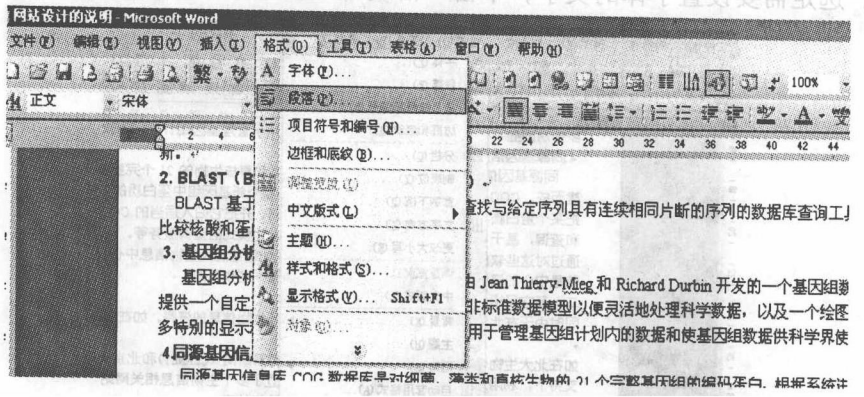


图 1.1.8　格式中的段落菜单

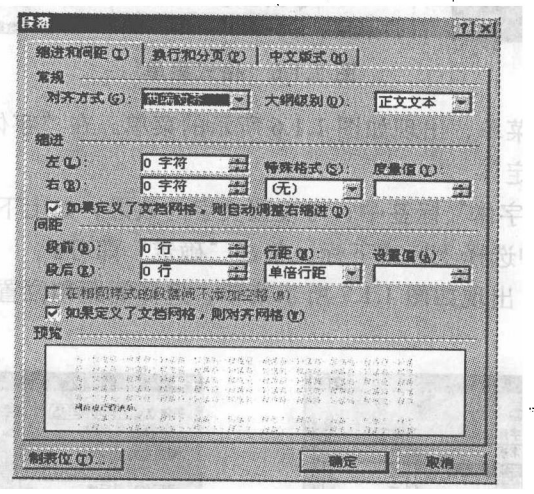


图 1.1.9　设置段落格式

3. 在文档中插入表格

在对文档进行编辑时，经常需要插入各种表格。

首先把鼠标移到文档中要插入的位置，然后用鼠标选择“表格”菜单，出现如图 1.1.10 所示菜单。

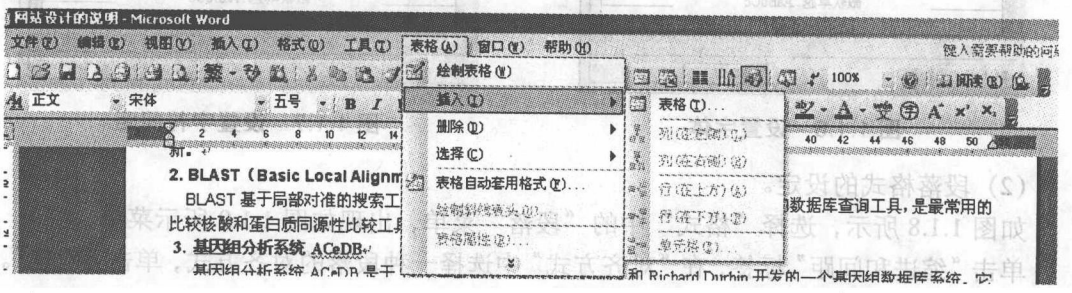


图 1.1.10　表格菜单

选择插入表格，出现如图 1.1.11 所示菜单：

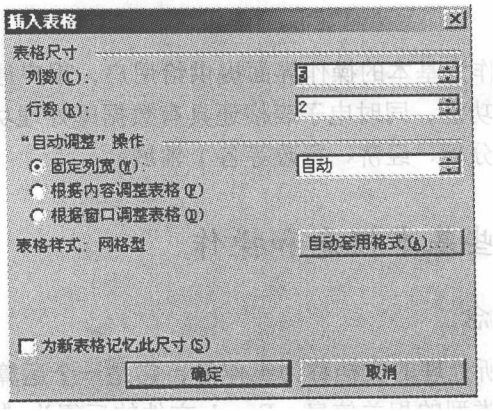


图 1.1.11　插入表格

在列数和行数文本框中设定表格的行数和列数，然后单击“确定”，回到 Word 文档编辑窗口，就会出现设定的表格。

4. 文档编辑、修改

在文档的编辑中，需要对编辑过的文章进行修改时，可选择菜单栏中的“编辑”进行快速修改，如图 1.1.12 所示。

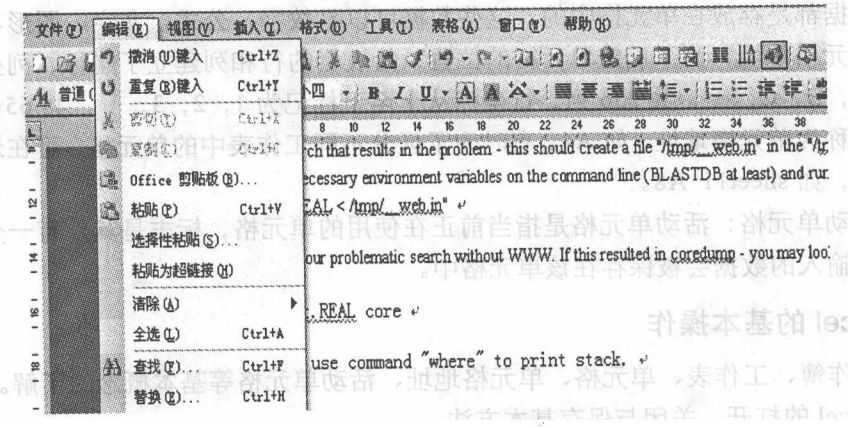


图 1.1.12　编辑菜单

选择“撤消键入”或“重复键入”菜单，可进行文章的快速修改。

练　习

(1) 了解 Word 具有哪些功能，与同类编辑软件相比有哪些特点。同时要注意与其他软件结合使用，熟练掌握编辑技巧。

(2) 用 Word 编辑一个 1 000 字左右的文档，要求进行字体设置、段落设置和添加表格。

第二节 Excel 的使用

Excel 是以二维表格作为基本的操作界面提供给用户，用户通过在表格中输入数据，由程序自动完成诸如计算等功能，同时电子表格还具有数据库处理功能。因此，电子表格被广泛应用于统计分析、财务分析、经济、行政等各个领域。

一、Excel 的一些基本概念和操作

1. Excel 的基本概念

(1) 工作簿：Excel 所处理的文档称为工作簿，它是一个运算和存储数据的文件，可以在一个工作簿中管理各种类型的相关信息。Excel 文件的后缀为“.xls”。

(2) 工作表：可以把 Excel 的工作簿理解为一个账本[①]，账本中的一页在 Excel 中称为一个工作表。Excel 2003 的工作表是由 65 536 行×256 列组成的巨型表格。工作表用来列出数据和分析数据，表中可以存放图表、对话框、程序等。

一个工作簿可以包括多个工作表，Excel 2003 默认为每个工作簿打开 3 个工作表，每个工作表有一个名字，称为工作表标签，系统把它们记为 sheet1，sheet2，…在一个工作簿中最多可以建立 255 个工作表。

(3) 单元格：工作表中的每一个表格称为单元格，它是组成工作表的最小单位。用户输入的任何数据都是存放在单元格中的，这些数据可以是数字、公式、文字、图形、声音等。

(4) 单元格地址：为了确定单元格的位置，给表格的行和列建立了坐标，列坐标从左至右标记为 A，B，C，…，共 256 列，行坐标从上至下标记为 1，2，3，…，共 65 536 行。单元格的坐标称为单元格地址，如 A4。有时为了区分不同工作表中的单元格，可在地址前加上工作表名称，如 sheet1！A8。

(5) 活动单元格：活动单元格是指当前正在使用的单元格，标志是其外有一个黑色的方框，这时所输入的数据会被保存在该单元格中。

2. Excel 的基本操作

(1) 工作簿、工作表、单元格、单元格地址、活动单元格等基本概念的了解。

(2) Excel 的打开、关闭与保存基本方法。

(3) 对工作簿、工作表和单元格的使用。

二、Excel 的使用步骤

1. Excel 的打开、关闭与保存

(1) 启动一个 Excel，用鼠标点击桌面 Excel 图标或从“开始”→“程序”→”Microsoft Office Excel”启动，运行 Excel 后选择“文件”菜单，点击“新建”，即可新建一个 Excel 文

① 根据《现代汉语词典》中字义的解释，本书文字内容中“账本”“账户”“账号”均使用“账”字，但计算机截图不便修改，仍保留“帐”字，特此说明。——责任编辑注

档，如图 1.2.1 所示。

（2）同样，如图 1.2.1 所示，若欲打开已存在的文件则选择“文件”菜单中的“打开”，选择要打开文件的存放路径。同时，单击，可以回到目前所处文件夹的上一层文件夹之中。单击工具(T)，可以打开[工具]下拉式框，使用[查找]、[删除]、[打印]等工具。

（3）同样，如图 1.2.1 所示，若要保存正在编辑的 Excel 文件，则选择“文件”菜单中的“保存”，选择该文件的保存路径和保存格式。在 Excel 中使用[保存]指令，如果文件已经保存过一次，那么单击[保存]指令时，文件就直接进行保存，而不出现[另存为]对话框了。当然，也可以把该文件另存为其他文件名或者文件格式。

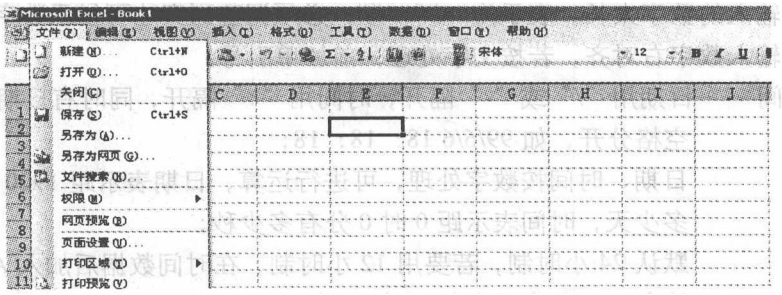


图 1.2.1　Excel 中的文件菜单

2. 工作簿、工作表和单元格的使用

（1）通过菜单插入单元格。

如图 1.2.2 所示，选择菜单栏中的“插入”，出现如下菜单。

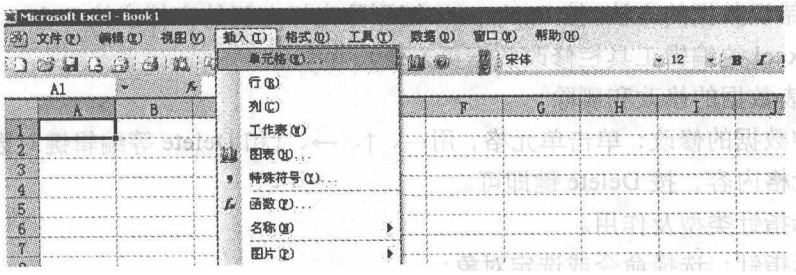


图 1.2.2　插入菜单

然后选择“单元格”，出现如图 1.2.3 所示菜单，选择插入的格式，单击“确定”即可。

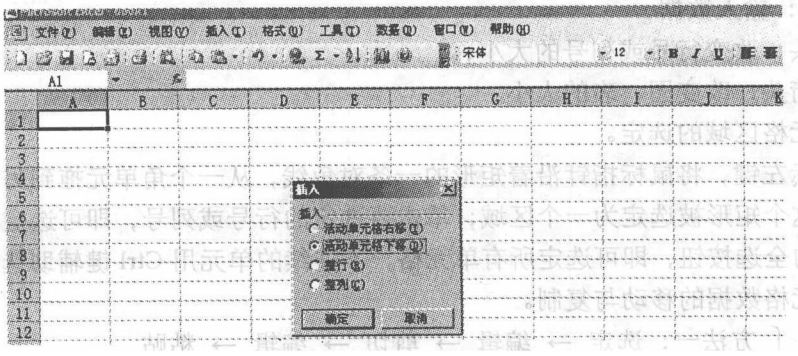


图 1.2.3　插入单元格

(2) 活动单元格：鼠标单击任何单元格，使其成为当前可输入内容的单元格，则该单元格即为活动单元格（每个单元格可容纳 32 000 个字符）。

(3) 在工作表中输入数据。

① 输入数据的格式：

文字 ——默认左对齐；

数字 ——输入正数，不用“+”号，加了也被省略；

用括号括住数字，表示输入的是负数，如（321）表示－321；

分数前加 0 与空格，否则表示为日期，如 1/2 表示 1 月 2 日；

输入的数字太长，表示为########，必须调整列宽才可显示数字；

输入数字右对齐，若按文字处理，前加单引号。

日期和时间 ——日期用“/”或“-”隔开，时间用“:”隔开，同时有日期和时间必须用空格分开，如 99/6/6 18：18：18；

日期、时间按数字处理，可进行运算，日期表示距 1900 年 1 月 1 日有多少天，时间表示距 0 时 0 分有多少秒；

默认 24 小时制，若要用 12 小时制，在时间数据后加入 AM 或 PM，如输入 6：18 表示上午 6 时 18 分，不是下午；

输入当前日期用 Ctrl＋;，当前时间用 Ctrl＋Shift＋:；

2 位年份 00-29 表示 2000—2029，30-99 表示 1930—1999。

② 输入数据的方法：单击单元格→键入数据。用箭头←、↑、→、↓或回车、Tab 来结束输入。

③ 改正错误数据的方法：按 Backspace 键删除光标左侧的字符或按 Delete 键删除光标右侧的字符（Excel 的编辑工具栏修改数据更方便）。

④ 工作表数据的修改和删除：

单元格中数据的修改：单击单元格，用←、↑、→、↓和 Delete 等编辑键对数据进行修改。

删除单元格内容，按 Delete 键即可。

(4) 鼠标指针类型及作用。

常规鼠标指针：选择命令或选定对象；

带阴影“十”字指针：选择单元格或区域；

黑“十”字指针：自动填充；

I 型指针：输入数据；

双向箭头：改变行号或列号的大小；

双头斜箭头：改变图、表的大小。

(5) 单元格区域的选定。

按下鼠标左键，将鼠标指针沿着矩形的一条对角线，从一个角单元拖到另一个角单元后再放开，则这个矩形被选定为一个区域；单击要选定的行号或列号，即可选定整行或整列；单击左上角的全选按钮，即可选定所有单元格；不连续的单元用 Ctrl 键辅助选定。

(6) 单元格数据的移动与复制。

数据移动 { 方法一：选定 → 编辑 → 剪切 → 编辑 → 粘贴
方法二：选定 →（鼠标指针变成一个箭头）拖动 → 目标位置放开

数据复制 { 方法一：选定 → 编辑 → 复制 → 编辑 → 粘贴
方法二：选定 →（鼠标指针变成一个“十”字箭头）拖动→目标位置放开

(7) 单元格的插入、删除及合并方式。

插入：在要插入空行的行号上单击右键，选择“插入”。

合并：将多个相邻的单元格选中，单击右键选“合并”。

删除：选择要删除的行，单击右键选“删除”。

(8) 多工作表中数据的编辑。

单击原工作表标签→选定单元格→复制/剪切→单击目标工作表标签→单击目标单元格→粘贴。

(9) 数据的类型。

Excel 是一个数据处理与分析的软件，所以数据的类型十分关键。

设置单元格数据类型：右键单击单元格→“设置单元格格式”→“数字”标签。我们一般使用的是常规类型，只有特殊的数据才需要改变单元格数据类型。

常见的数据类型有：常规、数值、时间、日期、百分比。

练　习

(1) 了解 Excel 具有哪些功能，并掌握 Excel 的一些基本操作。

(2) 编辑表 1.2.1 所示初学者成绩：

表 1.2.1　初学者成绩

学号	姓名	政治	语文	数学	英语
0301	王小红	84.5	91	88	85
0302	梁虹春	87	95	81	91
0303	冯立峰	77	84	79	74
0304	吕国华	80	88	91	78
0305	胡建军	91	86.5	74	80
0306	李静静	90	87	91	96
0307	马晓莉	82	93	97	89

第三节　Foxmail 的使用与设置

随着互联网的发展，电子邮件成为当今越来越常用的工具之一，而 Foxmail 这个免费的收发电子邮件的工具正好为我们提供了方便。

一、账户设置

首先打开 Foxmail 主界面，如图 1.3.1 所示，选择“账户”菜单下的“新建”命令，跳出

如图 1.3.2 所示向导，在向导中添加账户的名称和邮件的存放路径后点击“下一步(X)>”按钮，出现如图 1.3.3 所示窗口，要求用户输入“发送者姓名”（在发送的邮件中显示用户姓名，以便于邮件接收者识别邮件是由哪个用户发送过来的）与“邮件地址”（在发送的邮件中显示发送者的 E-mail 地址，以便于接收者回信）信息，完成后点击“下一步(X)>”按钮，出现如图 1.3.4 所示窗口，填写邮件服务器和密码（POP3 是用来接收邮件的服务器，SMTP 是用来发送和中转邮件的服务器，密码是指邮箱的密码。对于一些常用的免费邮箱如 163、新浪等，Foxmail 会自动填写正确的 POP3 和 SMTP 服务器地址。如果服务器地址填写不正确，就不能正常收/发邮件），最后选择是否需要进行 SMTP 验证和邮箱是否保留邮件，一般需要选择 SMTP 验证才可以发送邮件。

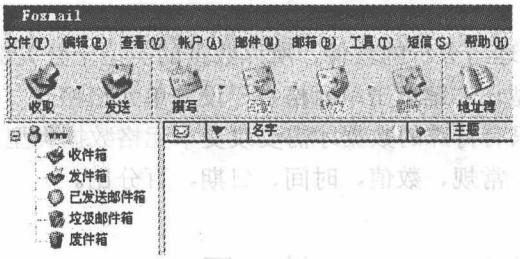


图 1.3.1 Foxmail 主界面

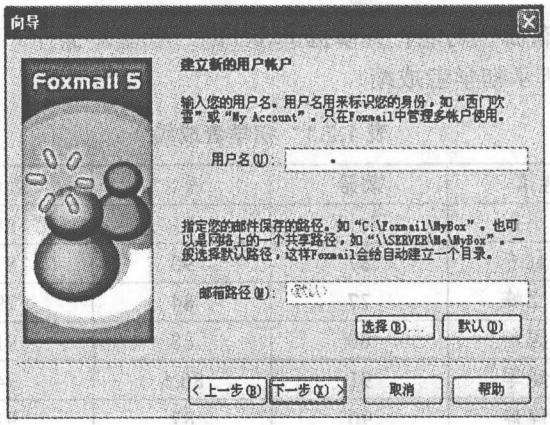


图 1.3.2 新建账户向导

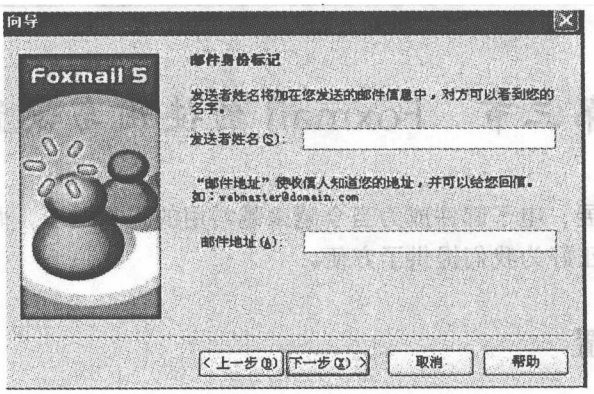


图 1.3.3 设置发送者姓名和邮件地址