# AN INTRODUCTION TO BIOINFORMATICS ALGORITHMS

# 生物信息学算法导论

[美] N.C.琼 斯 (N.C.Jones )
P.A.帕夫纳 (P.A.Pevzner) 著

王翼飞 等译

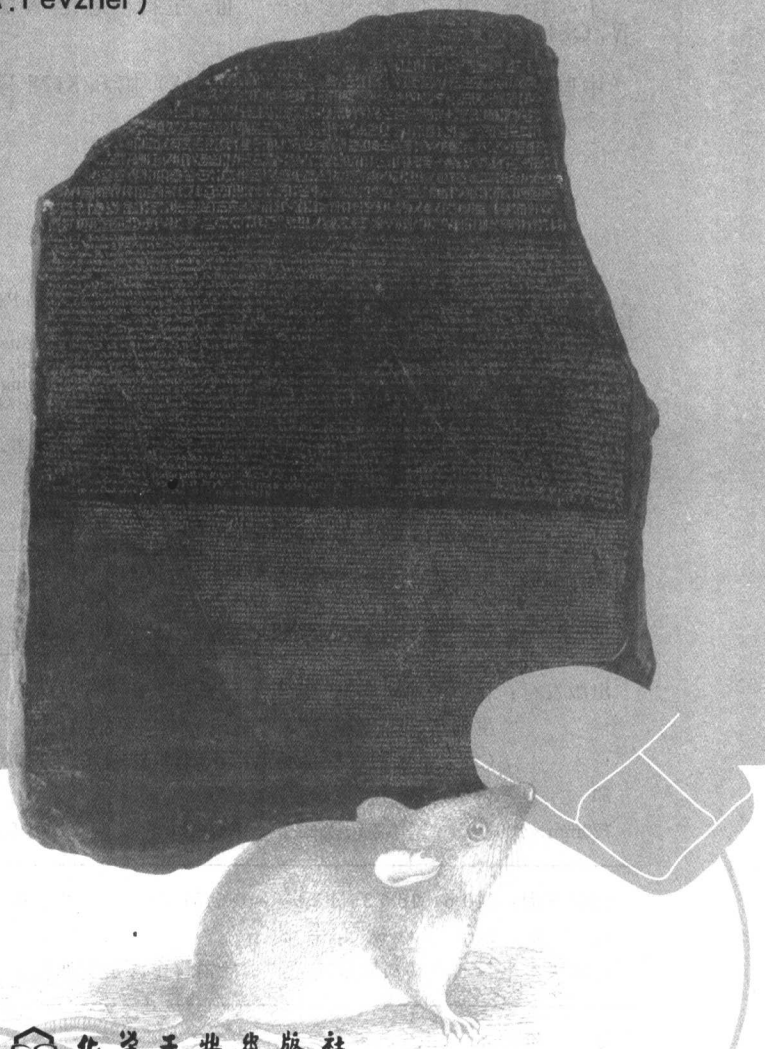
化学工业出版社
生物·医药出版分社

# AN INTRODUCTION TO BIOINFORMATICS ALGORITHMS

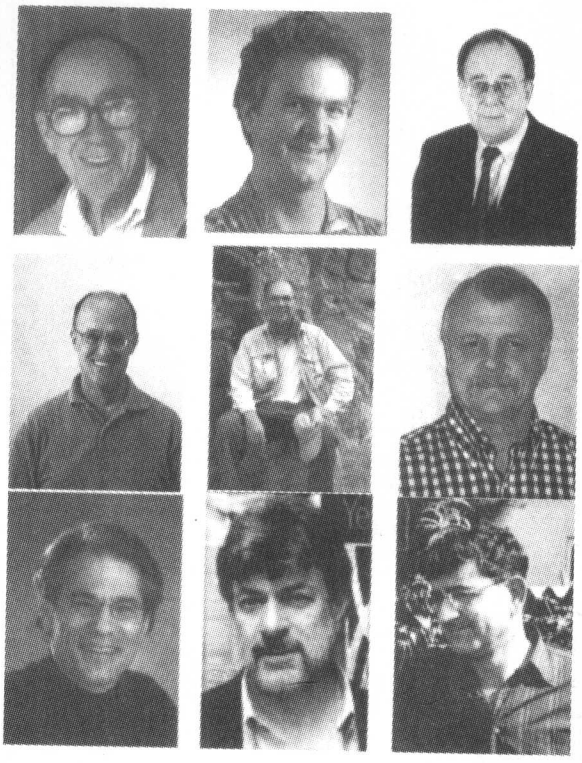
# 生物信息学算法导论

[美] N.C.琼 斯 (N.C.Jones)
P.A.帕夫纳 (P.A.Pevzner) 著

王翼飞 等译



化学工业出版社
生物·医药出版分社
·北京·



历史上重要的科学家

| | |
|---|---|
| Russell Doolittle | 第 3 章 |
| David Haussler | 第 11 章 |
| Richard Karp | 第 2 章 |
| Webb Miller | 第 7 章 |
| Gene Myers | 第 9 章 |
| David Sankoff | 第 5 章 |
| Ron Shamir | 第 10 章 |
| Gary Stormo | 第 4 章 |
| Michael Waterman | 第 6 章 |

# 译者序

很高兴《生物信息学算法导论》中译本和大家见面了。

生物信息学（bioinformatics）自20世纪80年代后期诞生以来发展迅速，不仅在于其自身的内涵和外延方面都有了明显的拓展，而且也确已成为当今生命科学研究不可或缺的重要支撑技术之一。生物信息学是一门涵盖数学、物理学、化学、生物学和计算机科学的多学科交叉的新兴学科。近年来，生物信息学研究在我国也引起了广泛的重视，许多高等院校和研究机构都开展了这一领域的研究，并开设了相关的研究生课程和本科生课程。虽然相关的书籍也出版了不少，但对生物信息学研究中用到的数学技术却很少有深入的剖析。其实，离开了数学技术的支撑，生物信息学的研究是走不了多远的。本书从基本的算法入手，系统而深入地展示了生物信息学研究中涉及到的数学技术问题，并对此作了活泼而不失严谨、深刻而又明晰的阐述。

本书作者Pavel A. Pevzner现在是美国加利福尼亚大学圣地亚哥分校（UCSD）的教授。他生于俄罗斯，数学专业出身，1985年开始转入计算分子生物学研究，20世纪90年代初赴美国。他长期从事生物信息学的教学和研究工作，无论在算法的理论方面还是在算法的实现方面都具有丰富的教学与研究经验。本书是他继《计算分子生物学：算法逼近》后的又一本力作。本书的另一作者Neil C. Jones则是由化学专业转向生物信息学研究的。作者们的自身学术经历，使他们对生物信息学的教与学有着深切的体会。他们懂得应当如何对有志于生物信息学研究的学生进行系统而有效的讲授和培训，使他们尽可能快地掌握生物信息学的基本知识和实用技巧，进而站到生物信息学研究的前沿。可以说，本书就是他们多年来从事生物信息学的学习、研究和教学的经验总结，是一本经过教学实践检验的、适合于本科生教学需要的、非常出色的入门教材。因此，本书的出版立即受到了广泛的重视，除欧美国家外，亚洲国家中，如印度、日本等国也将其引进为大学教材。我们认为，尽快出版本书的中译本不仅可为我国有志于生物信息学研究的青年学生提供一本优秀的入门读物，也可为我国的生物信息学本科层次的教学提供一本合适的教材。

上海大学数学系生物信息学实验室的研究生彭新俊、周文、张冬宁、刘阳、万虎、蔡传政、张家军、刘祥、沈称意、李冯、冯铁男、阎正楼、张玉滨、孟炜、张亮生、刘焕、王飞飞、赵洁苑、吕玉龙、沈青松、江浩、王晶等参加了本书的研讨和翻译，全书最后由王翼飞统稿并校订。

中国科学院上海生命科学院的丁达夫、李亦学、赵慕钧研究员，上海大学数学系史定华教授，上海师范大学郭本瑜教授，美国印地安那大学汤海旭博士，以及化学工业出版社的编辑等对本书的翻译出版都给予了热情的支持，并提供了宝贵的意见，原著作者 N. C. 琼斯和 P. A. 帕夫纳又特意为中译本写了序，在此一并致谢。

本书的翻译工作得到了上海市重点学科建设项目及国家 863 高技术研究发展计划项目（2006AA02Z190）的支持。

要将一本写作风格幽默、技术内涵精湛、适用于大学本科生的教材原汁原味地翻译成中文，不是一项轻松的工作，我们勉为其难。限于译者水平，对原著的理解和翻译难免有不妥甚至错误之处，敬请读者不吝批评指正。如能将译著中的纰漏之处用电子邮件方式通知我们（yifei wang@staff. shu. edu. cn），译者将不胜感激。

**王翼飞**

**2007 年 3 月于上海大学**

# 中译本序

当一位同事告诉我们《生物信息学算法导论》正在被翻译成中文以方便中国学生的学习时，起初我们不太相信（因为我们以为他在开玩笑），后来则很欣喜（因为我们发现他说的是真的）。尽管这本书的第一版在美国面世仅三年，但已被世界上十几所大学以多种形式采用，印度出版了一种供学生使用的特殊版本，而日本正将此书译为日文，现在又将以中文出版。当然，我们想声明，这本书如此成功应该完全归功于幽默的写作风格和精湛的技术内涵。但不幸的是，当下有一种不正确的说法：这本书的面世恰遇一波计算生物学这一热门领域的新学生潮，而读者贪婪的求知欲望又使这本书备受关注，但这超出了有关我们研究领域的任何特殊技术。来自使用者的反馈意见中既有赞美又有批评。有些教授抱怨我们一点都不涉及确定的专题（比如，支持向量机学习），或者我们提及的主题在一些事例中阐述得太过浅显（如 Baum-Welch 算法）。在亚马逊网站上（Amazon.com），一位评论者抱怨这不是一本“生物信息学”的书，而是一本“算法”的书。其实，对任何实际读到这个标题感到困惑的人而言，我们的说明是清楚的。即便如此，这些批评并没有令我们惊奇——我们故意绕开许多计算上的课题，是因为它们非常复杂，倘使对之讨论的话只会使手头的问题更混乱，这也就是对计算生物学而言多么需要理解它的算法基础的原因。此外，凭经验来讲，我们知道没有一位学生愿意携带一本十公斤重的参考书，而其中的问题只要一本一百克重的书就能解决。

尽管有读者抱怨，但来自学生和导师们的反馈意见对每个人来说绝对是积极而有益的。许多技术和语法上的错误在第二次印刷时被学生（其中一些学生并不就读任何大学，而仅仅是出于对生物信息学知识的执著的求知精神）及时地指出并纠正，并由此发现了其他一些错误。很高兴的是，发现错误的频率已明显地降低。本书及在线幻灯片中伪代码的含糊之处或混乱的来源引起了我们的注意，并在网站的在线论坛上作了必要的阐述。遗憾的是，我们没有把在线论坛（http://www.bioalgorithms.info）上的内容或幻灯片翻译成英语之外的其他语言，也没有提供这个演讲材料的国际化版本，但我们希望我们提供的资源在中国仍能成为教师的一个很好

的起点——当然我们愿意与任何具备必要的语言技巧的人公开合作，将这些内容翻译成更多语种使之广泛传播。

除了本书的有关工作之外，我们觉得承认中国科学界的努力是非常重要的。尽管十年前在国际刊物上中国人发表的生物信息学论文数量很少，但是这个数量现在增长很快，而且中国正在这一领域做出可观的贡献。像中国生物信息（biosino.org.cn）等机构推动了高通量数据（一般仅可利用英语）向更多研究团体的传播，且不仅仅局限于中国的科学界。北京（北京大学：华大基因研究中心）和上海的基因组中心为最近发表的许多突破性工作成果（包括基于本体论的数据整合、为蛋白质从头测序的肽作图等）提供了必要的环境和资源，也为水稻基因组计划做出了重要贡献。经费资助也正在得到改善：中国科学院为生物信息学研究拨了一笔等同于好几百万美元的经费，这些经费大都用来研究系统生物学、生物序列的功能分析以及药物发现。但是，古有谚语：先学走，后学跑。学习生物信息学算法也是这个道理，我们必须先懂得动态规划和强力搜索，然后才能编写算法去发现新的药靶。

因此，我们希望本教材中的知识能够帮助学生充分熟悉算法在计算生物学中的应用，使之可阅读——且创造出——具有高度科学价值的经得起同行评审的研究成果。希望大家高兴地阅读，并请不吝指出错误。

最后，我们非常感谢上海大学的王翼飞教授等人将本书的英文原版翔实地翻译为中文，使得更多的读者可以看到这本书，这些都不是我们力所能及的。

**N. C. 琼斯　P. A. 帕夫纳**
**(Neil C. Jones, Pavel A. Pevzner)**
**2007 年 1 月 10 日于 La Jolla**

# 前　言

20世纪90年代初，当我们中的一位第一次讲生物信息学课时，他还不能确定会有多少学生来听。虽然那时Smith-Waterman算法和BLAST算法已经被开发出来了，但在生物学家中还不像现在这样广为流传，甚至“bioinformatics（生物信息学）”一词还没有被创造出来。那时，大多数人认为DNA阵列只不过是尚未确定实际应用的智力玩具，只有少数狂热分子认为这是一项拥有巨大潜力的技术。为数不多的生物信息学专家开始为尚不存在的数据集发展新的算法思想：David Sankoff设立了一项基金，专门鼓励在缺乏基因次序资料时所作的基因组重排研究；Michael Waterman和Gary Stormo在几乎没有启动子样本可用时发展了基序（或称为模体，motif）查找算法；在还没有装配好细菌基因组时，Gene Myers就研发了一个精致的片段装配工具；当长为172 282个核苷酸的Epstein-Barr病毒还是GenBank中最长的序列时，Webb Miller已在梦想比对十亿核苷酸长度的DNA序列了。GenBank也恰好刚刚完成了自身的转变，从一系列（纸张!）装订的册子转变为记录在磁带上的电子数据库，从而可以发送给世界各地的科学家们了。

我们必须回到20世纪80年代中期和90年代早期去感谢在这十年中生物学所发生的变革。然而，生物信息学比生物学的影响更大——它对计算科学也产生了深远的影响。生物学已快速成为新算法和统计问题的一大数据源，而且应用在生物学上的算法之多也超越了其他任何一种基础科学。将计算机科学和生物学联系起来具有重要的教育意义，它不仅改变了我们教生物学家计算思想的方法，也教给了计算机科学家们懂得如何应用算法。

过去，只有计算机科学家们才学习计算机科学这门课程，而且仅局限于来自其他学科的少数学生。在20世纪90年代早期，几乎不可能（虽然欢迎）看到生物学专业学生去上算法课。但是现在情况改变了，许多生物学专业的学生都了解了一些浅显的算法知识。同时，好奇的计算机科学专业的学生也会学习一些遗传学和生物信息学的基础知识。虽然这样做的学生相对来说还是少数，但要知道在20世纪90年代初期几乎还没有生物信

息学课程，而现在看来数量已经不少了。我们预见，大学本科生的生物信息学课程将成为所有一流大学的固定课程。这是必然的趋势，而不是空想。

这是一本关于生物信息学算法及其计算思想的导论性教科书，近20年来，正是这些计算思想推动着生物信息学算法的发展。还有许多重要的概率和统计技术以及当前生物信息学专家们正在研究的重要课题尚未涉及到。我们故意不包括计算生物学的所有领域，例如，像蛋白质折叠这样重要的课题甚至也没讨论。最早的生物信息学教科书是由 Waterman 在1995年写的[108]，此书极好地介绍了 DNA 统计；而 Gusfield 在1997年写的书[44]可以说是字符串算法的大全。Durbin 等在1998年写的书[31]和 Baldi 与 Brunak 在1997年写的书[7]则着重于隐马氏模型和机器学习技术；Baxevanis 和 Ouellette 在1998年写的书[10]是一本优秀的生物信息学实用指南；Mount 在2001年写的书[76]很好地描述了生物学问题与生物信息学技术之间的联系；Bourne 和 Weissig 在2002年写的书[15]则将重点放在蛋白质生物信息学方面。网络上也有许多生物信息学课程笔记，我们从互联网（WWW）上由 Serafim Batzoglou、Dick Karp、Ron Shamir、Martin Tompa 等人提供的材料学到了许多有关生物信息学的教学法。

网站

与本书相伴，我们建立了网站 http://www.bioalgorithms.info。此网站为本书提供了大量的补充材料。例如，虽然书中没有词汇表，但是提供了一个索引，并在上述网址中有详细介绍。懂得技术的学生也可以下载实际的生物信息学练习题、书中算法的例子以及测试这些算法的数据。老师和学生们都可以从网站上得到预制的讲义。我们希望这个网站可作为有助于引导学生们进入生物信息学多变世界的信息宝库。

致谢

感谢为我们提供传记的专家，他们的深刻见解和真诚回应为本书增色不少。毋庸置疑，这些生动的故事和挑战性的观点将大大激发学生们探索未知世界的热忱。本来我们还想介绍更多的专家，但限于篇幅只得放弃。特别要感谢 Ethan Bier，是他激发了我们在本书中插入人物天地（专家小传）的灵感。

没有助教们在2003年、2004年冬季和秋季学期生物信息学课程授课期间的认真的教学辅导，我们也不可能完成这本书。他们是 Derren Barken，Bryant Forsgren，Eugene Ke，Coleman Mosley 和 Degui Zhi。感

谢他们指出本书中的技术错误，帮助精选实践练习题及设计每章后所附的问题。Helen Wu 和 John Allison 花了很多时间来制作技术性示意图，这是一项费力不讨好的工作。另外，我们还要感谢 Vagisha Sharma 通读了本书，提出了富有洞察力的建议并指出了伪码中的错误和不恰当之处。Steve Wasserman 以生物学家的眼光给我们提出了非常宝贵的意见，使得我们最终在本书中又增加了新的章节。Alkes Price 和 Haixu Tang 指出了书中概念模糊的地方，并帮助修正了有关图和聚类章节的内容。Ben Raphael 和 Patricia Jones 对前面几章反馈了意见，并帮助避免了一些潜在的错误理解。Dan Gilbert Art Group，Inc 的 Dan Gilbert 用 Triazzles 热忱地帮助我们解释了 DNA 序列装配问题。

我们还要特别感谢 www. kleemanandmike. com 网站的美术编辑 Randall Christopher，他帮助我们画了书中所有的插图，并且为一些生物信息学算法设计了许多独一无二的图形表示法。

与 MIT 出版公司的 Robert Prior 一起工作是非常愉快的，他怀着充分的耐心并通过有效的督促，设法使我们保持正常的写作过程。我们也很感谢 G. W. Helfrich 严谨的审稿。

最后，我们还要感谢 UCSD 大学中曾经听过生物信息学课程的本科生和研究生们，他们对本书的早期版本提出了很多有益的意见。

P. A. 帕夫纳诚挚地感谢曾经教授过他计算分子生物学各方面知识的人。Andrey Mironov 教给他的常识也许是任何应用研究中最重要的部分。当 P. A. 帕夫纳刚从莫斯科来到洛杉矶的那段时间，Mike Waterman 不管是在科研上还是在生活上都是他最好的老师。P. A. 帕夫纳还要感谢 Alexander Karzanov 教给他组合最优化，这也是他在计算生物学研究中最有用的技能。他还要特别感谢 Mark Borodovsky 在 1985 年鼓励他转到生物信息学的研究上来，尽管那时这门学科的前景还不明朗。

P. A. 帕夫纳也感谢他以前的学生、博士后和实验室的其他成员：Vineet Bafna，Guillaume Bourque，Sridhar Hannenhalli，Steffen Heber，Earl Hubbell，Uri Keich，Zufar Mulyukov，Alkes Price，Ben Raphael，Sing-Hoi Sze，Haixu Tang 和 Glenn Tesler。他们教给了他很多知识。

N. C. 琼斯真诚地感谢他大学时代的指导老师 Toshihiko Takeuchi、Harry Gray、John Baldeschwieler 和 Schubert Soares，他们对他耐心而严格的教导，使他懂得坚持是科学研究中最重要的品格。同时，他也要感谢加利

福尼亚大学圣地亚哥分校的招生委员会冒险同意他从化学家转行为计算机程序设计师，但愿这是最佳的决定。

N. C. 琼斯　P. A. 帕夫纳
2004 年于美国加利福尼亚州 La Jolla

# 目　录

# 1 绪 论

想像两个人物 Alice 和 Bob，还有两堆石头，每堆有十块。在一个无聊的星期六下午，他们开始玩一个游戏，游戏规则为：每个人每一轮允许从一堆石头中拿走一块或是从两堆石头中各拿走一块，石头拿走了就不许再放回；谁拿走了最后一块石头谁就赢得了比赛。Alice 先拿。

要想很快找出获胜策略并不容易，即使真的有这样的策略。先拿的人（或者后拿的人）总是有优势吗？Bob 试着分析这个游戏，他发现两堆石头、每堆各十块这个游戏（我们简称为 10＋10 游戏）中包含了太多的变量。利用还原论方法，他先试着找出简化的 2＋2 游戏的获胜策略。很快地，他发现后拿者——他自己能赢得各种情况的 2＋2 游戏，于是，他决定写下“获胜秘诀”：

如果 Alice 从每一堆中拿走一块石头，那么我将拿走剩下的石头并且赢得比赛。如果 Alice 拿走一块石头，那么我就从同一堆中拿走另一块石头。结果就只剩下一堆，而且有两块石头。那么 Alice 将只有一种选择，就是从中拿走一块石头。而我将拿走剩下的一块石头并且赢得游戏。

受这一分析的启发，Bob 的脑海中闪现了一种揣想：后拿者（即他自己）会在任何 $n \geqslant 2$ 的 $n+n$ 游戏中赢得胜利。当然，任何假设都必须由实验来证实，所以 Bob 与 Alice 玩了好几局。结果却是有时 Bob 赢，有时 Alice 赢。Bob 试图写出赢得 3＋3 游戏的策略，但这需要考虑很多种情况，使得取胜的策略很复杂。看起来要想写出赢得 10＋10 游戏的秘诀是不可能的了，因为 Alice 如何拿石头的情况实在太多了。

同时，Alice 也很快发现她总是会输掉 2＋2 游戏，但她并没有放弃寻找 3＋3 游戏获胜策略的希望。而且，她考虑到最简单的算法（Algorithms 101），明白 Bob 采用的编写秘诀的方式对她没有什么帮助：秘诀类型的说明性语言是不足以描述算法的。取而代之，她用符号↑、←、↖和＊画了一张表，如下所示。位置 $(i,j)$（即第 $i$ 行第 $j$ 列）上的符号表示 Alice 在 $i+j$ 游戏（A 和 B 堆中分别有 $i$ 块和 $j$ 块石头）中的拿法。A←表示她应该从 B 堆中拿走一块石头，A↑表示她应该从 A 堆中拿走一块石头，A↖表示她应该从两堆中各拿走一块石头，＊表示她不用再玩这个游戏了，因为她无论如何都不会赢的。

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | * | ← | * | ← | * | ← | * | ← | * | ← | * |
| 1 | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ |
| 2 | * | ← | * | ← | * | ← | * | ← | * | ← | * |
| 3 | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ |
| 4 | * | ← | * | ← | * | ← | * | ← | * | ← | * |
| 5 | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ |
| 6 | * | ← | * | ← | * | ← | * | ← | * | ← | * |
| 7 | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ |
| 8 | * | ← | * | ← | * | ← | * | ← | * | ← | * |
| 9 | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ | ↖ | ↑ |
| 10 | * | ← | * | ← | * | ← | * | ← | * | ← | * |

例如，对于3+3游戏，她发现第三行第三列是符号↖，这表明她应该从两堆中各拿一块石头，然后游戏就变成了2+2，并且是Bob先拿，由于标记了一个*，所以无论Bob怎么拿，Alice都会赢的。假设Bob从B堆中拿走一块石头，这样就变成了2+1游戏了，Alice参考位置（2，1）上的符号，决定从B堆中拿走一块石头，而剩下A堆中的两块石头；然而，如果Bob从A堆中拿走一块石头，就变成了1+2游戏，位置（1，2）的符号是↑，她又应该从A堆中拿走一块石头，剩下B堆中的两块。

根据上面的表，Bob可以学会如何利用它赢得10+10游戏。但是，Bob不知道如何给20+20游戏构建一个相似的表。问题并不是因为Bob笨，而是因为他没有学过算法。即使Bob凭运气赢得了20+20游戏，他也不敢自信地说他总是能赢Alice，当然他也写不出一般$n+n$游戏的获胜秘诀。更令Bob为难的是有三堆石头的10+10+10游戏，这对他来说简直是不可能做出的难题。

为了把自己从困境中拯救出来，有两件事是Bob可以做的。第一，他可以去上算法课，学习如何解决类似于拿石头那样的难题；第二，他可以记住Alice给他的适当大的表，然后利用它来玩这个游戏。现在我要问的是，作为一个生物学家你会怎么做？

当然，我们希望听到大多数理性的人回答说“我为什么要在乎两个假想的人玩的一个拿石头的游戏？我只是对生物学感兴趣，这个游戏对我来说没有意义。”事实上这是不对的：拿石头的游戏隐藏着的正是常见的序列联配问题。虽然一时很难看出DNA序列联配与拿石头游戏有什么关系，但是解决这两个问题的计算思想是一致的。Bob不知道这个游戏的策略表明他也不明白序列联配算法的工作原理。他也许不同意用联配算法或**BLAST**[1]做日常的基础工作，

[1] **BLAST**是一个数据库搜索工具——一个生物学序列的Google——本书后面会对其进行介绍。