“十一五”国家重点图书出版规划项目

生命科学实验指南系列

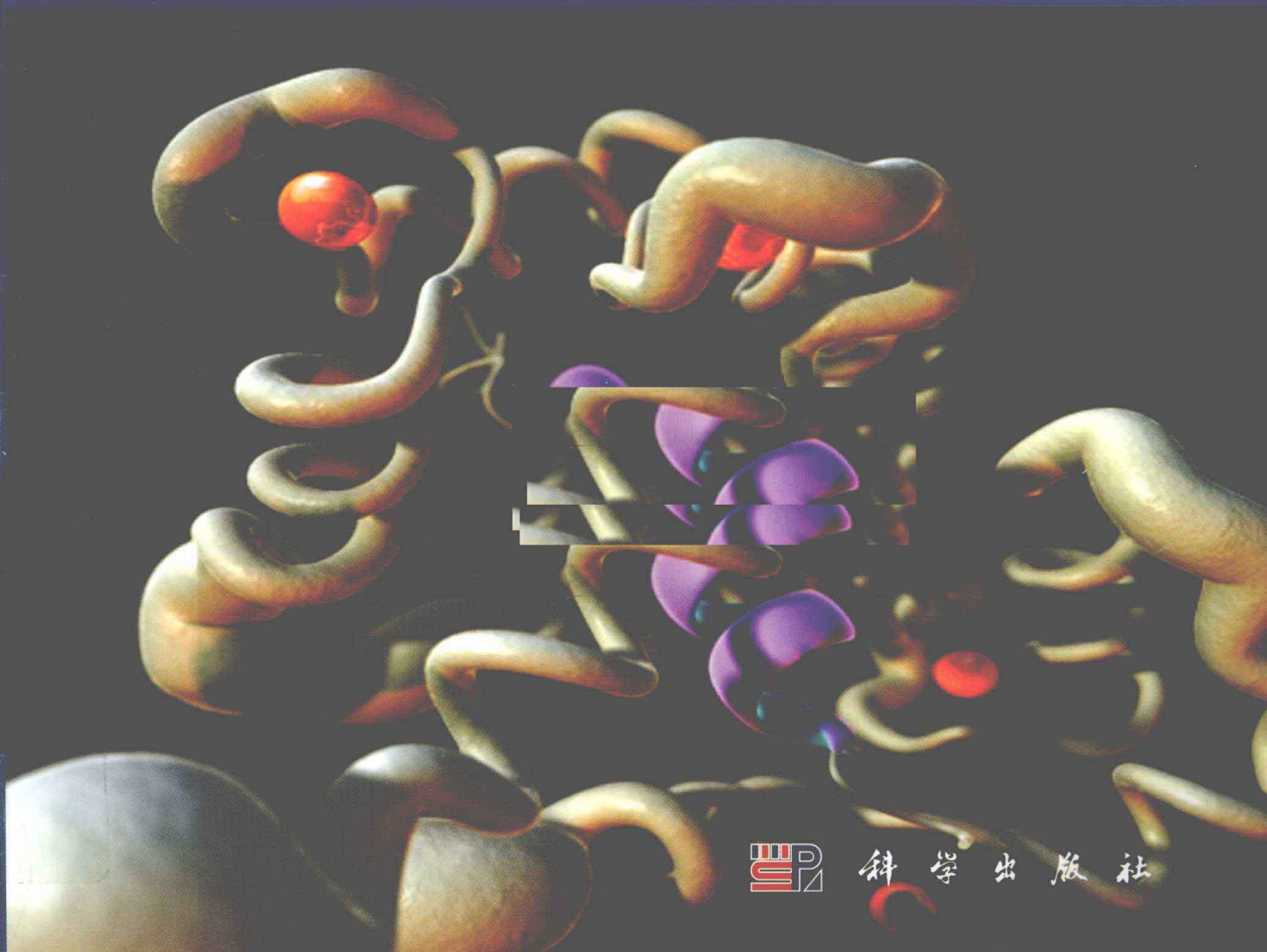
Protein Engineering Protocols

现代蛋白质工程实验指南

〔德〕K.M.阿恩特　K.M.米勒　编著

苏晓东　曾宗浩　杨　娜　译

苏晓东　高　嵘　校



科学出版社

图字:01-2008-3816

内 容 简 介

现代蛋白质工程是将分子生物学、蛋白质结构与功能分析、理论计算,以及生物化学有机结合的学科,其目标是快速及高效率地发展和改进实用的或有价值的蛋白质。本书分为两个部分,第一部分主要介绍了蛋白质理性设计的策略,包括理论计算方法,利用一些很有说服力的例证说明所设计蛋白质的全新特性,阐述了如何设计具有目标特性的蛋白质,并选择了很多如蛋白质-蛋白质相互作用、DNA 结合、抗体模拟,以及酶活性设计等具体实例;第二部分蛋白质的定向进化技术主要介绍了包括进化库设计的一般方法、进化库质量的统计评估、核酸混编的新方法,以及不同的选择筛选策略等。同时也给出了不同特性体外定向进化的一些实例,如蛋白质折叠类型、折叠热稳定性以及酶活性等。

本书适合蛋白质科学各个层次的科研工作者,特别是从事相关领域研究的高年级大学生及研究生参考使用。

图书在版编目(CIP)数据

现代蛋白质工程实验指南/(德)阿恩特(Arndt,K. M.)等编著;苏晓东等译. —北京:科学出版社,2011

(生命科学实验指南系列)

Protein Engineering Protocols

ISBN 978-7-03-030122-2

Ⅰ. ①蛋… Ⅱ. ①阿…②苏…③曾… Ⅲ. ①蛋白质-生物工程-实验-指南

Ⅳ. ①TQ93-33

中国版本图书馆 CIP 数据核字(2011)第 014243 号

责任编辑:罗 静 孙 青/责任校对:张 林

责任印制:钱玉芬/封面设计:耕者设计工作室

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

铭浩彩色印装有限公司 印刷

科学出版社发行 各地新华书店经销

*

2011 年 3 月第 一 版 开本:787×1092 1/16

2011 年 3 月第一次印刷 印张:15 1/2

印数:1—3 000 字数:342 000

定价:58.00 元

如有印装质量问题,我社负责调换

译 者 序

自20世纪六七十年代分子生物学及重组蛋白质技术建立及普遍应用以来，生物技术已经取得了非常广泛且令人瞩目的成果，“蛋白质工程”也似乎渐渐变成了一个古老的话题。然而，近二十年来，蛋白质工程的相关技术取得了长足的进展，特别是在蛋白质的定向进化(directed evolution)及理性设计(rational design)等方面。一年多以前，我们有幸得到英国剑桥大学 John M. Walker 教授(他因阐明了 ATP 合成酶的结构与机理而荣获 1997 年诺贝尔化学奖)主编的《分子生物学方法》*Methods in Molecular Biology* 系列丛书之一，由德国弗莱堡大学的 Katja M. Arndt 和 Kristian M. Müller 编著的 *Protein Engineering Protocols* 一书，此书由欧美及日本的多位蛋白质化学及分子生物学相关领域国际知名专家撰写，从理论到实践较为详细地介绍了现代蛋白质工程这两个最新前沿及其相关进展实例。

所谓定向进化，是指对蛋白质(通常是通过其编码核酸的随机合成及重组)引入大量的随机突变体，然后快速高效地选择或筛选具有相应特性的突变体蛋白的过程，定向进化可以通过体外或者体内实验进行，本书主要介绍的是体外定向进化方法。人们通过多轮的突变和筛选取得特定性质的突变体蛋白，可以说定向进化方法是在模拟自然界的进化及自然选择过程，只不过我们的选择压力是人为在实验室中施加的。另外，我们还可以模拟自然界有性生殖中的重组过程，通过各种 DNA 混编(DNA shuffling)技术将筛选得到的突变体进一步混合匹配以得到更优化的结果。一般说来，定向进化不需要预先知道蛋白质的详细三维结构信息，尽管这些信息有时对于设计定向进化实验很重要，也不需要预测一个突变体对蛋白质会产生怎样的功能及稳定性影响。事实上，很多定向进化的实验结果表明目标蛋白质特性的产生往往是由意想不到的突变造成的，通过定向进化实验我们能够学习到很多关于蛋白质结构与功能的新知识。

尽管目前蛋白质的折叠问题仍然没有取得根本性突破，我们还不能够一般地预测任意序列的蛋白质折叠类型及其三维结构，然而，随着结构基因组学及相关计算生物学的发展，蛋白质三维结构研究与预测已经取得了重大进展，截至 2010 年底，PDB 数据库(http://www.rcsb.org/pdb/home/home.do)已经收录了七万多个蛋白质三维结构坐标，并且统计数据表明从 2008 年以来，PDB 库中就没有收录到新的蛋白质折叠类型，这意味着当前 PDB 数据库已经覆盖了绝大部分已知蛋白质的折叠类型。在此基础上，根据目前已知蛋白质的三维结构与功能细节合理化设计新的蛋白质已经成为可能，本书介绍了这一方面的理论基础及应用实例。

本书在理论上从蛋白质工程中的设计计算策略和进化策略出发，讨论了采用计算方法的组合式蛋白质设计策略及蛋白质进化库的设计与筛选，并且介绍了一个行之有效的基于极性与非极性氨基酸的二元组图进行蛋白质设计的方法细节。在实践上本书详细介绍了蛋白质中非天然氨基酸的整体掺入方法；作为理性设计实例列举了基于钙调素与荧光蛋白融合的钙指示剂；人工锌指蛋白的设计与合成，以及基于纤连蛋白类型Ⅲ结构

域框架的抗体模拟等。在定向进化方法的应用方面，介绍了噬菌体展示技术、核糖体失活展示系统、利用核苷酸交换和剪切技术进行 DNA 碎裂和定向进化、简并寡核苷酸基因混编，以及其他酶的定向进化新方法。我们很高兴受科学出版社的委托将此书翻译成中文，介绍给我国广大蛋白质科研工作者，希望本书的中译本对于推动我国现代蛋白质工程学科的进一步发展做出贡献。

我们衷心感谢参与本书翻译及校正工作的很多同仁，他们包括中国科学院生物物理研究所的王大成院士，他和科学出版社编辑一起向我们推荐并介绍了此书。还要感谢北京大学生命科学学院的李兰芬老师，王娟、刘晓艳、雷剑博士，以及傅天民、王子曦及金坚石等同学。

译者

2011 年 2 月

前　言

蛋白质工程是将分子生物学、蛋白质结构分析、理论计算以及生物化学有机结合的学科，其目标是发展实用的或有价值的蛋白质。本书涉及蛋白质工程领域中两个普遍的但不互相排斥的策略。第一个策略是理性设计(rational design)，科学家根据蛋白质的结构与功能细节设计相应的蛋白质；第二个策略是所谓定向进化(directed evolution)，通过对蛋白质引入随机突变，然后选择或筛选具有相应特性的突变体。通过多轮突变和筛选的方法，可以说是模拟自然界的进化过程。另外，通过 DNA 混编(DNA shuffling)技术将筛选得到的突变体进一步混合匹配以得到更优的结果，从而模拟自然界有性生殖中的重组过程。

本书的第一部分介绍了蛋白质理性设计策略，包括理论计算方法，通过引入非天然氨基酸来扩展生物学字母表，以及一些很有说服力的例证说明所设计蛋白质的全新特性。尽管引入突变的策略已经成为常规，但预测和推断这些突变产生的效果仍是非常具有挑战性的，除了基本的蛋白质结构信息外，还需要对整个系统有深刻的理解。因此，这一部分主要讲述如何设计具有目标特性的蛋白质，并挑选了涵盖蛋白质工程大部分技术的实例，如蛋白质-蛋白质相互作用、DNA 结合、抗体模拟，以及酶活性设计等。

本书的第二部分主要讲述了进化技术。与理性设计不同的是，定向进化策略不需要预先知道蛋白质的三维结构信息，也不需要预测一个突变对蛋白质会产生怎样的影响。事实上，定向进化的实验结果表明目标特性的产生往往是由意想不到的突变造成的。定向进化策略成败的关键因素有：进化库的设计与质量、进化及 DNA 重组方法的选择，以及筛选方法的选择。因此，第二部分主要介绍了上述几个方面的内容，包括进化库设计的一般方法、进化库质量的统计评估、DNA 混编的新方法，以及不同的选择筛选策略等。同时也给出了不同特性进化的一些实例，如蛋白质折叠类型、热稳定性，以及酶活性等。

本书全面地介绍了蛋白质工程各个阶段所使用的方法，综合了完备的理论基础和具体的实验细节，适合该领域各个层次的科研工作者使用。感谢参与此书编著的所有人员作出的卓越贡献，特别是丛书系列主编 John M. Walker 教授(1996 年诺贝尔化学奖得主)在本书编撰过程中给予的耐心指导和帮助。

K. M. 阿恩特

K. M. 米勒

原作者及其单位

KATJA M. ARNDT • *Institut für Biologie III, Universität Freiburg, Freiburg, Germany*

JAMIE M. BACHER • *The Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, CA*

PETER L. BERGQUIST • *Biotechnology Research Institute, Macquarie University, Sydney, NSW, Australia; Department of Molecular Medicine and Pathology, Auckland University Medical School, Auckland, New Zealand*

LUKE H. BRADLEY • *Department of Chemistry, Princeton University, Princeton, NJ*

FRANÇOIS-XAVIER CAMPBELL-VALOIS • *Département de Biochimie, Université de Montréal, Montréal, Québec, Canada*

MICHEL DENAULT • *Department of Quantitative Methods, HEC Montréal, Montréal, Québec, Canada*

ANDREW D. ELLINGTON • *Institute for Cellular and Molecular Biology and Department of Biochemistry, University of Texas, Austin, TX*

BIRTE K. FELD • *Department of Chemistry and the Institute for Genomics and Bioinformatics, University of California, Irvine, CA*

PETER FRIEDHOFF • *Institut für Biochemie, Justus-Liebig-Universität, Giessen, Germany*

SATOSHI FUJITA • *Department of Chemistry and Biotechnology, School of Engineering, The University of Tokyo, Hongo, Tokyo, Japan; Research Institute for Cell Engineering, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan*

FARID J. GHADESSY • *MRC Laboratory of Molecular Biology, Cambridge, United Kingdom; Department of Oncology, University College Medical School, London, United Kingdom*

MORELAND D. GIBBS • *Biotechnology Research Institute, Macquarie University, Sydney, NSW, Australia*

MICHAEL H. HECHT • *Department of Chemistry, Princeton University, Princeton, NJ*

JOCHEN HECKY • *Institut für Biology III, Universität Freiburg, Freiburg, Germany*

PHILIPP HOLLIGER • *MRC Laboratory of Molecular Biology, Cambridge, United Kingdom*

MITSUHIKO IKURA • *Division of Molecular and Structural Biology, Ontario Cancer Institute and Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada*

AKIKO KOIDE • *Department of Biochemistry and Molecular Biology, The University of Chicago, Chicago, IL*

Shohei Koide · *Department of Biochemistry and Molecular Biology, The University of Chicago, Chicago, IL*

Hidetoshi Kono · *Computational Biology Group, Neutron Science Research Center, Quantum Beam Science Directorate, Japan Atomic Energy Agency, Kyoto, Japan*

Jody M. Mason · *Institut für Biologie III, Universität Freiburg, Freiburg, Germany*

Stephen W. Michnick · *Département de Biochimie, Université de Montréal, Montréal, Québec, Canada*

Kristian M. Müller · *Institut für Biologie III, Universität Freiburg, Freiburg, Germany*

Atsushi Miyawaki · *Laboratory for Cell Function and Dynamics, Advanced Technology Development Center, Brain Science Institute, RIKEN, Wako City, Saitama, Japan*

Wataru Nomura · *Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan*

Joelle N. Pelletier · *Département de Chimie, Université de Montréal, Montréal, Québec, Canada*

Alfred Pingoud · *Institut für Biochemie, Justus-Liebig-Universität, Giessen, Germany*

Jeffery G. Saven · *Makineni Theoretical Laboratories, Department of Chemistry, University of Pennsylvania, Philadelphia, PA*

Asako Sawano · *Laboratory for Cell Function and Dynamics, Advanced Technology Development Center, Brain Science Institute, RIKEN, Wako City, Saitama, Japan; Brain Science Research Division, Brain Science and Life Technology Research Foundation, Itabashi, Tokyo, Japan*

Sachdev S. Sidhu · *Department of Protein Engineering, Genentech Inc., South San Francisco, CA*

Sabine C. Stebel · *Institut für Biologie III, Universität Freiburg, Freiburg, Germany*

Yukio Sugiura · *Faculty of Pharmaceutical Sciences, Doshisha Women's University, Koudo, Kyotanabe, Japan*

Kazunari Taira · *Department of Chemistry and Biotechnology, School of Engineering, The University of Tokyo, Hongo, Tokyo, Japan; Gene Function Research Laboratory, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba Science City, Japan*

Peter Thumfort · *Department of Chemistry, Princeton University, Princeton, NJ*

Kevin Truong · *Institute of Biomaterials and Biomedical Engineering, Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada*

Wei Wang · *Makineni Theoretical Laboratories, Department of Chemistry, University of Pennsylvania, Philadelphia, PA*

Yinan Wei · *Department of Chemistry, Princeton University, Princeton, NJ*

Gregory A. Weiss · *Department of Chemistry and the Institute for Genomics and Bioinformatics, University of California, Irvine, CA*

CHRISTINE WURTH • *Department of Chemistry, Princeton University, Princeton, NJ*

JING-MIN ZHOU • *Department of Chemistry and Biotechnology, School of Engineering, The University of Tokyo, Hongo, Tokyo, Japan; Gene Function Research Laboratory, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba Science City, Japan*

目　　录

第一部分　蛋白质工程中的设计与计算策略

1 采用计算方法的组合式蛋白质设计策略

Hidetoshi Kono，Wei Wang，and Jeffery G. Saven

概要 计算方法一直在蛋白质设计中发挥重要作用。本工作主要集中在搜索蛋白质序列空间，以找到一条或数条与已知结构和功能相容的蛋白质序列。在期望的功能和结构限制下，概率性计算方法为所容许的氨基酸变化范围提供信息。这样的方法可用于指导建立蛋白质的单个序列或组合库。

关键词 全新蛋白质设计；组合库；蛋白质计算设计；偏好编码

1.1 介 绍

1.1 蛋白质设计

通过对蛋白质结构（包括那些具有特殊功能结构）的设计，研究者可以增进对决定蛋白质折叠状态特征的力和效应的理解。另外，对特定折叠结构设计的控制，可能得到新的具有生物效能和特异性的合成蛋白。这样的应用包括在医药、传感器、催化剂和材料等领域。甚至在尚未完全和定量地理解决定蛋白质结构的力的条件下，蛋白质设计仍有可能取得成功。

但是，由于决定蛋白质折叠态的相互作用的复杂性和精细性，蛋白质设计不是一件轻而易举的事。蛋白质是巨大的分子（含数十到数百个氨基酸残基），折叠态的给定需要许多的结构变量，其中包括序列、主链拓扑和侧链构象。即使主链结构已经给定，每个残基仍可能有多重构象。除了结构的复杂性之外，还有序列的复杂性。设计意味着从无数的可能的序列中，鉴别出可折叠的序列。在折叠蛋白中观察到的高度“一致性”引导这个搜索过程[1]。一般来说，处于折叠状态的蛋白质，在原子水平上，通过有利的范德华相互作用、疏水残基与溶剂隔离，并使大多数氢键作用都被满足而恰当地堆积。但这种一致性通常都是复杂的，可能没有什么能使问题简化的对称性。另外，精确地定量非共价相互作用属于最困难的一类问题，并且，估计残基替换或结构有序化的自由能，仍然是计算研究中最深奥的领域[2,3]。与预测能力相反，目前，我们还不能期望用详细的模拟估计自由能变化的方法，来确定大量序列的相对稳定性变化。尽管如此，从小分子和蛋白质数据库导出的分子位能，确实包含了已知的，对决定蛋白质结构起重要作用的，相互作用和力的部分信息。在某些情况下，这些位能的优化，已经在蛋白质设计中获得显著的成功[4]。这样的位能肯定是近似的，并且这样设计的任何序列，很可能对特定的位能和采用的目标结构敏感。作为另一种选择，在这些位能中包含的部分信息也可以做概率性的应用，以得到出现某种氨基酸的可能性。概率性方法也适合于确定可折叠到同一结构序列的完整可变性，因为似乎存在大量这样的序列——远

大于可以用序列搜索或列举方式所能处理的。

这样的概率性方法也特别适合于在蛋白质组合实验中的全新设计，这些实验能产生并快速测试许多序列。虽然组合方法能处理大量序列（$10^4 \sim 10^{12}$），但这些数量与可能的蛋白质序列数比较仍然是无穷小，如对 100 个残基的蛋白质，这个数是 $20^{100} \approx 10^{130}$（为了对 10^{130}这个数有多大有个一般的了解，假定合成出一条 100 个残基的蛋白质序列约需要10 000/N g 物质，$N \approx 6 \times 10^{23}$，是阿伏伽德罗常数，那么合成出 10^{130}条序列需要超过10^{107}kg的物质。这个数量远超过目前我们所知宇宙物质的总质量 10^{35} kg——译者注）。于是，即使是采用组合法，我们仍然必须集中于序列空间选定的一小部分。通过预先观察，在蛋白质中选定若干残基位点并在这些选定的位点允许残基的全部或部分可变性（全部可变性，即可以替换为 20 种氨基酸中任何残基；部分可变性，只替换为某些类型的残基——译者注）来实现对序列空间的限定。近来发展了可以在宽得多的范围内追踪序列可变性，并提供扫视和聚焦序列空间的定量计算方法。在这里，我们讨论序列设计的计算方法，把重点放在处理给定结构位点特异氨基酸可变性的概率方法。

1.1.2　蛋白质设计的定向方法

这里的“定向蛋白质设计”是指鉴定出一条（或一组）可能折叠为预先指定的主链结构的序列。然后可以用多肽合成或基因表达的办法，实验性地实现每一条这样的序列，以确认其折叠态及其他分子性质。早期的设计努力，是在已观察到的自然发生的结构和已被确认的蛋白质序列的指导下完成的，它们有着重要的二级结构，但并不必有明确的三级结构[5]。由于能够定量化并以表格列出残基间相互作用，计算方法已经极大地加速了蛋白质设计的成功率。典型情况下，这样的方法使序列搜索成为优化过程，在过程中改变氨基酸身份和侧链构象，以优化定量化序列结构相容性的打分函数，对所有 m^N个可能序列的完全搜索，只有仅仅少量的残基 N 是允许改变的，或允许改变的氨基酸数目显著地减少，如从 $m=20$ 减少到 $m=2$ 时才是可行的。为了达到从内部平均来看有利的原子间相互作用的合理堆积序列，必须搜索每个氨基酸的不同侧链构象（旋转异构态）（见参考文献 [6]）。结果，因为一个残基所可能有的状态数 m 增加 10 倍或更多——这取决于每个残基的旋转异构态数目和增加搜索的复杂性。如果只有很少的残基被允许改变，并且剩下的残基的构象受到限制，就可以对所有可能的组合完成完整数值计算，以鉴别出低能量的序列旋转异构态组合。由于（这个组合数）对链长和旋转异构态数目的指数依赖性，这样的完整数值计算在典型的情况下不能实现。在这样的情况下，序列空间可以用定向的方式取样以逐渐朝优化的（或部分优化的）序列方向移动。随机方法，如遗传算法和模拟退火，包含对序列空间的部分随机式搜索。在这种搜索中，搜索逐渐地移向高分（低能）序列[7~10]。这样的搜索有足够的“噪声”或重组，以允许越过序列-旋转异构地形图的局部极小值。当运用于精细到原子的表象中时，随机方法基本上集中于用疏水残基重新堆积结构的内部[9]，并已被用于 434 Cro[10]、泛素[11]、G 蛋白的 B1 结构域[12]、WW 结构域[4]和螺旋束[13,14]的野生型结构。虽然在许多情况下这些方法对鉴别出实验可行的序列[4,15]有帮助，但是随机搜索法不必鉴别出整体优化点[16]。对只含有位点和对相互作用的位能，通过排除法，如“死点排除法”，能找到整体优化点[16~20]。这样的方法可连续地移除不可能是整体优化点的氨基

酸-旋转异构态，直到再没有态可被移除。Mayo 研究组应用这一方法，已使拟 28 残基锌指蛋白[21]和在疏水和极性位点模式化之后的 51 残基同源域蛋白模体[22]的完整序列设计自动化。该组还重新设计了数种蛋白质内的部分残基组[23~25]。其功能性质，如结合金属或催化，也可以包括作为设计过程的元素[26~28]。蛋白质定向设计的要素和算法是最近一些综述的主体[4,29,30]。

尽管取得了某些惊人的成功，但序列定向设计的计算方法在鉴别折叠为特定结构的蛋白质序列特征上仍有局限性。随机方法可用于大蛋白，也容许多位点上的同时变化，但是，即使用于小蛋白，这样的计算消耗的机时和资源也是巨大的。定向方法对于使用的能量或打分函数必定是敏感的，因为它能鉴定能量函数的优化点。但是，所有这样的函数也必定是近似的，并且能量函数中的不确定因素可能不允许对整体优化点的搜索。许多天然存在的蛋白质是没有优化的。实际上，多数蛋白质只有微小的折叠稳定性，如 $\Delta G^{\circ}<10$kcal①/mol [31]。更有甚者，在功能上与其他分子结合的序列在结构稳定性上不必是整体优化的。重要的是发展与定向蛋白设计互补的方法，这些方法揭示可能折叠为某特定结构但又可能在结构上有未被优化的序列的特征。这样的技术可用于设计蛋白质序列。另外，这样的计算方法还可用于新型的蛋白质设计研究——组合式实验，即实验中大量蛋白质可以同时合成和筛选。

1.1.3 蛋白质设计的概率性方法

在蛋白质涉及的范围内，我们用定点氨基酸概率而不是特定的序列来描述“概率性蛋白质设计”。相对于定向的或决定论方法，概率性方法是常用于对问题只有部分信息场合的定量科学。对蛋白质设计，折叠过程的复杂性和不确定性促成了这样的概率性方法。蛋白质折叠是一个复杂的动态过程，有无数的相互作用规定折叠状态。每一个导致稳定的非共价键相互作用在大小上都是可以相互比较的，似乎没有哪一个具有压倒优势，以致于在折叠中起决定作用。定量化这些相互作用的办法，必定是近似的（见注 1）。概率性设计方法也直接提供非常有用的序列信息，特别是在结构上重要的氨基酸。氨基酸概率可以引导特定序列的设计，也能够凸显能容忍突变、对结构只有微小影响的位点；在几轮蛋白质设计之后，这样的位点可以成为用来改变的目标。

概率性方法可以以几种方式应用于蛋白质设计。序列应该以符合计算出的概率方式生成。首先，最直接的选择是一个共用序列，或在每一个位点用最可能的氨基酸组成的序列。在必要时，可以重复地计算，逐次地增加蛋白质中（已经确定的）的残基。用这样的方法，已得到 114 个残基的双核金属蛋白[32]和一个完整膜蛋白的可溶性变体[33]。其次，计算概率可用于引导对序列的搜索，已提出基于 Monte Carlo 的方法。在 Monte Carlo 轨道的每一点决定序列的接受或拒绝时，计算的氨基酸概率用作有倾向性的选择标准[34]。用这样的方法处理相关的氨基酸身份，但要付出用于搜索的计算运转开销，如果有信息可用于搜索，开销可以减少。最后，概率性方法可以用来定量地指导蛋白质组合库的设计[35]。

1.1.4 组合实验

组合的蛋白质实验可以用来研究序列结构相容性和发现折叠为特定结构的新序列。

① 1cal=4.1868J，后同。

在蛋白质组合设计实验中，筛选大量的序列（实物库），以找到以折叠为预先确定结构的迹象。取决于序列的离散性是如何产生和检验的，这类实验可以探测大量序列序列数可高达 10^{12}[36]。可以用选择性分析，如配体结合或催化活性筛选序列（实物）库。以序列离散程度受研究者控制的方式，这样的实验可以“超越蛋白质序列数据库”。由于去掉了与天然蛋白进化压力的耦合，可以研究对折叠（和其他生物学性质）重要的特征。组合方法已用于鉴别螺旋蛋白[37~39]、泛素变体[40]、单层自组装蛋白[41]、具有纤维样性质的蛋白[42]以及稳定的寡聚螺旋[43]。最近发表了几篇出色的关于组合实验和方法的综述[44~47]。

1.2 方　　法

蛋白质设计中的概率性方法是提供在一个特定的蛋白质结构中对某氨基酸（出现）在一指定位点的概率估计。这里，我们讨论几种估计这些概率的方法并将重点放在直接解出这些概率的基于熵的自洽公式。

1.2.1 关联序列的比对

蛋白质结构的序列可变性可用序列和结构数据库来探讨。已知折叠为非常相似结构的序列可以从蛋白质数据库或结构比对数据库中鉴别出来[48]。如果一个序列的结构已知，具有足够序列相似性的［如序列同一性（identity）大于 40%］可以认为共享同一结构。对这样结构相似蛋白质的多序列比对可以把氨基酸位点特异的概率简单地估计为比对中每一位置（出现）每一氨基酸的频度[49]。这样一组概率常称为序列剖面（sequence profile）。如果序列的数量不够，以致于某些氨基酸在某些特别位点上从未出现过，伪计数（psedocount）和其他方法可以用来规整化这些频度，以使它们在折叠为选定结构这点上更有代表性[50]。虽然如此，从这样的剖面得到的概率将使数据库中序列的性质产生严重偏差。因为存在大量相似性很低的序列折叠类似结构的例子，我们希望在更广的范围内对序列可变性得到完整的理解。从数据库导出的剖面也不适合于设计数据库中没有序列的新蛋白质结构。用一个给定的主链结构为模板，更普遍的计算方法可从头确定氨基酸概率。

1.2.2 建立剖面的定向搜索方法

定向搜索方法的重复应用可以估计一个序列整体的性质。对这类计算，通过给定主链原子坐标来选定一个目标结构。如果采用单个的蛋白质结构，几个最新的直接设计研究得到了与野生型序列相当相似的序列[51~54]。对一个给定结构，可以独立运行多序列搜索计算，以得到一组序列，这些序列的比对产生位点特异概率。Desjarlais 及其合作者，对与一特定折叠一致性的极相关的结构系统的每一个成员独立地运行了他们的序列预测算法[55]。对每一个结构，鉴别出一个优化的“成核”序列，随后，对整个结构探索序列/旋转异构。这个方法已被用于鉴定与小 β 片 WW 结构域[4,55]折叠相容的序列。对一特定折叠的 100 个微结构变体（1Å 均方差）的每一个，应用序列预测算法构建了比应用单个结构更为离散得多的计算剖面[56]。Xencor Inc. 的工作者对一优化序列（其中 β 内酰胺酶活性位点附近的残基被替换）进行了采样，采用了 Monte Carlo 采样

法[57]，找到了对一种抗生素的抗性增加了1000倍多的序列。但是，构建剖面的这些方法非常耗费计算资源。因为，为建立氨基酸位点特异的频度，要完成重复的定向搜索。

1.2.3 序列系综的统计理论

已经建立起统计的、基于熵的公式，对给定的主链结构鉴定出一组位点特异的氨基酸概率，而不只是最优的序列[58,59]。来源于统计学的理论被用来处理与主链结构相容的序列的数目和构成。这一理论也处理构成适合的整个空间，而不只是对实验和数值计算及取样可达到的小部分空间。亚优化序列的特性很容易检验。大蛋白结构（多于100个残基）计算起来很容易。这里的“熵”是指与目标结构相容的序列数。这个来源于热力学的概念被用来减少可能的序列数：对序列的限制减少了熵，并且伴随能量的降低熵也在减少。

方法中的输入是目标主链结构，以及定量化序列-结构相容性的能量函数。对于一个目标主链结构，该方法产生每一个氨基酸（出现）在每一个残基位点的概率（见注2）。在理论中整体的特性（如序列在该目标中的总能量）和局部特性（如在某特定位点所容许的氨基酸），两者可以作为限制包含在方法中。许多氨基酸概率的集合是可能的。这个方法用极大化有效熵的方法确定“最可几”（“最可能”）的这样的集合，借此，这种极大化是受限的。此方法有效地通过这样的限制来为系统提供手段，以减小需要搜索的序列空间体积，达到实验可及的水平。

在限制函数规定的具有期望的特性序列中，令 $w_i\ [\alpha,\ r_k\ (\alpha)]$ 表示氨基酸 α 出现在残基位置 i 并使得其侧链为一组离散构象——r_k（α）（旋转异构态；参考［6］和［60］）中的任何一个的概率。总的序列-构象熵——S_c（此处简单地称为“构象熵”）可以定义为

$$S_c=-\sum_{i,\alpha,k} w_i\ [\alpha,\ r_k\ (\alpha)]\ \ln w_i\ [\alpha,\ r_k\ (\alpha)]$$

求和遍及每一个序列位点 i 和所有可能的氨基酸 α。对每一种氨基酸求和也遍及 k 种可能的旋转异构态——r_k（α）。在限制条件 f_i 之下，通过极大化 S_c 来得到 $w_i\ [\alpha,\ r_k\ (\alpha)]$。极大化采用拉格朗日乘子法[61]。$w_i\ [\alpha,\ r_k\ (\alpha)]$ 的变分泛函 V 定义为

$$V=S-\beta_1 f_1-\beta_2 f_2-\cdots$$

一般来说，限制条件 f_i 也是概率 $w_i\ [\alpha,\ r_k\ (\alpha)]$ 的函数。在确定与特定的限制相容的状态概率时，第 m 个限制函数 f_m 被限定取值 f_m^o。确定概率的方程组和拉格朗日乘子的形式为（见注3）：

$$0=\partial V/\partial w_i\ [\alpha,\ r_k\ (\alpha)]$$

$$f_m^o=f_m\ \{w_i\ [\alpha,\ r_k\ (\alpha)]\}$$

这个大的耦合非线性方程组用求根法（root-finding）求解。虽然这样的方法有很多选择，我们找到一种可以广泛采用的整体收敛的方法[62]。

1.2.3.1 能量函数

在计算中考虑两种能量——构象能 E_c 和环境能 E_{env}，并在极大化构象熵中用作限制条件。

构象能 E_c 用基于原子的位能——AMBER力场[63]计算。E_c 包括范德华相互作用，带有与距离相关的介电常数（$4\varepsilon r_{ij}$）的静电相互作用，以及修正后的氢键项[64]。对一

个特定序列（α_1，…，α_N），其中氨基酸的构象态是［r_1（α_1），…，r_N（α_N）］，E_c是

$$E_c=\sum_i \varepsilon_i\ [\alpha,\ r_k\ (\alpha)]+\sum_{i,j>i}\varepsilon_{i,j}\ [\alpha,\ r_k\ (\alpha);\ \alpha',\ r_{k'}\ (\alpha')]$$

在考虑蛋白质能量函数的时候，单体项 ε_i［α，r_k（α）］包括主链和侧链原子的相互作用，以及氨基酸的参考能量（见 1.2.3.3 小节）。双体项 $\varepsilon_{i,j}$［α，r_k（α）；α'，$r_{k'}$（α'）］为对结构中两个不同位点的两个旋转异构态间相互作用的求和。对享有共同能量特性的大量序列，我们假定 E_c由于序列改变引起的围绕其平均值的涨落不大。那么，我们可以写出

$$E_c\approx\bar{E}_c=\sum_{i,\alpha,k}\varepsilon_i\ [\alpha,\ r_k\ (\alpha)]\ w_i\ [\alpha,\ r_k\ (\alpha)]+\sum_{\substack{i,\ j>i\\ \alpha,\ \alpha'\\ k,\ k'}}\varepsilon_{i,j}\ [\alpha,\ r_k\ (\alpha);\ \alpha',\ r_{k'}\ (\alpha')]\ w_i\ [\alpha,\ r_k\ (\alpha)]\ w_j\ [\alpha',\ r_{k'}\ (\alpha')]$$

作为另一个限制项，引入环境能 E_{env}以在统计理论内用等效的方式计入疏水效应[59]。这个位能考虑了氨基酸的表面暴露倾向。我们可以用氨基酸概率把 E_{env}写成

$$E_{env}\approx\bar{E}_{env}=\sum_{i,\alpha,k}\varepsilon_{env}\ [\alpha,\ r_k\ (\alpha)]\ w_i\ [\alpha,\ r_k\ (\alpha)]$$

式中，ε_{env}为在 1.2.3.2 小节定义的局部环境能量。需要注意的是这个能量不包含双体相互作用并且只取决于在每一个位置的氨基酸和旋转异构态。

1.2.3.2　溶解和疏水能

定量化疏水作用和其他溶液特性在蛋白质设计的方法上是一个重要的参数。用计算来检验序列中大量的变化是不切实际的，即使是计算溶液可及表面积，它常常与疏水倾向相关得很好，也可能要消耗大量计算资源。在用于统计计算一致的实用方法考虑溶液效应的努力中，作为每一个位点附近的β碳原子密度ρ的函数，引进了环境能[59]。一般来说，疏水残基倾向于定位在蛋白质的掩埋区，而亲水残基则倾向于定位在表面。因而，疏水残基倾向于有比亲水残基更高的β碳原子密度。通过 500 个不同的、已知结构的球蛋白，我们推导了计算氨基酸有效势能的通用“统计”势能方程

$$\varepsilon_{env}\ (\alpha,\ \rho)=-T_e\ln\frac{p\ (\alpha,\ \rho)}{p\ (\alpha)\ p\ (\rho)}$$

式中，p（α，ρ）为观察到残基 α 的局部β碳原子密度为ρ的次数；p（α）为观察到残基 α 在训练集中的次数；p（ρ）为不管残基类型，局域密度ρ被观察到的次数；T_e为有效温度；密度ρ为以残基某个特定取向为中心的“自由体积”内β碳原子的密度。自由体积即未被侧链排除的平均体积

$$\rho\ (\alpha)=\frac{n_\beta}{\frac{4}{3}\pi R^3-\langle V_{\text{access}}\ (\alpha)\rangle}$$

式中，n_β为从侧链质心起在距离 R 内（比如 8Å）的β碳原子数；$\langle V_{\text{access}}\ (\alpha)\rangle$ 为残基α的平均排斥体积，是对α的所有旋转异构态整体平均值的计算。我们注意局部密度依赖于残基的旋转异构态，所以 $\varepsilon_{env}\ \{\alpha,\ \rho\ [r_k\ (\alpha)]\}\equiv\varepsilon_{env}\ \{\alpha,\ r_k\ (\alpha)\}$。这种基于β碳原子密度的位能与其他的氨基酸疏水标度相关得很好[59]。对序列概率计算，E_{env}限制取值为一个具有同样结构的已知序列的值（如果有一个已知值的话），或者具有同