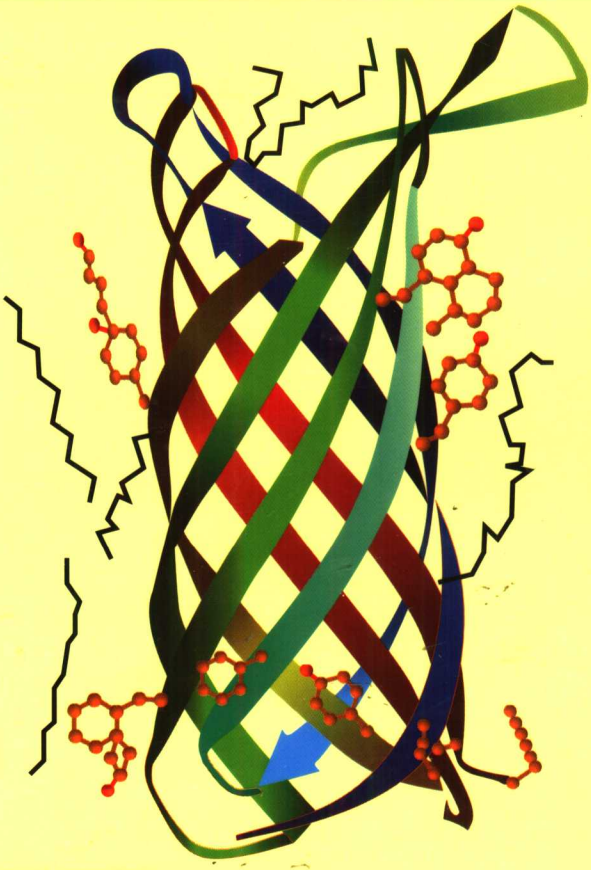
筒型外膜蛋白质生物信息学

孟朝晖　编著



国防工业出版社
National Defense Industry Press

筒型外膜蛋白质生物信息学

孟朝晖　编著

国防工业出版社

·北京·

内 容 简 介

简型外膜蛋白质是广泛存在于革兰氏阴性菌、线粒体和叶绿体外膜的结构类似简型的蛋白质，研究这种蛋白质的结构和功能对于了解细胞内物质和能量的运作机理十分重要。本书用生物信息学的方法研究简型外膜蛋白质，详细介绍了简型外膜蛋白质的基本结构和功能，包含结构模型、角度模型、侧链旋转异构体等，对目前已知三维结构的简型外膜蛋白质分为10种类型详细介绍，还研究了简型蛋白质OmpA的折叠过程。分别用神经网络模型、隐马氏模型及根据蛋白质的物理化学特性抽象出的几种结构辨识模型，分析了蛋白质的结构。为了方便不同学科的读者能够较为顺利地阅读本书，其中涉及到的相关基础知识均有适当介绍，包括蛋白质基础知识和蛋白质数据库PDB的相关内容，神经网络模型和隐马氏模型的基本内容也有简明阐述。

本书可作为生物学、计算机科学、信息科学、数学等领域的教师、学生和科研工作者的参考书。

图书在版编目(CIP)数据

简型外膜蛋白质生物信息学/孟朝晖编著.—北京：国防工业出版社，2007.4

ISBN 978-7-118-05034-9

Ⅰ.简… Ⅱ.孟… Ⅲ.蛋白质－生物信息论 Ⅳ.Q51 Q811.4

中国版本图书馆CIP数据核字(2007)第024335号

※

国防工业出版社出版发行

(北京市海淀区紫竹院南路23号 邮政编码100044)

国防工业出版社印刷厂印刷

新华书店经售

*

开本 710×960 1/16 **印张** 17¾ **字数** 338千字

2007年4月第1版第1次印刷 **印数** 1—2500册 **定价** 38.00元

(本书如有印装错误，我社负责调换)

国防书店：(010)68428422 发行邮购：(010)68414474

发行传真：(010)68411535 发行业务：(010)68472764

作者简介

孟朝晖，河海大学计算机及信息工程学院副教授，1968年生，1989年本科毕业、1992年硕士研究生毕业于西安交通大学数学系。从事计算机专业和通信专业的教学与科研工作，已发表9篇第一作者论文。目前的主要研究方向为：蛋白质结构分析，生物信息学。

地　　址：江苏省南京市西康路1号
　　河海大学计算机及信息工程学院
邮　　编：210024
学院网址：cies.hhu.edu.cn
个人信箱：mengzhaohui@hhu.edu.cn

前　言

蛋白质结构和功能的分析是目前生物学和生物信息学的基础研究领域。筒型外膜蛋白质是广泛存在于革兰氏阴性菌、线粒体和叶绿体外膜的结构类似筒型的蛋白质，研究这种蛋白质的结构和功能对于了解细胞内物质和能量的运作机理十分重要。但是，由于线粒体和叶绿体外膜蛋白质的结晶非常困难，所以，目前仍未有充分的三维结构数据用于研究筒型外膜蛋白质的功能。目前，已获得三维结构数据的筒型外膜蛋白质不到100个，且主要来自于革兰氏阴性菌。因此，采用生物信息学的方法，根据已有的少量数据，分析研究这类蛋白质的结构和功能的基础研究就显得十分重要。

本书主要取材于近10年尤其是近二三年的最新国际权威期刊的研究论文和综述，对筒型外膜蛋白质领域的研究成果做出了系统化的介绍和总结。在生物学文献数据库PubMed的搜索页面中输入关键词“Barrel Membrane Protein”（筒型外膜蛋白质）可以检索到1000多篇研究论文和近100篇综述，本书参考文献中的230多篇论文是与筒型外膜蛋白质研究领域紧密相关的论文，作者认真阅读了这些论文后，从中选择了一些有代表性的研究成果，经过总结加工，成为本书的核心内容。

第1章简要介绍蛋白质结构的基础知识，重点介绍β－折叠结构。第2章介绍蛋白质数据库PDB。PDB数据库是目前研究蛋白质的结构和功能的重要基础平台，目前，该数据库已收录了4万多个生物大分子（主要是蛋白质）的三维结构数据，本书中的蛋白质数据均取自该数据库。

第3章先概要性地介绍筒型外膜蛋白质的基本结构和功能，然后介绍目前对筒型外膜蛋白质结构模型研究的主要成果，包括结构模型、角度模型、侧链旋转异构体等内容。第4章将目前已知三维结构的近百个筒型外膜蛋白质按筒结构分为10种类型，分别加以详细介绍，对于每个有独立名称的蛋白质，对其氨基酸序列进行二级结构标注，并给出其结构示意图。

第5章介绍了筒型外膜蛋白质OmpA的折叠过程，介绍了在4种实验条件下，OmpA的折叠过程，其中有些折叠蛋白质的NMR结构测定结果已被收入PDB数据库。研究筒型外膜蛋白质折叠过程，对于认识了解蛋白质的结构和功能具有

重要意义。虽然在目前来看，蛋白质折叠过程的实验研究似乎与蛋白质信息学的关系不大，但是，蛋白质结构的信息计算模型研究必将从静态研究扩展到动态研究，从单纯结构研究过渡到结构与功能相结合研究，进一步的研究还要综合考虑蛋白质间相互作用以及蛋白质与其周围环境的相互作用。而建立蛋白质折叠过程的信息计算模型正是一个适合的研究切入点。第5章介绍的内容只是蛋白质折叠过程实验研究中比较初步的成果，参考文献中列出了50多篇这方面的研究论文，供读者参考。

第6章简要介绍了神经网络计算模型的基本原理和基本算法。第7章介绍了用神经网络计算模型预测β－筒型外膜蛋白质结构的研究成果。第8章介绍了隐马氏模型的基本概念和学习算法。第9章研究了3种针对β－筒型外膜蛋白质的特殊结构而设计的隐马氏模型。从第9章可以看出，隐马氏模型在模型结构设计上比神经网络模型具有更高的灵活性，更适合DNA序列、蛋白质氨基酸序列等生物大分子序列的分析研究。

第10章介绍了几种根据蛋白质的物理化学特性抽象出的结构辨识模型。这几种辨识模型需要较为深入的生物物理和生物化学知识，但其计算方法比较简单，数学模型直观易懂，并且其结果比较易于与蛋白质的真实结构对应起来。

本书用信息学的方法研究筒型外膜蛋白质，可以定位为生物信息学方面的深入专题的专著。蛋白质生物信息学是一个新兴的研究领域，既需要较为扎实的数学和计算机基础知识，又需要比较全面而深入的生物化学和分子生物学知识。本书在写作过程中注重生物基础知识和蛋白质实验数据的准确阐述，同时重视严谨的数学模型分析，尽量使得生物学背景的读者和数学、计算机科学背景的读者均能顺畅地阅读本书并从中获益。

感谢河海大学朱跃龙教授、王志坚教授、徐立中教授在本书的立项、写作和出版过程中给予的关怀和帮助。感谢河海大学李晓芳老师对本书的出版提供的帮助。

感谢河海大学“211工程”学科建设项目对本书出版的资助。

作者在写作过程中尽力做到每处内容均严谨有据，但本书涉及的知识面比较广泛，缺陷和不足在所难免，敬请读者批评指正。

孟朝晖

2006年12月于南京

目　录

第 1 章　蛋白质简介

1.1　氨 基 酸

蛋白质(Protein)主要由氨基酸(Amino Acid)组成，生物体中的蛋白质由 20 种氨基酸组成。氨基酸链接成肽链，再进一步折叠成复杂而多样的三维结构，从而使蛋白质具有了千差万别的结构和功能。

单个氨基酸为四面体结构，如图 1.1 所示，中心为一个碳原子，称为 C_α，C_α 通过 4 个共价键(Covalent)与氨基(Amino Group)、羧基(Carboxyl Group)、氢(Hydrogen)和侧链(Side Chain)相连接，羧基中的碳原子标记为 C，侧链中的碳原子依次为 C_β、C_γ、C_δ、C_ε等。侧链具有不同的结构，从而形成不同的氨基酸，在自然生物体中提取的氨基酸有 20 种。

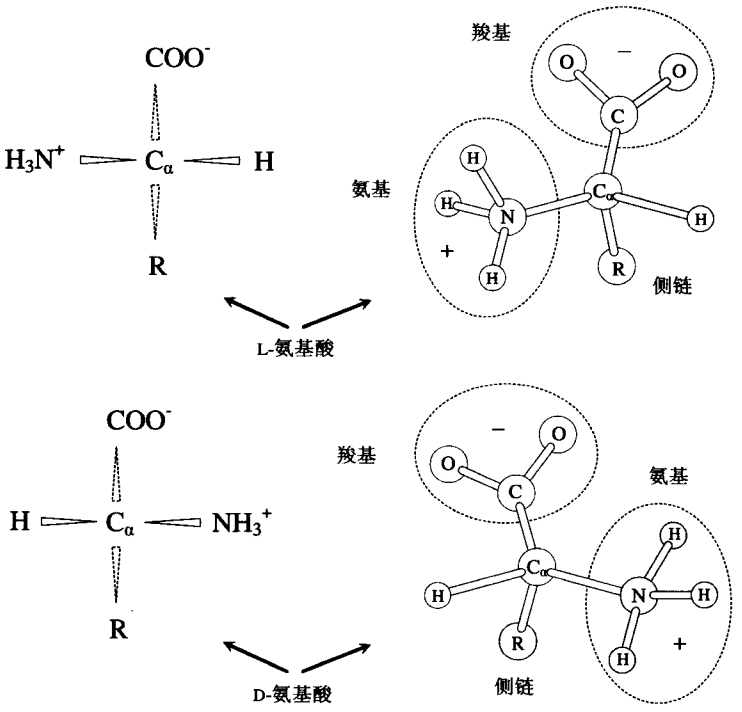


图 1.1　单个氨基酸的四面体结构与镜像异构体

在中性溶剂中(pH=7)，羧基的存在形式为 COO^-，氨基的存在形式为 NH_3^+。氨基酸为手性(Chiral)分子(甘氨酸除外)，即存在不能相互立体重叠的镜像异构体(Isomers)，根据旋光性的不同，分别称为 L-氨基酸和 D-氨基酸，如图 1.1 所示。注意图 1.1 中表示共价键的楔形符号，表示倾斜于纸面的方向性，如果标定 C_α 在纸面位置，则 H 和 NH_3^+在纸之上，R 和 COO^-在纸之下。L-氨基酸的判断法则为左手 C→R→N 规则，即以 C_α→H 为左手拇指方向，左手四指旋转方向为 C→R→N(COO^-→R→NH_3^+)或 R→N→C 或 N→C→R。蛋白质水解得到的氨基酸大多是 L-氨基酸，但在某些生物体内特别是细菌中还是有 D-氨基酸存在的。楔形共价键符号可以准确地表示氨基酸的立体结构，但在后面的叙述中，除非需要表示异构体，氨基酸分子式中将使用线段型共价键符号，并且单指 L-氨基酸。

根据侧链 R 的不同，通常将 20 种氨基酸分为 4 类：非极性氨基酸、极性不带电荷氨基酸、带负电荷的酸性氨基酸、带正电荷的碱性氨基酸。表 1.1 为 20 种氨基酸的名称和简记符号。

表 1.1　20 种氨基酸的名称和简记符号

分类	名称	英文名称	符号		化 学 名 称
非极性	亮氨酸	Leucine	Leu	L	α-氨基异己酸
	丙氨酸	Alanine	Ala	A	α-氨基丙酸
	缬氨酸	Valine	Val	V	α-氨基异戊酸
	脯氨酸	Proline	Pro	P	β-吡咯烷基-α-羧酸
	蛋氨酸	Methionine	Met	M	α-氨基-γ-甲硫基丁酸
	苯丙氨酸	Phenylalanine	Phe	F	α-氨基-β-苯基丙酸
	色氨酸	Tryptophan	Trp	W	α-氨基-β-吲哚基丙酸
	异亮氨酸	Isoleucine	Ile	I	α-氨基-β-甲基戊酸
极性不带电荷	甘氨酸	Glycine	Gly	G	氨基乙酸
	丝氨酸	Serine	Ser	S	α-氨基-β-羟基丙酸
	苏氨酸	Threonine	Thr	T	α-氨基-β-羟基丁酸
	半胱氨酸	Cysteine	Cys	C	α-氨基-β-巯基丙酸
	天冬酰胺	Asparagine	Asn	N	天冬酰胺
	谷氨酰胺	Glutamine	Gln	Q	谷氨酰胺
	酪氨酸	Tyrosine	Tyr	Y	α-氨基-β-对羟苯基丙酸
带负电荷酸性	天冬氨酸	Aspartic acid	Asp	D	α-氨基丁二酸
	谷氨酸	Glutamic acid	Glu	E	α-氨基戊二酸
带正电荷碱性	赖氨酸	Lysine	Lys	K	α,ε-二氨基己酸
	精氨酸	Arginine	Arg	R	α-氨基-δ-胍基戊酸
	组氨酸	Histidine	His	H	α-氨基-β-咪唑基丙酸

1. 非极性氨基酸(Nonpolar Amino Acids)

非极性氨基酸有 8 个，通常归为疏水(Hydrophobic)类氨基酸，包括 4 个含烷基(Alkyl)侧链的氨基酸(丙氨酸、缬氨酸、亮氨酸和异亮氨酸)、1 个脯氨酸、1 个含硫(Sulfur)的蛋氨酸(甲硫氨酸)、2 个侧链含芳香基(Aromatic)的氨基酸(色氨酸和苯丙氨酸)(图 1.2)。

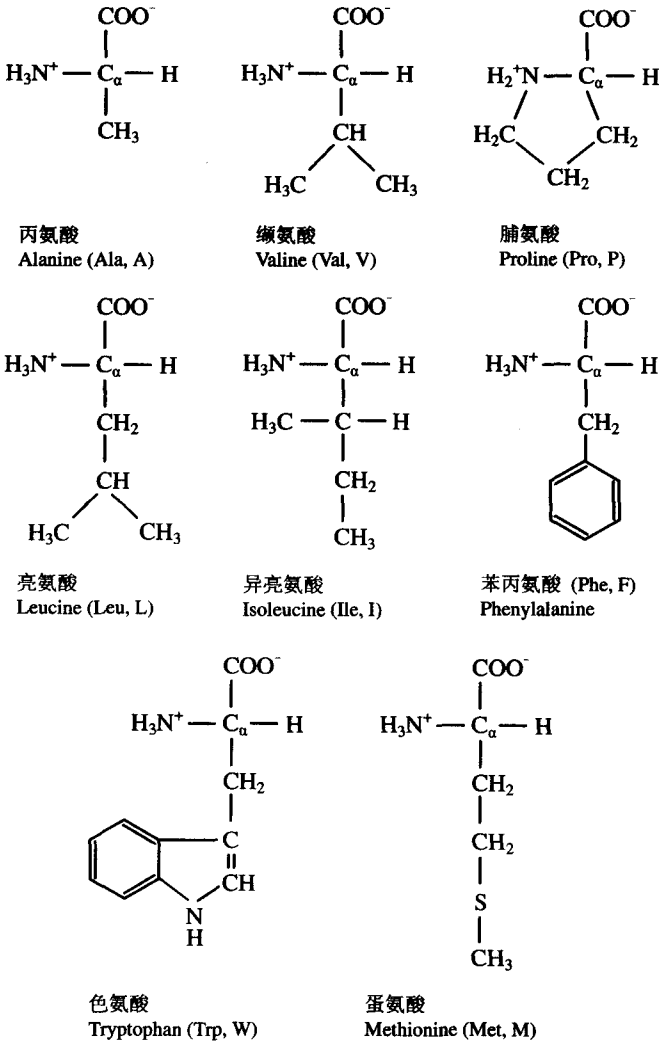


图 1.2　非极性氨基酸

色氨酸并不严格疏水，因为其吲哚环(indole ring)中的 N-H 会与水相互作用。脯氨酸不能算是标准的氨基酸，它没有自由的 α-氨基，是一种 α-亚氨基酸(α-Imino

Acid)。

2. 极性不带电荷氨基酸(Polar Uncharged)

极性不带电荷氨基酸包含 7 个氨基酸，除甘氨酸外，其它 6 个可以与水形成氢键，极性氨基酸相对于非极性氨基酸更能溶于水。天冬酰胺和谷氨酰胺的极性来自于其酰胺基(Amide)，酪氨酸、苏氨酸和丝氨酸的极性来自于其羟基(Hydroxyl)，半胱氨酸的巯基(Sulfhydryl)具有极性，酰胺基、羟基和巯基均能与水形成氢键。甘氨酸为最简单的氨基酸，其 R 侧链仅为 1 个氢原子，不易形成氢键，并且，甘氨酸不具有旋光性，为非手性分子。甘氨酸的水溶性主要源于其极性的氨基和羧基。甘氨酸有时也被归类为非极性氨基酸(图 1.3)。

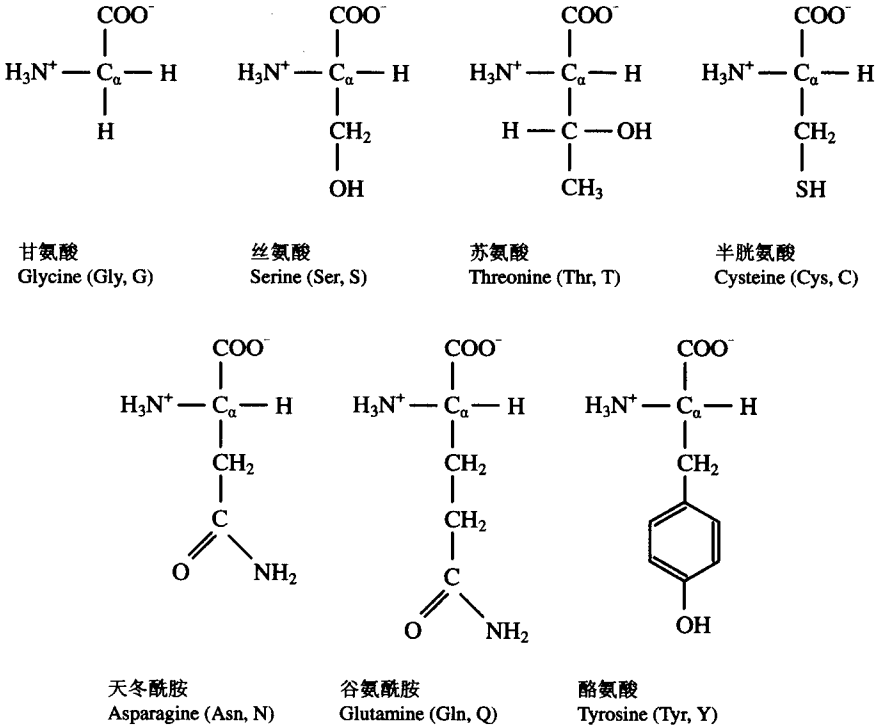


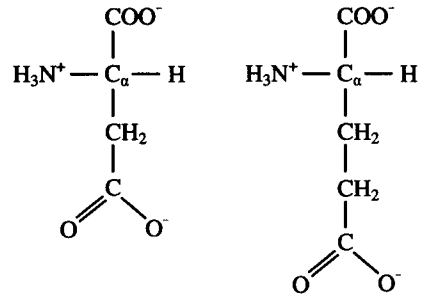
图 1.3　极性不带电荷氨基酸

3. 带负电荷的酸性氨基酸(Negative Charge Acidic)

酸性氨基酸有 2 个，天冬氨酸和谷氨酸，其 R 侧链含有羧基，在 pH=7 环境下，侧链羧基完全解离为—COO^-，因此带负电荷(图 1.4)。

4. 带正电荷的碱性氨基酸(Positive Charge Basic)

在中性溶剂中带正电荷的氨基酸为 3 个碱性氨基酸，赖氨酸、精氨酸和组氨酸。组氨酸的解离基(Ionized Group)为咪唑基(Imidazolium)，精氨酸的解离基为胍

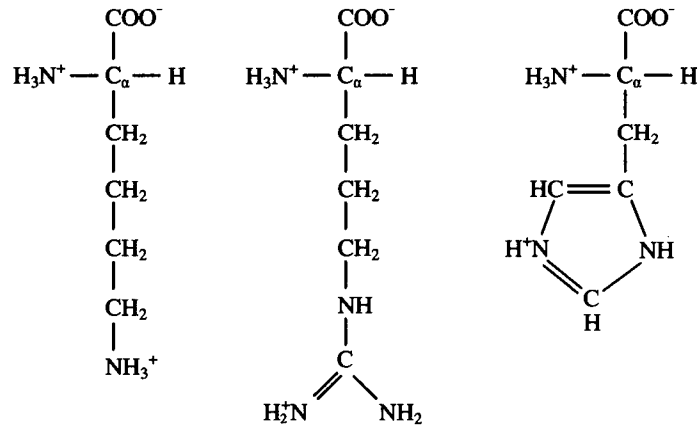


天冬氨酸 (Asp, D)
Aspartic acid

谷氨酸 (Glu, E)
Glutamic acid

图 1.4　带负电荷的酸性氨基酸

基(Guanidinium),均带正电荷，赖氨酸的脂肪链 C_ε 碳原子上有 1 个氨基，带正电荷。赖氨酸和精氨酸的侧链在 pH=7 基本上完全质子化，但组氨酸的侧链在 pH=7 时质子化分子不到 10%(图 1.5)。



赖氨酸
Lysine (Lys, K)

精氨酸
Arginine (Arg, R)

组氨酸
Histidine (His, H)

图 1.5　带正电荷的碱性氨基酸

1.2 肽　链

两个氨基酸中一个氨基酸的 α-羧基失去一个氧原子，另一个氨基酸的 α-氨基失去两个氢原子，合计失去一个水分子，以致在两个氨基酸之间形成一个酰胺键(Amide Linkage)，也称为肽键(Peptide Bond)。多个氨基酸可以由肽键相连而形成多肽链(polypeptide chain)，简称肽链，蛋白质即由一条或多条肽链所组成一肽链

为无分支结构的一维链，其主链(Backbone)的重复单元为—N—C_α—C—，故肽链也称为蛋白质的一级结构(primary structure)。如图 1.6、图 1.7 为肽键形成的示意图和反应式。

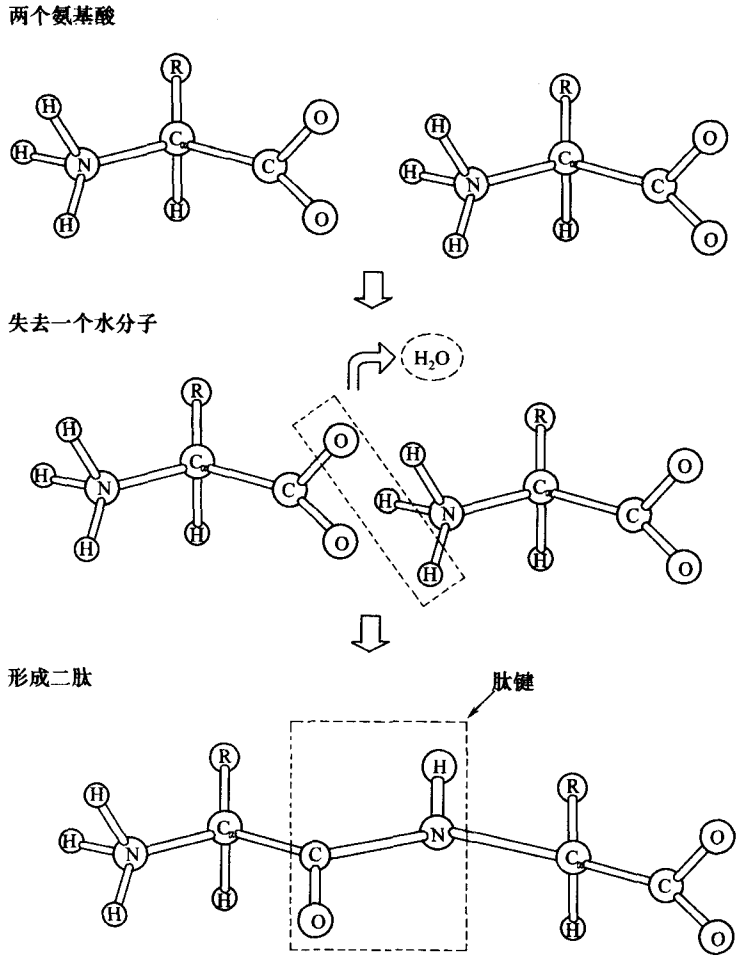


图 1.6　肽键形成的示意图

肽链的基本重复单元为合计失去一个水分子的氨基酸剩余部分，称为氨基酸残基(Amino Acid Residue)，但两端的氨基酸合起来失去一个完整的 H_2O，其中只失去一个氧的末端氨基酸有完整的氨基 NH_3^+，称为 N 末端残基，失去两个氢的末端氨基酸有完整的羧基 COO^-，称为 C 末端残基。在肽链上没有完整的氨基酸，所以肽链也称为氨基酸残基序列、残基序列，也可称为氨基酸序列。如图 1.8 所示。

肽键也称肽键平面。在肽链的分子式中，肽键的羰基碳和酰胺氮常表示为单键 C—N，羰基氧和碳表示为双键 C═O，即图 1.9 中的结构 1 的形式，单键 C—

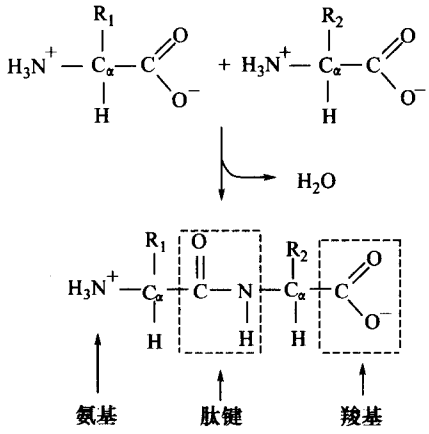


图 1.7　肽键形成的反应式

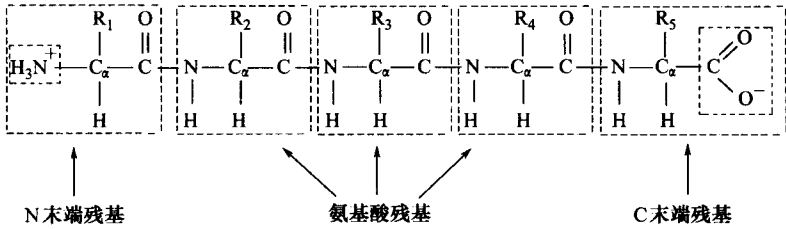


图 1.8　肽链和残基示意图

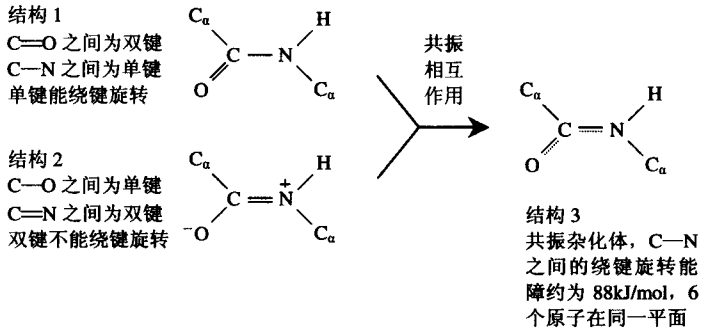


图 1.9　肽键的结构

N 是可以任意旋转的，即一个肽键上构成氨基酸主链的 3 个键，C_α—C、C—N 和 N—C_α，均可以自由旋转。但实际上，这只是肽键的一个极端结构，称为共振结构(Resonance Structure)，而且还有另一种极端结构，即图 1.9 中的结构 2，羰基碳和酰胺氮为双键 C═N，羰基氧和碳为单键 C—O，这时，氮原子带一正电荷，氧原子带一负电荷，肽键上氨基酸主链中只剩 2 个键，C_α—C 和 N—C_α，可以自由旋转。按照电子轨道理论[6]和共振理论[7]，肽键的实际结构是结构 1 和结构 2 的共振杂化体(Resonance Hybrid)，即图 1.9 中的结构 3 的形式，结构 3 是介于结构 1 和结构 2 之间的中间体。

共振杂化体的 6 个原子位于同一个平面，如图 1.9 中结构 3 所示，称为酰胺键平面(Amide Plane)，也可称为肽键平面。正常 C—N 单键的键长是 0.148nm，C═N 双键的键长是 0.127nm，而肽键中的 C—N 键的键长为 0.133nm，介于单键键长和双键键长之间，说明肽键中的 C—N 键具有双键性质，使得 6 个原子基本位于同一个平面，并且 C—N 键具有绕键旋转能障(Energy Barrier of Bond Rotation)，约为 88kJ/mol，此能障在室温环境下足以阻止 C—N 键的旋转，从而维持酰胺键的平面结构。图 1.10 为二肽的肽键平面，图 1.11 为肽键平面中的键角。

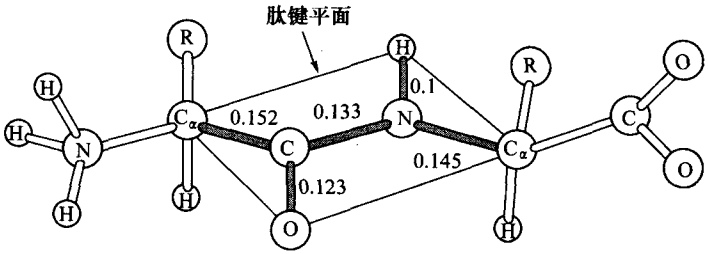


图 1.10　二肽的肽键平面

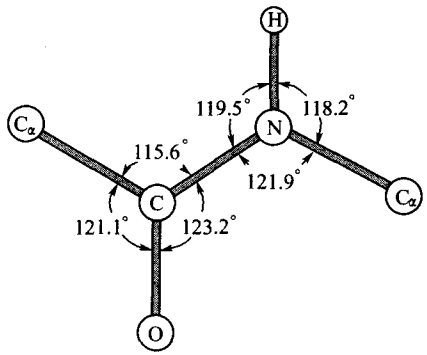


图 1.11　肽键平面的键角

注意，图 1.10 中的肽键中的羰基氧(Carbonyl Oxygen)与酰胺氢(Amide Hydrogen)为反式结构，即 O 和 H 在 C—N 的两边。若 O 和 H 在 C—N 的同一边，则称为顺式。反式结构在蛋白质中最为常见，因为其较少引起非键原子间的立体位阻现象(Steric Hindrance)。

1.3　肽键的空间结构与 β-折叠

1.3.1　肽键平面之间的空间关系

肽键平面是氨基酸链中相当稳定的局部结构，氨基酸链就是由主链 C_α 链接在