



## 第 1-1 章 生物信息类型及其产生途径

### 第一节 生物信息的类型

#### 1. 核苷酸序列数据

常见的核苷酸序列数据由脱氧核糖核酸(DNA)和核糖核酸(RNA)数据组成。DNA的组成单位为四种脱氧核苷酸,分别为腺嘌呤(A)、鸟嘌呤(G)、胸腺嘧啶(T)和胞嘧啶(C)。而RNA的组成单位为四种核糖核苷酸,分别为腺嘌呤(A)、鸟嘌呤(G)、尿嘧啶(U)和胞嘧啶(C)。核苷酸序列就是指DNA或者RNA中这四种碱基的排列顺序。核苷酸序列数据的测定以及目前较权威的核苷酸序列数据库将在本章第二节和下一章中阐述。

#### 2. 蛋白质序列和结构数据

蛋白质序列是指20种氨基酸的排列顺序(也就是蛋白质的一级结构)。这20种氨基酸的名称及缩写符号见表1-1.1(详细说明见第4-1章)。

表 1-1.1 20 种氨基酸的名称和缩写符号

名称	单字符号	名称	单字符号
丙氨酸	A	亮氨酸	L
精氨酸	R	赖氨酸	K
天冬氨酸	D	甲硫氨酸	M
半胱氨酸	C	苯丙氨酸	F
谷氨酰胺	Q	脯氨酸	P
谷氨酸	E	丝氨酸	S
组氨酸	H	苏氨酸	T
异亮氨酸	I	色氨酸	W
甘氨酸	G	酪氨酸	Y
天冬酰胺	N	缬氨酸	V

蛋白质结构数据主要是蛋白质的三级结构信息。蛋白质的三级结构是蛋白质的多肽链在各种二级结构的基础上,再进一步盘曲或者折叠形成的具有一定规律的三维空间结构。目前蛋白质的三级结构数据的主要来源是通过实验(X射线晶体衍射、核磁共振等)来测定,该内容将在本章第四节详细阐述。

#### 3. 其他类型数据

##### (1) 分子标记数据

分子标记(molecular marker)是遗传标记的一种。遗传标记是指在染色体上位置已知的一个基因或者一段DNA序列,可被用于鉴定生物个体或者物种,包括形态标记

(morphological marker)、细胞学标记(ctological marker)、生化标记(biochemical marker)和分子标记四种类型。其中,分子标记指能反映生物个体或种群间基因组中某种差异特征的 DNA 片段,它直接反映基因组 DNA 间的差异。

分子标记大多以电泳谱带的形式表现,大致可分为三大类,见表 1-1.2。

表 1-1.2 分子标记的分类

类别	分子标记技术名称	简称
以分子杂交为核心的分子标记技术	限制性片段长度多态性标记 restriction fragment length polymorphism	RFLP
	DNA 指纹技术 DNA finger printing	—
	原位杂交 <i>in situ</i> hybridization	—
以聚合酶链式反应(PCR)为核心的分子标记技术	随机扩增多态性 DNA 标记 random amplification polymorphism DNA	RAPD
	简单序列重复标记 simple sequence repeat	SSR
	简单序列长度多态性 simple sequence length polymorphism	SSLP
	扩展片段长度多态性标记 amplified fragment length polymorphism	AFLP
	序标位 sequence tagged sites	STS
新型的分子标记	序列特征化扩增区域 sequence characterized amplified region	SCAR
	单核苷酸多态性 single nucleotide polymorphism	SNP
	表达序列标签 expressed sequences tags	EST

## (2) 生物芯片数据

生物芯片(biochip 或 bioarray)技术起源于核酸分子杂交。该技术根据生物分子间特异相互作用的原理,将生化分析过程集成于芯片表面,实现生物信息的存储和集成,从而实现了对 DNA、RNA、多肽、蛋白质以及其他生物成分的高通量快速检测。

生物芯片按其成分可以分为基因芯片(gene chip)、蛋白质芯片(protein chip 或 protein microarray)、细胞芯片(cell chip)和组织芯片(tissue chip)。其中基因芯片又称为 DNA 芯片(DNA chip)或 DNA 微阵列(DNA microarray),是将 cDNA 或寡核苷酸固定在微型载体上形成微阵列。蛋白质芯片是将蛋白质或抗原等一些非核酸生命物质固定在微型载体上形成微阵列。细胞芯片是将细胞按照特定的方式固定在载体上,用来检测细胞间相互影响或相互

作用。组织芯片是将组织切片等按照特定的方式固定在载体上,主要用来对免疫组织等组织内成分差异进行研究。

### (3) 生物表型数据

生物表型(phenotype)数据是指生物体的个体形态、外观、生理、功能等相关的一些指标数据,如身高、肤色、血型、酶活力、药物耐受力乃至性格等。一般情况下通过常规的测量和检测就能够得到相应的数据集。

## 第二节 DNA 测序技术

### 一、第一代测序技术

#### 1. 双脱氧链终止法(Sanger 法)

第一代 DNA 测序技术主要为 1977 年由桑格(Sanger)等提出的双脱氧链终止法(dideoxy sequencing technique),也称为 Sanger 法。Sanger 法的核心原理是:双脱氧核糖核苷酸(ddNTP)的 2' 和 3' 位置都不含羟基(图 1-1.1 右),因此 ddNTP 在 DNA 的合成过程中不能形成磷酸二酯键,从而中断 DNA 的合成反应;在 4 个 DNA 合成反应体系中分别加入一定比例的带有放射性同位素标记的 ddATP、ddCTP、ddGTP 和 ddTTP,通过凝胶电泳和放射自显影后可根据电泳条带的位置确定待测分子的 DNA 序列(图 1-1.2)。

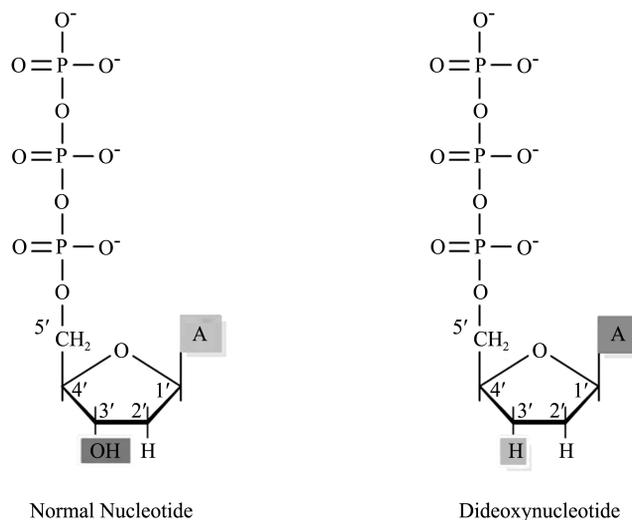


图 1-1.1 脱氧核苷酸(左)和双脱氧核苷酸(右)

左图为正常的脱氧核苷酸(dNTP),其 3' 位置含有羟基;右图为双脱氧核苷酸(ddNTP),其 2' 和 3' 位置都不含羟基。

如图 1-1.2 所示,Sanger 法测序的具体步骤:

①分离待测核酸模板,在 4 支试管中分别加入适当的引物、模板、DNA 聚合酶和 4 种脱氧核糖核苷酸(dNTP),再在上面 4 支试管中分别加入一定浓度的带有放射性同位素标记的 ddATP、ddCTP、ddGTP 和 ddTTP。

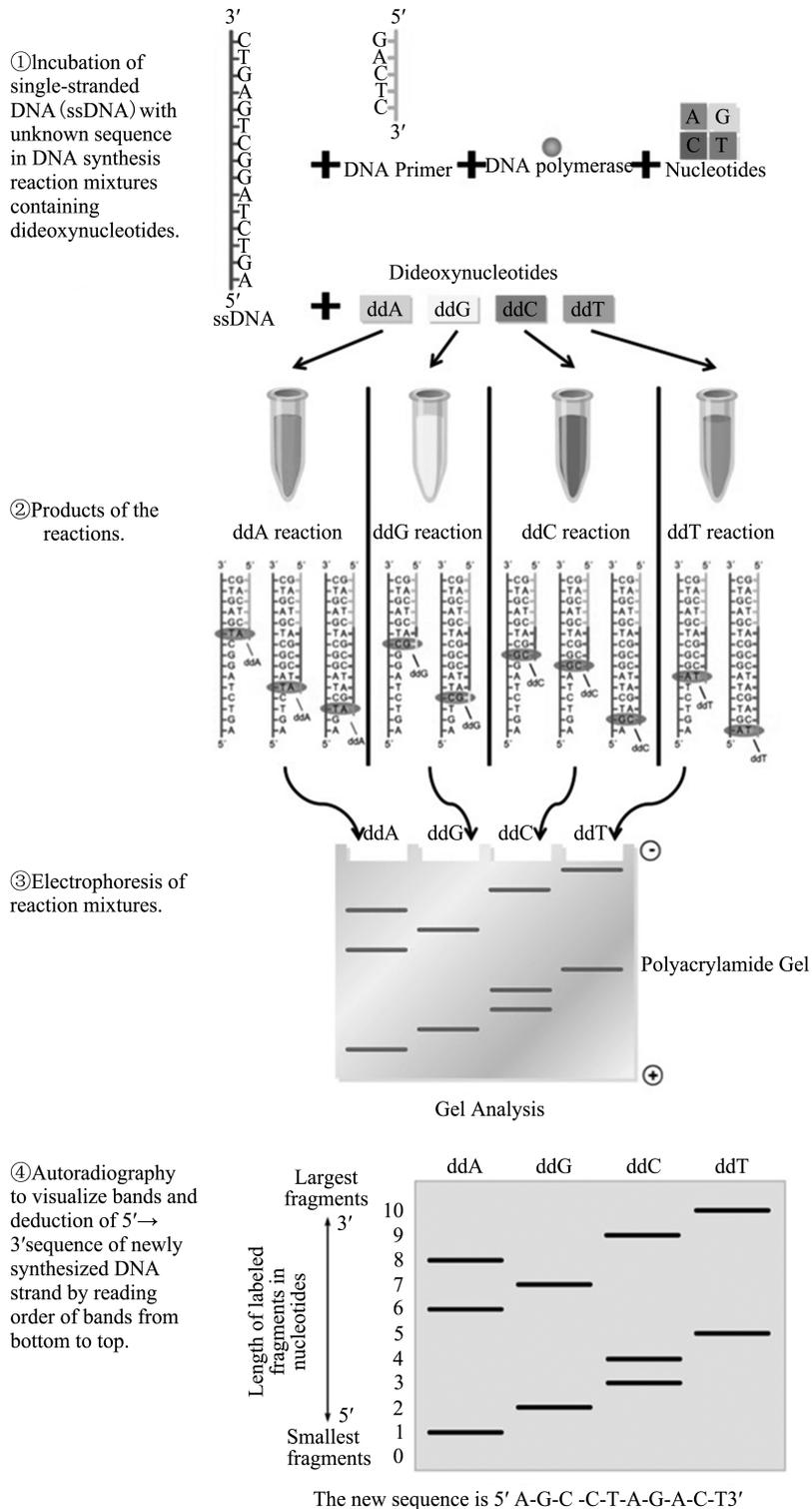


图 1-1.2 双脱氧链终止法 (Sanger 法) 测序原理

①反应混合物的制备;②DNA 合成反应;③凝胶电泳;④放射自显影

②进行 DNA 合成反应。加入的引物在 DNA 聚合酶作用下从 5' 端向 3' 端进行延伸反应。当 ddNTP 掺入时,由于它在 3' 位置没有羟基,故不能与下一个 dNTP 结合,从而使链延伸终止。由于 ddNTP 在不同位置掺入,因而产生一系列不同长度的新的 DNA 链。

③用变性聚丙烯酰胺凝胶电泳同时分离 4 只反应管中的反应产物,由于每一反应管中只加一种 ddNTP(如 ddATP),则该管中各种长度的 DNA 都终止于该种碱基(如 A)处,所以凝胶电泳中该泳道不同带的 DNA 的 3' 末端都为同一种双脱氧碱基。

④放射自显影。根据四泳道的编号和每个泳道中 DNA 带的位置直接从自显影图谱上读出与模板链互补的新链序列。

## 2. 化学降解法

几乎在双脱氧链终止法出现的同时,Maxam 和 Gilbert 在 1977 年提出了化学降解法。化学降解法测序的基本原理:

①对待测 DNA 末端进行放射性标记。

②通过 5 组(或 4 组)相互独立的化学反应分别得到部分降解产物,其中每一组反应特异性地针对某一种或某一类碱基进行切割,如表 1-1.3。因此,产生 5 组(或 4 组)不同长度的放射性标记的 DNA 片段,每组中的每个片段都有放射性标记的共同起点,但长度取决于该组反应针对的碱基在原样品 DNA 分子上的位置。

③各组反应物通过聚丙烯酰胺凝胶电泳进行分离。

④通过放射自显影检测末端标记的分子,并直接读取待测 DNA 片段的核苷酸序列。

表 1-1.3 化学降解法涉及的 5 种化学反应

碱基体系	化学修饰试剂	化学反应	断裂部位
G	dimethyl sulphate(硫酸二甲酯)	甲基化	G
A+G	piperidine formate(哌啶甲酸), pH2.0	脱嘌呤	G 和 A
C+T	hydrazine(肼,联氨 NH <sub>2</sub> NH <sub>2</sub> )	打开嘧啶环	C 和 T
C	hydrazine+NaCl(1.5M)	打开胞嘧啶环	C
A > C(可选)	90°C, NaOH(1.2M)	断裂反应	A 和 C

## 3. 双脱氧链终止法与化学降解法的比较

在双脱氧链终止法与化学降解法这两种测序方法刚被提出的时候,化学降解法不仅重复性高,而且只需要简单的化学试剂和一般的实验条件,易为普通实验室和研究人员所掌握。而链终止法需要单链模板、特异的寡核苷酸引物和质量的大肠杆菌 DNA 聚合酶 I 大片段(Klenow 片段),这在 20 世纪 80 年代一般的实验室很难做到。但随着 M13mp 系列载体的发展、DNA 合成技术的进步及 Sanger 法测序反应的不断完善,至今 DNA 测序已大都采用 Sanger 法进行,例如将在接下来“第二代测序技术”部分提到的 Roche 公司的 454 测序技术、ABI 公司 SOLID 技术,他们的核心手段都是利用了 Sanger 法中的可中断 DNA 合成反应的脱氧核苷酸。

当然,化学降解法不需要进行酶催化反应,可对合成的寡核苷酸进行测序,可以分析 DNA 甲基化修饰情况,还可以通过化学保护及修饰等干扰实验来研究 DNA 的二级结构和 DNA 与蛋白质的相互作用,这些仍然是化学降解法所独具的特点。

## 二、第二代测序技术

随着研究的不断深入,第一代测序技术由于其成本高、通量低等缺点,越来越满足不了日益发展的生物研究需求,并且难以实现大规模的应用。经过不断的技术研发和改善,同时具备成本低、通量高、速度快等特点的第二代测序技术应运而生。

第二和三代测序技术(见下节)又称下一代测序(next generation sequencing, NGS)技术,或者高通量测序(high-throughput sequencing, HTS)。自 2005 年第一台二代测序技术平台出现后,不断有新测序技术出现,测序通量不断提高;2010 年第一台三代测序平台出现,该技术在保持测序通量的同时,显著提高了读序长度(图 1-1.3)。三代测序读序平均长度达到 5~8kb,最长的可以达到几十 kb;但它的一个致命缺陷是测序错误率很高(详见下节)。

目前,二代测序技术的主要有 Illumina 公司的 Solexa 基因组分析仪(Illumina Genome Analyzer)、Roche 公司的 454 测序仪(Roche GS FLX sequencer)和 ABI(Applied BioSystem)公司的 SOLiD 测序仪(ABI SOLiD sequencer)。这三个平台的测序原理各不相同,分别在通量、读长、准确度、速度和成本方面各具优势,并且都在基因组 *de novo*、重测序、转录组、表观遗传学研究中发挥了重要作用。接下来将详细阐述以上三种测序技术的特点和差异。

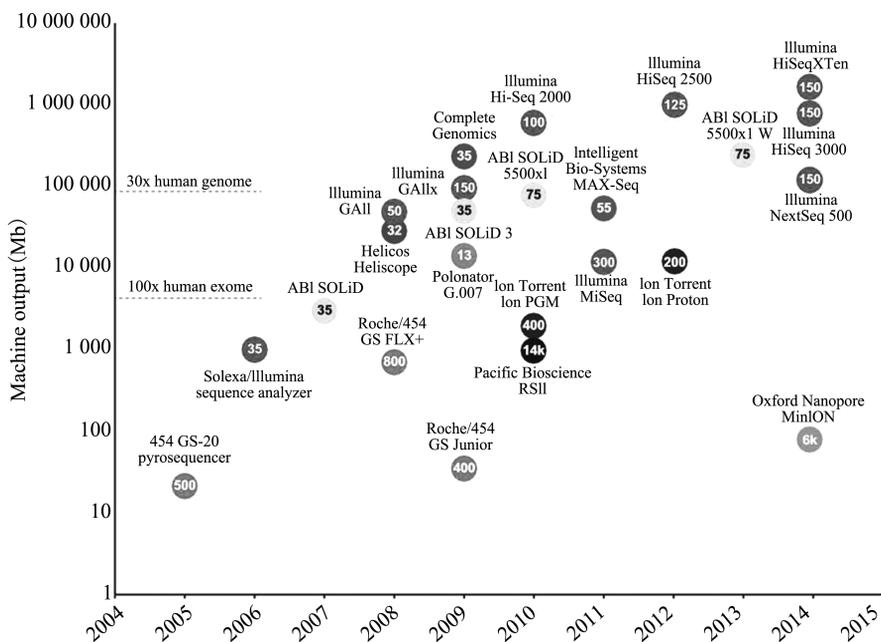


图 1-1.3 高通量测序技术出现年代及其测序通量 (Router 等, 2015)

### 1. Roche 公司的 454 测序技术

2005 年,美国 454 生命科学(Life Sciences)公司提出了焦磷酸测序的方法,并且于 2005 年底推出了基于此方法的高通量基因组测序系统:Genome Sequencer 20 System。2007 年 454 生命科学公司被 Roche 正式收购,并且推出了性能更优的第二代基因组测序系统:Genome Sequencer FLX System。2008 年 10 月,全新的 GS FLX Titanium System,让 GS FLX 的通量一下子提高了 5 倍,准确性和读长也进一步提升。

这三种测序系统的测序原理一样,接下来将详细介绍 GS FLX 的测序原理。

GS FLX 高通量测序技术是一种依靠生物发光进行 DNA 序列分析的方法,在 DNA 聚合酶 (polymerase)、ATP 硫酸化酶 (sulfurylase)、荧光素酶 (luciferase) 和双磷酸酶 (apyrase) 的协同作用下,将引物上每一个 dNTP 聚合与一次荧光信号释放偶联起来,通过检测荧光的释放和强度,达到实时测定 DNA 序列的目的(图 1-1.4)。此技术不需要荧光标记的引物或核酸探针,也不需要进行电泳,具有分析结果快速、准确、灵敏度高和自动化的特点。

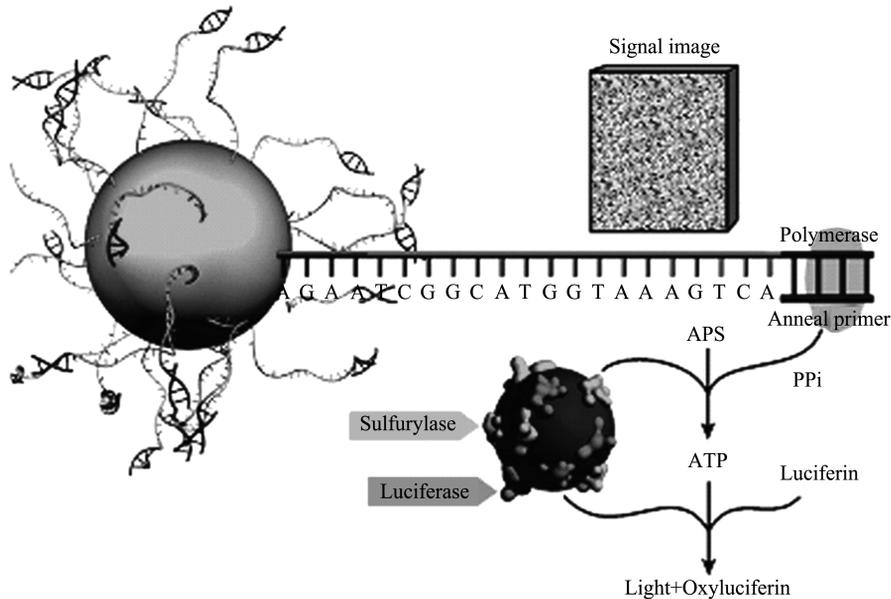


图 1-1.4 454 测序原理简要示意图 (GS FLX) (Mardis, 2008)

图左部分为包含有数百万个来自单个 DNA 片段同源扩增片段的磁珠。具体说明见文中。

GS FLX 高通量测序的具体流程如图 1-1.5 所示。

①DNA 文库制备(图 1-1.5a)。根据样品的来源和实验目的,利用喷雾法将待测 DNA 或者 cDNA 片段打断成 300~800bp 长的小片段;并在这些小片段的两端连上特异性接头,经过提纯可以获得单链模板 DNA (single-stranded template DNA, sstDNA) 文库。

②emPCR, 即 emulsion RCR, 通常被译为乳液 PCR(图 1-1.5b)。特定比例的单链 DNA 文库被固定在特别设计的 DNA 捕获磁珠上(含有与接头互补的 DNA 序列,直径约 28 $\mu\text{m}$ ), 使大部分磁珠携带了一个独特的单链 DNA 片断。

磁珠结合的文库被扩增试剂乳化,形成了油包水的混合物(油为矿物油,水即包含 PCR 所有反应成分的水溶液)。每个独特的片断在自己的微反应器里进行独立的扩增,而不受其他的竞争性或者污染性序列的影响。

整个片段文库的扩增平行进行。扩增后产生了几百万个相同的拷贝。随后,乳液混合物被打破,扩增后仍结合在磁珠上的片断可被回收纯化用于后续的测序实验。

③测序(图 1-1.5c)。测序前需要先用一种聚合酶和单链结合蛋白处理带有 DNA 的磁珠,接着将磁珠放在一种 PTP (picotiter plate) 平板上。这种平板上特制有许多直径约为 44 $\mu\text{m}$  的小孔,每个小孔仅能容纳一个磁珠,通过这种方法来固定每个磁珠的位置,以便检测接下来的测序反应过程。



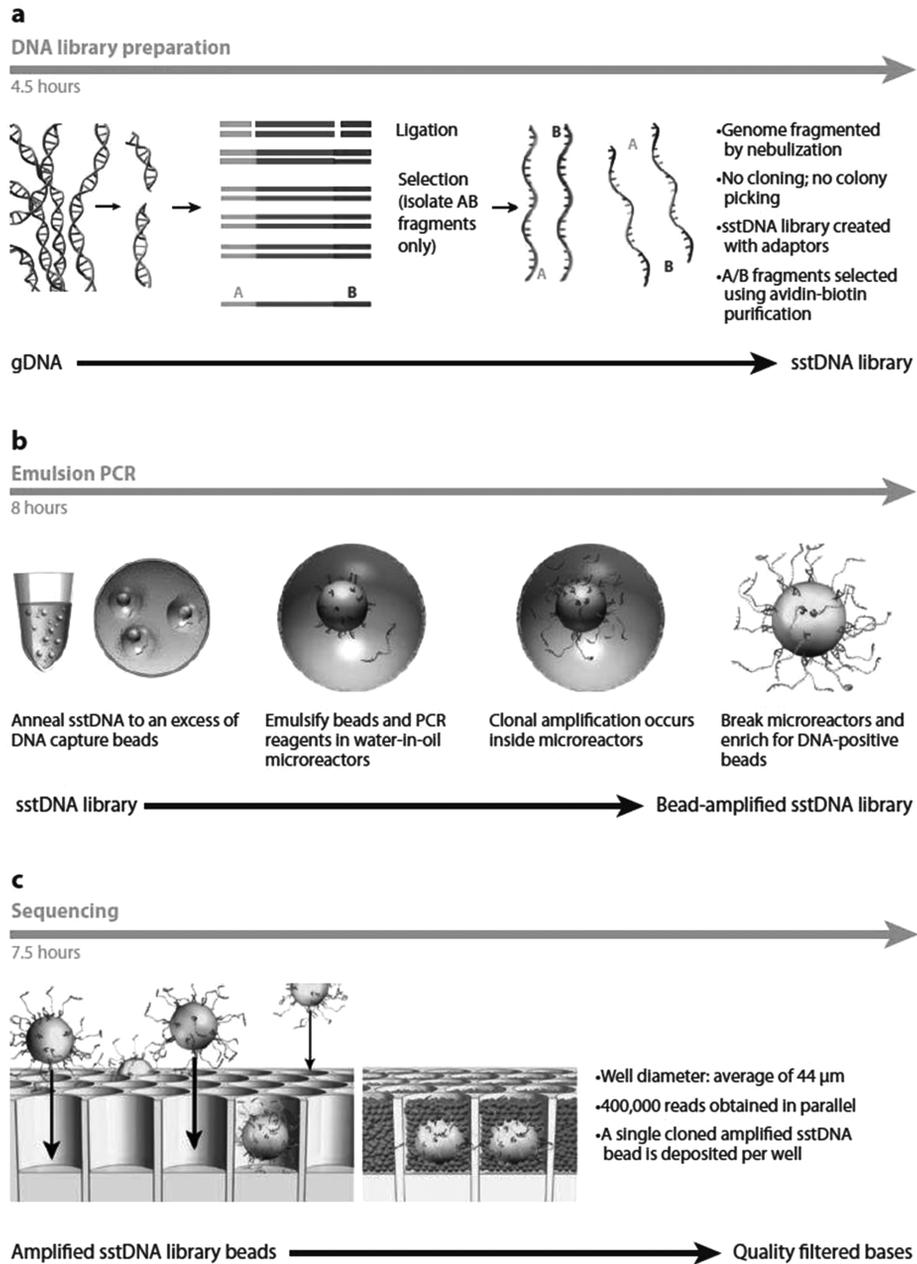


图 1-1.5 454 测序的具体流程(GS FLX) (引自 Mardis, 2008)

a. DNA 文库制备;b. 乳液 PCR(emPCR);c. 测序

测序方法采用的是焦磷酸测序法(pyrosequencing)。测序反应以磁珠上大量扩增出的单链 DNA 为模板,每次反应加入一种 dNTP 进行合成反应。如果 dNTP 能与待测序列配对,则会在合成后释放焦磷酸基团。释放的焦磷酸基团会与反应体系中的 ATP 硫酸化学酶反应生成 ATP。生成的 ATP 和荧光素酶共同氧化使测序反应中的荧光素分子释放并发出荧光,同时由 PTP 板另一侧的 CCD(charge-coupled device,电荷耦合元件)照相机记录,最后通过计算机进行光信号处理而获得最终的测序结果。由于每一种 dNTP 在反应中产生的荧光

颜色不同,因此可以根据荧光的颜色来判断被测分子的序列。反应结束后,游离的 dNTP 会在双磷酸酶的作用下降解 ATP,从而导致荧光淬灭,以便使测序反应进入下一个循环。

④数据分析。GS FLX 系统在 10 小时的运行当中可获得 100 多万个读长,读取超过 4~6 亿个碱基信息,通过 GS FLX 系统提供的生物信息学工具对测序数据进行分析。

454 技术最大的优势在于其能获得较长的测序读长(400bp);但它最主要的一个缺点是无法准确测量同聚物的长度,如当序列中存在类似于 PolyA 的情况时,测序反应会一次加入多个 T,而所加入的 T 的个数只能通过荧光强度推测获得,这就有可能导致结果不准确。也正是由于这一原因,454 技术会在测序过程中引入插入和缺失的测序错误。

## 2. Illumina 公司的 Solexa 和 Hiseq 技术

Illumina 公司创立之初,是一家主要销售微阵列芯片的公司。2006 年,Illumina 宣布收购 Solexa,获得新一代高通量测序技术并开始进军大规模测序市场,渐渐地该测序技术发展成为市场上的主流技术,Illumina 也在测序技术的发展历程中成了不可或缺甚至是推动其不断进步的重要力量。就目前来说,Illumina 公司的 Solexa 和 Hiseq 系统应该是全球使用量最大的第二代测序仪。

Solexa 和 Hiseq 这两个系列的技术核心原理是相同的。这两个系列的测序仪采用的都是边合成边测序(sequencing-by-synthesis, SBS)的方法,测序的具体流程如图 1-1.6。

①DNA 文库制备(图 1-1.6①)。利用超声波把待测的 DNA 片段打断成 200~500bp 长的小片段,并在这些小片段的两端连上特异性接头,构建出单链 DNA 文库。

②Flow Cell 杂交(流动槽杂交,图 1-1.6②)。Flow Cell 的表面结合着一层 oligo 接头,是用于吸附流动 DNA 片段的槽道。当单链 DNA 文库建好后,这些文库中的 DNA 在通过 Flow Cell 的时候会杂交到 Flow Cell 表面的 oligo 引物上。DNA 片段杂交到 Flow Cell 表面后,oligo 引物就会在聚合酶的作用下延伸。

③桥式 PCR 扩增与变性(图 1-1.6③和④)。合成 DNA 双链之后,双链分子变性分开,其中的模板链被洗掉,新合成的单链以共价键的形式紧紧连接在 FlowCell 表面。另一方面,新合成的单链弯曲杂交在相邻的 oligo 引物上形成一个桥式结构。杂交之后,引物在聚合酶的作用下延伸,形成双链的桥式结构。双链的桥式结构变性打开,形成 2 个以共价键结合在 FlowCell 表面的单链模板。单链再弯曲杂交在相邻的 oligo 引物上形成桥式结构,杂交之后再延伸。由此,桥式扩增一直循环重复,直至形成 5 000~10 000 个拷贝。拷贝足够多之后,双链 DNA 桥变性分开,DNA 反链被剪切后洗掉,仅留下由正链组成的簇,并且游离的 3' 端被封闭,防止不必要的 DNA 延伸(图 1-1.6④)。最后,测序引物被杂交到接头序列上,便于之后的测序。

④测序(图 1-1.6⑤)。测序方法采用的是边合成边测序(SBS)的方法。向反应体系中同时添加 DNA 聚合酶、接头引物和带有碱基特异荧光标记的 4 种 dNTP(如同 Sanger 测序法)。这些 dNTP 的 3' 端羟基被化学方法所保护,因而每次只能添加一个 dNTP。在 dNTP 被添加到合成链上后,所有未使用的游离 dNTP 和 DNA 聚合酶会被洗脱掉。接着,再加入激发荧光所需的缓冲液,用激光激发荧光信号,并有光学设备完成荧光信号的记录,最后利用计算机分析将光学信号转化为测序碱基。这样荧光信号记录完成后,再加入化学试剂淬灭荧光信号并去除 dNTP 3' 端羟基保护基团,以便能进行下一轮的测序反应。

Illumina 测序技术每次只添加一个 dNTP 的特点能够很好的地解决同聚物长度的准确

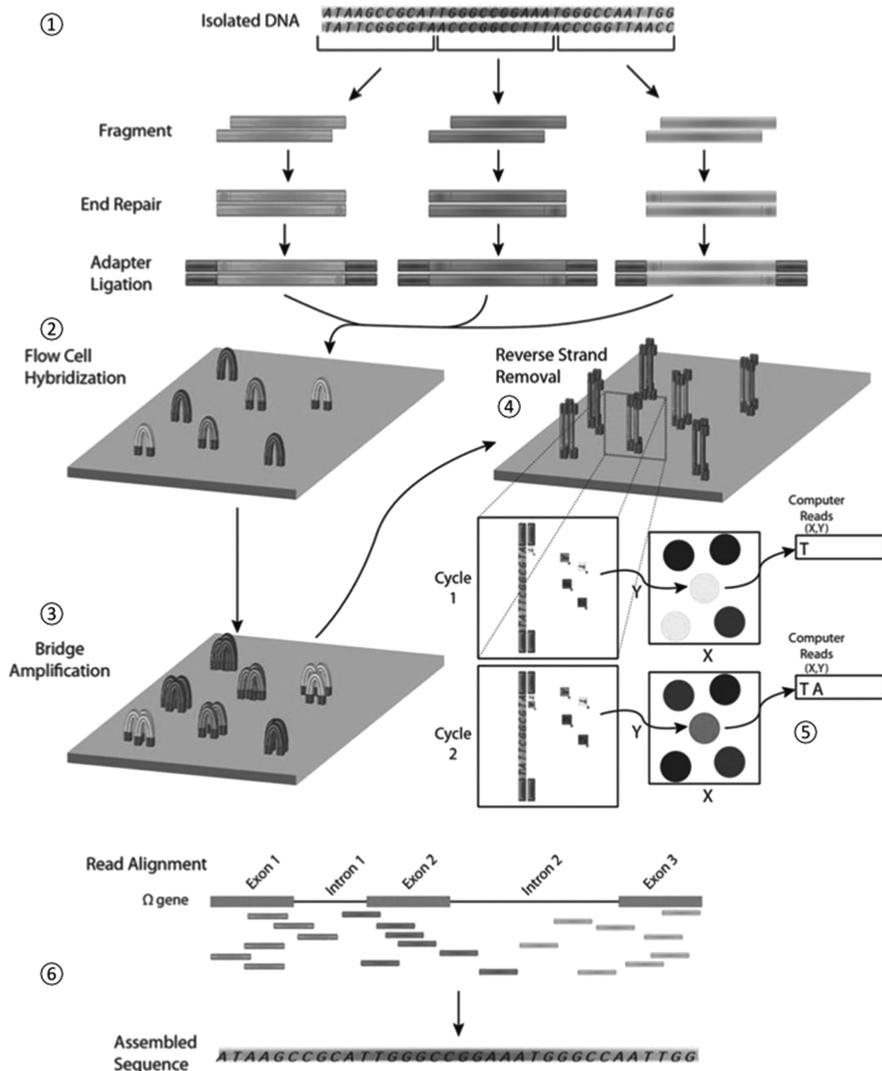


图 1-1.6 Illumina 测序的具体流程

①DNA 文库制备;②Flow Cell 杂交;③桥式 PCR 扩增;④洗掉 DNA 反链;⑤测序;⑥序列组装

测量问题,但它的主要测序错误来源是碱基的替换。

目前,世界上采用最广泛的第二代测序平台之一是 Illumina 公司的 HiSeq 2000 系统。可以说 HiSeq 2000 系统是具有革命性的存在。HiSeq 2000 能够在单次运行中产生 600Gb 的数据,每天最高产生 55Gb。该系统使用两个流动槽和一种新颖的双表面成像方法,能够使测序通量及实验灵活性提升到新的水平。在费用方面,研究人员以 30 倍的覆盖度对两个人类基因组进行测序,每个基因组的费用不到 1 万美元。

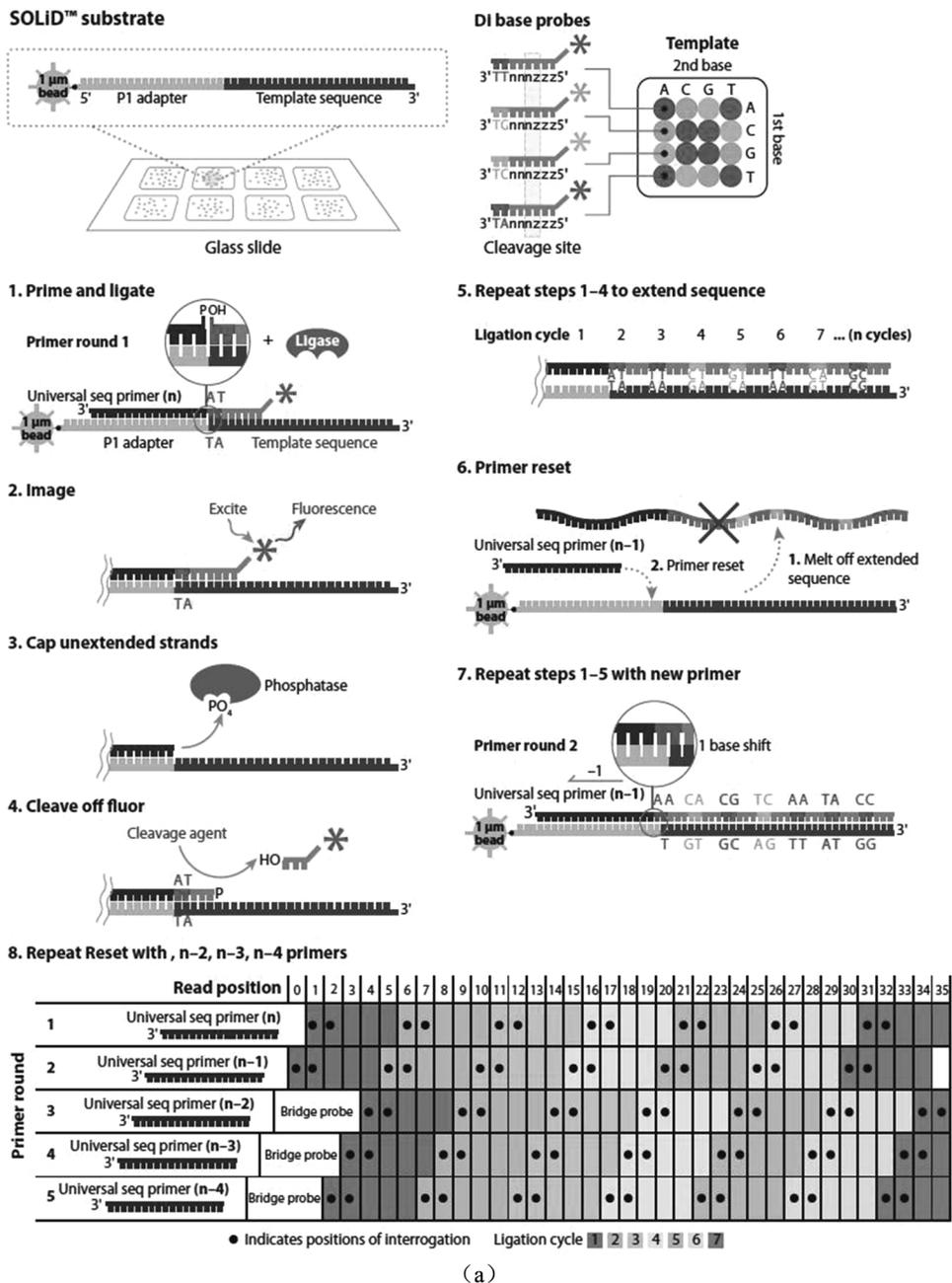
除了 HiSeq 2000 系统,还有 HiSeq 1500/2500 系统是在 HiSeq 2000 系统的基础上优化的仪器。2014 年,Illumina 公司又发布了 HiSeq X Ten 系统,这是针对大规模人群全基因组测序的系统,它可以实现临床期待已久的真正的 1 000 美金人基因组测序。另外, MiSeq 系统是唯一一台在单个仪器上整合了扩增、测序和数据分析的测序仪,每次运行最多能产生超过 7Gb 的数据。革命性的流程和无可比拟的准确性,这让 MiSeq 成为快速高效的测序平台,

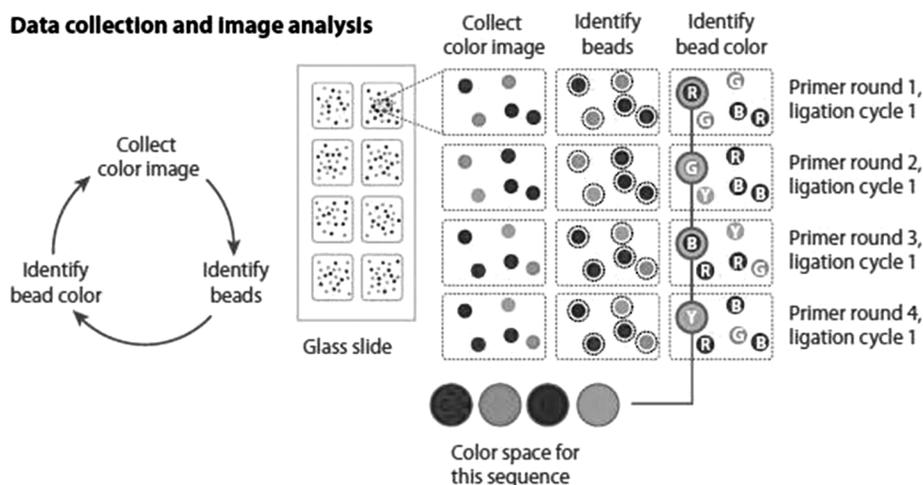
适合广泛的应用。

### 3. ABI 公司的 Solid 技术

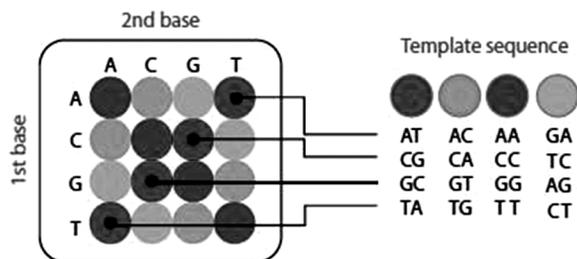
过去几十年间,美国应用生物系统公司(Applied Biosystems, ABI)在测序方面一直占据着垄断地位。直到2005年,454推出了FLX焦磷酸测序平台,ABI的领先地位开始有些动摇。之后,ABI迅速收购了Agencourt Personal Genomics测序公司,并在2007年底推出了SOLiD(supported oligo ligation detetion)新一代测序平台。

Solid 测序技术基于连接酶法,即利用DNA连接酶在连接过程之中测序。它的原理以及具体流程如图1-1.7。



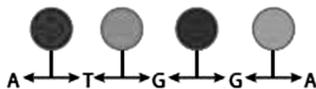


### Possible dinucleotides encoded by each color

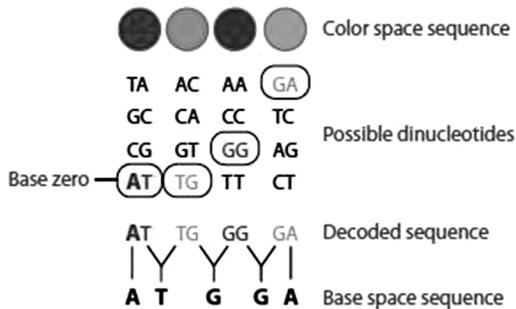


### Double Interrogation

With 2 base encoding each base is defined twice



### Decoding



(b)

图 1-1.7 SOLiD 系统测序原理(引自 Mardis, 2008)

a. 测序过程; b. 数据收集和图像分析。该图下半部分展示了双碱基编码技术, 具体描述可见正文部分。

①DNA 文库制备。SOLiD 系统能支持两种测序模板: 片段文库(fragment library)或配对末端文库(mate-paired library)。片段文库就是将基因组 DNA 打断, 两头加上接头, 制成文库。片段文库适用于转录组测序、RNA 定量、miRNA 探索、重测序、3'/5'-RACE、甲基化分

析、ChIP 测序等。配对末端文库是将基因组 DNA 打断后,与中间接头连接,再环化,然后用 EcoP15 酶切,使中间接头两端各有 27bp 的碱基,再加上两端的接头,形成文库。配对末端文库适用于全基因组测序、SNP 分析、结构重排/拷贝数等。

②乳液 PCR。SOLiD 系统的 emPCR 过程与 454 的 GS FLX 系统基本相同。在微反应器中加入测序模板、PCR 反应元件、微珠和引物,进行乳液 PCR。PCR 完成之后,变性模板,富集带有延伸模板的微珠,去除多余的微珠。微珠上的模板经过 3' 修饰,可以与玻片共价结合,沉积在玻片上。这些微珠与 454 系统的相比要小得多,只有 1 $\mu$ m。SOLiD 系统最大的优点就是每张玻片能容纳更高密度的微珠,在同一系统中轻松实现更高的通量。

③连接酶测序。这一步是 SOLiD 系统区别于其他测序技术的独特之处。它的独特之处在于没有采用惯常的聚合酶,而用了连接酶。SOLiD 连接反应的底物是 8 碱基单链荧光探针混合物(这里将其简单表示为 3'XXnnzzz5')。连接反应中,这些探针按照碱基互补规则与单链 DNA 模板链配对。探针的 5' 末端分别标记了 CY5、Texas Red、CY3、6-FAM 这 4 种颜色的荧光染料(图 1-1.7 上双碱基编码矩阵规定了该编码区 16 种碱基对和 4 种探针颜色的对应关系)。探针 3' 端 1~5 位为随机碱基,可以是 ATCG 四种碱基中的任何一种碱基,其中第 1、2 位构成的碱基对是表示探针染料类型的编码区,而 3~5 位的“n”表示随机碱基,6~8 位的“z”指的是可以和任何碱基配对的特殊碱基。

单向 SOLiD 测序包括五轮测序反应,每轮测序反应含有多次连接反应。第一轮测序的第一次连接反应由连接引物“n”介导,由于每个磁珠只含有均质单链 DNA 模板,所以这次连接反应掺入一种 8 碱基荧光探针,SOLiD 测序仪记录下探针第 1、2 位编码区颜色信息,随后的化学处理断裂探针 3' 端第 5、6 位碱基间的化学键,并除去 6~8 位碱基及 5' 末端荧光基团,暴露探针第 5 位碱基 5' 磷酸,为下一次连接反应作准备。因为第一次连接反应使合成链多了 5 个碱基,所以第二次连接反应得到模板上第 6、7 位碱基序列的颜色信息,而第三次连接反应得到的是第 11、12 位碱基序列的颜色信息……通过这种测序方法,每次测序的位置都相差 5 位(图 1-1.7a 第 8 幅小图中的引物 n)。

几个循环之后,引物重置,开始第二轮的测序。由于第二轮连接引物 n-1 比第一轮错开一位,所以第二轮得到以 0、1 位起始的若干碱基对的颜色信息(图 1-1.7a 第 8 幅小图中的引物 n-1)。第二轮测序完成,依此类推,直至第五轮测序,最终可以完成所有位置的碱基测序,并且每个位置的碱基均被检测了两次。

由于 SOLiD 系统采用了双碱基编码技术,在测序过程中对每个碱基判读两遍,从而减少原始数据错误,提供内在的校对功能。因此该测序技术是目前新一代基因分析技术中准确度最高的。但由于双碱基编码规则中双碱基与颜色信息的简并特性(一种颜色对应 4 种碱基对),前面碱基的颜色编码直接影响紧跟其后碱基的解码,所以一个错误颜色编码就会引起“连锁解码错误”,改变错误颜色编码之后的所有碱基。

#### 4. 三种高通量测序技术的比较

以上介绍的三种测序技术是目前使用最多的三类第二代测序技术。虽然分别来自不同的测序公司,但是他们在原理上还是有很多共同之处,例如:将待测 DNA 片段打断成小片段;单个小片段 DNA 分子结合到固相表面;单分子独立扩增;每次只复制一个碱基并检测信号;高分辨率的成像系统等。

另一方面,这三种测序方法也各具特色,表 1-1.4 详细比较了这三种测序技术的不同

同点。

表 1-1.4 三种第二代测序技术的不同之处

测序技术	测序方法	大约读长 (碱基数)	相对优势和局限性以及应用
Roche 公司 454 测序技术	焦磷酸测序法	400	<p>优势:在第二代中最高读长;比第一代的测序通量大,成本低;准确度高。</p> <p>局限性:样品制备较难;难于处理重复和同种碱基多聚区域;试剂冲洗带来错误累积;仪器昂贵。</p> <p>应用:适合 <i>de novo</i> 测序、转录组测序、宏基因组研究等。</p>
Illumina 公司 Solexa 和 HiSeq 技术	合成测序法	125~150	<p>优势:很高测序通量;广泛的应用灵活性;高质量数据。</p> <p>局限性:仪器昂贵;数据分析较困难,且费用高。</p> <p>应用:适用于基因组、表观基因组和转录组研究。</p>
ABI 公司 Solid 技术	连接测序法	50~75	<p>优势:很高测序通量;在广为接受的几种第二代平台中,准确度最高。</p> <p>局限性:测序运行时间长;读长短,数据分析困难和基因组拼接困难;仪器昂贵。</p> <p>应用:适于基因组重测序和 SNP 检测。</p>

### 三、第三代测序技术

第二代测序技术具有通量大、时间短、精确度高和信息量丰富等优点,但是仍不能满足日益深入的研究工作,因此第三代测序技术应运而生。

第三代测序(third generation sequencing, TGS)是基于单个分子信号检测的 DNA 测序,也被称为单分子测序(single molecule sequencing, SMS)。目前,第三代测序的新技术包括 Helicos 的 tSMS、PacBio 的 SMRT、Oxford 的 Nanopore 以及其它一些尚处于实验室阶段的技术,如电镜测序、蛋白质晶体管测序等。

#### 1. tSMS(true single molecule sequencing)

美国 Helicos Bioscience 于 2008 年推出 HeliScope 单分子测序平台,该平台被认为是第一个商品化的第三代测序仪。其测序原理 tSMS 是一种利用光学信号进行 DNA 碱基识别的边合成边测序技术,与二代测序中的 Illumina 公司的 Solexa 测序有类似之处,但该技术无需对样本进行 PCR 扩增,简化了测序文库的构建过程,也避免了 DNA 扩增中出现的错误。

HeliScope 的文库制备相对简单,首先将待测 DNA 随机打断成约 200bp 大小的片段,然后在 3' 末端加上 50bp 带有荧光标记的 polyA 尾巴。文库退火形成单链,与芯片上固定的 Oligo dT 探针结合,利用 polyA 上的荧光标记进行精确定位。接下来依次加入 4 种 Cy5 荧光染料标记的单核苷酸,在 DNA 聚合酶的作用下与模板互补配对并延伸一个碱基,ICCD (Intensified CCD,增强电荷耦合元件)相机采集荧光信号。最后通过化学剪切去除荧光基团并清洗,进行下一轮反应。原理如图 1-1.8(左)。

tSMS 与 Illumina 测序原理较为相似,所不同的在于 tSMS 采集的是一条 DNA 模板合成时所发出的荧光,而 Illumina 检测的信号来自于桥式 PCR 扩增得到的 DNA 簇合成时发出的

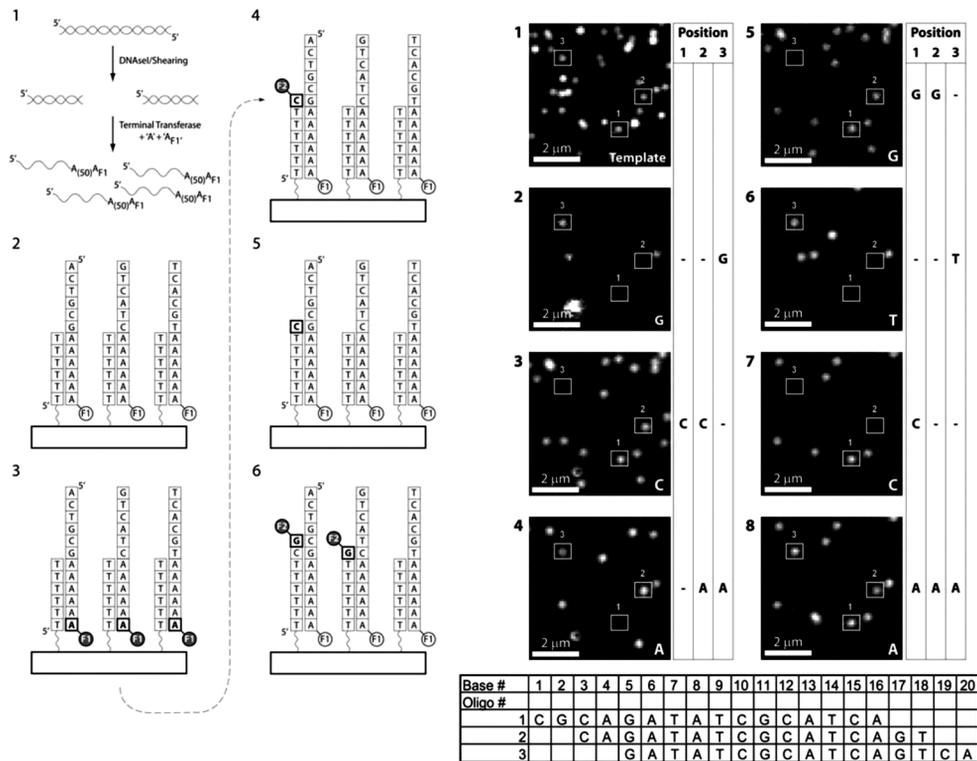


图 1-1.8 HeliScope 测序原理(左)和成像过程(右)(引自 Harris 等,2008)

右图 1-8 均为同一位置的成像结果。面 1 是利用 polyA 上的荧光标记进行定位,每一个光点均代表一条被固定在芯片上的文库模板。面 2-8 显示了 7 轮反应的结果,依次加入 G-C-A-G-T-C-A,根据图像可识别出位置 1 延伸的前 4 个碱基是 5'-CGCA-3'。

荧光。因此,tSMS 技术能够实现单分子测序,主要依赖于高分辨率的 ICCD 相机,能够对单个分子产生的荧光信号进行识别。但是较弱的信号强度导致测序的读长较短,错误率偏高,尽管通过两次测序能够降低错误率,但同时也提高了测序成本和运行时间。HeliScope 可同时运行两个芯片,平均读长约为 35bp,一次运行的数据产量可达 30Gb 左右。该测序仪的售价和运行成本相对较高,一个人类基因组的测序成本约为 5 万美元。不过,该公司由于经营不善等原因目前已经破产了。

## 2. SMRT(single molecule real-time)sequencing

美国 Pacific Biosciences 公司提出的单分子实时测序是采用四色荧光标记的 dNTP 和被称为零级波导(zero-mode waveguides, ZMW)的纳米结构对单个 DNA 分子进行测序。这些 ZMW 是直径 50~100nm、深度 100nm 的孔状纳米光电结构,通过微加工在二氧化硅基质的金属铝薄层上形成微阵列,光线进入 ZMW 后会呈指数级衰减,从而使得孔内仅有靠近基质的部分被照亮。DNA 聚合酶被固定在 ZMW 的底部,模板和引物结合之后被加到酶上,再加入四色荧光标记的 dNTP。当 DNA 合成进行时,连接上的 dNTP 由于在 ZMW 底部停留的时间较长(约 200ms),其荧光信号能够与本底噪音区分开来,从而被识别。荧光基团被连接在 dNTP 的磷酸基团上,因此在延伸下一个碱基时,上一个 dNTP 的荧光基团被切除,从而保证了检测的连续性,提高了检测速度。SMRT 的测序原理如图 1-1.9。



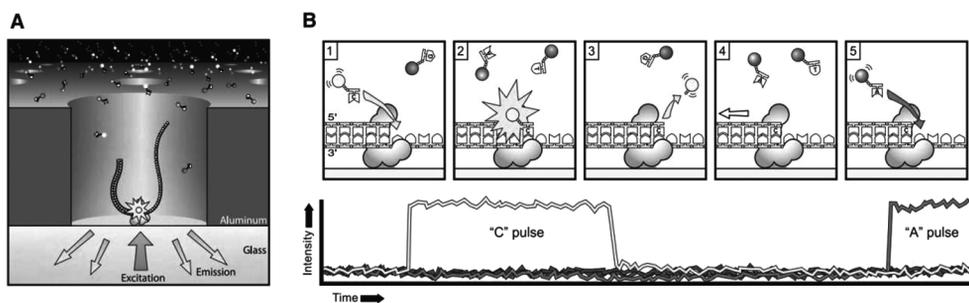


图 1-1.9 SMRT 的测序原理(引自 Eid 等, 2009)

图 A 显示的是 ZMW 的结构, 激光进入 ZMW 后呈指数级衰减, 仅能照亮靠近底部的约 30nm 区域, 因此大部分游离的荧光标记 dNTP 不会被激发, 只有结合到 DNA 聚合酶上的 dNTP 其荧光基团被激光照亮, 激发荧光。图 B 显示的是 DNA 合成过程中检测到的荧光信号及持续时间。结合到酶上的 dNTP 停留时间较长, 信号呈脉冲式激发, 因而能够与噪音区分。

SMRT 的一大优势是超长的读长, PacBio RS II 测序平台能够得到的最大读长为 30Kb, 平均读长约 8.5Kb, 是目前所有商品化测序仪中读长最长的。但是, 与 tSMS 类似, 因为单分子的荧光信号较弱, SMRT 的单碱基准确率仅有 87.5%, 但由于错误是随机产生的, 通过多重测序和校正, 在 10 倍覆盖度的条件下, 准确率可提高到 99.9%。

### 3. Nanopore

利用纳米孔进行核酸序列的识别在 20 世纪 90 年代已有报道。其基本原理为: 当单链 DNA 或 RNA 分子经过纳米级的小孔时, 由于碱基形状大小不同, 引起孔内电阻变化, 在小孔两端保持一个恒定的电压, 则能够检测到通过小孔的电流变化情况, 通过测到这些特征电流, 就能够识别出通过小孔的 DNA 分子上的碱基排列。该方法具有检测速度快, 成本低, 准确率高等特点, 但也面临 DNA 易位速率过快, 电流变化幅度较小, 制备纳米孔材料的稳定性等问题。

英国 Oxford Nanopore Technologies 公司应用这一原理开发了两种纳米孔测序技术: 外切酶测序(exonuclease sequencing)和链测序(strand sequencing)。外切酶测序是将  $\alpha$ -溶血素和环化糊精组成的纳米孔固定在脂质双分子膜上, 两侧为浓度不同的 KCl 溶液, 并加以 160mV 的电压。DNA 单链在 *E. coli* 核酸外切酶 I 的作用下被依次剪切为单核苷酸, 通过记录单核苷酸分子经过纳米孔时引起的电流变化进行 DNA 测序(图 1-1.10)。链测序则是利用 DNA 解旋酶将 DNA 双链解旋为单链, 并通过纳米孔, 进行连续测序。

纳米孔测序的主要特点是: 读长很长, 大约在几十至 100kb; 错误率目前介于 1% 至 4%, 且是随机错误, 而不是聚集在读序的两端; 数据可实时读取; 通量很高; 起始 DNA 在测序过程中不被破坏; 样品制备简单又便宜。理论上, 它也能直接测序 RNA。另外, 纳米孔测序还有另一特点, 它能够直接读取甲基化的胞嘧啶, 而不必像传统方法那样对基因组进行重亚硫酸盐(bisulfite)处理。这对于在基因组水平直接研究表观遗传相关现象有极大的帮助, 并且该方法的测序准确性可达 99.8%, 而且一旦发现测序错误也能较容易地进行纠正。

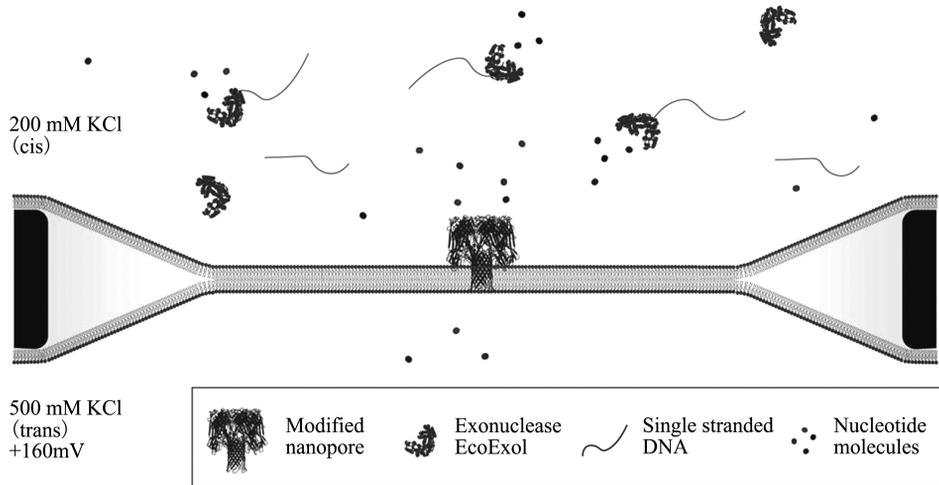


图 1-1.10 纳米孔测序原理(外切酶测序)(Clarke 等,2009)

脂质双分子膜两侧有浓度不同的 KCl 溶液,并有 160mV 电压。DNA 单链在外切酶作用下被剪切为单核苷酸,单核苷酸通过纳米孔时引起电流变化,据此可进行 DNA 测序。

#### 4. 其他技术

利用电子显微镜对 DNA 分子直接进行测序的设想在 20 世纪六七十年代已经出现并经过实验验证,其主要原理是利用重原子如汞、钨、金等对单核苷酸进行标记,PCR 得到带标记的 DNA 双链,将之固定在基质上并拉伸成直线,用透射电镜(transmission electron microscope, TEM)观察,从而利用重原子标记识别 DNA 上的碱基排列。TEM 测序的优势在于长片段读取(10-20Kb),且没有光学信号错误,但重原子标记是其目前亟需解决的问题,PCR 过程可能引入错误的标记信号,同时 TEM 的使用和维护成本较高也是制约该技术的主要因素。

2013 年台湾国立交通大学的 Chen 等(2013)利用 DNA 聚合酶、抗体和纳米金颗粒制备蛋白质晶体管(protein transistor),并用于 DNA 测序的研究。研究人员首先在硅基质上制备了一对宽 50nm,间距 10nm 的金电极,在原子力显微镜(atomic force microscope, AFM)下将两个直径 5nm 的金颗粒分别放到两个电极的边缘,最后采用微流控的方法将抗体和 DNA 聚合酶自组装到预定位置,组成单分子蛋白质晶体管。通过检测 DNA 聚合酶合成 DNA 时电导率的变化,识别出 DNA 序列。据报道,利用该结构的蛋白质晶体管,能够检测到 DNA 聚合酶在合成 DNA 时由于电导率变化引起的电流改变,变化幅度约为 3pA,且不同 dNTP 合成时会产生不同的电流特征,因此可以利用这一现象区分 DNA 的碱基排列。该方法的测序速率约为每秒 22nt,单碱基准确度达到 99.99%以上。

### 第三节 高通量测序技术的应用

#### 一、DNA/RNA 相关测序

##### 1. 基因组及基因组重测序

基因组是一个细胞或者生物体所携带的一套完整的单倍体序列,包括编码序列和非编码

序列在内的全部 DNA 分子。核基因组是单倍体细胞核内的全部 DNA 分子,线粒体基因组是一个线粒体所包含的全部 DNA 分子,叶绿体基因组则是一个叶绿体所包含的全部 DNA 分子。

基因组重测序是对已知基因组序列的物种进行不同个体的基因组测序,并在此基础上对个体或群体进行差异性分析。基因组重测序的个体,通过序列比对,可以找到大量的单核苷酸多态性(SNP)位点、插入/缺失(Indel)位点、结构变异(SV,包括染色体的缺失、重复、倒位、易位)位点和基因拷贝数变异(CNV,也称拷贝数目多态,是一种大小介于 1Kb 至 3Mb 的 DNA 片段的变异)等。

## 2. 目标区域捕获测序

目标序列捕获测序,其原理是将感兴趣的基因组区域定制成特异性探针与基因组 DNA 在序列捕获芯片(或溶液)进行杂交,然后将目标基因组区域的 DNA 片段进行富集后,再利用第二代测序技术进行测序。目标序列捕获测序的具体做法是将基因组 DNA 随机打断成片段,通过与序列捕获芯片上特异性探针杂交,以富集目标基因组区域,然后通过第二代测序技术对捕获的序列进行测序。

目前的序列捕获系统有:

① NimbleGen Sequence Capture Array,固相捕获,385K 序列捕获芯片可捕获多达 5Mb 的基因组区域,2.1M 序列捕获芯片可捕获多达 30Mb 的基因组区域,序列捕获后的测序平台为 Solexa。

② Agilent SureSelect DNA Capture Array,固相捕获,120-mer 探针,244K 序列捕获芯片可捕获多达 1287 个独立的基因组区域,序列捕获后的测序平台为 Solexa/SOLiD/Roche 454。

③ Agilent SureSelect Target Enrichment System,120-mer 探针,可根据 eArray 在线设计工具自行设计 SureSelect 探针混合物,杂交过程发生在液体环境中,序列捕获后的测序平台为 Solexa/SOLiD/Roche 454。

相对于以往只能将基因突变锁定在染色体某一片段区域的检测方法,目标区域测序技术是一个非常好的进一步检测手段。目标区域测序技术可以将经过连锁分析锁定了目标范围或经过全基因组筛选的特定基因或区域进行更深一层的研究,是解决连锁分析无法发现致病基因的有效手段。目标区域测序技术对于已知基因突变的筛查具有明显优势,可以快速、全面地检测出目标基因突变。同时,由于目标区域受到了限制,测序范围大幅度减少,测序时间和费用相应降低。

## 3. 转录组及表达谱测序

转录组测序(RNA-Seq)的研究对象为特定细胞在某一功能状态下所能转录出来的所有 RNA 的总和,主要包括 mRNA 和非编码 RNA(non-coding RNA)。

表达谱测序是直接对某一物种或特定细胞在某一功能状态下产生的 mRNA 进行高通量测序,可以用来研究基因的表达差异情况。该技术结合了转录组测序建库的实验方法,与转录组测序相比,基因表达谱测序要求的读长更短,测序通量更小,但仅可用于基因表达差异的研究。

## 4. 小 RNA 测序

小 RNA(small RNA)是一类长度在 20~30nt 的 RNA 分子,主要包括 miRNA(micro RNA)、siRNA(small interfering RNA)等,参与调控基因表达、生长发育、非生物胁迫和病原菌

的侵害等过程。小 RNA 测序是对目标物种小 RNA 进行大规模测序分析,能够快速全面地鉴定该物种在特定状态下的小 RNA 和发现新的小 RNA。小 RNA 测序为研究小 RNA 的种类、结构和功能及此物种的基因调控机制提供了有力工具。

### 3. 降解组测序

降解组测序(degradome sequencing)主要针对 miRNA 介导的剪切降解片段进行测序,从实验中筛选 miRNA 作用的靶基因,并结合生物信息学分析,确定降解片段与 miRNA 精确的配对信息。

降解组测序的原理:绝大多数的 miRNA 是利用剪切作用调控靶基因的表达,且剪切常发生在 miRNA 与 mRNA 互补区域的第 10-11 位核苷酸上。靶基因经剪切产生两个片段,5' 剪切片段和 3' 剪切片段。其中 3' 剪切片段,包含有自由的 5' 单磷酸和 3' polyA 尾巴,可被 RNA 连接酶连接,连接产物可用于下游高通量测序;而含有 5' 帽子结构的完整基因,或是含有帽子结构的 5' 剪切片段或是其他缺少 5' 单磷酸基团的 RNA 是无法被 RNA 酶连接,因而无法进入下游的测序实验。最后对测序数据进行深入地比对分析,可以直观地发现在 mRNA 序列的某个位点会出现一个波峰,而该处正是候选的 miRNA 剪切位点。

## 二、蛋白质-DNA/RNA 互作测序

### 1. ChIP-Seq

研究蛋白质与 DNA 互作的主要技术是染色质免疫共沉淀技术(chromatin Immunoprecipitation, ChIP),也称结合位点分析法。该技术通常用于转录因子结合位点或者组蛋白特异性修饰位点的研究。将 ChIP 与第二代测序技术相结合的 ChIP-Seq 技术,能够高效地在全基因组范围内检测与组蛋白、转录因子等互作的 DNA 区段。

ChIP-Seq 的原理是:首先通过染色质免疫共沉淀技术(ChIP)特异性地富集目的蛋白结合的 DNA 片段,并对其进行纯化与文库构建;然后对富集得到的 DNA 片段进行高通量测序。通过将获得的数百万条序列标签精确定位到基因组上,从而获得全基因组范围内与组蛋白、转录因子等互作的 DNA 区段信息。

### 2. CLIP-Seq

CLIP-Seq(cross-linking immunoprecipitation and high-throughput sequencing),即紫外交联免疫沉淀结合高通量测序,是一项在全基因组水平揭示 RNA 分子与 RNA 结合蛋白相互作用的技术。其主要原理是基于 RNA 分子与 RNA 结合蛋白在紫外照射下发生耦联,以 RNA 结合蛋白的特异性抗体将 RNA-蛋白质复合体沉淀之后,回收其中的 RNA 片段,经添加接头、RT-PCR 等步骤,对这些分子进行高通量测序,再经生物信息学的分析和处理、总结,挖掘出其特定规律,从而深入揭示 RNA 结合蛋白与 RNA 分子的调控作用及其对生命的意义。

## 三、甲基化/宏基因组测序

### 1. MeDIP-Seq

MeDIP-Seq(methylated DNA immunoprecipitation sequencing),即甲基化 DNA 免疫共沉淀测序,是基于免疫富集原理进行的全基因组 DNA 甲基化研究方法,它是通过 5'-甲基胞

嘧啶(5mC)抗体将基因组中的 DNA 甲基化区域富集后进行二代高通量测序,从而检测基因组上高 CpG 区域的甲基化位点。通过 MeDIP-Seq 技术可以快速准确地寻找样品间相对 DNA 甲基化差异区域,进行不同细胞、组织及疾病样本间的 DNA 甲基化修饰模式的差异分析,并可对发现的靶点区进行甲基化特异性 PCR 验证,特别适用于大样本量的疾病表观研究。该方法可以高效、经济地比较大样本量细胞、组织等样品间的相对甲基化差异,且成本较低,适合多样品 DNA 表观遗传学的研究。

## 2. Bisulfite-Seq

Bisulfite 处理作为表观遗传学研究的经典方法,能够将基因组中未发生甲基化的 C 碱基转换成 U,经 PCR 扩增后变成 T,与原本具有甲基化修饰的 C 碱基区分开来。Bisulfite-Seq (bisulfite sequencing,也称 bisulphite sequencing)即是将 Bisulfite 处理与高通量测序技术相结合,从而绘制单碱基分辨率的 DNA 甲基化图谱,用于研究特定 DNA 区域甲基化与特定表型之间的关联,为疾病发生、治疗相关的研究提供研究基础。

## 3. 宏基因组测序

宏基因组是指特定环境中全部生物(微生物)遗传物质的总和。宏基因组测序即利用测序技术对环境样品中全部微生物的基因组进行测定,以分析微生物群体的基因组成及功能,解读微生物群体的多样性和丰度,探索微生物与环境及宿主之间的关系,发掘和研究新的具有特定功能的基因等。

目前,第二代高通量测序技术在宏基因组的研究上已被广泛应用。与传统方法相比,基于高通量测序的宏基因组研究无需构建克隆文库,这避免了文库构建过程中利用宿主菌对样品进行克隆而引起的系统偏差,简化了实验操作,提高了测序效率,从而极大地促进了宏基因组学的发展。详细介绍见第 2-6 章。

# 第四节 蛋白质序列及其结构测定

由于蛋白质相关内容将在第 2-7 章详细阐述,本节仅对相关概念进行简单地阐述。

## 一、蛋白质序列与蛋白质互作测定

### 1. 蛋白质序列测定

测定蛋白质序列,常用的是蛋白质谱技术。蛋白质谱技术简单来说就是一种将质谱仪用于研究蛋白质的技术。目前,它的基本原理是蛋白质经过蛋白酶的酶切消化后成肽段混合物,在质谱仪中肽段混合物电离形成带电离子,质谱分析器的电场、磁场将具有特定质量与电荷比值(即质荷比,  $M/Z$ )的肽段离子分离开来,经过检测器收集分离的离子,确定每个离子的  $M/Z$  值。经过质量分析器可分析出每个肽段的  $M/Z$ ,得到蛋白质所有肽段的  $M/Z$  图谱,即蛋白质的一级质谱峰图。离子选择装置自动选取强度较大肽段离子进行二级质谱分析,输出选取肽段的二级质谱峰图,通过一级质谱峰图和二级质谱峰图进行比对可鉴定蛋白质。

### 2. 蛋白质互作测定

#### 1) 酵母双杂交系统(yeast two hybrid)

酵母双杂交系统是当前广泛用于蛋白质相互作用组学研究的一种重要方法。其原理是当靶蛋白和诱饵蛋白特异结合后,诱饵蛋白结合于报道基因的启动子,启动报道基因在酵母细胞内的表达,如果检测到报道基因的表达产物,则说明两者之间有相互作用,反之则两者之间没有相互作用。将这种技术微量化、阵列化后则可用于大规模蛋白质之间相互作用的研究。在实际工作中,人们根据需要还发展了单杂交系统、三杂交系统和反向杂交系统等。

#### 2) 噬菌体展示技术(phage display technology)

在编码噬菌体外壳蛋白基因上连接一单克隆抗体的 DNA 序列,当噬菌体生长时,表面就表达出相应的单抗,再将噬菌体过柱,柱上若含目的蛋白,就会与相应抗体特异性结合,这被称为噬菌体展示技术。此技术不仅有高通量及简便的特点,还具有直接得到基因、高选择性的筛选复杂混合物、在筛选过程中通过适当改变条件可以直接评价相互结合的特异性等优点。

目前,用优化的噬菌体展示技术,已经展示了人和鼠的两种特殊细胞系的 cDNA 文库,并分离出了人上皮生长因子信号传导途径中的信号分子。

#### 3) 表面等离子共振技术(surface plasmon resonance, SPR)

表面等离子共振技术已成为蛋白质相互作用研究中的新手段。它的原理是利用一种纳米级的薄膜吸附上诱饵蛋白,当待测蛋白与诱饵蛋白结合后,薄膜的共振性质会发生改变,通过检测便可知这两种蛋白的结合情况。SPR 技术的优点是不需标记物或染料,反应过程可实时监控。测定快速且安全,还可用于检测蛋白与核酸及其它生物大分子之间的相互作用。

#### 4) 荧光共振能量转移(fluorescence resonance energy transfer, FRET)

荧光共振能量转移是指在两个不同的荧光基团中,如果一个荧光基团(donor)的发射光谱与另一个基团(acceptor)的吸收光谱有一定的重叠,当这两个荧光基团间的距离合适时(一般小于  $100\text{\AA}$ ),就可观察到荧光能量由供体向受体转移的现象,即前一种基团的激发波长激发时,可观察到后一个基团发射的荧光。

荧光共振能量转移广泛用于研究分子间的距离及其相互作用,与荧光显微镜结合,可定量获取有关生物活体内蛋白质、脂类、DNA 和 RNA 的时空信息。随着绿色荧光蛋白(GFP)的发展,FRET 荧光显微镜有可能实时测量活体细胞内分子的动态性质。

#### 5) 蛋白芯片技术

蛋白芯片技术的出现给蛋白质组学研究带来新的思路。蛋白质组学研究中一个主要的内容就是研究在不同生理状态下蛋白水平的量变,微型化、集成化、高通量化的抗体芯片就是一个非常好的研究工具,它也是芯片中发展最快的一种,而且在技术上已经日益成熟。这些抗体芯片有的已经在向临床应用上发展,比如肿瘤标志物抗体芯片等。

#### 6) 免疫共沉淀技术(co-immunoprecipitation)

免疫共沉淀是以抗体和抗原之间的专一性作用为基础的用于研究蛋白质相互作用的经典方法,是确定两种蛋白质在完整细胞内生理性相互作用的有效方法。用免疫共沉淀方法得到的目的蛋白是在细胞内与兴趣蛋白天然结合的,符合体内实际情况,得到的结果可信度高。这种方法常用于测定两种目标蛋白质是否在体内结合;也可用于确定一种特定蛋白质新的作用搭档。

### 7) GST Pull-down 技术

GST Pull-down 技术又叫做蛋白质体外结合实验(binding assay in vitro),是一种在试管中检测蛋白质之间相互作用的方法(这里 GST 为谷胱甘肽硫基转移酶)。其基本原理是将靶蛋白-GST 融合蛋白亲和和固化在谷胱甘肽亲和树脂上,作为与目的蛋白亲和的支撑物,充当一种“诱饵蛋白”,目的蛋白溶液过柱,可从中捕获与之相互作用的“捕获蛋白”(目的蛋白),洗脱结合物后通过 SDS-PAGE 电泳分析,从而证实两种蛋白间的相互作用或筛选相应的目的蛋白。“诱饵蛋白”和“捕获蛋白”均可通过细胞裂解物、纯化的蛋白、表达系统以及体外转录翻译系统等方法获得。此方法简单易行,操作方便。

## 二、蛋白质结构测定

在蛋白质结构数据中,接近 90%的蛋白质结构是用 X 射线晶体衍射学的方法测定的。大约 9%的已知蛋白结构是通过核磁共振技术来测定的。下面将详细介绍这两种方法。

### 1. X 射线晶体衍射

X 射线晶体衍射可以利用电子对 X 射线的散射作用,获得晶体中电子密度的分布情况,再从中分析获得原子的位置信息,即晶体结构。

利用 X 射线测定蛋白质结构的具体做法是:获得可供衍射的蛋白质单晶之后,用 X 射线打到晶体上,产生衍射,并记录衍射数据。衍射数据(包括衍射点的位置和强度)的记录多采用像板或 CCD 探测器,通过对衍射数据的分析可得晶体结构。

### 2. 核磁共振

在强磁场中,原子核发生自旋能级分裂,当吸收外来电磁辐射时,将发生核自旋能级的跃迁,产生核磁共振现象。核自旋能级的共振频率和原子的类型有关,且受到所处化学结构微环境的影响。由此核磁共振能够提供分子的结构信息,并被广泛应用于化学领域的研究。

## 习 题

1. 生物信息的类型有哪些?
2. 简述 Sanger 法测序的原理。
3. 第二代测序有哪些主流的测序技术,简述这些测序技术的原理。
4. 第二代测序技术相比于第一代测序技术有何不同?
5. 第三代测序技术有哪些?
6. 什么是基因组重测序、转录组测序、小 RNA 测序?
7. 检测蛋白与蛋白互作的方法有哪些?