Invited paper published in *Chinese Bulletin of Life Sciences* **15** (2003) 65-69

结构基因组学中的衍射相位问题
The diffraction-phase problem in structural genomics

范海福
中国科学院物理研究所

梁栋材
中国科学院生物物理研究所

摘要
本文简要介绍结构基因组学研究中，用于测定蛋白质结构的 X一射线分析在解决衍射相位问题方面的最新进展。

关键词：结构基因组学，蛋白质结构分析，晶体学中的直接法

Abstract
New developments are briefly described on solving the phase problem of X-ray structure analysis of proteins in structural genomics.

Keywords: structure genomics, protein structure analysis, direct methods in crystallography

作者简介：

范海福，男，生于1933年8月。1956年毕业于北京大学化学系。现任中国科学院物理研究所研究员。1991年当选中国科学院院士。2000年当选第三世界科学院院士。研究方向：蛋白质晶体结构分析方法；非公度调制结构分析(材料科学)；粉晶衍射分析方法；电子显微学中的图像处理；晶体学中的电子计算机软件设计。

梁栋材，男，生于1932年5月。1955年毕业于中山大学化学系。现任中国科学院生物物理研究所研究员。1980年当选中国科学院院士。1985年当选第三世界科学院院士。研究方向：蛋白质晶体结构分析，生物大分子三维结构与功能，结构生物学研究以及结构基因组等工作。

前言

人类基因组测序的基本完成标志着“后基因时代”的开始。其特征之一就是“结构基因组学”研究的兴起。蛋白质是生命活动的关键物质。基因序列只决定蛋白质分子的氨基酸顺序（一维结构)。但是，蛋白质的功能却与其三维空间结构密切相关。 要想在原子、分子的层次上去认识生命过程，就需要从基因组序列出发，系统地研究相应的基因表达产物一蛋白质的三维结构与功能。“结构基因组学” 研究就是在这一背景下，于 1998 年由美国的国家科学管理机构 NIGMS (National Institute of General Medical Sciences) 发起的[1] 。它是结构生物学迅猛发展的必然结果，是二十一世纪的一个重大国际性科研计划。世界上许多发达国家乃至发展中国家都参预了这一计划。结构基因组计划需要大规模测定并研究蛋白质的三维结构，特别是测定和研究与已知结构的蛋白质没有序列同源性的，以及属于新结构类型的蛋白质结构。由于要求大规模和高速度地测定蛋白质的三维结构。因

而人们称结构基组计划为“高产出”(high-throughput)的结构测定，甚至称之为蛋白质结构工厂(Structural factory)。结构基因组计划的研究对象是：在整个人类基因组、各种生物体(病毒、细菌、昆虫、动物的某一组织细胞等)和许多重要疾病基因等等的基因序列基础上，测定其各个基因表达产物–蛋白质的三维结构。结构基因组学的研究对于结构生物学的发展将有深远的影响。另一方面它还有广阔的实用前景。它将为预防、诊断、和治疗人类一些重要疾病指出根本的途径。就其复杂程度而言，如果说基因组测序要解决的是一维的问题，那么结构基因组学面临的则是三维乃至四维的问题。美国在其结构基因组学计划中所提出的第一个目标，是要用十年时间测定出一万个独立的非同源蛋白质三维结构，而人类至今只测定了两千多个独立的非同源蛋白质三维结构。为了实现这第一个目标，提高蛋白质三维结构分析的效率是一大关键。目前 NIGMS 设置了九个结构基因组学研究基地，其中有八个基地强调要发展高产出 (high throughput) 的测定蛋白质三维结构的技术和方法。可以认为，是否在高效测定蛋白质三维结构的方法和技术上有所创新，在某种程度上决定着我国是否有能力参预结构基因组学研究的国际合作和竞争，并作出有分量的贡献。

蛋白质的 X一射线衍射分析和衍射相位问题

测定蛋白质三维结构的主要手段有三种。即单晶体的 X-射线衍射分析、多维核磁共振、以及电子晶体学。其中使用最为广泛的是单晶体 X一射线衍射分析。迄今为止，百分之八十以上已知的蛋白质三维结构都是由它测定的。这种情况在今后十年内恐怕很难改变。蛋白质晶体结构的 X一射线衍射分析包含：试样制备、数据采集、相位推定、模型建立和精修等四个主要部分。每一部分都有急需改进的关键性环节，都关系到整体的效率。每一部分又都与其它部分密切相关。某一个部分取得进展可能减轻另一个部分的负担；某一个部分的“瓶颈”也可以转化为另一个部分的“瓶颈”。所谓“衍射分析”就是要从物体对波的衍射效应推算出物体内部的结构。这是物理学中一个既古老、又不断发展的逆问题。“衍射相位问题” 是这个逆问题的关键环节。 因为衍射效应包含衍射波的振幅和相位，而目前能够普遍使用的实验方法只能记录到衍射波的振幅。要想推算出物体的结构，就必须先找回“丢失”了的相位。X一射线衍射测定蛋白质晶体结构，用于解决相位问题的主要方法有三种，即分子置换法、多对同晶型置换法(MIR)、以及多波长异常衍射法(MAD)。第一种方法用于测定有已知同源类似物的蛋白质结构。后两种方法用于测定未知的蛋白质结构。在结构基因组学研究中，用于测定蛋白质结构的主流手段，可以归纳为“同步辐射＋硒(Se)代＋MAD (Multi-wavelength Anomalous Diffraction)”。同步辐射提供波长连续可调的、高强度的 X一射线，用于数据采集；硒代是试样制备的一个环节，是用硒原子取代蛋白质分子中甲硫氨酸的硫原子以产生便于利用的 X一射线异常衍射；MAD 是在上述基础上推定衍射相位的算法。尽管目前 MAD 方法被广泛采用，但是它并非万灵的，更不是终极的方法。为了提高测定蛋白质结构的效率，国际上有大量的研究工作致力于探索各种解决衍射相位问题的新途径。下面是比较活跃的几个研究方向。

一、多数蛋白质都含硫原子，是否可以不经过硒代，直接利用硫的异常散射效应来测定蛋白质的晶体结构？如果能实现这一目标，就可以大量节省实验工作，简化数据分析过程。但要达到这个目标，需要解决两方面的问题。一方面，由于硫原子的散射能力较弱，需要设法提高数据采集的精度。另一方面，以硫原子作为异常散射原子，不便于使用 MAD(多波长异常衍射) 方法，而须以 SAD (单波长异常衍射) 方法代替。但是 SAD 有“相位双解” 问题，即每一个衍射都有两个可能的相位，孰真孰伪难以区分。这就需要发展一套破析 SAD 相位双解的方法。我国在二十世纪六十年代就提出用晶体学中的直接法破析 SAD 的相位双解 [2]。自二十世纪八十年代至今，这一研究一直是国际“直接法”研究的热点之一。我国在这方面有相当的积累，可望在竞争中取得优势。

二、有的蛋白质不含硫，或者虽含硫而不易硒代。要想方便地利用异常衍射效应，就需要设法在蛋白质晶体中引入原子序较大的“重原子”。浸泡法是常用的方法，但是效率不够理想。新近有利用高压将惰性气体原子引入到母体蛋白质晶体的方法。也有利用重原子络合物与蛋白质分子起作用以引入重原子的方法。这些方法所产生的含重原子蛋白质

晶体不一定便于使用 MAD 方法。这时 SAD 或 SIR (单对同晶型置换)，或二者的结合将是首选的方法。使用上述新方法制备重原子衍生物，辅以破析 SAD 或 SIR 相位双解，是一个很有前途的研究方向 [3]，[4] 。

三、在目前的条件下 MAD 是无可争议的主流方法。如何进一步提高 MAD 的功效，从同样一套多波长衍射数据获得质量更好的电子密度图 ，使之便于诠释以获得分子结构模型，这是许多蛋白质晶体学家非常关心的问题。当前有关的研究可分作两类。第一类着眼于过程的前端，设法提高推定衍射相位的精度 [5] ；第二类着眼于过程的末端，设法更合理地利用相位信息以提高所得电子密度图的质量 [6] 。两者是相辅相成的。

四、除了 MAD 方法以外，MIR (多对同晶型置换) 方法也是常用的测定蛋白质晶体结构的方法。和 MAD 方法一样，如何从一套实验的 MIR 数据获得质量更好的电子密度图 ，也是许多蛋白质晶体学家非常关心的问题。上面提到有关改进 MAD 方法的研究结果，很容易推广到 MIR 方法。我国已经有成功的试验 [7] 。

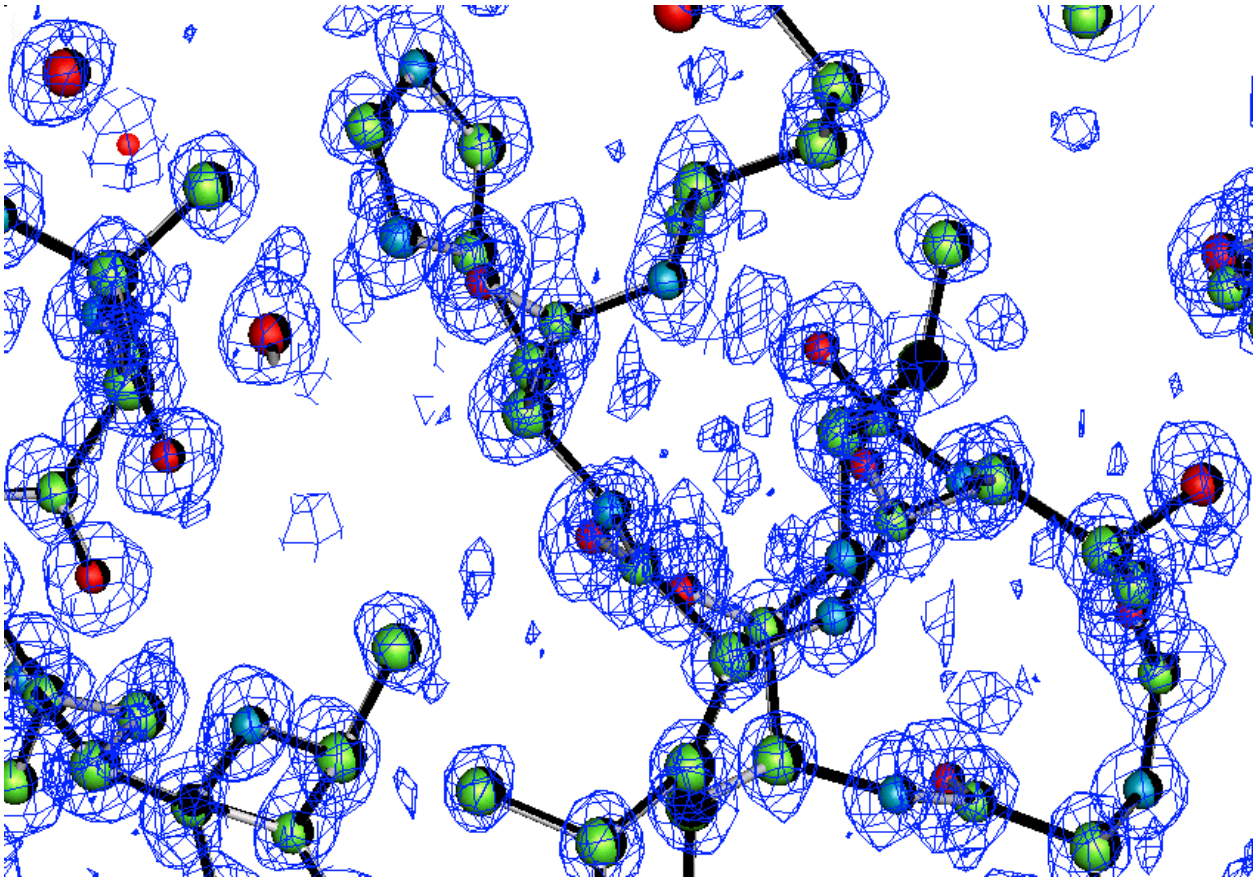
五、上述所有方法都基于异常衍射或者同晶置换。能否像测定小分子晶体结构那样，无需异常衍射或者同晶置换，只用一套由母体蛋白质晶体产生的衍射数据来测定其结构呢？这是蛋白质晶体学家梦寐以求的。它是近十年来晶体结构分析方法学研究的一个热点。对于达到“原子分辨率” (分辨极限约为 1.2Å) 的衍射数据，目前已经可以直接测定含有几千个独立原子的蛋白质晶体结构 [8]，[9]，[10] 。由于能满足“原子分辨率”要求的蛋白质晶体为数不多，今后的发展方向是使之适用于更低分辨率的衍射数据。

直接法在测定蛋白质晶体结构中的应用

上面提到的研究方向都和晶体学中的直接法有密切关系。直接法是要从一组衍射振幅直接推定相应的衍射相位。它创始于 1947 年。起初，由于直接法本身尚不完善，又由于当时采集衍射数据的精度还不高，直接法从诞生至六十年代初的十几年间，基本上是在纸上谈兵。以 H. Hauptman 和 J. Karle 为代表的一批人把二十世纪五十年代用于建立直接法的理论体系。在此基础上，I. L. Karle 和 J. Karle 于 1964 年在实际应用上取得重大突破。稍后， M. M. Woolfson 等人在发展直接法的新算法，并使之标准化和自动化方面，取得了革命性的进展。及至二十世纪八十年代初，直接法终于在小分子的单晶体结构分析中占据了统治地位。直接法使测定小分子晶体结构的周期从几个月缩短为几天；使能够测定的、不含重原子的晶体结构从二、三十个独立原子（不算氢原子）提高到一百多个独立原子。它有力地推动了结构化学的发展，并为基于小分子结构的药物设计提供了必不可少的实验基础。为此，直接法的两位先驱 H. Hauptman 和 J. Karle 于 1985 年获得诺贝尔化学奖。有关直接法截至二十世纪八十年代中期的发展，M. M. Woolfson 写过一篇综述 [11]。直接法并未因获得诺贝尔奖而止步。从二十世纪八十年代起，直接法迅速开拓新的应用领域。其中成效最为卓著的，当推直接法用于测定蛋白质晶体结构。这方面的工作可以分为两大类。第一类用于达到“原子分辨率”的衍射数据。现在，直接法已经可以从一套极限分辨率约为 1. 2Å或者更好的母体蛋白质衍射数据解出含几千个独立原子的蛋白质结构。就其“解题能力”而言，比二十世纪八十年代初提高了几十倍。但是，能够提供“原子分辨率”衍射数据的蛋白质晶体只占当前用于晶体结构分析的蛋白质晶体的百分之五左右。因此，这类方法的适用范围很小。第二类方法以“低分辨率”（分辨率极限约为 2～3Å）的衍射数据为对象。这类方法是将直接法与传统的蛋白质晶体学方法（如 SAD/MAD 以及 SIR/MIR 等）相结合。其适用范围可达当前全部蛋白质晶体结构分析的百分之八十以上，因而具有全面提高蛋白质结构分析效率的巨大潜力。我国在这方面拥有多个国际上最佳的试验结果和应用实例。下面是直接法用于测定蛋白质晶体结构的一些例子。

一、从“原子分辨率”衍射数据直接测定蛋白质晶体结构

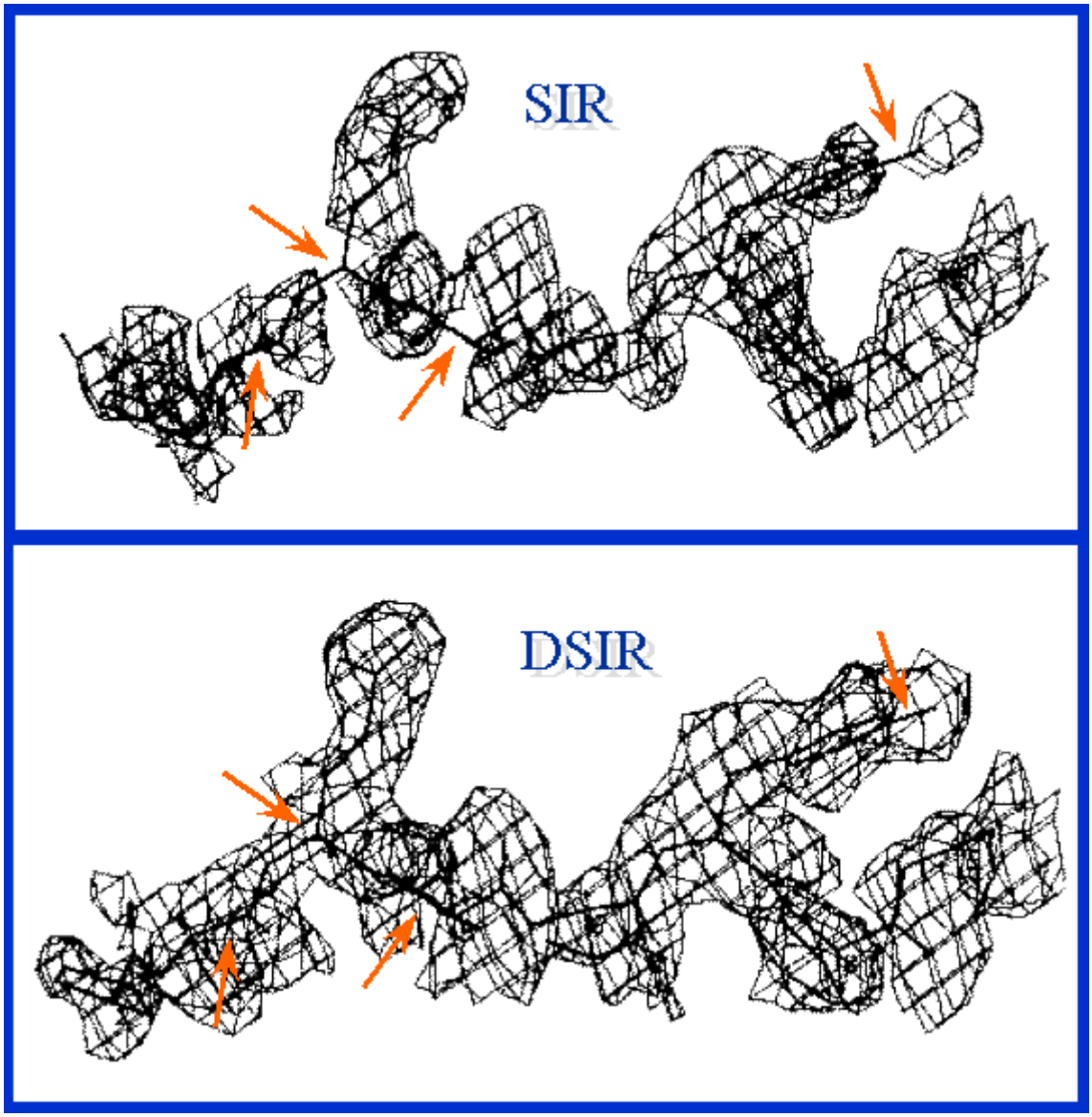
图一示出由 ACORN 程序 [10] 测定未知蛋白质 metalloproteinase deuterolysin 晶体结构的实例 [12]。该蛋白质晶体含有约 1700 个独立的非氢原子，其中有一个 Zn 原子。空间群为 $P2_1$。晶胞参数 a = 38.4, b = 34.8, c = 60.3Å; β = 106.0°。衍射数据的分辨率极限是 1.0Å。整个分析过程由 ACORN 程序自动执行。在其输出的 E—图 (图一) 中大多数原子清晰可辨。



图一 、蛋白质 metalloproteinase deuterolysin 晶体结构的局部 E—图和结构模型。

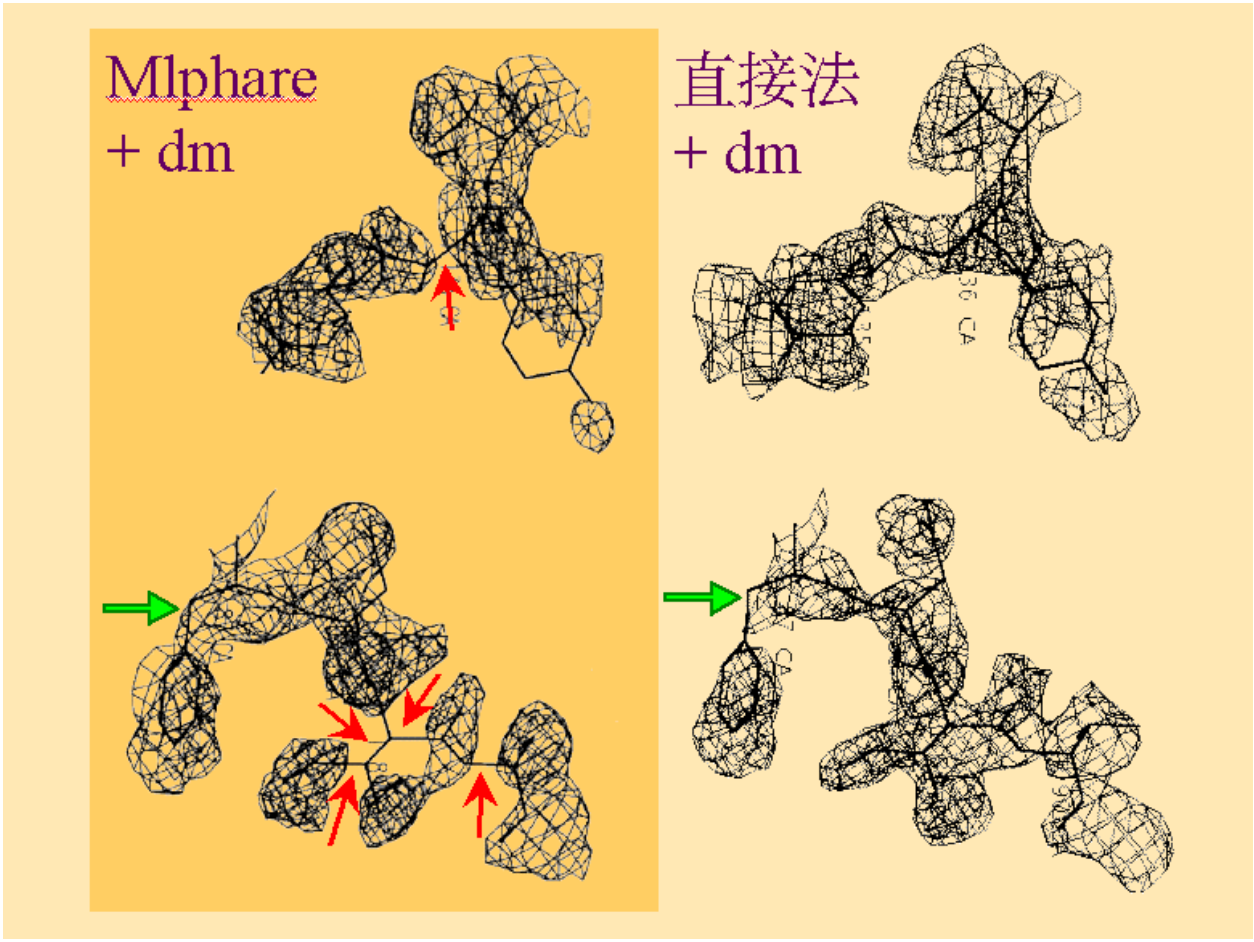
二、从 SIR 或 SAD 数据测定蛋白质晶体结构

单对同晶型置换 (SIR) 方法可以免去制备多个重原子衍生物的麻烦。但是利用 SIR 衍射数据，会遇到衍射相位双解的问题。尤其是当置换原子属中心对称分布时，很难破析相位双解。图二示出用直接法破析这种相位双解的试验结果 [13]。试样是已知结构的蛋白质 aPP (avian pancreatic polypeptide) 的母体及其 Hg 衍生物。该蛋白质晶体属 C2 空间群。晶胞参数 a = 34.18, b = 32.92, c = 28.22Å; β = 105.3^{o}。衍射数据的分辨率极限是 2.0Å 图中所示是晶胞中电子密度分布的一个局部。上半部是由常规的 SIR 方法所得的结果。如箭头所指，电子密度在蛋白质分子的主链上有多处断缺。下半部是直接法与 SIR 相结合 (DSIR) 的结果。所示区域与上半部相同。其电子密度与最终模型的吻合明显改善。



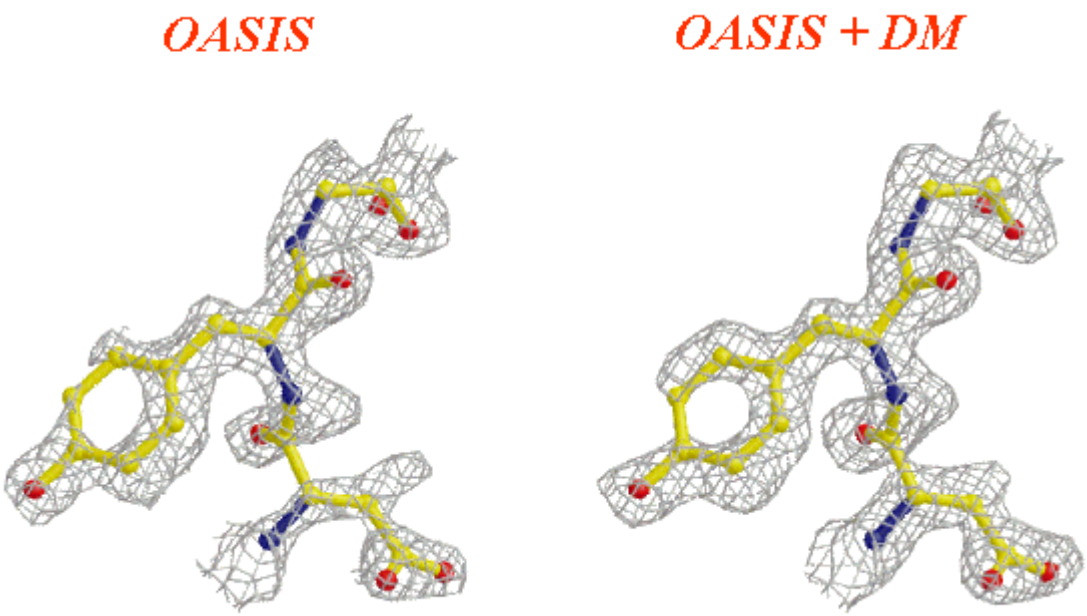
图二、蛋白质 aPP (avian pancreatic polypeptide) 晶体结构的局部电子密度图和结构骨架。上半部是常规单对同晶型置换 (SIR) 方法的结果；下半部是直接法和单对同晶型置换法相结合 (DSIR) 的结果。

用直接法破析单波长异常衍射 (SAD) 的相位双解以测定蛋白质晶体结构，是二十世纪八十年代至今国际直接法研究的一个热点。因为它是代替 MAD 方法的首选。我国率先用这种方法测定出一个原属未知的蛋白质晶体结构，从而证明这一方法的实用性 [14]。Rusticyanin 的晶体属 $P2_1$ 空间群。晶胞参数 $a = 32.43, b = 60.68, c = 38.01$Å; $\beta = 107.82^{o}$。衍射数据的分辨率极限是 2.1Å。母体蛋白含有 Cu 原子。用于采集衍射数据的 X—射线波长选在 Cu 的吸收边附近 (1.376Å)　此时 $\Delta f'' = 3.87$　本工作使用由中科院物理所提出的方法 [15]　[16] 以及由物理所编写的电子计算机程序。该程序后经修改，定名 OASIS [17]　已纳入国际上普遍使用的蛋白质晶体学程序库 CCP4 [18]　由直接法相位所得的 rusticyanin 电子密度图经 CCP4 程序库中的密度修饰程序 dm 处理后，得到可以跟踪的电子密度图。其中一部分示于图三的右半部。作为对比，在其它条件完全相同的情况下，用常规方法替换直接法，所得结果示于图三的左半部。红箭头所指之处，直接法远优于常规方法；绿箭头所指之处，直接法则略逊。总体而言，直接法明显胜过常规方法。事实上，当时用常规方法所得的电子密度图无法测定出 rusticyanin 的晶体结构。



图三、蛋白质 rusticyanin 晶体结构的局部电子密度图和结构骨架。左半部是由常规方法 (使用CCP4 程序库中的 Mlphare 程序) 推定衍射相位并用 CCP4 中的电子密度修饰程序 dm 处理所得的结果；右半部是由直接法推定衍射相位并用 CCP4 中的电子密度修饰程序 dm 处理所得的结果。

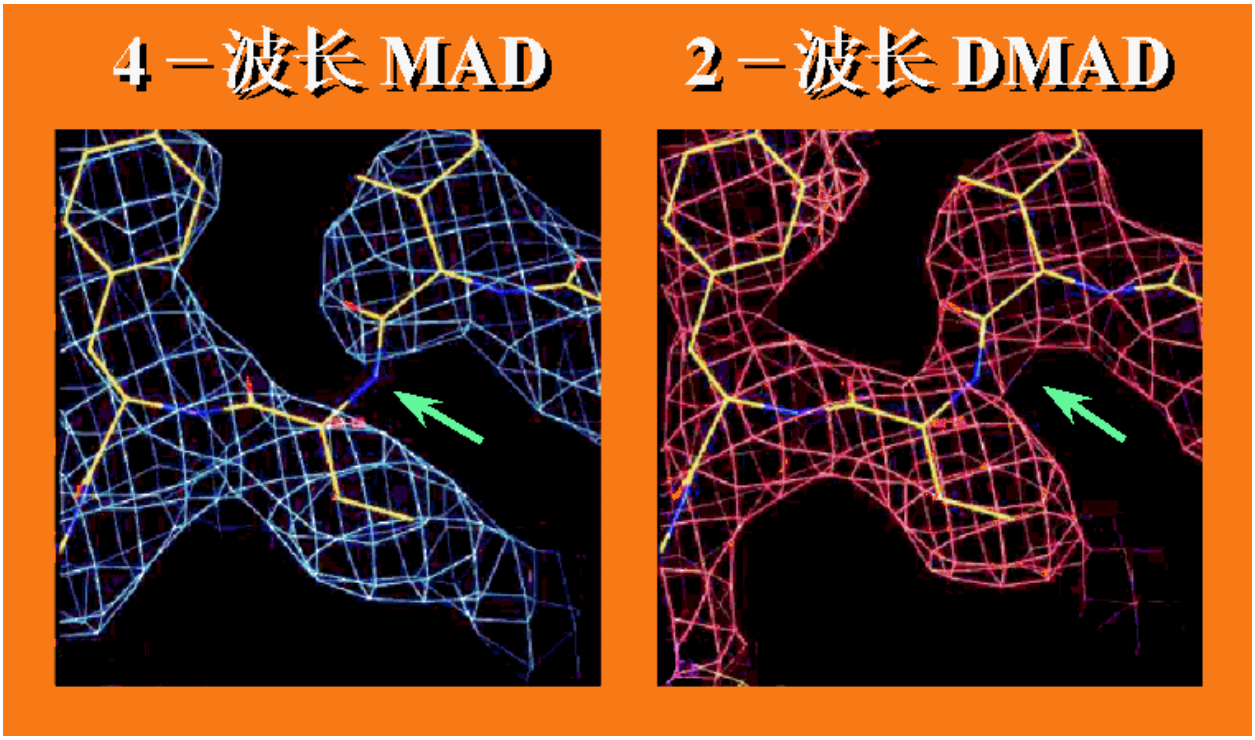
国外的同行在试验一种新的重原子衍生物制备方法时，由于所得衍生物不便于采集 MAD 数据，于是使用我们的 OASIS 程序求解蛋白质晶体结构 [4]。所得结果非常满意，见图四。试样是蛋白质 hen egg-white lysozyme 的 Gd-HPDO3A 衍生物。晶体所属空间群是 $P4_32_12$。晶胞参数 a = 77.25, c = 38.66Å。用于采集 SAD 数据的 X-射线是由 RU-200 转靶 X-射线发生器产生的 Cu $K\alpha$ 射线 (波长 1.5418Å)。相应的 Gd 的 $\Delta f''$ = 12。衍射数据的分辨率极限是 1.7Å



图四、蛋白质 Gd-HPDO3A tetragonal hen egg-white lysozyme 晶体结构的局部电子密度图及分子结构模型。左半部是用 OASIS 程序推定的直接法相位计算的电子密度图；右半部是左边的电子密度图经过 CCP4 程序库中的 DM (电子密度修饰) 程序处理后的结果。

三、直接法同 MAD 相结合

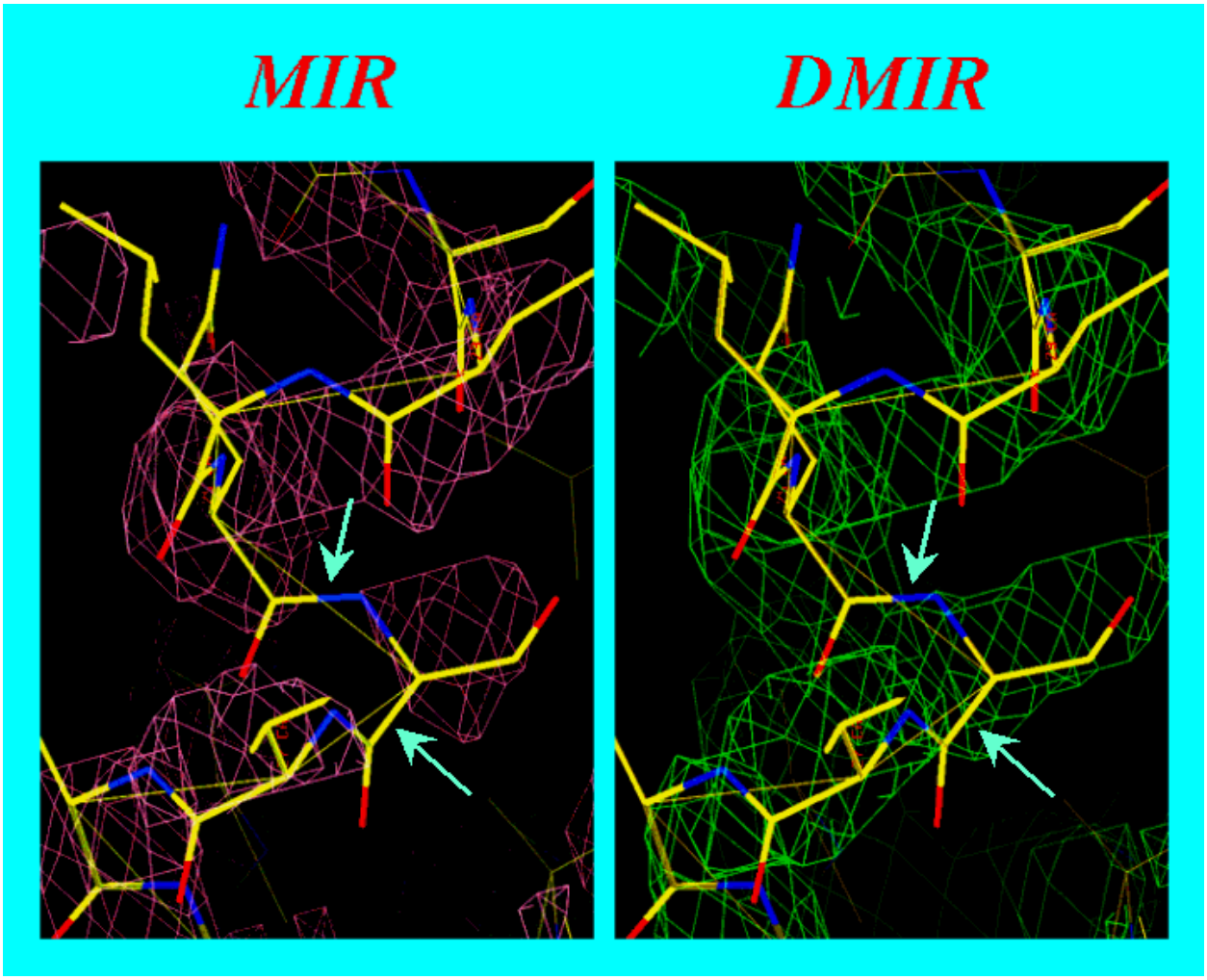
多波长异常衍射 (MAD) 方法，是结构基因组学研究中用于测定蛋白质结构的首选方法。如果能够用同样的、甚至更少的衍射数据获得质量更好的电子密度图，这将是一种有很有价值的方法。直接法可以在不增加实验数据的情况下提供独立于异常散射效应的相位信息。因此，将直接法与常规的 MAD 方法相结合，很有希望达到上述目的。中科院物理所利用美国 Cornell 大学提供的、已知结构的蛋白质晶体 human adenosine kinase 的衍射数据证实了这一设想 [19]。结果显示，对于同一套衍射数据，用直接法与 MAD 相结合的 DMAD 方法来处理，其结果明显优于使用常规的 MAD 方法。利用另一个已知结构的蛋白质 yeast Hsp40 晶体的衍射数据所作的试验 [5]，得到更进一步的结果 (见图五)。该蛋白质晶体的空间群为 $P4_12_12$。晶胞参数 a = 73.63, c = 80.76Å。衍射数据的分辨率极限是 3.0 Å。晶胞中含有 1380 个独立的非氢原子，其中有一个Se原子。图中左半部是用常规 MAD 方法处理四种不同波长衍射数据的结果。箭头所指之处，电子密度在蛋白质分子主链上发生断缺。这很容易在构建分子模型时导致错误的结果。图中右半部是用 DMAD 方法只处理两种不同波长衍射数据的结果。尽管所用数据少了一半，但是结果却比常规的 MAD 方法好得多。



图五、蛋白质 yeast Hsp40 晶体结构的局部电子密度图及结构模型。左半部是用常规的 MAD (多波长异常衍射) 方法从4种不同波长的衍射数据所得的结果；右半部是只用2种不同波长的衍射数据，由直接法和 MAD 相结合的方法 (DMAD) 推定衍射相位所得的结果。

四、直接法同 MIR 相结合

多对同晶型置换 (MIR) 也是测定未知蛋白质晶体结构的主要方法。如果能够用同样的衍射数据获得质量更好的电子密度图，将是有很有意义的工作。直接法可以在不增加实验数据的情况下提供独立于同晶型置换的相位信息。因此，直接法与 MIR 的结合有希望达到这一目的。这一设想由中科院生物物理所首先提出，并由中科院物理所和生物物理所合作完成了首例试验。图六是用已知结构的蛋白质 R-phycoerythrin 所作试验的结果。该蛋白质晶体属 R3 空间群。晶胞参数 a = 189.8, c = 60.1Å。衍射数据的分辨率极限是 3.0Å。晶胞中含有 682 个独立的氨基酸残基。这是目前用直接法成功地作过相位推定试验的、最复杂的蛋白质晶体。图中左半部是用常规 MIR 方法处理四对同晶型置换数据所得电子密度图的一部分。箭头所指之处，电子密度在蛋白质分子主链上发生断缺。右半部是用直接法同 MIR 相结合的方法 (DMIR) 处理同一套数据的结果。图中可见电子密度显著地改善。



图六、蛋白质 R-phycoerythrin 晶体结构的局部电子密度图及结构模型。左半部是用常规的 MIR (多对同晶型置换) 方法所得的电子密度图；右半部是用直接法同 MIR相结合 的方法 (DMIR) 所得的结果。

结语

晶体学中的直接法自诞生之日至二十世纪八十年代初，花了三十多年的时间，从纸上谈兵发展至主导小分子晶体结构分析，并荣获诺贝尔奖。又过了将近二十年，直接法闯进了蛋白质晶体学的领域，从无人问津变得炙手可热，新的辉煌正等待着有志者去创造。作为二十一世纪国际生命科学重大研究项目之一的结构基因组学将会由此受益。

本文有关中科院物理所和生物物理所的研究工作，得到中国科技部973项目和中国科学院知识创新工程的支持。

参考文献

[1] 见美国国家卫生研究院的网页 http://www.nigms.nih.gov/funding/psi.html
[2] 范海福，物理学报，21, 1114-1118 (1965)。
[3] Q. Hao (郝权)，*CCP4 Newsletter* 40 (2002)。
[4] É. Girard, L. Chantalat, J. Vicat & R. Kaln, *Acta Cryst.* D58, 1-9 (2002)。
[5] Y. X. Gu (古元新), F. Jiang (江凡), B. D. Sha (沙炳东) & H. F. Fan (范海福)，*Z. Krist.* 217, 710-714 (2002)。
[6] T. Terwilliger, *Acta Cryst.* D57, 1755-1762 (2001)。
[7] Y. X. Gu (古元新), W. R. Chang (常文瑞), T. Jiang (江涛), C. D. Zheng (郑朝德), & H. F. Fan（范海福）, *Acta Cryst.* A58, 547-551 (2002)。
[8] H. Xu, H. A. Hauptman & C. M. Weeks, *Acta Cryst.* D58, 90-96 (2002)。
[9] G. M. Sheldrick, *Z. Krist.* 217, 644-650 (2002)。
[10] J. X. Yao(姚家星), M. M. Woolfson, K. S. Wilson & E. J. Dodson, *Z. Krist.* 217, 636-643 (2002)。
[11] M. M. Woolfson, *Acta Cryst.* A43, 593-612 (1987)。
[12] K. McCauley, J. X. Yao, E. J. Dodson, J. Lehmbeck, P. R. Østergard, K. S. Wilson, *Acta Cryst*. D57**,** 1571-1578 (2001)。
[13] Y. D. Liu (刘玉东), Y. X. Gu (古元新), C. D. Zheng (郑朝德), Q. Hao (郝权) & H. F. Fan (范海福), *Acta Cryst*. D55, 846-848 (1999)。
[14] Y. D. Liu (刘玉东), I. Harvey, Y. X. Gu (古元新), C. D. Zheng (郑朝德), Y. Z. He (何亦宗), H. F. Fan (范海福), S. S. Hasnain & Q. Hao (郝权), *Acta Cryst.* D55, 1620-1622 (1999)。
[15] H. F. Fan (范海福) & Y. X. Gu (古元新), *Acta Cryst.* A41, 280-284 (1985)
[16] H. F. Fan (范海福), Q. Hao (郝权), Y. X. Gu (古元新), J. Z. Qian (千金子), C. D. Zheng (郑朝德), & H. Ke (柯衡明), *Acta Cryst.* A46, 935-939 (1990)。
[17] Q. Hao (郝权), Y. X. Gu (古元新), C. D. Zheng (郑朝德) & H. F. Fan (范海福), *J Appl. Cryst.* **33**, 980-981 (2000)。
[18] Collaborative Computational Project, Number 4, *Acta Cryst.* D50, 760-763 (1994)。
[19] Y. X. Gu (古元新), Y. D. Liu (刘玉东), Q. Hao (郝权), S. E. Ealick & H. F. Fan (范海福), *Acta Cryst.* D57, 250-253 (2001)。