

无序蛋白质的判定及其结构、功能和进化特征 *

马 冲^{1, 2)} 杨 冬^{2)**} 姜 颖²⁾ 钟儒刚¹⁾ 贺福初²⁾

(¹⁾ 北京工业大学生命科学与生物工程学院, 北京 100022;

²⁾ 军事医学科学院放射与辐射医学研究所, 北京蛋白质组研究中心, 蛋白质组学国家重点实验室, 北京 102206)

摘要 固有无序蛋白质(intrinsically disordered proteins, IDPs)是天然条件下自身不能折叠为明确唯一的空间结构, 却具有生物学功能的一类新发现的蛋白质. 这类蛋白质的发现是对传统的“结构-功能”关系认识模式的挑战. 本文首先总结了无序蛋白质的实验鉴定手段、预测方法、数据库; 并介绍了无序蛋白质结构(包括一级结构、二级结构、结构域无序性及变构效应)和功能特征; 然后重点总结了无序蛋白质在进化角度研究的进展, 包括无序区域产生的进化机制、进化速率, 蛋白无序性的进化在蛋白质功能进化及生物学复杂性增加等方面的重要作用; 最后展望了无序蛋白质在医药方面的应用前景. 本文对于深入认识无序蛋白质的形成机制、结构和功能特征及其潜在的临床应用前景具有重要意义.

关键词 固有无序蛋白质, 序列, 结构, 功能, 进化

学科分类号 Q51

DOI: 10.3724/SP.J.1206.2014.00084

关于蛋白质的结构和功能, 一般认为“序列决定结构, 结构决定功能”. 这里的“结构”, 传统观念上一般指有序的三维结构. 但近 50 多年来, 研究者们不断地发现一类新的蛋白质, 它们没有明确的三维结构却具有特定生物学功能. 并且, 对其中的一部分早有研究, 例如多肽类激素^[1] (polypeptide hormones), 研究发现其在溶液状态是无序的. 最近十多年, 人们所发现的此类蛋白质数量呈快速增长趋势^[2]. 这类蛋白质的全部或部分区域, 自身不能折叠为明确唯一的空间结构, 这些区域就是无序区域. 在生命活动中, 这类蛋白质发挥着重要的生物学功能, 当和其他蛋白质结合时, 可以折叠为有序结构, 因而存在一系列快速互变的构象. 我们把这类具有无序区域的蛋白质称为固有无序蛋白质(intrinsically disordered proteins, IDPs), 也称为固有非结构化蛋白质(intrinsically unstructured proteins, IUPs). 无序蛋白质依据无序区域的范围可以分为完全无序蛋白质(全序列无序)和部分无序蛋白质(局部含有超过 30 个残基的无序

区域).

无序蛋白质在生物体内是普遍存在的, 在不同的生物中其占蛋白质组的比例差异明显. Ward 等^[3] 分析发现, 在原核生物中, 蛋白质组中 0.9%~7% 的蛋白质含有长无序区域(连续超过 30 个残基); 而在真核生物中, 含长无序区域的蛋白质占整个蛋白质组的 27.5%~36.6%. 由此可见, 无序蛋白质在真核生物蛋白质组中占据着重要的地位. 对这类蛋白质的结构、功能、进化特征的认识, 有助于我们更深层地理解无序蛋白质的功能重要性及其参与重要生理病理过程的分子机制. 本文将介绍无序蛋白质的预测、实验方法、数据库, 并对其序列、结构、功能及进化方面的研究进展进行总结(研究历程中的主要事件见图 1).

* 国家国际科技合作专项(2014DFB30020)和国家重大科学研究计划(2014CBA02001)资助项目.

** 通讯联系人.

Tel: 010-80727777-1140, E-mail: yangdongbprc@163.com

收稿日期: 2014-03-24, 接受日期: 2014-06-17

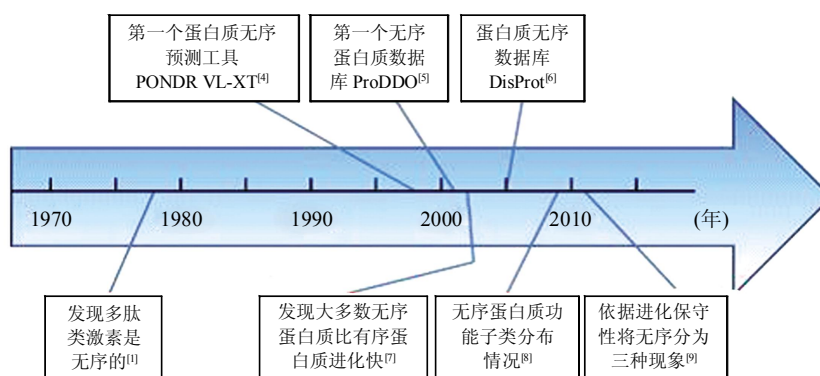


Fig. 1 Milestones during the research history of intrinsically disordered proteins

图 1 无序蛋白质研究历程中的重要事件

1 固有无序蛋白质的实验测定、预测方法及数据库

1.1 检测固有无序蛋白质的实验方法

目前有几种物理化学方法可以检测固有无序蛋白质, 包括 X 光散射技术^[10]、核磁共振技术^[11] (nuclear magnetic resonance spectroscopy, NMR)、近紫外圆二色光谱法^[12]、远紫外圆二色光谱法^[13]、傅里叶变换红外法^[14]、拉曼光谱和光活性法^[15]、蛋白质降解率法^[16]、流体动力学技术[包括凝胶过滤、黏度测定法、小角度 X 射线散射法(SAXS)、小角度中子散射法(SANS)]^[17]等。现在常用的是前两种, 尤其 NMR 是研究无序蛋白质最为重要的手段。多种方法综合利用可以更好地区分无序蛋白质和结构化蛋白质。

1.2 蛋白质无序区域的预测方法

尽管利用上述方法可以对蛋白质的无序特征进行测定, 但是, 通过实验手段确定蛋白质的结构, 不但成本高、耗时长, 而且实验中还会遇到一些目前无法解决的技术困难^[8]。因而迄今为止, 无序蛋白质结构实测数据不易得到, 尤其是大规模地进行无序蛋白质结构测定更是困难。因此, 无序区域的预测方法快速发展起来。无序区域预测软件的开发为人们更好地了解无序区域的功能和特点创造了条件。

目前, 无序区域预测方法大体分为两类^[18]: 一类是依赖人工神经网络(artificial neural networks, ANNs)、支持向量机(support vector machines, SVMs)

等机器学习方法开发的算法。例如 1997 年 Romero 等^[4]开发的第一个无序区域预测的工具 PONDRL VL-XT, 它是基于 PDB 数据库中 67 个无序区域(1 340 个残基)和一些有序区域(16 543 个残基)建立的一种“双层前馈式神经网络”, 首次表明单纯从氨基酸序列可以预测无序区域。之后利用计算技术开发了一系列的算法, 如 PONDRL VL3^[19]、DISOPRED2^[3]、POODLE-L^[20]等。第一类算法的缺点是不能很好地揭示潜在的序列性质。第二类是基于序列氨基酸的理化性质开发的算法, 这类算法的基本原理是无序区域中氨基酸的高净电荷导致的静电排斥, 以及低疏水性使得蛋白质无法形成紧密结构^[14], 如果这样的氨基酸聚集在蛋白质的某些片段或整个蛋白质, 那么这些片段或整个蛋白质就是无序的。有很多算法都是依据此原理开发的, 例如: IUPred^[21-22]、FoldUnfold^[23]、CH-plot^[14]和 FoldIndex^[24]等。

评估无序预测软件的准确率主要是通过每两年举行一次的蛋白质结构预测比赛(critical assessment of structure prediction, CASP), 从 CASP5^[25]开始加入了对无序蛋白质的预测, 目前已经举行到 CASP10^[26]。

目前进一步提高预测软件精度的困难主要在于如何消除结构化蛋白质和无序蛋白质训练集的噪音(noise level of training sets)^[8]。很多无序蛋白质能够形成结构化的复合物, 当从 PDB 数据库中筛选结构化蛋白质组成结构化蛋白质训练集时, 那些独立状态时是无序蛋白, 但与其他蛋白质结合形成复合物时变为结构化的蛋白质可能会被筛选到训练集

中, 这样会增加结构化蛋白质训练集的噪音。

1.3 数据库

第一个收录无序蛋白质区域的公共资源是由 Sim 等^[9]建立的 ProDDO。但 ProDDO 数据库是很不完善的, 它只局限于 PDB 数据库中收录的条目, 而且既没有提供无序的类型, 也没有无序区域的功能。

这些局限在 DisProt^[6]数据库得到了很好的解决, DisProt 数据库收录了实验表征的无序蛋白、无序区域和它们的生物学功能。因此, DisProt 数据库系统地将无序蛋白质和无序区域的结构与功能联系了起来。此数据库通过收集和整理无序蛋白质的实验验证信息及功能关联信息, 用以推进无序蛋白质的研究。在 2004 年 2 月的首发版本中, DisProt 数据库包含 154 种蛋白质(190 个无序区域), 而在 2006 年 8 月发布的版本中包含 460 种蛋白质(1 103 个无序区域)。截止到 2013 年 5 月 24 日, 共包含 694 个固有无序蛋白和 1 539 个固有无序区域。DisProt 数据库为研究无序蛋白质的序列、结构、生物学功能提供了很好的数据来源, 是研究无序蛋白质的重要信息基础, 数据库访问网址为 <http://www.disprot.org>。

Fukuchi 等^[27]建立了一个收集实验验证过的固有无序蛋白质的数据库 IDEAL (intrinsically disordered proteins with extensive annotations and literature)。为我们提供了一个全面了解无序蛋白质结构信息的工具, 由于有实测证据支持, 更具有参考价值。截止到 2013 年 8 月 30 日, 已收录的条目达到 340 条。

2 无序蛋白质的结构特征

2.1 无序蛋白质的一级结构特征

无序序列的鲜明特征是其复杂性低^[28], 即氨基酸组成种类很少, 无序区域中含有大量重复序列, 且序列组成具有偏性: 疏水氨基酸残基(缬氨酸、亮氨酸、异亮氨酸、甲硫氨酸、苯丙氨酸、色氨酸、酪氨酸)含量较低, 带电氨基酸残基和极性氨基酸残基(谷氨酰胺、丝氨酸、苏氨酸、脯氨酸、谷氨酸、赖氨酸)含量较高^[29]。另外, 无序蛋白的序列中富含丝氨酸和苏氨酸残基, 因此磷酸化与脱磷酸化修饰与其功能的发挥密切相关^[30]。

2.2 无序蛋白质的空间结构特征

2.2.1 无序蛋白质中的二级结构

Fuxreiter 等^[31]在研究与伙伴蛋白结合的无序蛋

白质的结构特征时发现, 无序蛋白质的有些区域具有很强的形成 α 螺旋的倾向, 这些局部的二级结构通常是“蛋白质-蛋白质”以及“蛋白质-DNA”相互作用时的结合部位。

2.2.2 蛋白质结构域的无序性

结构域是蛋白质结构、功能、进化的基本单位, 在通常的理解中, 结构域是蛋白质结构化的区域, 但在研究中发现, 一些蛋白质的结构域也是无序的, 例如转录因子环腺苷酸应答因子结合蛋白(transcription factor cAMP response-element binding protein, CREB)中的激酶诱导结构域(kinase inducible domain, KID)^[29]。

Daughdrill 等^[32]用核磁共振技术研究了来自 5 个物种的同源蛋白质 RPA70 中连接结构域(intrinsically unstructured linker domain, IULD)的结构, 发现它们是无序的。RPA70 是真核蛋白质, 对复制、重组、DNA 修复至关重要^[33], IULD 是连接 RPA70 两个亚基的结构域。

Buljan 等^[34]研究发现, 蛋白质在进化中新增的结构域大体上无序序列比例较高, 平均来说, 新增结构域中有 21%残基是无序的, 而其他结构域只有 5%是无序的。无序序列比例最高的结构域是由那些已有外显子扩展增加的。

2.2.3 无序蛋白质的变构效应

与结构化蛋白质相比, 无序蛋白质具有结构柔性。由于无序蛋白质的结构柔性大, 构象变化的能量势垒低, 所以容易实现各种构象的变化, 即变构效应。

无序蛋白质的变构效应主要体现在发生蛋白质相互作用的过程中。很多无序蛋白在结合到它们的靶标上时, 转变成更有序的状态或折叠成稳定的二级或三级结构, 也就是说, 它们的折叠和结合过程是耦联在一起的。这种折叠和结合的耦联过程可能涉及几个残基或者整个蛋白质结构域。

无序蛋白质的变构效应是其实现功能复杂性的机制之一。例如, Ferreon 等^[35]发现 E1A (adenovirus early region 1A)可以和一小部分蛋白质发生相互作用, 如 pRb (retinoblastoma protein)、CBP/p300 (CREB 结合蛋白及其旁系同源蛋白 p300), 组成一系列的中心复合物。其他复合物再和中心复合物依据相互作用偏好发生相互作用, 形成不同的分子构象。

由于无序蛋白质具有变构效应, 因而在溶液状态下, 无序蛋白质是多种构象共存的体系。通常描

述无序蛋白质的结构采用“系综”(ensemble)的概念。系综是指在一定的宏观条件下, 大量性质和结构完全相同的、处于各种运动状态的、各自独立的系统的集合。这里是指描述出无序蛋白质所有存在的构象。

3 无序蛋白质的生物学功能

固有无序蛋白质的实例不断涌现, 使得人们对其功能有了深入的认识。无序蛋白质在很多方面都发挥着重要的功能, 包括调节转录、翻译、细胞信号转导、蛋白质磷酸化、小分子存储, 以及对大的多蛋白复合体(如细菌鞭毛及核糖体)自组装的调控。Liu 等^[36]研究发现无序蛋白质多为高亲和力结合蛋白或转录和信号转导蛋白。DisProt 数据库介绍了 32 种与无序蛋白质相关的功能子类, 并对每一子类涉及的无序蛋白质种类进行了统计^[8], 其中, “蛋白质与蛋白质相互结合”这一功能子类的数目最多, 但仍有 123 种无序蛋白质是不知道它们功能的。这里的统计只反映了 DisProt 数据库收集的无序蛋白质各功能子类的数目, 而不能作为所有无序蛋白质的功能分类特征。

关于无序蛋白质的功能, Janin 等^[37]有不同的观点, 他们认为无序蛋白质是没有功能的, 有功能的是其在体内形成复合物时形成的有序状态的蛋白。无序是一种瞬态, 不是固有性质。这种观点比较新颖, 尚未得到很多的支持。

4 无序蛋白质的进化特征及其进化生物学效应

4.1 无序区域在不同物种中的分布及进化机制

无序蛋白质在生物体内是普遍存在的, 在不同的生物中其占蛋白质组的比例差异明显。在原核生物蛋白质组中无序蛋白质比率低, 而在真核生物蛋白质组中无序蛋白质比率较高, 即在进化中无序蛋白质的比率有增加的趋势。

Yruela 等^[38]研究拟南芥和水稻中染色体上无序蛋白的分布发现, 其分布是非随机分布的, 在大多数染色体上, 大于 30 个残基的无序片段的分布频率和基因重组率及 G+C 含量呈正相关。这些分析表明, 植物蛋白中无序片段的的存在与编码基因的基因重组率相关, 并且高的重组率能够在进化中诱导无序片段的出现。

Schad 等^[39]研究发现高水平的无序区域是由对称外显子编码的, 而对称外显子也更偏向于编码无

序区域。大多数无序蛋白的基因是来自经常出现内在串联重复的对称外显子。在人类基因进化中结构无序促进复杂基因的模块化组装。对称外显子是指处于两同相位内含子的外显子, 其核苷酸数为 3 的整数倍, 它可以被成功复制, 不会造成阅读框的推移。

4.2 无序蛋白质进化速率

一般来说, 无序区域的进化速率比有序区域更快。Brown 等^[7]对无序蛋白质进化速率的研究发现, 他们所比较的 26 个蛋白质家族中, 5 个家族的有序区域和无序区域的遗传距离没有明显的区别, 而 19 个家族的无序区域进化相比于有序区域更快; 仅剩下的 2 个家族, 无序区域的进化速率比有序区域慢。

不同进化速率的无序区域, 倾向于具有不同类型的生物学功能。上述 Brown 等^[7]的研究发现, 19 个无序区域进化速率较快的家族中, 其无序区域功能包括“蛋白质、DNA、RNA 的结合位点”, “柔性的连接区域”, 还有些区域的功能是不清楚的; 无序区域进化速率较慢的 2 个家族, 其无序区域的功能主要是 DNA 结合位点。Bellay 等^[9]对此问题进行了更全面、精细的分析, 将“无序”划分为三个截然不同的生物现象: 灵活的无序、有约束的无序和非保守的无序。“灵活的无序”是指无序区域是保守的, 但氨基酸序列是快速演变的; “有约束的无序”是指无序区域和氨基酸序列都是保守的; “非保守的无序”是指无序区域不是保守的。研究表明, “灵活的无序”和“有约束的无序”有不同的功能特性, “灵活的无序”与信号通路和多功能相关, 而“有约束的无序”有明显的不同功能属性, 主要参与 RNA 结合和蛋白质分子伴侣。最后, “非保守的无序”基于对文章的分析缺乏清晰的功能特征。

Daughdrill 等^[32]首次从实验角度验证了无序蛋白质动态特性的进化保守, 论证了氨基酸序列的保守性对于动态特性和分子功能的保守不是必需的。文章研究了一个特定无序蛋白质种类的动态特性的保守性, 此特定无序蛋白质种类是一类连接侧翼折叠结构域的连接结构域(intrinsically unstructured linker domains, IULD)。通过用核磁共振方法分析了 5 个 IULDs 的骨架柔性, 这 5 个 IULDs 代表着 3 个界(kingdom), 2 个 IULDs 是动物界的, 1 个 IULD 是真菌界的, 2 个 IULDs 是植物界的, 结果表明骨架柔性水平是相似的。而与此相反, 它们的

序列却没有相似性。为了研究 IULDs 的动态特性保守而序列不保守, 文章又研究了 9 个哺乳类 IULDs 的进化速率, 结果显示很多 IULD 的位点是进化中立的, 这表明动态特性在缺乏自然选择时可以维持下去。

Siltberg-Liberles^[40]使用系谱树重建、蛋白质结构无序预测和二级结构预测等方法研究了包含直系同源和旁系同源基因的蛋白质家族中结构无序的进化动力学, 结果表明, 无序区域在进化的时间尺度上通常是非保守的, 并且无序区域通常具有更强的构象上的柔性, 蛋白质结构在空间构象柔性方面的进化, 能改变蛋白质家族蛋白间同源区域的二级结构倾向性。由于无序区域经由翻译后修饰会参与蛋白质-蛋白质交互, 无序倾向的变化表明旁系同源的功能上的不同。

蛋白质的进化不仅通过点突变, 也可通过插入和缺失影响蛋白质长度。Light 等^[41]在研究蛋白质长度进化机制时, 发现长度的差异和固有无序残基数目的差异之间有着很强的关系, 3/4 以上的长度差异可以通过固有无序残基数目的改变来解释。进一步分析发现, 在插入和缺失中无序是很常见的; 插入缺失事件不会引起无序, 而已经无序的区域会引起插入缺失, 暗示对于在固有无序区域出现的插入缺失有低选择压力。

无序蛋白质快速进化的特点, 也使无序区域蛋白质修饰位点的进化呈现出特有的特征。例如, 对于蛋白质磷酸化修饰来说, 无序蛋白质快速进化和容易发生插入缺失的特点, 导致位于无序区域的磷酸化位点与位于有序区域的磷酸化位点在进化方面有所不同^[42]。在哺乳动物中, 磷酸化位点比非磷酸化位点更保守, 但无序区域的磷酸化位点比有序区域的磷酸化位点进化更为迅速^[43]。

4.3 进化生物学效应

4.3.1 无序蛋白质与生物学功能进化的关系

进化过程中无序蛋白质比率的增加, 促进了蛋白质生物学功能的进化。Tompa 等^[44]利用 PONDR VSL1 和 IUPred 两种预测方法预测了大肠杆菌和酵母蛋白质组的无序水平, 证明无序蛋白质比率在真核生物中水平较高: 超过 20% 的残基落入局部无序区域, 所有蛋白质中的 50%~60% 有至少一个长无序区域(>30 个残基)。酵母蛋白质组中包含长无序区域(>30 个残基)的蛋白质是大肠杆菌的 3 倍。从基因功能的角度分析, 发现与调控相关的蛋白质含有的无序比例远远高于与新陈代谢、生物催化等

相关的蛋白质, 因而无序蛋白与“调控”更为相关, 而不是“代谢”; 即蛋白质进化中无序区域比例的增加, 促进了真核生物“调控”类生物学功能的进化。

在进化历程中, 病原体效应蛋白中的无序特征也被保留下来, 并代表着细菌性病原体对抗植物先天免疫系统的一种机制。Marín 等^[45]对此进行了研究, 预测出植物病原体的效应蛋白含有大量的较长无序区域(>50 个残基)。效应蛋白中含有大量固有无序区域, 暗示了对这种结构特性的正向的进化选择。研究发现, 这种结构柔性对以下功能是必不可少的: a. 效应蛋白转移; b. 逃避先天免疫系统; c. 对宿主功能的模拟。提示病原体效应蛋白进化过程中保持无序性, 有助于其特定生物学功能的演化。

Chen 等^[46]研究发现, 与较多高无序蛋白发生相互作用的蛋白质进化更慢, 这也解释了之前的研究结果, 蛋白质-蛋白质相互作用的数目与进化速率呈负相关。

从蛋白质相互作用组水平研究发现, 无序蛋白质在蛋白质相互作用组的进化中起着重要的作用。结构无序在蛋白质-蛋白质相互作用网络和连接度高的蛋白质的进化中扮演着关键的角色。通过比较人、果蝇、酵母的蛋白质相互作用组, Mosca 等^[47]发现, 无序的相互作用网络没有结构化的相互作用网络保守。进一步分析显示, 无序蛋白质在进化中有更强的能力重新选择它们的相互作用“邻居”。无序区域有着明显的进化优势, 能够促进进化中相互作用对象的改变。

4.3.2 无序蛋白质与生物学复杂性的关系

蛋白质结构特征与生物学复杂性密切相关。如: 进化过程中, 多结构域蛋白质比率的增加及新结构域的出现, 大大促进了生物学复杂性的增加^[48]。近年来, 随着蛋白质无序性这一结构特征研究的深入, 研究者们发现蛋白质无序性与生物学复杂性之间也存在密切关系。

Schad 等^[49]研究生物学复杂性与蛋白质无序性的关系发现, 无序蛋白质占蛋白质组的比例从原核生物到真核生物明显增加, 但在真核生物进化过程中, 不再发生显著增加。蛋白质组中无序区域的结合位点数目随着生物学复杂性的增加而增加。此外发现, 随着人体组织复杂性的增加, 蛋白质无序比率也有显著增加, 即在复杂的组织中蛋白质无序比率更高。上述结果表明, 蛋白质无序性是生物学复

杂性相关的促成因素之一。

可变剪切导致生物复杂性增加的重要因素, 蛋白质无序性与可变剪切间也有着紧密的联系。Romero 等^[50]分析了 46 个人类编码蛋白基因, 研究表明 75 个可变剪切片段中的 81% 是无序的, 其中 57% 是全无序, 24% 是部分无序片段, 即可变剪切的片段倾向于编码无序区域。

5 无序蛋白质与疾病、药物开发

固有无序蛋白质在信号转导和转录过程中发挥重要的作用, 固有无序区域也是许多与疾病相关的位点^[29]。发生在编码无序区域的染色体易位保持了折叠结构域的完整, 从而产生一种功能异常的融合蛋白。相比之下, 发生在编码完全结构化结构域的基因区域发生易位, 肯定会导致生成一个错误折叠的蛋白质, 那么将会在细胞降解机制的作用下迅速被分解, 也就不会产生疾病表型。

并且, 固有无序蛋白质丰度的改变伴随着几种疾病的发生。例如, TC-1 的过表达或 Arf 与 p27 的低表达和各种类型的癌症有关。相似的, 核突触蛋白的过表达增加了聚合形成的风险, 和帕金森症和阿兹海默症有关联。

Vavouri 等^[51]通过检测酵母中过表达表型, 发现蛋白质固有无序是剂量敏感性的重要决定因素。当基因过表达时, 蛋白质浓度升高, 而无序区域更倾向于使分子间发现相互作用, 这可能是产生病理的原因。此结论在果蝇黑腹和秀丽隐杆线虫这两种生物上得到了证实。在小鼠和人中, 蛋白质无序性与剂量敏感性的致癌基因有很强的关联, 这样质量驱动分子相互作用可能是癌症的常见原因。

由于无序蛋白质常常与疾病相关联, 因而, 在药物开发中, p53、 α 突触核蛋白等无序蛋白质成为首选靶标。因为无序蛋白质的结合位置就像酶的活性区域, 无序蛋白质的结合伙伴就是目标蛋白。利用这种方法开发的抗癌药物——nutlins, 它能抑制 p53 与 MDM2 的相互作用, 恢复癌细胞中 p53 通路^[52]。

6 小 结

关于蛋白质的结构和功能, “序列 - 结构 - 功能”模式深入人心, 但无序蛋白质的发现使人们对这种模式产生了怀疑。无序蛋白质或无序区域的存在丰富了蛋白质实现其功能的形式, 与有序蛋白质共同配合, 使得复杂的生命活动能够有条不紊地实

现。无序蛋白质在序列、结构、功能和进化上与有序蛋白质有很大不同(图 2), 预示着“‘非结构’生物学时代”的到来^[53]。

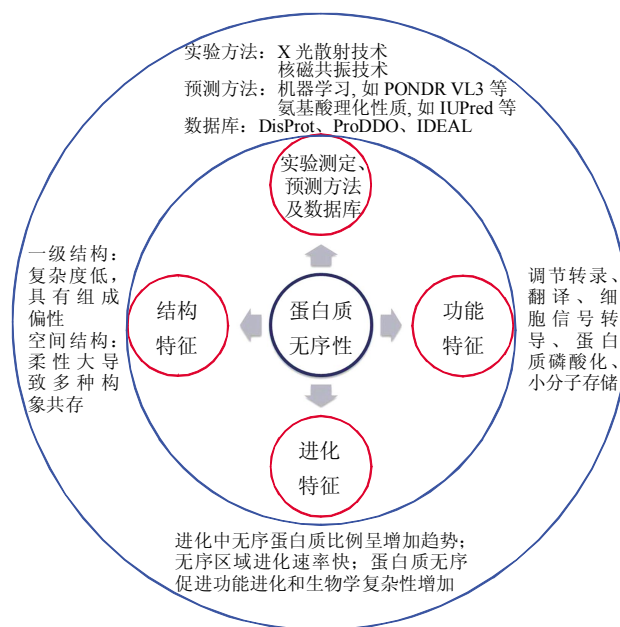


Fig. 2 Current understanding of disordered proteins

图 2 目前对无序蛋白质的认识

蛋白质无序性在很大程度上决定于其固有的序列特征, 是蛋白质的“内在属性”之一。目前, 基于蛋白质“内在属性”(如基因/蛋白质长度、结构域数目、进化速率等)与“分子表型特征”(如基因表达丰度、表达广度等)关联的分析已取得丰硕成果^[54-55], 对“蛋白质组构成”^[56]、“基因表达调控”^[57]等诸多基本生物学问题形成了规律性的认识。将蛋白质无序性引入此类研究中, 将使得人们对蛋白质“内在属性”与“分子表型特征”的规律性关系得到更深入、全面的认识。这种规模化的分析, 目前主要依赖于在全蛋白质组范围内进行蛋白质无序性的预测, 因此是具有一定局限性的。为真正实现上述分析目的, 需要对蛋白质无序性及其动态变化进行精确的测定或更准确的预测。

但我们对蛋白质无序性的精确测定和准确预测目前仍然存在着很多障碍。从无序蛋白质结构的实验测定方法来讲, 目前的实验费用昂贵, 花费时间长, 并且需要大量的专业知识, 因而还不适用于在蛋白质组范围内进行大规模研究。只有结构测定的技术和方法能够实现突破性进展, 才能打破这方面的障碍。从蛋白质无序性的预测方面来看, 现阶段

存在的主要问题包括：如何提高短无序区域以及蛋白质 N 端和 C 端的无序性预测准确率。预测准确率的提高依赖于对无序蛋白质结构更清晰的认识，进而开发出更优的预测方法。相信随着今后相关技术方法的突破和蛋白质无序性研究的深入，我们对无序蛋白质在生命活动中所扮演的角色会有更为全面、深刻的认识。

参 考 文 献

- [1] Boesch C, Bundi A, Oppliger M, *et al.* 1H nuclear-magnetic-resonance studies of the molecular conformation of monomeric glucagon in aqueous solution. *Euro J Biochem*, 1978, **91**(1): 209–214
- [2] Dunker A K, Oldfield C J, Meng J, *et al.* The unfoldomics decade: an update on intrinsically disordered proteins. *BMC genomics*, 2008, **9**(Suppl 2): S1
- [3] Ward J J, Sodhi J S, McGuffin L J, *et al.* Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*, 2004, **337**(3): 635–645
- [4] Romero P, Obradovic Z, Kissinger C R, *et al.* Identifying disordered regions in proteins from amino acid sequences. *IEEE Int Conf Neural Netw*, 1997, **1**: 90–95
- [5] Sim K L, Uchida T, Miyano S. ProDDO: a database of disordered proteins from the Protein Data Bank (PDB). *Bioinformatics*, 2001, **17**(4): 379–380
- [6] Vucetic S, Obradovic Z, Vacic V, *et al.* DisProt: a database of protein disorder. *Bioinformatics*, 2005, **21**(1): 137–140
- [7] Brown C J, Takayama S, Campen A M, *et al.* Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol*, 2002, **55**(1): 104–110
- [8] He B, Wang K, Liu Y, *et al.* Predicting intrinsic disorder in proteins: an overview. *Cell Res*, 2009, **19**(8): 929–949
- [9] Bellay J, Han S, Michaut M, *et al.* Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol*, 2011, **12**(2): R14
- [10] Ringe D, Petsko G A. Study of protein dynamics by X-ray diffraction. *Methods Enzymology*, 1986, **131**(14): 389–433
- [11] Dyson H J, Wright P E. Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance. *Adv Protein Chem*, 2002, **62**: 311–340
- [12] Fasman G D. Circular Dichroism and the Conformational Analysis of Biomolecules. New York: Plenum Press, 1996
- [13] Adler A J, Greenfield N J, Fasman G D. Circular dichroism and optical rotatory dispersion of proteins and polypeptides. *Methods Enzymology*, 1972, **27**(1): 675–735
- [14] Uversky V N, Gillespie J R, Fink A L. Why are "natively unfolded" proteins unstructured under physiologic conditions?. *Proteins: Structure, Function, and Bioinformatics*, 2000, **41**(3): 415–427
- [15] Smyth E, Syme C D, Blanch E W, *et al.* Solution structure of native proteins with irregular folds from Raman optical activity. *Biopolymers*, 2001, **58**(2): 138–151
- [16] Hubbard S, Eisenmenger F, Thornton J. Modeling studies of the change in conformation required for cleavage of limited proteolytic sites. *Protein Science*, 1994, **3**(5): 757–768
- [17] Uversky V. A multiparametric approach to studies of self-organization of globular proteins. *Biochem Biokhimiia*, 1999, **64**(3): 250–266
- [18] Dosztányi Z, Mészáros B, Simon I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Briefings in Bioinformatics*, 2010, **11**(2): 225–243
- [19] Radivojac P, Obradovic Z, Brown C J, *et al.* Prediction of boundaries between intrinsically ordered and disordered protein regions. *Pac Symp Biocomput*, 2003, **8**: 216–227
- [20] Hirose S, Shimizu K, Kanai S, *et al.* POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics*, 2007, **23**(16): 2046–2053
- [21] Dosztányi Z, Csizmok V, Tompa P, *et al.* The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol*, 2005, **347**(4): 827–839
- [22] Dosztányi Z, Csizmok V, Tompa P, *et al.* IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 2005, **21**(16): 3433–3434
- [23] Galzitskaya O V, Garbuzynskiy S O, Lobanov M Y. FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics*, 2006, **22**(23): 2948–2949
- [24] Prilusky J, Felder C E, Zeev-Ben-Mordehai T, *et al.* FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, 2005, **21**(16): 3435–3438
- [25] Melamud E, Moulton J. Evaluation of disorder predictions in CASP5. *Proteins: Structure, Function, and Bioinformatics*, 2003, **53**(S6): 561–565
- [26] Kryshchuk A, Barbato A, Fidelis K, *et al.* Assessment of the assessment: evaluation of the model quality estimates in CASP10. *Proteins: Structure, Function, and Bioinformatics*, 2014, **82**(Suppl 2): 112–126
- [27] Fukuchi S, Sakamoto S, Nobe Y, *et al.* IDEAL: intrinsically disordered proteins with extensive annotations and literature. *Nucleic Acids Res*, 2012, **40**(D1): D507–D511
- [28] Romero P, Obradovic Z, Li X, *et al.* Sequence complexity of disordered protein. *Proteins: Structure, Function, and Bioinformatics*, 2001, **42**(1): 38–48
- [29] Dyson H J, Wright P E. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*, 2005, **6**(3): 197–208
- [30] Dafforn T R, Smith C J. Natively unfolded domains in endocytosis: hooks, lines and linkers. *EMBO Reports*, 2004, **5**(11): 1046–1052
- [31] Fuxreiter M, Simon I, Friedrich P, *et al.* Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J Mol Biol*, 2004, **338**(5): 1015–1026
- [32] Daughdrill G W, Narayanaswami P, Gilmore S H, *et al.* Dynamic behavior of an intrinsically unstructured linker domain is conserved in the face of negligible amino acid sequence conservation. *J Mol*

- Evol, 2007, **65**(3): 277–288
- [33] Wold M S. Replication protein A: a heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism. *Annu Rev Biochem*, 1997, **66**(1): 61–92
- [34] Buljan M, Frankish A, Bateman A. Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol*, 2010, **11**(7): R74
- [35] Ferreon A C M, Ferreon J C, Wright P E, *et al.* Modulation of allostery by protein intrinsic disorder. *Nature*, 2013, **498** (7454): 390–394
- [36] Liu J, Faeder J R, Camacho C J. Toward a quantitative theory of intrinsically disordered proteins and their function. *Proc Natl Acad Sci USA*, 2009, **106**(47): 19819–19823
- [37] Janin J, Sternberg M J. Protein flexibility, not disorder, is intrinsic to molecular recognition. *F1000 Biology Reports*, 2013, **5**(1): 2–8
- [38] Yruela I, Contreras-Moreira B. Genetic recombination is associated with intrinsic disorder in plant proteomes. *BMC Genomics*, 2013, **14**(1): 772
- [39] Schäd E, Kalmar L, Tompa P. Exon-phase symmetry and intrinsic structural disorder promote modular evolution in the human genome. *Nucleic Acids Res*, 2013, **41**(8): 4409–4422
- [40] Siltberg-Liberles J. Evolution of structurally disordered proteins promotes neostructuralization. *Mol Biol Evol*, 2011, **28**(1): 59–62
- [41] Light S, Sagit R, Sachenkova O, *et al.* Protein expansion is primarily due to indels in intrinsically disordered regions. *Mol Biol Evol*, 2013, **30**(12): 2645–2653
- [42] Iakoucheva L M, Radivojac P, Brown C J, *et al.* The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*, 2004, **32**(3): 1037–1049
- [43] Landry C R, Levy E D, Michnick S W. Weak functional constraints on phosphoproteomes. *Trends in Genetics*, 2009, **25**(5): 193–197
- [44] Tompa P, Dosztanyi Z, Simon I. Prevalent structural disorder in *E. coli* and *S. cerevisiae* proteomes. *J Prot Res*, 2006, **5**(8): 1996–2000
- [45] Marín M, Uversky V N, Ott T. Intrinsic disorder in pathogen effectors: protein flexibility as an evolutionary hallmark in a molecular arms race. *Plant Cell Online*, 2013, **25**(9): 3153–3157
- [46] Chen S C C, Chuang T J, Li W H. The relationships among microRNA regulation, intrinsically disordered regions, and other indicators of protein evolutionary rate. *Mol Biol Evol*, 2011, **28**(9): 2513–2520
- [47] Mosca R, Pache R A, Aloy P. The role of structural disorder in the rewiring of protein interactions through evolution. *Mol Cell Proteomics*, 2012, **11**(7): M111. 014969
- [48] Yang D, Zhong F, Li D, *et al.* General trends in the utilization of structural factors contributing to biological complexity. *Mol Biol Evol*, 2012, **29**(8): 1957–1968
- [49] Schäd E, Tompa P, Hegyi H. The relationship between proteome size, structural disorder and organism complexity. *Genome Biol*, 2011, **12**(12): R120
- [50] Romero P R, Zaidi S, Fang Y Y, *et al.* Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci USA*, 2006, **103**(22): 8390–8395
- [51] Vavouri T, Semple J I, Garcia-Verdugo R, *et al.* Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell*, 2009, **138**(1): 198–208
- [52] Vassilev L T, Vu B T, Graves B, *et al.* *In vivo* activation of the p53 pathway by small-molecule antagonists of MDM2. *Science*, 2004, **303**(5659): 844–848
- [53] Tompa P. Unstructural biology coming of age. *Curr Opin Struct Biol*, 2011, **21**(3): 419–425
- [54] Yang D, Jiang Y, He F. An integrated view of the correlations between genomic and phenomic variables. *J Genet Genomics*, 2009, **36**(11): 645–651
- [55] He F. Lifeomics leads the age of grand discoveries. *Sci China Life Sci*, 2013, **56**(3): 201–212
- [56] Zhong F, Yang D, Hao Y, *et al.* Regular patterns for proteome-wide distribution of protein abundance across species. *PloS One*, 2012, **7**(3): e32423
- [57] Alemu E Y, Carl J W, Bravo H C, *et al.* Determinants of expression variability. *Nucleic Acids Res*, 2014: gkt1364

The Identification of Intrinsically Disordered Proteins and Their Structural, Functional, Evolutionary Features*

MA Chong^{1,2)}, YANG Dong^{2)**}, JIANG Ying²⁾, ZHONG Ru-Gang¹⁾, HE Fu-Chu²⁾

⁽¹⁾ College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100022, China;

²⁾ State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing 102206, China)

Abstract Intrinsically disordered proteins (IDPs) do not shape into a stable and well-structured three-dimensional fold, while they are biologically active. The discovery of IDPs is in contradiction to the traditional "structure-function" relationship. In this review, the experimental and computational methods for the identification of IDPs, and the corresponding databases were summarized. Then, we introduce the structural features (including primary structure, secondary structure, disorder of protein domain and the allosteric effect) and functional features of IDPs. We also specially focused on the evolutionary researches of IDPs. The evolutionary mechanisms of the formation of IDPs and the evolutionary rates of disordered regions were described. And the important roles of IDPs' evolution in the evolution of biological function and the increasing of biological complexity were summarized. Finally, we discussed the prospects of IDPs in medical applications. This review is of great significance for the further understanding of IDPs' formation mechanism, structural and functional characteristics and their potential prospects in clinical application.

Key words intrinsically disordered proteins, sequence, structure, function, evolution

DOI: 10.3724/SP.J.1206.2014.00084

* This work was supported by a grant from International Science & Technology Cooperation Program of China(2014DFB30020) and Chinese State Key Projects for Basic Research (2014CBA02001).

**Corresponding author.

Tel: 86-10-80727777-1140, E-mail: yangdongbprc@163.com

Received: March 24, 2014 Accepted: June 17, 2014