PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2009, 36(3): 280~287
www.pibb.ac.cn

结构域相互作用数据库的产生、发展与应用

欧阳玉梅 *
(伊犁师范学院奎屯校区文理系计算机教研室，奎屯 833200)

摘要 结构域是进化上的保守序列单元，是蛋白质的结构和功能的标准组件．典型的两个蛋白质间的相互作用涉及特殊结构域间的结合，而且识别相互作用结构域对于在结构域水平上彻底理解蛋白质的功能与进化、构建蛋白质相互作用网络、分析生物学通路等十分重要．目前，依赖于对实验数据的进一步挖掘和对各种不同输入数据的计算预测，已识别出了一些相互作用 / 功能连锁结构域对，并由此构建了内容丰富、日益更新的结构域相互作用数据库．综述了产生结构域相互作用的 8 种计算预测方法．介绍了 5 个结构域相互作用公共数据库 3DID、iPfam、InterDom、DIMA 和 DOMINE 的有关信息和最新动态．实例概述了结构域相互作用在蛋白质相互作用计算预测、可信度评估，蛋白质结构域注释，以及在生物学通路分析中的应用．

关键词 蛋白质相互作用(PPI)，蛋白质结构域，结构域相互作用(DDI)，数据库
学科分类号 Q51，Q503
DOI: 10.3724/SP.J.1206.2008.00437

蛋白质 - 蛋白质相互作用(protein-protein interactions, PPI)在很多的生命过程和细胞活动中都扮演着非常重要的角色，是生物网络大规模特性研究、预测未知 PPI、疾病机理研究、药物设计等的宝贵资源．

结构域是蛋白质进化上的保守序列单元，是蛋白质结构和功能的标准组件．它携带了许多特征，包括催化活性、底物结合、结构特征，而且包括介导蛋白质间或蛋白质与其他分子间物理相互作用的分子连接物．典型的蛋白质含有多种结构域，当然，一种结构域在一个蛋白质中也可能有多个拷贝、相同的结构域在不同的蛋白质中可能处在不同的位置．两个蛋白质间的相互作用涉及特殊结构域间的结合(即一个蛋白质中的一个结构域与另一个蛋白质中的一个结构域或较小的缩氨酸模体的结合)，而且识别相互作用结构域对于理解 PPI 十分重要．结构域 - 结构域相互作用(domain-domain interactions, DDI)的模式在同种生物体内可重复，并且可以存在于不同的生物体中，说明这些模式在生物界中被广泛保留了下来．这些模式组成了“蛋白质识别码”，成为破译 PPI 的密码．

然而，当前的生物技术(不论是小规模实验，还是酵母双杂交系统等高通量实验)都只能够检测到 PPI，不能够提供结构域水平上的相互作用特异性洞察[1]，即实验技术还不能够揭示关于相互作用界面和蛋白质复合物形成的结构细节．而提取 DDI 又是一项单调乏味和艰辛的工作．所以一直以来对于结构域水平上的关系，可使用的主要数据库很少．近年来，随着更多物种的完整基因组被测定，以及对蛋白质序列、结构、功能、家族、结构域及 PPI 研究的逐步深入和相关数据库[2]的发展，DDI 提取研究已出现了良好的开端．这项复杂技术，首先进行数据收集，然后设计运算法则(用公式来解决高度复杂、特异的问题)，最后应用当代强大的计算能力(in silico)挖掘目标数据．目前，依赖于对实验数据的进一步挖掘和对各种不同输入数据的计算预测，已识别出了一些相互作用 / 功能连锁(functionally linked)结构域对，并由此构建了多个内容丰富、日益更新的 DDI 数据库．它们是在结构域水平上彻底理解蛋白质功能与进化、构建 PPI 网络、分析生物学通路(pathway)的重要资源．如何高效地存储、管理和发展这些数据，如何深入分析、充分利用这些数据，已成为生物信息学的一个崭新课题．

* 通讯联系人.
Tel: 0992-3288820, E-mail: fang9352@163.net
收稿日期：2008-06-16，接受日期：2008-09-11

1 预测结构域相互作用的计算方法

近年来，已提出了一些计算方法[3]致力于挖掘基因组水平上的尚未识别的 DDI. 它们是最大似然估计与贝叶斯网络、结构域融合分析、结构域系统发育谱、结构域对排除分析、随机决策森林结构、依赖简约原则的线性规划、结构域对相关共进化和消息传递算法等. 由于许多方法的输入数据是实验数据，所以事实上它们是实验技术与计算技术不同程度的联合. 一些方法推断了潜在的蛋白质结构域间的功能连锁(如结构域融合分析、结构域系统发育谱等)，而不能确定是否是直接的物理相互作用. 一些方法既可用于 DDI 预测，也可用于 PPI 预测(如序列共进化、系统发育谱等). 一些方法是从有结构域注释的蛋白质序列出发的，还有许多方法是在高通量的 PPI 数据集上训练出来的.

DDI 的计算预测是一项庞大的工程. 例如，扫描各物种的所有已收录蛋白质序列(SWISS-PROT55.5 包含的蛋白质序列条目已达 389 046 个)鉴定结构域融合. 对 460 个已完全测序基因组运用结构域系统发育谱方法推断结构域功能连锁等. 用于推断的输入数据可能是生物信息数据库中各物种的基因组，蛋白质序列、三维结构、结构域，PPI 和蛋白质复合物等. 算法设计涉及数据挖掘、模式识别、统计学等领域. 程序实现面临海量数据、高维特征、时空复杂度问题. 结果验证，通常考察与已知相互作用结构域对(例如从实验得到的蛋白质结构数据中提取的 DDI) 的比较、与高质量的 PPI 的吻合以及与其他预测方法的覆盖等.

1.1 最大似然估计与贝叶斯网络

最大似然估计(maximum likelihood estimation，MLE)是统计推断中参数估计的一种基本方法. 它利用样本分布密度构造似然函数，通过求似然函数的最大值得到总体的未知参数的最大似然估计值. 贝叶斯网络(bayesian network models，BNM)是用来表示变量间连接概率的图形模式，它提供了一种自然的表示因果信息的方法，用来发掘数据间的潜在关系. 它可以处理不完整和带有噪声的数据集，能够利用已有的知识和观测数据进行学习和预测.

Deng 等[4](2002 年)首先应用 MLE 方法从一个给出的 PPI 网络中推断 DDI. 这种方法假设 DDI 是彼此独立的，即两个结构域发生或不发生相互作用与其他结构域无关，而且只考虑蛋白质间单结构域对的相互作用. 在求出已观测到的 PPI 数据的似然函数的最大值后，估计每对结构域间的相互作用概率. 使用期望最大化算法(expectation maximization, EM)解决缺失数据问题，同时还通过处理 PPI 中的假阳率与假阴率优化概率.

Lee 等[5](2006 年)则用贝叶斯方法整合了使用 MLE、GO(gene ontology)注释和结构域融合信息的多基因组特征的 DDI 预测. 用 MLE 方法估计贝叶斯模型参数，在更多物种(酵母、线虫、果蝇和人类)的 PPI 网络上，基于似然率得分，提取到高可信度的 DDI.

1.2 结构域融合分析

观察发现，在一个物种中的一些相互作用蛋白质对在另外一个物种中可能存在融合成一个单一蛋白质肽链的同源物. 结构域融合分析(domain-fusion analysis, DFA)鉴定在一个物种中的两个成分蛋白质(component proteins)组成另一物种的融合蛋白质(fusion proteins)的情况. 这个方法又被称为 Rosetta Stone 方法. 对于那些在第一个物种中单独存在而在第二个物种中融合在一起的蛋白质或结构域，可认为它们在第一个物种中是一个功能相关的通路的一部分，存在相互作用. 因为融合后的物种由于共调控和两个蛋白质可同时拥有较高的局部浓度而具有进化上的优势. 通过比较各物种的蛋白质肽链成分，结构域融合分析寻找这样的结构域对来推断 DDI.

Ng 等[6](2003 年)应用这个方法扫描了当时的 SWISS-PROT 数据库中 7 000 个以上物种的所有收录的蛋白质序列，产生出 4 792 个推断的 DDI. 显然，该方法受到融合事件发生频率低的限制，综合其他方法，Ng 等推出了第一个 DDI 数据库 InterDom.

1.3 结构域系统发育谱

系统发育谱(phylogenetic profiles，PP)是为预测蛋白质间的功能相关和相互作用而提出的一种方法. 通过分析一组完全测序的基因组的直系同源簇同时存在或不存在的模式来实现预测. 这种存在或不存在的模式被称作系统发育谱，它依赖于蛋白质相关进化. 因为具有相似系统发育谱的蛋白质趋向于功能相关，所以可以根据系统发育谱进行蛋白质聚类来获得未知蛋白质的功能信息和相互作用信息，即当未知蛋白质与一个或更多的功能已知的蛋白质归为一组(簇)时，则认为它们是功能连锁的.

Pagel 等[7,8] (2004 年，2007 年)提出的结构域系统发育谱(domain phylogenetic profiles，DPP)是结

构域水平上的而不是基因组水平上的系统发育谱，即用结构域代替了完整蛋白质．该方法的基本步骤如下：a．选定待研究的一组蛋白质 P_1、P_2、…、P_m，它们至少应具有一个或一个以上的已知结构域信息，整理出它们涉及的所有不同的结构域 D_1、D_2、…D_n．b．选定一组完全测序的基因组 G_1、G_2、…、G_q，它们是编码这些蛋白质不同组合的．c．按照结构域在某基因组中出现得分为 1，不出现得分为 0，得到结构域系统发育谱，对结构域系统发育谱按某给定阈值(例如阈值为 2表示最多允许两个二进制位不同，超过的认为没有相互作用)进行聚类，相同组的被认为是功能连锁的，由此推断出 DDI 网络．d．由蛋白质所包含的结构域信息和结构域系统发育谱可进一步得到完整蛋白质的系统发育谱，同样地，按某给定阈值对蛋白质的系统发育谱进行聚类也可得到 PPI 网络．显然，这种方法的准确性依赖于完全测序的基因组数目和系统发育谱方法的可靠性．Pagel 等推出了 DDI 数据库 DIMA．

1.4 结构域对排除分析

Riley 等[9](2005 年)提出的结构域对排除分析(domain pair exclusion analysis，DPEA)是一种通过已知 PPI 预测 DDI 的方法．该方法的基本步骤如下：a．对于已经证实的多个物种的 PPI 数据集，在蛋白质结构域数据库中为每个蛋白质寻找到它的所有结构域，定义每对共同出现在相互作用蛋白质中的结构域为潜在的相互作用结构域对；b．计算含有结构域 i 的蛋白质和含有结构域 j 的蛋白质 PPI 的频率 S_{ij}；c．把 S_{ij} 作为初始假设，利用期望最大化算法评估每种潜在 DDI 的倾向性 θ_{ij}；d．评估一个给定的 DDI 被排除后对 PPI 似然率产生的影响，得到每个推断的 DDI 的证据 E_{ij}，通过较高的 E_{ij} 预测出高可信度的 DDI 对．

1.5 随机决策森林结构

各种分类方法已成功应用于预测 PPI 与 DDI．这些方法使用已知数据源训练分类器区别正、负样本，再用构造好的分类器对待识别的样本进行分类决策．众多探索表明，随机决策森林结构(random decision forest framework，RDFF)是最优分类器，支持向量机(support vector machines，SVM)位居第二[3]．

决策树(decision trees) 是以实例为基础的归纳学习算法．传统的决策树分类方法(如 ID3 和 C4.5)对于相对较小的数据集是很有效的，但是，应用于大规模的数据处理时，其有效性就显得不足．

RDFF 采用了一种提升(boosting)方法．它构造了多个决策树．这些决策树是在从原始训练集中随机挑选出的多个不同子集上生长出来的．由于子集空间的特征维数比原始训练集的低得多，森林中的每一棵树被允许无修剪地生长到可能的最大程度．训练集中的每个子集被分配了一个权值，权值表示该子集对于分类器的重要性．构造好的 RDFF 以加权投票方式对测试集进行分类决策．它是一种适合于高维数据库的组合多分类器算法，在大规模的数据处理中减少了对系统资源的占用，分类速度快、准确率高．

Chen 等[10](2005 年)首先将 RDFF 用于 DDI 预测．对于 PPI 这个两类分类问题，每一蛋白质对是一个样本，由组成它们的结构域来描述．若由 n 个样本构成的样本集中所有蛋白质涉及的 Pfam 结构域总数目为 m(例如，m=4293)，则每个样本可以用一个与 m 个结构域对应的 m 维特征向量来表示．每维属性的值可取 0、1、2．在某个样本中，如果两个蛋白质都不包含这个结构域，则属性值取 0；如果仅有一个包含，则属性值取 1；如果两个都包含，则属性值取 2．这样，每个样本被编码为由 0、1、2 构成的长度为 m 的代码串．基于相互作用蛋白质对训练集构造出 RDFF，构造好的 RDFF 用来对测试集进行分类决策．对于每一个预测正确的 PPI 对，可以通过在生成这个分类的正确决策树分类器中追踪其产生的分枝或路径来提取涉及的结构域，从而推断出 DDI．这种方法不仅可以用于推断单结构域对相互作用，也可以用于推断多结构域对相互作用．

1.6 依赖简约原则的线性规划

线性规划(linear programming，LP)是一种优化方法，是对由线性等式或不等式约束的线性函数的极大化或极小化．可以用它们作为适合的线性规划函数的约束条件来解线性不等式组．

Guimarães 等[11](2006 年)首先使用了依赖简约原则(parsimony principle)的线性规划方法预测 DDI．提出的简约原则解释如下：通过识别结构域对的最小加权集来预测 DDI 伴体，只要该加权集对于证实一个给出的 PPI 网络是有效的．该方法首先给出一个 PPI 网络，然后对于可能证实两个蛋白质间相互作用的每一结构域对，用线性规划计算出一个取值在(0，1)之间的线性规划得分(LP-score)．PPI 网络中的假阳性通过一个概率结构(P-scores)来

处理．结构域对的 LP 得分在某一域值以上的被认为存在相互作用．

Guimarães 等[12](2008 年)后来又提出了广义的简约原则解释．调整了结构域定义粒度以适应输入数据集的粒度，允许 DDI 具有不同的代价，即允许优先选择那些可能介导蛋白质间相互作用的所谓的“共生结构域”．

1.7　结构域对相关共进化

如果说几个蛋白质存在于同一个大分子复合物中，或几个蛋白质共同参与一个代谢或信号传导途径，那么，这些蛋白质的关系被称为“功能连锁”．所谓共进化，指的是进化的选择压力常常使功能连锁蛋白质的进化相互影响．这个概念背后的假设是相互作用伴体必须共进化，以便在一个蛋白质的对接面中的改变能在它的伴体界面中被补足．

Jothi 等[13](2006 年)研究发现，两个相互作用蛋白质间的那些相互作用结构域对的共进化程度比非相互作用结构域对的高，由此提出预测 DDI 的结构域对相关共进化(relative co-evolution of domain pairs，RCDP)方法．这是一个在序列水平上理解 DDI 的尝试．在这种分子系统发生分析中，首先对一套物种集进行每对蛋白质 / 结构域的多重序列比对，比对结果用来构建系统发生树和相似性矩阵．两个结构域的共进化程度是通过计算两个相似矩阵的线性相关系数(这实际上是暗中比较了两个结构域的进化历史)来度量的．与已知相互作用结构域对的相关得分比较，较高得分的推断为相互作用对，反之为非相互作用对．

1.8　消息传递算法

Iqbal 等[14](2008 年)提出的消息传递算法(message-passing algorithms，MPA) 是一种从 PPI 数据预测 DDI 的方法．基于消息传递的置信传播(belief propagation，BP)算法是一个强大的和广泛使用的推断方法．它是一个因子图模型，其中各个节点之间传递的消息是概率密度函数或置信信息．在迭代过程中，消息不断更新直到收敛，常按某些迭代终止策略来终止消息传递和更新．该方案中的贝蒂自由能提供了量化 PPI 实验数据噪声和查明那些最有疑问的数据的途径，因为是它们导致了 DDI 模式中的矛盾．该方法输入的是蛋白质相互作用集以及有关蛋白质的所有结构域信息，输出的是每对结构域间相互作用的概率表．

2　结构域相互作用数据库

自 2003 年出现第一个综合在线 DDI 数据库 InterDom 之后，现已陆续有了几个特点各异的数据库，见表 1．

Table 1　Databases of domain-domain interactions[1)]

表 1　结构域-结构域相互作用数据库 [1)]

数据库	类型 [2)]	相互作用数目	涉及结构域数目	网址	引用
3DID	S	5 038	3 496	http://gatealoy.pcb.ub.es/3did	[15]
iPfam	S	4 030	2 500	http://www.sanger.ac.uk/Software/Pfam/iPfam	[16]
InterDom	F	148 938	8 957	http://interdom.i2r.a-star.edu.sg	[17]
DIMA	F, S	28 870	7 038	http://mips.gsf.de/genre/proj/dima2	[18]
DOMINE	F, S	20 513	4 036	http://domine.utdallas.edu/cgi-bin/Domine	[19]

[1)] 数据收集 2008 年 8 月 26 日; [2)] 类型说明 S：结构数据；F：功能预测．

早在 2003 年，可用于构建综合 DDI 数据库的结构数据还很少，Ng[6]等开创性地尝试了从其他数据源计算预测 DDI 来构建推断的数据库．由于其中的 DDI 是计算预测得到的，InterDom 不同于从结构数据中提取 DDI 的数据库．3DID 与 iPfam 是从结构数据中提取的，被称为已观测到的 DDI 或结构数据 DDI，可用来检验计算预测方法的准确度．DIMA 和 DOMINE 目前是联合多种计算预测方法、整合实验与预测资源的综合数据库．毕竟，由于目前已知蛋白质序列数目(38 万以上)远大于已知蛋白质结构数目(5 万多)，有限的结构数据 DDI 在大规模应用时存在很大的局限性．

另外，还有许多蛋白质相互作用界面数据库，诸如 Pibase、ProtCom、CBM、InterPare、SCOPPI. 它们多数是从结构数据中提取的，这些界面包含 DDI，但也涉及其他机制，例如，线性模体等．可以相信这些数据库是关于相互作用物理界面的．还有一些数据库其中也包含相互作用结构域，例如，IntAct 生物分子相互作用开放资源数据库就是如此．本文仅介绍表 1 中的 5 个 DDI 数据库．

依赖于蛋白质家族、蛋白质结构分类等各种数据库，蛋白质结构域有多种定义，如 Pfam、SCOP、CDD、CATH、Interpro 定义等．在 PPI 界面研究中较多地使用了 SCOP、CDD 或 CATH 定义，Interpro 定义试图整合多个数据库的资源．本文介绍的 5 个 DDI 数据库均采用的是 Pfam 定义．

2.1 3DID 数据库

3DID(3D interacting domains)数据库收集了已知高分辨率三维结构的蛋白质 DDI．各蛋白质和复合物的三维结构来自 PDB(Protein Data Bank)数据库[20]．PDB 是著名的生物大分子结构数据库，收录了由 X 射线晶体衍射和核磁共振等实验测定的蛋白质三维结构档案，记录有原始结构数据：原子坐标、配基的化学结构和晶体结构的描述等．至 2008 年 4 月 8 日，PDB 结构档案已达到它 37 年来历史上的一个重要里程碑，超过了 5 万个．

2005 年 1 月发布第一版 3DID．2008 年 6月发布的基于 Pfam22.0 的当前版，在考察 120 250 个已知三维结构的蛋白质肽链(不独立)产生的总数达 120 524 个(不独立)DDI 的基础上，聚合整理得到涉及 3 496 个 Pfam 结构域的 5 038 个独立的 DDI，其中分子链内 788 个，分子链间 3 501 个，既在分子链内又在链间观察到的 749 个．

3DID 开发了结构信息，提供对于理解相互作用如何发生所必须的标准的分子详细资料．提供结构域的查询、浏览，相互作用的查询、下载．有基于结构域、序列、GO、SCOP 类的查询；有多种视图的相互作用输出页，这些视图是结构域、相互作用、路径、PDB、序列等视图．由欧洲分子生物学实验室(EMBL)管理的 3DID 实现了每逢周日更新的诺言．

2.2 iPfam 数据库

首先介绍 Pfam(Protein families database)数据库[21] (http://pfam.sanger.ac.uk)．Pfam 是基于序列多重比对和隐马尔可夫模型的蛋白质家族和结构域的大型综合数据库．Pfam 23.0 (2008 年 7 月)不仅基于 UniProtKB 序列数据库，而且基于 NCBI GenPept 和从微生物环境基因组学 (metagenomics)项目中挑选的序列，目前已包括 10 340 个蛋白质家族．

iPfam(interaction Pfam)数据库是 Pfam 数据库的子数据库．它描述在 PDB 中收录的已观测到的 DDI．以 PDB 中储存的大量的多结构域蛋白质和蛋白质复合物为基础，通过计算结构上足够接近并能形成相互作用的残基之间的距离和原子之间的键(范德华力、侧链和主链上的氢键、盐桥和二硫化物等)识别相互作用．

2005 年 7 月推出了基于 Pfam12.0 的第一版 iPfam，收录涉及 2 500 个独立 Pfam 结构域的 4 030 个 DDI．可从 iPfam 中获取或浏览每一个结构域的家族页面．并可从不同细节水平上和从结构或序列视角上获取 iPfam 中的数据信息．可以观看 DDI 图，也可从原子水平上进行较深层次的分析，并允许在图形界面进行关系数据表下载．由 Sanger 研究院生物信息学团队 (http://www.sanger.ac.uk/bioinfo)管理的 iPfam 目前正在进行基于最新 Pfam 版本的更新．

2.3 InterDom 数据库

InterDom(database of interacting domains)数据库是从多种数据资源计算推断的蛋白质 DDI 数据库．2003 年首次发布．2007 年 7 月发布的 InterDom2.0，推断出潜在的 DDI 为 148 938 对，涉及 8 957 个 Pfam 结构域．其中从蛋白质复合物(来自 PDB)中推断出 7 718 个(此为计算推断，非结构数据提取)，从 PPI (来自 DIP 与 BIND)中推断出 143 820 个，从结构域融合分析(来自 SWISS-PROT)中推断出 4 631 个．InterDom 基于它使用的资源数据库 PDB、DIP 与 BIND、SWISS-PROT 的更新而更新．通过采用综合策略的计算方法，为来自不同数据源的多种方法独立提取的 DDI 赋予高可信度得分，提高了计算机模拟筛选(in silico)的质量．

InterDom 提供了多种形式的网络服务：为探测的或预测的 PPI 提供的验证；基于蛋白质、结构域、序列的查询；依赖结构域名、证据类型、可信度得分等的浏览；所有 InterDom DDI 对的表格形式的下载等．

2.4 DIMA 数据库

DIMA(domain interaction map) 数据库是 MIPS (munich information center for protein sequences)数据库[22]的子数据库．MIPS 由慕尼黑蛋白质序列信息中心建立，主要提供来自全基因组蛋白质的分析与注释．

2004 年推出的第一版 DIMA1.0，是一个基于结构域系统发育谱的简单网络服务．2008 年 1 月发布的 DIMA2.0 在基因组覆盖、结构域范围以及预测方法方面有了很大的改进．它联合了 DDI 方面的实验结构数据与来自两种不同算法(结构域系统发育谱、结构域对排除分析)的预测数据，已逐

渐发展成为一个综合预测资源．收录了 28 870 个 DDI 对，涉及超过 460 个原核生物与真核生物的已测序完整基因组的 7 038 个 Pfam 结构域．提供了基于 Interpro ID 和 Pfam ID 的 DDI 查询、网络拓扑图形绘制及基于 Pfam ID 的 DDI 下载．DIMA 还为每个结构域提供了外部数据库 Pfam、Interpro 的交叉链接．

2.5 DOMINE 数据库

DOMINE(database of protein domain interactions)数据库是一个已知的和预测的蛋白质 DDI 数据库．2008 年 2 月发布的 DOMINE1.1 包括在 PDB 中的已知相互作用结构域对和那些使用 Pfam 结构域定义由 8 种不同计算方法预测的相互作用对．它整合了近年来的多项研究成果，从 10 种不同资源提取 DDI 数据，包括 iPfam、3DID 数据库以及最大似然估计、相关序列共进化、统计方法、结构域融合(InterDom1.2，2004 年)、依赖简约原则的线性规划、结构域对排除分析、随机决策森林结构和系统发育谱(DIMA1.0, 2004 年)方法预测的结果．

DOMINE 收录了涉及 4 036 个结构域的 20 513 个 DDI．其中 4 349 个相互作用是从 PDB 数据库收录的实验结构数据推断得到的，17 781 个是至少由一种计算方法预测得到的．在计算方法预测的 17 781 个相互作用中，有 3 143 个是高可信度的(使用多种信息源预测得到或至少有两个显著不同的方法预测得到)，729 个是中等可信度的(具有相同的 GO 项)，其余的 13 909 个是低可信度的．

DOMINE 提供了 DDI 的查询、浏览与下载，下载文件夹还提供了结构域的 Interpro ID 和 Pfam ID 的对照表．它为每个结构域提供了外部数据库 Pfam、Interpro、GO 的链接，用户只要一个单击就能获得有关这个结构域的更多相关信息．

3 结构域相互作用应用举例

3.1 在蛋白质相互作用预测中的应用

PPI 预测的关键问题是特征提取与算法设计．决定 PPI 的因素多种多样，反映 PPI 的现象也多种多样．由于 PPI 的动态性、瞬时性、多样性、复合性，导致从高通量蛋白质组学数据中提取生物相关特征仍然是一大挑战．因为蛋白质结构域是蛋白质的功能单位，而 PPI 主要通过 DDI 来完成，在结构域水平上的 PPI 建模与分析可能具有更好的广泛性和深刻性，即预测 PPI 的高相关信息来自它们的蛋白质结构域的结构．

Singhal 等[23](2007 年)提出了一个基于结构域的预测方法(DomainGA，结构域遗传算法)．它是一个量化 DDI 的遗传算法类的机器学习方法．该算法产生把 DDI 划分为高、低和模糊三大类的得分集．用此得分集预测一对蛋白质是否发生相互作用．

Kim 等[24] (2007 年)提出了使用结构域信息预测 PPI 的贝叶斯方法．该方法通过从几种生物整合数据来同时评估 DDI 概率、高通量数据的假阳率和假阴率，而且还建立了一个模型，把反映每个蛋白质中结构域数量的 DDI 概率与 PPI 概率联系在一起．

3.2 在评估大规模 PPI 可信度中的应用

随着基因组规模的高通量实验鉴定技术和计算预测方法的发展，出现了大量 PPI 数据．但大规模 PPI 数据中较高比例的假阳性影响了相互作用数据的质量．评估 PPI 可信度的一个关键问题就是特征选择与综合属性抽取．综合多种特征将得到更显著的效果，目前各种探索正在不断出现．如果蛋白质对之间存在潜在的 DDI 则 PPI 存在的可能性较大．

Ramírez 等[25] (2007 年)计算分析了人类 PPI 蛋白质相互作用网络．基于 GO 功能注释、结构上的蛋白质家族的 DDI、似然率和网络拓扑参数，比较评估了预测得到的数据、来自酵母双杂交的高通量实验结果和从文献中挖掘得到的 PPI 数据．

3.3 在蛋白质结构域注释中的应用

理论上，如果确定了构成蛋白质的那些结构域的功能，则蛋白质的功能可以被直接推定．尽管结构域在功能基因组中扮演着非常重要的角色，到目前为止，已有注释的结构域并不多，而且自动完成预测结构域功能的工作也寥寥无几．与 PPI 广泛用于预测蛋白质功能一样，DDI 也能提供一个推断结构域功能的途径．

Zhao 等[26](2008 年)提出了两种预测结构域功能的方法，基于阈值的分类方法和支持向量机方法．他们综合了多种信息资源：蛋白质 - 结构域映射特征、DDI、结构域共存特征．这项研究不仅提高了预测精度、可信度，而且为注释结构域开采了 DDI 信息．

3.4 在通路研究中的应用

细胞中众多蛋白质功能的信息可以被整合在数据库中，并且以蛋白质网络图(protein network map)的形式显示出来．一个通路就是一个相互联系在一

起的生化反应的集合．将生化反应组织成通路的目的是为了更好地展现细胞中复杂的生物过程．PPI数据和蛋白质中特定结构域之间的相互作用信息可以用于通路作图．

Ng 等[27] (2008 年)探索了在通路研究中 DDI 的应用．使用结构域联合对方法从 7 个物种的 PPI 中提取推定的蛋白质 DDI．为了确定预测性能，把提取结果与 InterDom 数据库的数据进行了比较，发现平均匹配率达到了大约 76%．基于预测的 DDI 结果重构了几个真实的 PPI 通路，如趋化性通路、血凝通路、MAPK 通路等．更进一步地，引入了一个被称作“AP-order index”的新量为 6 个 PPI 通路预测调控次序．发现新方法有很好的预测准确度，“AP-order index”是一个确定 PPI 调控次序的非常可信的参数．

4 结　　语

结构域相互作用数据库是近几年才开始被研究者发掘并利用的生物学工具．已在联合多种方法、整合不同资源、以网络形式可视化分析方面迈出了可喜的步伐．未来可做的工作还有许多：增加文献挖掘、添加新方法、随着资源的更新而更新、设计联合得分方案、使用多种数据资源提高预测准确度、评估可信度等．另外，也应看到，由于实验验证的极度匮乏以及方法上的偏好(蛋白质结构数据除外)，目前还几乎不能建立起一个评判 DDI 的“黄金标准”．由于许多计算预测的输入数据是实验数据(例如 PPI 数据)，实验数据的不完全、有噪声、种属性和静态性也必然影响预测的质量．针对这些 DDI 数据库发展的难点和重点，已提出一些解决方案并在逐步执行，但其任务还很艰巨，需要计算生物学家和实验生物学家的进一步共同努力．相信随着 DDI 数据库的不断完善，其应用领域也将越来越深入而广泛．

参 考 文 献

1 Shoemaker B A, Panchenko A R. Deciphering protein-protein interactions. Part Ⅰ. Experimental techniques and databases. PLoS Comput Biol, 2007, **3** (3): e42

2 余鑫煜, 许正平. 蛋白质相互作用数据库及其应用. 中国生物化学与分子生物学报, 2008, **24**(3): 189～196
Yu X Y, Xu Z P. Chin J Biochem Mol Biol, 2008, **24**(3): 189～196

3 Shoemaker B A, Panchenko A R. Deciphering protein-protein interactions. Part Ⅱ. Computational methods to predict protein and domain interaction partners. PLoS Comput Biol, 2007, **3**(4): e43

4 Deng M, Mehta S, Sun F, *et al*. Inferring domain-domain interactions from protein-protein interactions. Genome Res, 2002, **12**(10): 1540～1548

5 Lee H, Deng M, Sun F, *et al*. An integrated approach to the prediction of domain-domain interactions. BMC Bioinformatics, 2006, **7**: 269

6 Ng S K, Zhang Z, Tan S H. Integrative approach for computationally inferring protein domain interactions. Bioinformatics, 2003, **19**(8): 923～929

7 Pagel P, Wong P, Frishman D. A domain interaction map based on phylogenetic profiling. J Mol Biol, 2004, **344**(5): 1331～1346

8 Pagel P, Strack N, Oesterheld M, *et al*. Computational prediction of domain interactions. Methods Mol Biol, 2007, **396**: 3～15

9 Riley R, Lee C, Sabatti C, *et al*. Inferring protein domain interactions from databases of interacting proteins. Genome Biol, 2005, **6**(10): R89

10 Chen X W, Liu M. Prediction of protein-protein interactions using random decision forest framework. Bioinformatics, 2005, **21**(24): 4394～4400

11 Guimarães K S, Jothi R, Zotenko E, *et al*. Predicting domain-domain interactions using a parsimony approach. Genome Biol, 2006, **7**(11): R104

12 Guimarães K S, Przytycka T M. Interrogating domain-domain interactions with parsimony based approaches. BMC Bioinformatics, 2008, **9**: 171

13 Jothi R, Cherukuri P F, Tasneem A, *et al*. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. J Mol Biol, 2006, **362**(4): 861～875

14 Iqbal M, Freitas A A, Johnson C G. Message-passing algorithms for the prediction of protein domain interactions from protein-protein interaction data. Bioinformatics, 2008, **24**(18): 2064～2070

15 Stein A, Russell R B, Aloy P. 3did: interacting protein domains of known three-dimensional structure. Nucl Acid Res, 2005, **33** (Database issue): D413～D417

16 Finn R D, Marshall M, Bateman A. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. Bioinformatics, 2005, **21**(3): 410～412

17 Ng S K, Zhang Z, Tan S H, *et al*. InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. Nucl Acid Res, 2003, **31**(1): 251～254

18 Pagel P, Oesterheld M, Tovstukhina O, *et al*. DIMA 2.0—predicted and known domain interactions. Nucl Acid Res, 2008, **36**(Database issue): D651～D655

19 Raghavachari B, Tasneem A, Przytycka T M. DOMINE: a database of protein domain interactions. Nucl Acid Res, 2008, **36**(Database issue): D656～D661

20 Berman H, Henrick K, Nakamura H, *et al*. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucl Acid Res, 2007, **35**(Database issue): D301～D303

21 Finn R D, Tate J, Mistry J, *et al*. The Pfam protein families database. Nucl Acid Res, 2008, **36**(Database issue): D281～288

22 Mewes H W, Dietmann S, Frishman D, *et al*. MIPS: analysis and

annotation of genome information in 2007. Nucl Acid Res, 2008, **36** (Database issue): D196～201

23 Singhal M, Resat H. A domain-based approach to predict protein-protein interactions. BMC Bioinformatics, 2007, **8**: 199

24 Kim I, Liu Y, Zhao H. Bayesian methods for predicting interacting protein pairs using domain information. Biometrics, 2007, **63** (3): 824～833

25 Ramírez F, Schlicker A, Assenov Y, *et al*. Computational analysis of human protein interaction networks. Proteomics, 2007, **7**(15): 2541～2552

26 Zhao X M, Wang Y, Chen L. Protein domain annotation with integration of heterogeneous information sources. Proteins, 2008, **2** (1): 461～473

27 Ng K L, Huang C H, Liu H C, *et al*. Applications of domain-domain interactions in pathway study. Comput Biol Chem, 2008, **32** (2): 81～87

The Birth, Development and Applications of Domain-Domain Interaction Databases

OUYANG Yu-Mei*

(*Department of Liberal Arts & Science, Yili Normal University, Kuitun* 833200, China)

Abstract Domains are evolutionarily conserved sequence units and they are structural and functional building blocks of proteins. Interaction between two proteins typically involves binding between specific domains, and identifying interacting domain pairs is an important step towards thoroughly understanding protein function and evolution, constructing protein-protein interaction (PPI) networks, and analyzing pathway at the domain level. A number of interacting and/or functionally linked domain pairs have been identified and the information was organized and hosted in many domain-domain interactions (DDI) databases with the help of further mining experimental data and computational predictions from various input data. First, the 8 computational predicting methods used to acquire DDI data will be introduced. Then the introduction of DDI public databases, including 3DID, iPfam, InterDom, DIMA and DOMINE will be given. And finally, some examples are described about applications of DDI in computational predicting interacting protein pairs, assessment of the reliability for PPI, protein domain annotation, and in pathway study.

Key words protein-protein interactions (PPI), protein domain, domain-domain interactions (DDI),database
DOI: 10.3724/SP.J.1206.2008.00437

*Corresponding author.
Tel: 86-992-3288820, E-mail: fang9352@163.net
Received: June 16, 2008 Accepted: September 11, 2008