

人类蛋白质结构互作网络 ——结构域对网络拓扑与蛋白质功能的影响*

陈丽娜^{1)**} 王倩^{1)**} 尚玉奎^{1)**} 张良才¹⁾ 孙钊²⁾ 何伟明³⁾
赵研¹⁾ 李琬¹⁾ 王宏¹⁾ 何月涵¹⁾ 李霞¹⁾

(¹⁾ 哈尔滨医科大学生物信息科学与技术学院, 哈尔滨 150086; (²⁾ 哈尔滨医科大学 2005 级七年制, 哈尔滨 150086;

(³⁾ 哈尔滨工业大学光电子技术研究所, 哈尔滨 150086)

摘要 高通量的蛋白质互作数据与结构域互作数据的出现, 使得在蛋白质组学领域内研究人类蛋白质结构互作网络, 进一步揭示蛋白质结构与功能间的潜在关系成为可能. 蛋白质上广泛分布的结构域被认为是蛋白质结构、功能以及进化的基本功能单元. 然而, 结合蛋白质的结构信息(例如蛋白质结构域数目、长度和覆盖率等)来研究这些表象后的内部机制仍然面临着挑战. 将蛋白质分为单结构域蛋白质与多结构域蛋白质, 并进一步结合蛋白质互作信息与结构域互作信息构建了人类蛋白质结构互作网络; 通过与人类蛋白质互作网络进行比较, 研究了人类蛋白质结构互作网络的特殊结构特征; 对于单结构域蛋白质与多结构域蛋白质, 分别进行了功能富集分析、功能离散度分析以及功能一致性分析等. 结果发现, 将结构域互作信息综合考虑进来后, 人类蛋白质结构互作网络可以提供更多的单纯的蛋白质互作网络无法提供的细节信息, 揭示蛋白质互作网络的复杂性.

关键词 结构域, 蛋白质结构互作网络, 蛋白质功能

学科分类号 Q50

DOI: 10.3724/SP.J.1206.2009.00640

随着高通量蛋白质互作数据的积累, 蛋白质互作信息已经被广泛地应用到功能模块的挖掘^[1]、蛋白质复合物的鉴别^[2]以及进化网络的分析^[3-4]中. 然而, 单纯利用蛋白质互作数据的分析方法仍然面临着以下几个问题: a. 由传统的高通量技术所产生的蛋白质互作数据存在大量的假阳性或假阴性信息^[5]; b. 蛋白质互作网络中存在大量的噪声信息干扰了蛋白质的功能注释; c. 多数的蛋白质都由功能不同的结构域组合而成, 而蛋白质的功能并不是结构域功能的简单叠加^[6]. 因此, 结合其他信息对蛋白质互作网络进行分析, 有利于揭示蛋白质互作网络的细节信息^[7-8].

结构域是蛋白质维持其结构特征的基本单元, 并且通常与特定的功能相关, 例如, 催化活性或者与其他生物分子的相互作用^[9-10], 而结构域间的相互作用直接反映了宿主蛋白质之间的物理互作. 幸运的是, 现有的一些实验技术与预测方法已经可以成功地获取多物种的基因、蛋白质等序列信息, 构

建识别分子结构域结构的合理模型^[11-12], 同时一系列相关的结构域数据库业已建立, 比如 Pfam^[13], CATH^[14], SCOP^[15]等.

最近, 有多篇文章研究了结构域区域或结构域互作与蛋白质互作网络间的关系^[16-19], 发现结构域或者结构域互作关系对蛋白质互作网络的进化、模块性有相应的影响. 然而, 在大多数情况下, 结构域对蛋白质互作网络拓扑性质与宿主蛋白质功能特征的影响被忽略了. 而充分考虑结构互作构建的人类蛋白质结构互作网络, 可以揭示蛋白质互作网络

* 哈尔滨医科大学研究生创新基金资助项目(HCXS2009007), 黑龙江省自然科学基金项目(D2007-48, JC200711), 哈尔滨医科大学医学基础学科创新群体基金项目和国家自然科学基金资助项目(30571034).

** 共同第一作者.

*** 通讯联系人.

Tel: 0451-86674768, E-mail: chenlina_2004@yahoo.com.cn

收稿日期: 2009-10-30, 接受日期: 2010-01-29

的潜在复杂性并有助于进一步理解结构域对宿主蛋白质的影响。

在本文中, 我们的主要目标是研究结构域对人类蛋白质结构互作网络的拓扑性质与宿主蛋白质功能特征的影响。首先, 根据蛋白质上结构域的数目将蛋白质分为单结构域蛋白质与多结构域蛋白质, 并结合结构域互作信息构建了人类蛋白质结构互作网络; 同时, 与蛋白质互作网络进行比较, 分析了人类蛋白质结构互作网络的拓扑性质; 然后分别对单结构域蛋白质与多结构域蛋白质进行了功能富集; 最后, 研究了上述两类蛋白质不同的互作模式并进一步讨论了不同的互作模式的功能一致性。综上所述, 人类蛋白质结构互作网络能够提供更多的分子互作的细节信息, 有助于理解蛋白质的物理互作。

1 材料与方法

1.1 数据与预处理

整合了几个不同的高通量蛋白质互作数据库^[20-23]的人类蛋白质互作数据, 构建了包含 8 049 个节点与 31 476 条边的蛋白质互作网络。然后利用 iPfam 数据库^[24]构建了结构域互作网络, 并将互作的结构域映射到蛋白质互作网络中, 只保留蛋白质相应的结构域也互作的蛋白质互作关系。

1.2 拓扑性质分析

利用 Cytoscape^[25]中的 Network Analyzer^[26]插件对人类蛋白质结构互作网络的基本拓扑性质进行分析, 主要包括网络平均聚类系数、网络直径与最大连通子图。此外, 还分析了单纯由单结构域蛋白质与多结构域蛋白质所构成的网络的度分布与网络密度。

网络平均聚类系数是指网络中所有节点的聚类系数的均值; 对于互作邻居小于两个的节点, 其聚类系数设置为 0。网络直径是指节点间最大的距离, 对于不完全连通的网络来说, 网络直径指的是连通子图之间的最大距离。一个复杂的网络中一般含有多个连通子网络, 最大连通子图即可以简单地理解为包含最多数目节点的子网络。在一个无向网络中, 某个节点的连接度即为与该节点相连的节点的数目, 而网络的度分布则表示为具有连接度 $k(k=0, 1, 2, \dots)$ 的节点数目的分布。网络密度指的是标准化后的网络中节点平均的连接度。

1.3 超几何分布

对于每一个结构域, 利用超几何分布^[27], 分别

计算了它在宿主蛋白质集合中出现的显著性 P 值。假设在一个包含 C 个蛋白质的蛋白质集合中, 有 $k \leq C$ 个蛋白质包含结构域 X , 则结构域 X 在这个集合中的显著性即为

$$P = 1 - \sum_{i=0}^k \frac{\binom{C}{i} \binom{G-C}{n-i}}{\binom{G}{n}} \quad (1)$$

最终的 P 值表示了某个结构域在不同的集合中的富集程度, 其中 G 代表了网络中蛋白质的总数, $n(n \leq G)$ 代表了网络中含有结构域 X 的蛋白质数目。

1.4 功能多样性与连接度

为了比较单 / 多结构域蛋白质的功能多样性与连接度之间的关系, 我们从 Gene Ontology (GO)^[28] 数据库中下载了基因的功能注释文件。由于大部分基因都有相应的功能注释, 因此可以构建一个 $M \times N$ 的基因 - 功能矩阵, 其中 M 代表了所有的功能条目, N 分别代表了单 / 多结构域蛋白质的数目。在矩阵中, 1 代表了某个基因 - 功能的匹配, 也就是某个基因具有某种功能, 0 代表了某个基因 - 功能的缺失或者基因功能未知, 所以网络中的每个基因在矩阵中都有一列由 1 或者 0 组成的列向量与其对应。然后分别统计了每一连接度下单 / 多结构域蛋白质的功能数目, 将基因 - 功能矩阵 $M \times N$ 转换为功能 - 连接度矩阵 $M \times D_s$ 与 $M \times D_m$ (s 代表单结构域蛋白质, m 代表多结构域蛋白质), D 代表了网络中蛋白质的最大连接度。

由于单 / 多结构域蛋白质数目的不同, 每一连接度下所对应的单 / 多结构域蛋白质的数目也不相同。因此, 定义了“平均功能类”来校正这一偏倚,

$$\bar{T} = \frac{T_d}{P_d} \quad (2)$$

其中, $\bar{T}$ 代表了“平均功能类”, P_d 代表了不同连接度下蛋白质的总数, T_d 代表了对应连接度下功能条目的总数。

1.5 功能离散度

为了构建“功能离散度”这一特征, 我们定义了 $G_{P-G}=(P, G, E_{P-G})$ 作为一个连接蛋白质与功能条目的二分图, $P=\{p_1, p_2, \dots, p_{|P|}\}$ 表示了蛋白质集合, $G=\{g_1, g_2, \dots, g_{|G|}\}$ 为功能条目集合并且 $E_{P-G} \subseteq P \times G$ 。对于每一个功能条目 $g \in G$, 根据 g 在蛋白质上的分布, 构建了 E_{P-G} 的关联关系。进一步, 功能离

散度可以表示如下：

$$\begin{cases} x = \frac{k}{P} \\ y = G \end{cases} \quad (3)$$

P 代表了单结构域蛋白质或者多结构域蛋白质的集合，其中 $k(k \leq P)$ 个蛋白质具有某个功能条目 X ， G 代表可以映射到同样数目蛋白质上的功能条目数。

例如， P_1, P_2 与 P_3 是单结构域蛋白质，并且都注释到同一功能条目 X 上。假设单结构域蛋白质的数目为 3 000，而可以映射到 3 个单结构域蛋白质的功能条目数为 50，则 x 的值即为 0.001； y 值即为 50。

2 结 果

2.1 人类蛋白质结构互作网络

我们首先整合构建了人类蛋白质互作网络，包含 8 049 个节点与 31 476 条边。在进行结构域互作映射后，人类蛋白质结构互作网络包含 3 617 个节点与 6 994 条边，其中单结构域蛋白质为 2 409 个，多结构域蛋白质为 1 208 个，共包含结构域 975 个及结构域之间的互作关系 2 387 对。

2.2 拓扑性质分析

构建人类蛋白质结构互作网络后，根据蛋白质

的分类将上述网络分为单结构域蛋白质网络与多结构域蛋白质网络。人类蛋白质结构互作网络的平均聚类系数为 0.126，网络密度为 0.001；而多结构域蛋白质网络的密度为 0.003，单结构域蛋白质网络的密度为 0.001。结果显示，多结构域蛋白质之间的连接性要比单结构域蛋白质之间的连接性高。

度分布是互作网络的重要属性。分析了蛋白质互作网络与人类蛋白质结构互作网络的度分布。这两种网络都呈现 power-law 分布。这说明人类蛋白质结构互作网络具有生物网络的一般属性并且具有潜在的生物学意义。

2.3 结构域分布

为了分析两类蛋白质上不同的结构域分布，将结构域分为 3 类：单结构域蛋白质特异性结构域、多结构域蛋白质特异性结构域与混合型结构域。其中单结构域蛋白质特异性结构域仅出现在单结构域蛋白质中，多结构域蛋白质特异性结构域仅出现在多结构域蛋白质中，混合型结构域同时存在于两类结构域中。

对每一个结构域都计算了它在两类蛋白质中出现的显著性(见 1.3)，并且使用错误发现率(FDR)作为后续的多重检验校正方法。结果如图 1 所示。

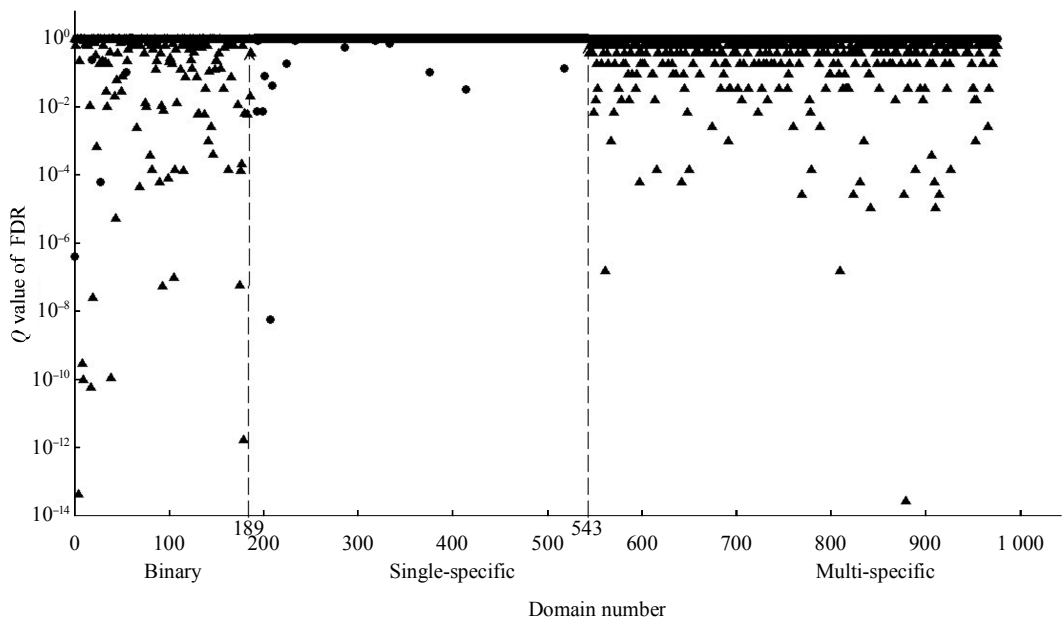


Fig. 1 Enrichment of domains

The x -axis represents each domain and the y -axis represents enrichment significance. The left 189 are binary domains; the middle ones (354) are single-specific and the right (432) are multi-specific. The black horizontal line represents $P=0.05$ significance. ● : Single-domain protein; ▲ : Multi-domain protein.

从图 1 中, 可以发现, 在 $P \leq 0.05$ 的阈值下, 有少数的结构域在单结构域蛋白质集合或者多结构域蛋白质集合中显著富集. 换句话说, 这些结构域的出现频率远远高于期望值. 表 1 中列出了显著性最高的 20 个结构域. 其中, 在单结构域蛋白质集中仅有 Ras family 结构域(Pfam acc. PF00071)与七次跨膜受体结构域(Pfam acc. PF00001)呈现富集性, 这说明单结构域蛋白质上的结构域具有特异性. 而富集显著性最高的结构域: zf-C4 结构域(Pfam acc. PF00105), Pkinase_Tyr 结构域(Pfam acc. PF07714)、

PH 结构域(Pfam acc. PF00169), 以及 Pkinase_C 结构域和 EGF 结构域 (Pfam acc. PF00433 and Pfam acc. PF00008) 都与重要的功能相关, 例如 DNA 绑定、DNA 修复以及细胞凋亡等.

同时我们发现, 这些结构域的宿主蛋白质也倾向于较高的连接度(表 1), 而且宿主蛋白质间的相互作用也比人类蛋白质结构互作网络紧密(网络密度: 0.007). 这说明, 单 / 多结构域蛋白质上不同的结构域分布会影响宿主蛋白质间的相互作用.

Table 1 The 20 most enriched protein domains

	Pfam ID	Domain annotation	P-value of hyper geometric	Q value of FDR ¹⁾
	PF00105	Zinc finger	0	0
★	PF07714	Pkinase_Tyr	0	0
★	PF00169	PH domain	0	0
	PF00433	Pkinase_C	1.11E-16	2.71E-14
★	PF00008	EGF-like domain	2.22E-16	4.33E-14
★	PF07679	I-set	1.02E-14	1.66E-12
★	PF00041	Fibronectin type III domain	4.09E-13	5.70E-11
★	PF00018	SH3 domain	7.96E-13	9.70E-11
	PF00104	Hormone_receptor	1.00E-12	1.08E-10
★	PF00017	SH2 domain	2.96E-12	2.89E-10
★ #	PF00071	Ras family	5.76E-12	5.62E-09
★	PF00047	Immunoglobulin domain	2.85E-10	2.53E-08
★	PF00560	Leucine rich repeat	6.56E-10	5.33E-08
★	PF07645	EGF_CA	7.78E-10	5.84E-08
	PF00621	RhoGEF domain	1.35E-09	9.40E-08
	PF08441	Integrin_alpha2	2.46E-09	1.50E-07
	PF01839	FG-GAP repeat	2.46E-09	1.50E-07
#	PF00001	7 transmembrane receptor	8.36E-10	4.08E-07
	PF00130	C1 domain	9.18E-08	5.27E-06
★	PF00271	Helicase conserved C-terminal domain	2.04E-07	1.05E-05

¹⁾ FDR: False discovery rate. The 20 most of enriched protein domains in human proteins, listed in a descending order. Domains preceded with ★ showed that the degree >10. Domains preceded with # showed that they are enriched in single-domain protein set.

2.4 互作模式

从研究中发现, 大多数的结构域都位于多结构域蛋白质上并且超过 70% 的结构域存在自身互作现象(689 个结构域). 因此可能存在这样一个假设: 多结构域蛋白质更倾向于与其他多结构域蛋白质相互作用并且多结构域蛋白质中自身互作的现象应该更加明显.

为了验证上述假设, 计算了每一个蛋白质的直接互作邻居中的单结构域蛋白质的比例, 相应的结果见图 2. 我们发现, 单结构域蛋白质与多结构域蛋白质的直接互作邻居具有显著的差异, 也就是说这两类蛋白质具有不同的互作倾向性(Wilcoxon rank sum test, $P \leq 0.05$): 大多数单结构域蛋白质的直接互作邻居也是单结构域蛋白质(Wilcoxon rank

sum test, $P \leq 0.05$), 而大多数多结构域蛋白质倾向于与多结构域蛋白质相互作用(Wilcoxon rank sum test, $P \leq 0.05$). 更进一步的分析发现, 多结构域蛋

白质呈现出更强的自身互作现象(Chisq test $P < 0.05$). 这说明在不同种类的蛋白质上的结构域分布可以影响宿主蛋白质的互作模式.

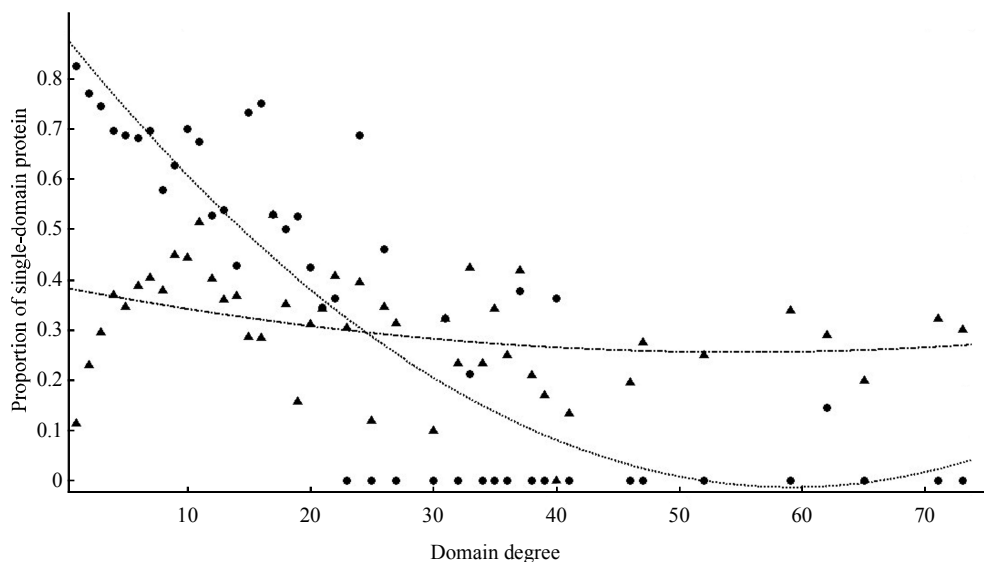


Fig. 2 Interacting pattern tendency

The x -axis represents the degree and the y -axis represents the proportion of single-domain protein of the direct neighbors. ●: Single-domain protein; —: Single-domain protein; ▲: Multi-domain protein; ---: Multi-domain protein.

2.5 功能多样性与连接度

在人类蛋白质结构互作网络中, 分析了蛋白质的功能与连接度之间的联系. 首先分别对单/多结

构域蛋白质进行功能注释, 并计算在相应的连接度下蛋白质的功能条目的总数(图 3a). 我们发现, 随着连接度的增加, 功能条目的数量在减少, 这可能

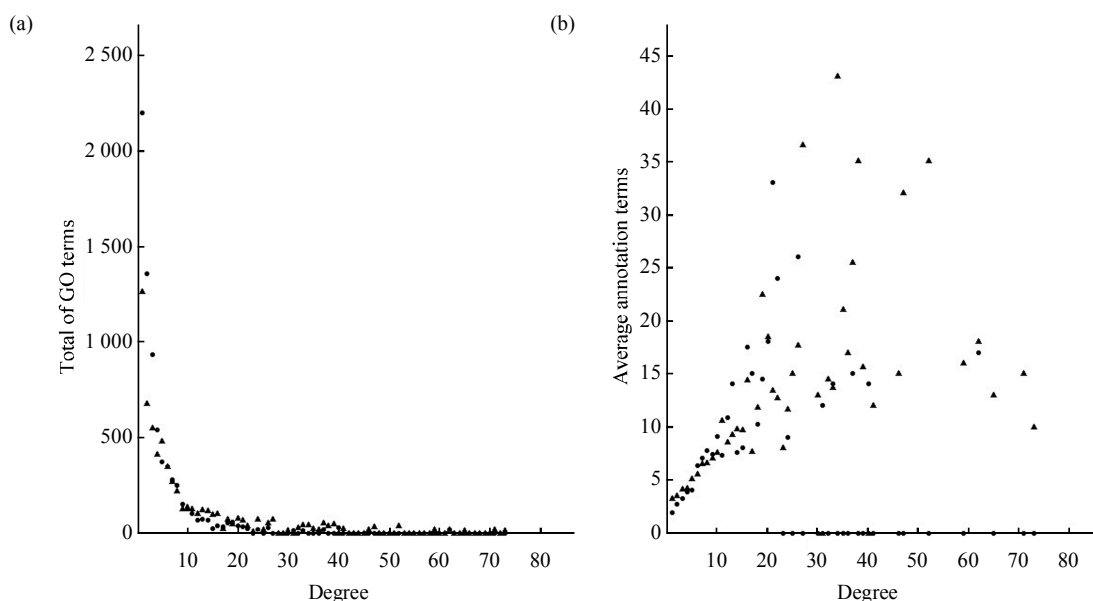


Fig. 3 Connectivity and function diversity

(a) It represents the total annotation terms of each connectivity. (b) The relationship of the average annotation terms to the connectivity for single-/multi-domain proteins. ●: Single-domain protein; ▲: Multi-domain protein.

是由于连接度较高的蛋白质数目较少造成的。因此，蛋白质数目对结果造成的影响应当被考虑进来。然后我们计算了每一连接度下蛋白质的“平均功能类”(图 3b)。结果显示，“平均功能类”的数目与蛋白质的连接度呈正相关(线性回归，单结构域蛋白质： $n=0.28d+5.994$ ，多结构域蛋白质： $n=0.23.3d+8.004$)。这说明，随着连接度的增加，蛋白质的功能呈现多样化的趋势，并且这一趋势在单结构域蛋白质与多结构域蛋白质中同样存在，没有差异。

从图 3b 中可以看出，当连接度增加到一定程度时(60~73)，蛋白质的平均功能类急剧下降。推断这可能是由于以下两种原因造成的：a. 较高连接度的蛋白质数目较少；b. 这些蛋白质可能在某些蛋白质复合物中起到核心的作用。为了证实我们的推断，首先，计算了连接度大于 45 的蛋白质的数目。结果发现，除了连接度为 62 的蛋白质有多个以外，其他的连接度都只对应了一个蛋白质。其次，分析了 MIPS 数据库^[29]提供的蛋白质复合物数据。Swiss-Prot: P26993 (degree=73) 与 Swiss-Prot: P29353 (degree=62) 都参与了 SLP-76- Cbl-Grb2-Shc 复合物的形成；Swiss-Prot: P19174 (degree=65) 与 Swiss-Prot: P27986 (degree=59) 参与了 PDGFRA-PLC-gamma-1-PI3K-SHP-2 复合物与 LAT-PLC-

gamma-1-p85-GRB2-SOS 复合物的形成。高连接度的蛋白质倾向于参与蛋白质复合物的形成，并行使相同或相似的功能。由于以上两点，在连接度较高的时候，功能多样性反而降低了。

2.6 功能离散度

虽然在连接度与功能多样性的关系上，单结构域蛋白质与多结构域蛋白质没有显著差异，但是由于结构域分布与互作模式的不同，这两类蛋白质在其他的功能特征上仍然可能存在差异性。我们引入“功能离散度”的概念来表征这两类蛋白质在功能参与度上的不同。

在一个确定的蛋白质 - 功能二分图 $G_{P-GO}=(P, G, E_{P-GO})$ 中，如果有大量的功能条目仅与少量的蛋白质相连，那么可认为这个蛋白质集合的功能离散度较高。

分别计算了单结构域蛋白质集合与多结构域蛋白质集合的功能离散度，结果见图 4。我们发现，单结构域蛋白质的功能离散度更高，这也与单结构域蛋白质上的结构域分布更具特异性有关。此外，多结构域蛋白质更倾向于组成相互作用的蛋白质复合物共同执行相同或者相似的功能，而单结构域蛋白质不具有这个特性(Chisq test, $P < 0.05$)，这点也与我们之前的研究结果相吻合。

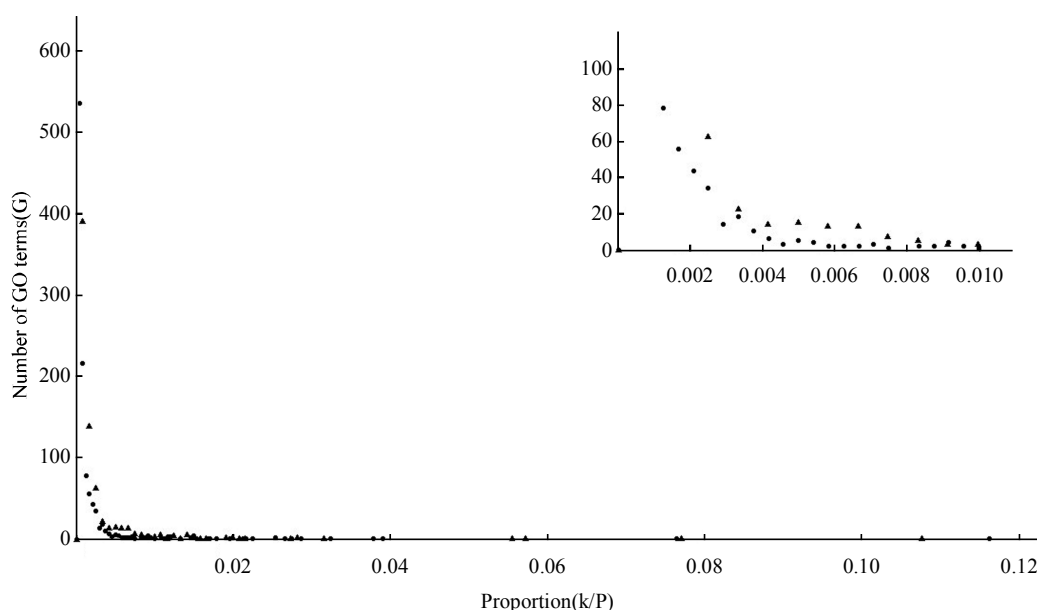


Fig. 4 Function discrete

● : Single-domain protein; ▲ : Multi- domain protein.

2.7 功能一致性

2.7.1 蛋白质集合功能分析. 对单 / 多结构域蛋白质进行功能注释后, 对这两类蛋白质集合进行了 Kappa^[30]检验来评价两类蛋白质集合中蛋白质功能一致性的高低. 发现这两类蛋白质集合的功能一致性分值较低(Kappa Statistics, $K=0.213$), 说明这两类蛋白质的功能存在差异性. 又进一步选择所有蛋白质作为背景参考集合, 对单结构域蛋白质与多结构域蛋白质分别进行功能富集 / 非富集分析(Over-/Under-Representation, ORA)^[31]. 大多数多结构域蛋白质在与分子绑定相关的功能类上呈现富集状态, 而单结构域蛋白质除在电压门控通道活性(voltage-gated ion channel activity)功能外, 没有富集现象出现.

2.7.2 互作模式功能分析.

我们分析了不同互作模式的蛋白质相互作用间的功能一致性. 根据连接的蛋白质性质, 将蛋白质互作关系分为 3 类: a. 混合类型. 一个是单结构域蛋白质另一个是多结构域蛋白质 (1 971 pairs).

b. 单结构域互作类型. 两个都是单结构域蛋白质 (2 663 pairs). c. 多结构域互作类型. 两个都是多结构域蛋白质 (2 360 pairs). 然后, 统计了互作的两个蛋白质共享的公共的功能条目. 结果发现, 第一种互作类型的功能一致性最低(图 5a). 这可能是由于单结构域蛋白质与多结构域蛋白质的功能差异性造成的. 单结构域互作类型与多结构域互作类型的功能一致性明显增高, 但这两者间没有明显差异.

接下来, 针对第二种与第三种互作类型, 我们同样采用了 Kappa 检验来衡量互作蛋白质间的功能一致性. 有趣的是单结构域蛋白质互作对间的功能一致性明显高于预期. 一个可能的合理解释就是在单结构域互作类型中自身互作的绝对值要大于多结构域互作类型(在单结构域互作类型中, 自身互作对为 909, 多结构域互作类型中, 自身互作对为 650). 在去除自身互作后, 多结构域互作类型的功能一致性明显高于单结构域互作类型(t test, $P < 0.05$)(图 5b).

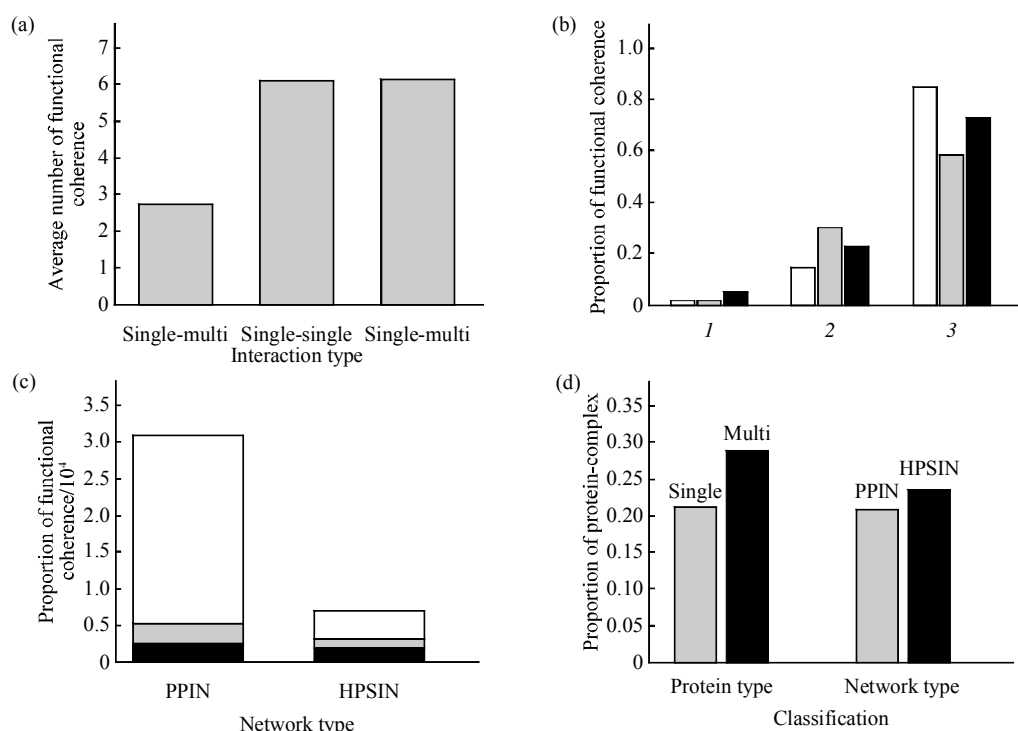


Fig. 5 Function coherence and protein complex

(a) Average number of functional coherence. (b) Proportion of coherent function. In this Figure, the different color represented the different interaction type. The different groups represented the different coherence. The left group indicated the coherence is higher than 0.75 (Kappa Statistics, $K > 0.75$), the middle segment represents coherence is between 0.4 and 0.75 and the right one represent the coherence is lower than 0.4. 1: High-coherence; 2: Medium-coherence classification of coherent function; 3: Low-coherence. □: Single-multi; □: Single-single; ■: Multi-multi. (c) The proportion of coherent function of PPIN and HPSIN. ■: High coherence; □: Medium coherence; □: Low-coherence. (d) The left group describes the proportion of single-domain protein and multi-domain protein, and the right segment indicates the proportion of proteins, which participate in the protein-complexes, in PPIN and HPSIN.

2.7.3 网络功能分析. 我们比较了人类蛋白质结构互作网络与人类蛋白质互作网络中互作蛋白质的功能一致性. 结果显示, 人类蛋白质结构互作网络的功能一致性分值明显高于蛋白质互作网络(图 5c). 此外, 与蛋白质互作网络中的蛋白质与单结构域蛋白质比较, 人类蛋白质结构互作网络中的蛋白质与多结构域蛋白质更倾向于参与蛋白质复合物的形成 (Chisq test, $P \leq 0.05$, 图 5d). 上述结果表明, 在滤除冗余信息后构建的人类蛋白质结构互作网络比单纯的蛋白质互作网络更好地反应了互作蛋白质间的功能一致性.

3 讨 论

为了证明文章中结果的稳定性, 我们将上述方法应用到其他模式物种的蛋白质互作网络中, 由于线虫、果蝇与小鼠的结构域信息尚不完备, 因此选择了对酵母的结构互作网络进行进一步的分析. 首先我们利用酵母的蛋白质互作网络与结构域信息, 构建了酵母蛋白质结构互作网络并进行了上述分析, 结果显示文章中的结果与结论在酵母生物网络中同样成立.

如结果所述, 我们发现单结构域蛋白质与多结构域蛋白质之间存在显著的差异. 例如, 多结构域蛋白质之间的相互作用比单结构域蛋白质明显紧密, 因此多结构域蛋白质在网络拓扑性质的维持上起到了更为重要的作用, 而且多结构域蛋白质之间的联系也会比单结构域蛋白质广泛. 此外, 多结构域蛋白质更倾向于自身互作、参与蛋白质复合物的形成, 因此多结构域蛋白质更倾向于执行相同或者相似的功能, 它们之间的功能一致性分值也更高.

这些差异性都反映了结构域对网络拓扑性质与宿主蛋白质功能的影响. 那么为什么结构域会影响这些特征呢? 研究显示, 蛋白质上结构域的数目是造成这些差异的原因之一. 单结构域蛋白质只有一个结构域, 这在很大程度上限制了单结构域蛋白质与其他蛋白质的相互作用; 而多结构域蛋白质有两个或者两个以上的结构域, 它们则可以同时与多个蛋白质相互作用. 而相互作用的蛋白质数目直接影响了信息传递与功能交互的广泛性, 单结构域蛋白质在同一时间内只能与一个蛋白质相互作用而执行某种特定的功能, 多结构域蛋白质可以同时与多个蛋白质相互作用共同执行某种功能, 因此, 多结构域蛋白质之间的功能一致性更高. 我们注意到, 研

究结果可能会由于现有的蛋白质结构数据的不完备性存在一定的限制, 但是以人类蛋白质结构互作网络为基础, 可以与其他一些类型的数据相结合进行更全面的分析, 例如共表达信息、序列信息以及进化信息等, 这样就可以为揭示细胞内复杂的分子间的相互作用提供更好的平台.

综上所述, 在将结构域互作信息引入到蛋白质互作网络中后, 蛋白质结构互作网络揭示了生物互作网络潜在的复杂性(例如互作模式、功能离散与功能一致性). 而这些复杂性单纯地由蛋白质互作网络是无法揭示的.

参 考 文 献

- [1] Rives A W, Galitski T. Modular organization of cellular networks. *Proc Natl Acad Sci USA*, 2003, **100**(3): 1128–1133
- [2] Spirin V, Mirny L A. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA*, 2003, **100** (21): 12123–12128
- [3] Fraser H B, Hirsh A E, Steinmetz L M. Evolutionary rate in the protein interaction network. *Science*, 2002, **296**(5568): 750–752
- [4] Brown K R, Jurisica I. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol*, 2007, **8**(5): R95
- [5] Titz B S, Schlesner M, Uetz P. What do we learn from high-throughput protein interaction data?. *Expert Rev Proteomics*, 2004, **1**: 111–121
- [6] Pandey J, Koyutürk M, Subramaniam S. Functional coherence in domain interaction networks. *Bioinformatics*, 2008, **24**(16): i28–34
- [7] Kiel C, Beltrao P, Serrano L. Analyzing protein interaction networks using structural information. *Annu Rev Biochem*, 2008, **5**: 5.1–5.27
- [8] Lee I, Date S V, Adai A T. A probabilistic functional network of yeast genes. *Science*, 2004, **306**(5701): 1555–1558
- [9] Costa S, Cesareni G. Domains mediate protein-protein interactions and nucleate protein assemblies. *Handb Exp Pharmacol*, 2008, (186): 383–405
- [10] Pawson T, Nash P. Assembly of cell regulatory systems through protein interaction domains. *Science*, 2003, **300**: 445–452
- [11] Veretnik S, Bourne P E, Alexandrov N N. Toward consistent assignment of structural domains in proteins. *J Mol Biol*, 2004, **339**(3): 647–678
- [12] Holland T A, Veretnik S, Shindyalov I N. Partitioning protein structures into domains: why is it so difficult?. *J Mol Biol*, 2006, **361**(3): 562–590
- [13] Finn R D, Mistry J, Schuster-Böckler B. Pfam: clans, web tools and services. *Nucleic Acids Res*, 2006, **34**(Database issue): D247–251
- [14] Greene L H, Lewis T E, Addou S. The CATH domain structure database: new protocols and classification levels give a more

- comprehensive resource for exploring evolution. *Nucleic Acids Res*, 2007, **35**(Database issue): D291–297
- [15] Murzin A G, Brenner S E, Hubbard T. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 1995, **247**(4): 536–540
- [16] Itzhaki Z, Akiva E, Altuvia Y. Evolutionary conservation of domain-domain interactions. *Genome Biol*, 2006, **7**(12): R125
- [17] Xia K, Fu Z, Hou L. Impacts of protein-protein interaction domains on organism and network complexityPPI domains and organism/network complexity. *Genome Res*, 2008, **18**: 1500–1508
- [18] Schuster-Bockler B, Bateman A. Reuse of structural domain-domain interactions in protein networks. *BMC Bioinformatics*, 2007, **8**: 259
- [19] Kim P K, Lu L J, Xia Y. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 2006, **314**(5807): 1938–1941
- [20] Stark C, Breitkreutz B J, Reguly T. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*, 2006, **34**(Database issue): D535–539
- [21] Salwinski L, Miller C S, Smith A J. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, 2004, **32** (Database issue): D449–451
- [22] Persico M, Ceol A, Gavrila C. HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics*, 2005, **6**(Suppl 4): S21
- [23] Mishra G R, Suresh M, Kumaran K. Human protein reference database--2006 update. *Nucleic Acids Res*, 2006, **34** (Database issue): D411–414
- [24] 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
- [25] Shannon P, Markiel A, Ozier O. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 2003, **13**(11): 2498–2504
- [26] Assenov Y, Ramírez F, Schelhorn S E. Computing topological parameters of biological networks. *Bioinformatics*, 2008, **24** (2): 282–284
- [27] King A D, Przulj N, Jurisica I. Protein complex prediction *via* cost-based clustering. *Bioinformatics*, 2004, **20**(17): 3013–3020
- [28] Ashburner M, Ball C A, Blake J A. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 2000, **25**(1): 25–29
- [29] Pagel P, Kovac S, Oesterheld M. The MIPS mammalian protein-protein interaction database. *Bioinformatics*, 2005, **21** (6): 832–834
- [30] Byrt T, Bishop J, Carlin J B. Bias, prevalence and kappa. *J Clin Epidemiol*, 1993, **46**(5): 423–429
- [31] Backes C, Keller A, Kuentzer J. GeneTrail--advanced gene set enrichment analysis. *Nucleic Acids Res*, 2007, **35** (Web Server issue): W186–192

Human Protein Structural Interaction Network: Domain Effects on Network Topology and Protein Function*

CHEN Li-Na^{1)***}, WANG Qian^{1)**}, SHANG Yu-Kui^{1)**}, ZHANG Liang-Cai¹⁾, SUN Zhao²⁾,
HE Wei-Ming³⁾, ZHAO Yan¹⁾, LI Wan¹⁾, WANG Hong¹⁾, HE Yue-Han¹⁾, LI Xia¹⁾

¹⁾College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, China;

²⁾2005,7-Programs, Harbin Medical University, Harbin 150086, China;

³⁾Institute of Opto-electronics, Harbin Institute of Technology, Harbin 150086, China)

Abstract Availability of high-throughput protein-protein interaction and domain-domain interaction information make it possible to study human structural interaction network, and uncover the inner relationship between structure and function of proteins in proteomics area. The widely-distributed domains are thought to affect structure and function of proteins. However, it is still a challenge to investigate potential mechanisms of these effects combining the structural information (e.g domain number, domain length and domain coverage). The proteins were classified into single- or multi- domain proteins, then human protein structural interaction network was constructed with classification information by integrating the protein-protein interaction and the domain-domain interaction data. Furthermore, comparing with the protein-protein interaction network, specific structure characteristics of human protein structural interaction network were studied. And with respect to single-/multi-domain proteins, function enrichment analysis was carried out for their functions. With domain-domain interaction considered, human protein structural interaction network could provide more detailed information distinct from networks based on protein-protein interaction datasets only, and might reveal the underlying complexity of protein-protein interaction network from the perspective of protein classification. Human protein structural interaction network could exploit domain information to provide additional and crucial protein interaction details necessary for understanding what human structural interactions imply.

Key words domain, protein structural interaction network, protein function

DOI: 10.3724/SP.J.1206.2009.00640

*This work was supported by grants from The Master Innovation Funds of Harbin Medical University (HCXS2009007), The Natural Science Foundation of Heilongjiang Province (D2007-48, JC200711), Population Innovation Funds of Harbin Medical University and The National Natural Science Foundation of China(30571034).

**These authors contributed equally to this work.

***Corresponding author.

Tel: 86-451-86674768, E-mail: chenlina_2004@yahoo.com.cn

Received: October 30, 2009 Accepted: January 29, 2010