综述与专论

转录组与蛋白质组比较研究进展

吴松锋 朱云平* 贺福初*
(军事医学科学院放射医学研究所，北京 100850)

摘要 转录组和蛋白质组比较研究发现，总体而言其间的相关性不高. 根据数据的类型不同可以将现有的研究分为 4 类：单点比较、两点差异比较、多点时序比较和多点非时序比较. 对其差异原因的研究和分析表明：除了由实验系统及数据类型不同导致的差异外，转录后蛋白质合成各步骤所受到的限制，以及在此过程中的分子调控也对其有重要的影响；而且不同基因，不同组织和细胞在不同状态下可能也会有差异. 因此，结合转录组和蛋白质组的表达谱研究倾向于利用蛋白质组和转录组研究的差异和互补性，同时对生物体特定状态下的基因和蛋白质表达水平进行全方位度量，以获得表达谱的全景图，并挖掘受到转录后调控的基因.

关键词 转录组，蛋白质组，相关性，翻译调控，密码子偏性
学科分类号 Q50

在生物学研究中以表征全部基因的表达情况为目的的研究被称为转录组表达谱研究，相应的以表征全部蛋白质表达情况为目的的研究被称为蛋白质组表达谱研究. 生物在遭受某种刺激或病变时往往会伴随着某些基因或蛋白质表达量的变化，其中有些变化是引起生物体变化的原因，有些则是生物体变化的结果[1]. 在医学研究中，对这些基因或蛋白质表达量及其变化情况的测定为寻找病因，发掘相应的药物提供了直接的依据.

在生物系统中，基因组是遗传信息的储存体，mRNA (转录组) 是基因表达的中间体，功能性蛋白质 (蛋白质组) 是基因功能的执行体[2]. 因此基因组和转录组的不对应关系是理所当然的，而 mRNA 和蛋白质是基因表达的两个不同阶段的产物，它们之间到底是什么关系却难以下定论.

真核生物中从基因组转录出 mRNA 是由转录因子调控的，而在 mRNA 翻译出蛋白质的过程中，虽然没有和转录因子相似的调控，但也有诸多因素影响整个翻译过程以及翻译后的情况，比如密码子偏性等. 在众多因素影响下，转录组和蛋白质组究竟在多大程度上是对应的？转录组在多大程度上可以推测出其蛋白质组？存在此差异的原因和机理是什么？这些问题自转录组和蛋白质组的研究开始以来逐渐受到关注.

1 转录组和蛋白质组比较研究

转录组和蛋白质组一般都是对特定状态下特定的生物、组织、细胞或细胞器的 mRNA 和蛋白质表达水平的度量，鉴定结果包括所研究的生物样品表达的基因、蛋白质类型，以及其丰度. 为了讨论方便，我们把对处于特定状态下的特定生物、组织、细胞或细胞器的研究称为一个“点”. 根据此定义，现有的关于转录组和蛋白质组比较的研究可以分为 4 种类型 (表 1).

1.1 单点比较

单点比较是指针对某个时间点的样品同时进行转录组和蛋白质组的研究，此类研究耗费较少，数据相对容易得到，研究也较早.

1996 年日本研究者在肝脏 bodymap 数据库的基础上，讨论了人类肝脏的转录物和其分泌到血液中蛋白质丰度 (共 10 个分泌蛋白) 的关系，发现其间有较好的正相关关系[3]. 随后 Anderson 等[4]对此数据进行重新分析发现，其总体相关性达到 0.96. 但如此高的相关性在很大程度上是被高丰度蛋白所夸大的.

1997年，Anderson 等[4]对人肝脏 19 个基因进行 mRNA 和蛋白质丰度的比较，发现只有中等程度的相关，Pearson 相关系数只有 0.48. 蛋白质的丰度有 70 倍的变化范围，而 mRNA 的丰度只有 16 倍的变化范围，表明转录后基因表达的调控在高等生物中较明显.

*通讯联系人.
贺福初. Tel: 010-66931246, E-mail: hefc@nic.bmi.ac.cn
朱云平. Tel: 010-66932248, E-mail: zhuyp@hupo.org.cn
收稿日期：2004-09-02，接受日期：2004-10-28

表 1 转录组和蛋白质组比较研究概况

Table 1 Summary of comparison studies between transcriptome and proteome

研究类型 *	物种及组织、细胞	转录组数据来源	蛋白质组数据来源	对应的基因数	相关性	发表时间	参考文献
Ⅰ	人 - 肝/血浆	Bodymap	[5]	10	r_p=0.96	1996	[3, 4]
Ⅰ	人 - 肝	Microarray	2DE	19	r_p=0.48	1997	[4]
Ⅰ	酵母	SAGE	2DE	106	r_p=0.935/0.356	1999	[6]
Ⅰ	酵母	SAGE/microarray	2DE	148	r_s=0.74	1999	[7]
Ⅰ	大肠杆菌	Microarray	2DE[8]	80	r_p=0.67	2000	[9]
Ⅱ	酵母	Microarray	ICAT	289	r=0.61	2001	[10]
				245	r_s=0.21	2002	[11]
Ⅱ	酵母	Microarray	MudPIT	678	r_s=0.45	2003	[12]
Ⅱ	人腔上皮细胞	Microarray	2DE	43	r=0.81	2004	[13]
Ⅲ	大肠杆菌	Microarray	2DE	57	7个基因有差异	2003	[14]
Ⅳ	人肺腺癌 (85 个样品)	Microarray	2DE	98	−0.025	2002	[15]
Ⅳ	人 NCI-60 细胞系 (60 个样品)	cDNA/oligo microarray	Protein microarray	19	0.52/0.40	2003	[16]
Ⅳ	小鼠线粒体 (4 种组织)	Microarray	定性鉴定	569	426个相关 (75%)	2003	[17]

*研究类型：Ⅰ:单点比较，Ⅱ: 双点差异比较，Ⅲ: 多点时序比较，Ⅳ: 多点非时序比较.

Gygi 等[6]于 1999 年首次用较大规模的数据 (106 个基因) 比较了酵母蛋白质组和转录组的关系. 研究发现，一般情况下，随着 mRNA 丰度的上升，蛋白质丰度也随之升高. 对 106 个基因计算 Pearson 相关系数得到的相关性为 0.935，但这结果与很少量的极高丰度蛋白质有很大的关系. 进一步对每细胞中含有 10 个 mRNA 拷贝以下的基因 (106 个基因中有 73 个，占 69%) 进行研究，发现其相关系数只有 0.356[6].

紧接着，Futcher 等[7]也发表他们对酵母转录组和蛋白质组比较研究的结果，但所得到的结论几乎和 Gygi 的结果相反. Futcher 的研究结果表明，mRNA和蛋白质丰度的 Spearman 秩相关系数 r_s=0.74，而丰度对数值的 Pearson 相关系数为 r_p=0.76. 用不同的统计学方法得到了非常相似的结果，表明 mRNA 和蛋白质水平有较强的相关性. Futcher[7]认为，转录组和蛋白质组的丰度数据不能用 Pearson 相关系数衡量，而应该用 Spearman 秩相关系数衡量，因而利用 Gygi 的实验数据进行了重新分析，发现其 mRNA 和蛋白质的相关性 r_s=0.59. 因此，转录组和蛋白质组呈现中等程度的相关性.

2000年 Arfin 等[9]在对 *E.coli* 转录组研究的基础上，将其结果和 1996 年 Vanbogelen 等[8]的 2DE 结果进行了比较，也得到了 mRNA 和蛋白质表达水平相关性较好 (r_p=0.67) 的结论.

虽然影响 mRNA 和蛋白质表达水平差异的原因很多，但有些基因的 mRNA 和蛋白质表达水平很明显是不对应的. Anderson 等[4]发现 β 和 γ 肌动蛋白同源性很高，而且等电点很接近，但其 mRNA 和蛋白质表达水平却相反 (分别对应的 mRNA 丰度为 0.189%和 0.215%，而蛋白质丰度为 1.41%和 0.65%). Gygi 等[6]的研究发现，对于某些基因，mRNA 丰度一致，但蛋白质丰度可以相差 20 倍以上，而某些蛋白质含量一致，mRNA 水平的差别却可以相差 30 倍；Futcher 等[7]的研究也发现某些 mRNA 丰度相同的基因，蛋白质水平有 10 倍差别.

1.2 两点差异比较

两点差异比较是对两个不同时间点 (或样品点) 的 mRNA 丰度比值和蛋白质丰度比值进行比较. 这种方法可以很好地消除实验系统的误差，可以较明确地描述在状态改变时基因和蛋白质表达的变化情况. 但这类研究在消除系统误差的同时也有可能会把生物系统自身的限制因素 (比如密码子偏性等)

消除了.

最早大规模进行两点差异比较探索的是 2001 年 Ideker 等[10]的工作. Ideker 对在有和无半乳糖的条件下酵母 mRNA 和蛋白质表达水平的比较研究发现，对于在网络中具有物理相互作用的基因，基因和蛋白质表达水平差异的相关性明显较强. Griffin 等[11]对之进行了进一步的分析，得到两者的 Spearman 秩相关系数为 0.21，表明 mRNA 变化和蛋白质的变化呈弱正相关. 对糖、乙醇代谢和能量产生相关基因的 mRNA 和蛋白质水平变化的分析，发现不同通路的 mRNA 和蛋白质表达差异的相关性是不同的.

Washburn等[12]对在营养充足和限制的介质中培养的酵母进行分析，发现 mRNA 改变量和蛋白质改变量的 Spearman 秩相关系数只有 0.45. 其中甲硫氨酸生物合成途径相关蛋白几乎完全是正相关的，表明 mRNA 和蛋白质表达水平的对应关系研究不应该是所有基因的分析，而应该是基于通路中各基因的分析.

对用 ErbB 特异的生长因子 heregulin b1 刺激后，ErbB-2 受体酪氨酸激酶过表达的哺乳动物腔上皮细胞系的研究表明，其 43 个对应基因的蛋白质和 mRNA 表达差异在统计学上表现出显著的强相关性 (r=0.81，P<0.001). 说明 mRNA 表达水平的改变以及蛋白质表达水平改变有较强的正相关性[13].

此外，Lian 等[18]虽然研究的是鼠骨髓分化程序中的多个时相，但用于比较的只有两个点，因而也可以算是差异的比较研究. 其研究结果表明，有 80% 的基因 mRNA 和蛋白质丰度的变化方向是相同的，定量相关性达到 r_p=0.58 (51 个基因).

从总体上来说，转录组和蛋白质组表达水平差异的相关性较好，大部分基因转录组和蛋白质组变化方向是相同的，而在特定代谢通路，细胞过程的基因以及在网络中具有物理相互作用的基因，转录组和蛋白质组的相关性更高[9]. 但也有少量报道发现相关性很低[19].

同样，研究表明某些基因的转录组和蛋白质组表达水平差异是明显不同的. 比如，Griffin 等[11]对酵母的研究发现，定位在线粒体上的蛋白质所对应的基因及蛋白质合成相关的基因呈现负相关.White 等[13]在研究人类上皮细胞系时虽然得到了相关系数达到 0.81 的结果，但也发现有些基因 mRNA 和蛋白质丰度的变化方向是相反的，如 EGFR 和 YWHAB 基因.

1.3 多点时序比较

时序研究一般是着眼于在某种外界刺激或某种状态变化过程中，不同时间点的 mRNA 和蛋白质表达水平的变化过程.

2003年 Yoon 等[14]对 *E.coli* 细胞密度增大过程中 mRNA 和蛋白质表达的动态变化进行了详细的研究. 所研究 57 个基因的 mRNA 和蛋白质表达水平的动态变化中，mRNA 和蛋白质表达水平的变化大部分是相似的，只有 7 个基因有差异.

大部分时序变化研究结果表明，转录组和蛋白质组的变化趋势比较相似[14, 20]，但在肺癌研究中得到了蛋白质组和转录组差异较大的结果[21].

1.4 多点非时序比较

除了上述 3 类基于同类型的组织、细胞等单个点、差异点或时序的研究外，还有另外一个类型的研究. 这类研究关注点在于相关细胞、细胞器的比较. 由于这类研究没法按差异或时序的方式考虑，因此，其分析方式比较特殊. 而且，由于这方面的研究一般样本量较大，因此可以根据这些数据求出单个基因的转录组和蛋白质组的相关性，再进行综合分析.

2002年，Chen 等[15]对 76 个肺腺癌和 9 个正常肺组织进行了转录组和蛋白质组的平行研究，结果发现，在所研究的 2DE 胶上 165 个点 (98 个基因) 中，有 28 个点 (21 个基因) 具有统计学上显著的相关性 (r>0.2445, P<0.05). 对这 85 个不同样品的 mRNA 丰度和蛋白质丰度正态化后，计算 Spearman 相关系数得到相关性在–0.467~0.442 之间. 为了求总体相关性，先求出这 85 个样品每个基因的平均 mRNA 和蛋白质的丰度，由此算出总体相关性只有–0.025；而对前面提到的 28 个具有统计学上显著的相关性蛋白用这方法分析，得到的相关性也只有–0.035. 有些 mRNA 丰度相似，但蛋白质水平却相差 24 倍，有些蛋白质丰度相似，但 mRNA 差异达到 28 倍. 因此，结论是转录组和蛋白质组的相关性极差.

2003年，Nishizuka 等[16]利用 cDNA 和 oligo 基因芯片以及反相溶解物芯片 (reverse-phase lysate microarrays) 对人类 NCI-60 的 60 个癌细胞系进行了转录组和蛋白质组的研究，得到 19 个在 cDNA 和 oligo 基因芯片中相关性很好的基因. 这 19 个基因的 cDNA/protein 总体相关性为 0.52，oligo/protein 的总体相关性为 0.40 .

Mootha 等[17]在对小鼠心、脑、肾、肝 4 种组织的线粒体蛋白质组研究的基础上，将其蛋白质的鉴定结果和基因芯片得到的对应基因丰度进行比较，发现在 569 个基因中有 426 个是相关的.

2 差异原因

上述各研究结果表明，不同的研究得到的转录组和蛋白质组相关性结果不完全一致，但不管相关性如何，都存在某些基因有着明显差异. 这些差异一方面是由于生物体本身因素导致，另一方面还受到非生物因素的限制.

2.1 非生物学因素

2.1.1 实验系统的差异.

由于转录组和蛋白质组的研究都有多种不同的技术体系，而且即使是同样的技术体系，不同实验室得到的结果可能也会有差异. 因此，在转录组和蛋白质组相关性分析中，这是导致相关性较差的原因之一.

大规模的转录组研究所用的技术一般是 SAGE 或 cDNA/oligo 微阵列技术. 然而 SAGE 技术其实相当于抽样测序，因此对于低丰度的蛋白质往往会由于抽样的局限性无法准确定量. 而 cDNA/oligo 芯片技术对于高丰度基因往往有饱和作用，会低估这些蛋白质的表达量[7]. 研究表明，SAGE 和微阵列技术鉴定的基因其丰度间的 Spearman 相关系数为 0.425，如果忽略低丰度的基因，则其相关性可达到 0.657[22]. 蛋白质组技术现在一般只有 2DE、ICAT 及蛋白质芯片可能得到定量信息，而且很难检测到极性蛋白 (极酸、极碱和强疏水性蛋白等). 这些不同系统的差异是导致转录组和蛋白质组间差异的原因之一.

Futcher 等[7]将自己的酵母蛋白质组数据和 Gygi 的酵母蛋白质组数据做了比较，扣除了由于生物学因素导致的差异，发现其间的相关性也只有 0.88，原因可能是蛋白质丰度测量方法不同导致的.

2.1.2 数据类型和方法的差异.

不同类型的研究会相应地产生不同类型的数据，其所受到的影响因素也是不同的. 单点比较无法消除系统误差，但能观测到生物体自身的限制所造成的差异 (如密码子偏性等). 两点差异比较可以较好地消除实验系统误差，但生物体自身的限制也会被抹掉，也即无法考查密码子偏性等因素对翻译效率的影响.

在数据分析时所使用统计方法的不同也会导致结果的差异. Gygi 在计算单点转录组和蛋白质组相关性时用的是 Pearson 相关性的计算方法，但 Futcher 等[7]认为对于丰度数据是不适于用 pearson 相关性计算方法的，原因是 Pearson 方法要求数据是正态分布的，但转录组和蛋白质组的丰度远远不符合正态分布的要求. 因此如果要用 Pearson 方法计算的话必须先将数据进行正态化处理 (如用 Box-Cox 算法)，然后再计算 Pearson 相关性. 或者直接计算 Spearman 相关系数[7].

此外，对多点非时序数据在计算单个基因转录组和蛋白质组表达水平的相关系数时，所用的计算方法一般比较特殊. Chen 等[15]在分析肺癌时用的是类似于 SAM (significance analysis of microarray) 的计算方法，利用域值 $\Delta=0.115$ 判断 mRNA 和蛋白质表达水平的显著相关性. Nishizuka 等[16]在研究 NCI-60细胞系的转录组和蛋白质组关系时用的是 Pearson 相关系数计算方法. Mootha 等[17]在分析小鼠脑、心、肾和肝 4 种组织的线粒体转录组和蛋白质组时用的是自己定义的一种统计方法：由于实验所得到的蛋白质组数据为定性数据 (即有或无)，因此在分析时相当于把检测到的蛋白质当做丰度较高的蛋白质，没有检测到的作为丰度较低的蛋白质，以此和转录组的定量数据比较. 这些数据类型的差异和比较方法的差异也会导致在结果的描述上有所差别.

另外一个不可忽视的影响因素是高丰度蛋白所导致计算得到相关性的偏性. 早在 1997 年，Anderson 等[4]就发现了高丰度蛋白对相关系数具有明显的夸大作用. Gygi 等[6]对其得到的蛋白质利用丰度排序，然后逐步加入丰度较高的蛋白质，分别计算 Pearson 相关系数，发现在低表达的 40 到 95 个蛋白质中其相关系数稳定在 0.1~0.4，而在加入最后 11 个高丰度蛋白时，每加一个相关系数都能得到很大的提高. 虽然对于单个点的计算高丰度蛋白有很强的夸大作用，但蛋白质丰度和单个基因转录组和蛋白质组相关性的值不存在相关性[15].

2.2 生物学因素

2.2.1 翻译过程的限制.

基因转录成 mRNA 后的首要问题是 mRNA 的稳定性. 有研究表明，mRNA 的稳定性和 mRNA 的 3′ UTR 的特定核苷酸序列以及其相应的结合蛋白有关系[23].

随后，mRNA 可以和多个核糖体结合形成多核糖体结构 (polysome) 成活性的翻译形式，或隐匿

于 mRNP (messager ribonucleoprotein)，或只结合一个核糖体成单体 (monosome) 而变为非翻译活性形式，这两类 mRNA 可以用蔗糖密度梯度离心分离[24]. Zong 等[24]研究了静息的和具有有丝分裂活性的纤维原细胞，分别进行活性和非活性 mRNA 的研究，发现在所研究的 1 200 个基因中有 1%受到了翻译的调控. 随后的研究发现，细胞内总 RNA 的组成和多核糖体浓度有很强的正相关性，对于大部分的 mRNA，这两种不同的 RNA 存在形式变化方向相同. 因此，活性的 mRNA 虽然不能直接由总 mRNA 丰度代表，但和总 mRNA 丰度是相关的[25].

mRNA 翻译的起始还与翻译起始点上下游的几个核苷酸序列有关. 研究发现在翻译起始密码子上下游的几个核苷酸是非随机的，这些非随机的核苷酸可能和核糖体在滑行过程中寻找并锚定到 AUG 上有关[26]. 由于这现象最早由 Kozak 发现，因此被称为 Kozak 规则. 研究发现丰度低的蛋白质一般倾向于使用次优的翻译起始上下游核苷酸[7].

紧接着，mRNA 进入了翻译阶段，蛋白质的翻译速度受到其所用特定密码子的 tRNA 丰度影响，也即受到密码子偏性的影响. 遗传密码的简并性使在编码或翻译同一个氨基酸序列时可以用几种不同的密码子，但基因对于特定氨基酸所使用的密码子并不随机，受到简并密码子使用系统偏性的影响，这就叫密码子偏性[27]. 生物体中基因所使用的密码子和 tRNA 的丰度有着强的正相关性[28]. 研究表明，高表达的基因一般使用较常用的密码子[27]. 对酵母密码子偏性的计算发现 CAI (codon adaptation index) 值和蛋白质丰度的相关性 r_s=0.80 (P<0.0001)，证明密码子偏性和蛋白质丰度间是强相关的[6, 7]，但密码子偏性在 0.8~1.0 时丰度的变异很大[6].

蛋白质寿命对蛋白质丰度也有影响. 蛋白质寿命基本上是由蛋白质的氨基端残基决定的，这种现象被称为“N 端规则”[29]，虽然还有不少短寿命蛋白质并不依赖于 N 端规则[30]. 在对酵母蛋白的分析中发现，蛋白质组鉴定的大部分蛋白质是属于长寿命的蛋白质[6]，而很多丰度较低的蛋白质是不稳定的[7]. 有明确的证据表明，转录组和蛋白质组的相关性和蛋白质的降解有明显的关系[31].

蛋白质翻译后的修饰 (如酶切、磷酸化、糖基化等) 也可能是 mRNA 和蛋白质丰度不对应的因素之一. 如酵母细胞 Yef3 蛋白在 2DE 胶上是呈现多点分布，导致其蛋白质 /mRNA 丰度的比率很低[7]. 此外，蛋白质合成后的定位也可能是个不可忽视的影响因素.

2.2.2 基因的差别.

不同基因可能受到不同的转录后调控的影响，因而会导致不同基因的转录组和蛋白质组的关系存在一定的差异.

研究发现，转录组鉴定的基因和蛋白质组鉴定的蛋白质所代表的功能类别是对应的[20]. 而对饥饿状态和正常状态下酵母代谢途径中 mRNA 和对应的蛋白质丰度的差异研究发现，不同通路存在着不同的关系：a. mRNA 和蛋白质丰度差异基本吻合 (如糖酵解途径的基因)；b. 变化方向相同但变化幅度相差很大 (半乳糖代谢基因)；c. 变化方向不同 (定位在线粒体的基因，蛋白质合成的基因等)[11].

在分析多点非时序数据时由于样本量较大，因而可以对单个基因进行 mRNA 和蛋白质表达水平的相关性分析. 对 NCI-60 细胞系转录组和蛋白质组表达水平差异的研究发现，在 19 个蛋白质中，7 个结构相关蛋白相关系数较高，12 个非结构相关蛋白其相关性较低[16].

Chen 等[15]对肺癌转录组和蛋白质组的分析发现，不同的可变剪接体具有不同的相关性，其间可能存在着某些特殊的调控机制.

因此，不同的可变剪接体、不同类型基因、不同的通路其转录组和蛋白质组的相关性是不同的.

2.2.3 物种、组织、细胞、细胞器及其状态的差异.

对于不同的物种、组织或细胞，转录组和蛋白质组对应关系是比较相似还是有较大不同，还没有很明确的报道，但由于现有的对不同样品的研究结果得到的转录组和蛋白质组对应关系差别很大，因而这可能也是转录组和蛋白质组对应关系存在差别的一个原因.

此外，细胞发育、分化或其他状态的变化也会导致转录组和蛋白质组相关性有差异. 研究发现，在 DC 细胞分化时有约 4%的基因或蛋白质表达，其转录组和蛋白质组间表现出很好的相关性. 但在 DC 细胞成熟时，蛋白质水平发生改变的基因其转录水平很少发生变化，进一步分析提示，DC 细胞的成熟在很大程度上受控于转录后和翻译后调控[32]. 而对两种角蛋白阳性的肺细胞系 DLKP 和 H82 在溴脱氧尿苷 (BrdU，bromodeoxyuridine) 处理前后的比较发现，其 K8 和 K18 基因的 mRNA 水平不变，但蛋白质水平明显升高[33]，这也提示了在不同的细胞状态下，某些基因受到的调控是不一

样的.

Chen等[15]对肺癌中表达的21个基因进行了详细研究，发现其中有16个在一期肺癌和三期肺癌中没有什么差别，但另外5个基因则在这两种不同期癌症中有显著差别. 这结果预示着不同阶段的肺癌某些基因的转录组和蛋白质组相关性是不同的.

3 转录组和蛋白质组的不完全性和互补性

虽然转录组和蛋白质组在实验方法上差异很大，但由于这两种方法的首要目的都是获得基因的表达情况，其间存在着某种共同之处. 从生物学角度上看，mRNA水平代表了基因表达的中间状态，能代表着潜在的蛋白质表达情况. 然而蛋白质是直接的功能执行体，因而，对蛋白质表达水平的度量有着不可取代的优势. 但由于转录后调控的影响，转录组并不能完全代表蛋白质组，而且在体内转录组和蛋白质组间的相关性可能更差[7]. 而在实验方法上，转录组能在较低消耗下实现较高的通量[1]，并能在某种程度上提供较详细的信息[34]，但由于样品的原因，至今对很多疾病的转录情况难以测量[1]，而且蛋白质组对偏性蛋白、膜蛋白的度量也有一定的难度. 因此，转录组和蛋白质组研究分别有自己的优势和缺陷.

最近的文献也明确报道了转录组和蛋白质组的部分不相关或负相关的结果，并且用统计方法证明了这种显著差异是由生物学因素造成的，而不仅仅是噪音. 说明了基因表达情况不能单纯用转录组的方法解决[35].

而从逻辑上看，由于转录组代表了基因表达的中间状态，蛋白质组代表了基因表达的最终形式，也即基因功能执行体的最终形式. 因此，生物体为了尽可能节约资源，一般会实现这两种表达水平的对应，但另外一方面生物体也完全可以充分利用这环节，并将它作为一个基因表达调控步骤. 因此，总的来说，转录组和蛋白质组应该是大部分相关的，只有少数基因由于受到调控而导致其不相关. 而且在不同的组织、细胞中，由于需要被调控的基因可能是不同的，因而不同的组织、细胞的不对应基因 (受调控基因) 一般是不同的. 现有的研究虽然没有明确揭示这些规律，但总体上反应了这种趋势.

由于这两种不同的表达谱研究手段的不完全性和互补性，现有的研究倾向于综合转录组和蛋白质组的研究，获得一个表达谱的“全景图”，并实现其间的互补和整合. 这种研究方式在基础研究上已经有不少报道，此外，在癌的分子分析、分类、早期检测、预防，以及在药物开发中也得到了应用. 整合转录组和蛋白质组以及其他各种组学的研究也在进行中，这些研究将有望为获得新的代谢途径，调控网络提供基 础[36]. 由于转录组和蛋白质组的比较研究能揭示基因表达的转录后调控状态，因此转录组和蛋白质组之间的关系很可能将是未来的系统生物学[37]研究中不可忽略的一部分.

参 考 文 献

1 Hegde P S, White I R, Debouck C. Interplay of transcriptomics and proteomics. Curr Opin Biotechnol, 2003, **14** (6): 647~651

2 钱小红, 贺福初. 蛋白质组学: 理论与方法. 北京: 科学出版社, 2003. 8~10
Qian X H, He F C. Proteomics: Theory and Mothod. Beijing: Science Press, 2003. 8~10

3 Kawamoto S, Matsumoto Y, Mizuno K, *et al*. Expression profiles of active genes in human and mouse livers. Gene, 1996, **174** (1): 151~158

4 Anderson L, Seilhamer J. A comparison of selected mRNA and protein abundances in human liver. Electrophoresis, 1997, **18**: 533~537

5 Putnam F W. Progress in plasma proteins. In: Putnam F W ed. The Plasma Proteins: Structure, Function, and Genetic Control. 2nd. Orlando: Academic Press, 1986. 1~44

6 Gygi S P, Rochon Y, Franza B R, *et al*. Correlation between protein and mRNA abundance in yeast. Mol Cell Biol, 1999, **19** (3): 1720~1730

7 Futcher B, Latter G I, Monardo P, *et al*. A sampling of the yeast proteome. Mol Cell Biol, 1999, **19** (11):7357~7368

8 VanBogelen R A, Abshire K Z, Pertsemlidis A, *et al*. *Escherichia coli* and Salmonella Cellular and molecular biology. American Society for Microbiology, Washington D C. 1999

9 Arfin S M, Long A D, Ito E T, *et al*. Global gene expression profiling in *Escherichia coli* K12. The effects of integration host factor. J Biol Chem, 2000, **275** (38): 29672~29684

10 Ideker T, Thorsson V, Ranish J A, *et al*. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science, 2001, **292** (5518): 929~934

11 Griffin T J, Gygi S P, Ideker T, *et al*. Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. Mol Cell Proteomics, 2002, **1** (4): 323~333

12 Washburn M P, Koller A, Oshiro G, *et al*. Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA, 2003, **100** (6): 3107~3112

13 White S L, Gharbi S, Bertani M F, *et al*. Cellular responses to ErbB-2 overexpression in human mammary luminal epithelial cells: comparison of mRNA and protein expression. Br J Cancer, 2004, **90** (1): 173~181

14 Yoon S H, Han M J, Lee S Y, *et al*. Combined transcriptome and proteome analysis of *Escherichia coli* during high cell density

culture. Biotechnol Bioeng, 2003, **81** (7): 753~767
15 Chen G, Gharib T G, Huang C C, *et al*. Discordant protein and mRNA expression in lung adenocarcinomas. Mol Cell Proteomics, 2002, **1** (4):304~313
16 Nishizuka S, Charboneau L, Young L, *et al*. Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. Proc Natl Acad Sci USA, 2003, **100** (24): 14229~14234
17 Mootha V K, Bunkenborg J, Olsen J V, *et al*. Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. Cell, 2003, **115** (5): 629~640
18 Lian Z, Kluger Y, Greenbaum D S, *et al*. Genomic and proteomic analysis of the myeloid differentiation program: global analysis of gene expression during induced differentiation in the MPRO cell line. Blood, 2002, **100** (9): 3209~3220
19 de Nobel H, Lawrie L, Brul S, *et al*. Parallel and comparative analysis of the proteome and transcriptome of sorbic acid-stressed *Saccharomyces cerevisiae*. Yeast, 2001, **18** (15): 1413~1428
20 Bro C, Regenberg B, Lagniel G, *et al*. Transcriptional, proteomic, and metabolic responses to lithium in galactose-grown yeast cells. J Biol Chem, 2003, **278** (34): 32141~32149
21 Kim C H, Kim do K, Choi S J, *et al*. Proteomic and transcriptomic analysis of interleukin-1 beta treated lung carcinoma cell line. Proteomics, 2003, **3** (12): 2454~2471
22 Kim H L. Comparison of oligonucleotide-microarray and serial analysis of gene expression (SAGE) in transcript profiling analysis of megakaryocytes derived from $CD34^+$ cells. Exp Mol Med, 2003, **35** (5): 460~466
23 Waggoner S A, Liebhaber S A. Regulation of alpha-globin mRNA stability. Exp Biol Med (Maywood), 2003, **228** (4): 387~395
24 Zong Q, Schummer M, Hood L, *et al*. Messenger RNA translation state: the second dimension of high-throughput expression screening. Proc Natl Acad Sci USA, 1999, **96** (19): 10632~10636
25 Preiss T, Baron-Benhamou J, Ansorge W, *et al*. Homodirectional changes in transcriptome composition and mRNA translation induced by rapamycin and heat shock. Nat Struct Biol, 2003, **10**(12): 1039~1047
26 Pesole G, Gissi C, Grillo G, *et al*. Analysis of oligonucleotide AUG start codon context in eukariotic mRNAs. Gene, 2000, **261** (1): 85~91
27 Kurland C G. Codon bias and gene expression. FEBS Lett, 1991, **285** (2): 165~169
28 Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol, 1985, **2** (1): 13~34
29 Bachmair A, Finley D, Varshavsky A. *In vivo* half-life of a protein is a function of its amino-terminal residue. Science, 1986, **234** (4773): 179~186
30 Varshavsky A. The N-end rule. Cell, 1992, **69** (5): 725~735
31 Beyer A, Hollunder J, Nasheuer H P, *et al*. Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. Mol Cell Proteomics, 2004, **3** (11): 1083~1092
32 Richards J, Le Naour F, Hanash S, *et al*. Integrated genomic and proteomic analysis of signaling pathways in dendritic cell differentiation and maturation. Ann N Y Acad Sci, 2002, **975**: 91~100
33 McBride S, Walsh D, Meleady P, *et al*. Bromodeoxyuridine induces keratin protein synthesis at a posttranscriptional level in human lung tumour cell lines. Differentiation, 1999, **64** (3): 185~193
34 Eymann C, Homuth G, Scharf C, *et al*. Bacillus subtilis functional genomics: global characterization of the stringent response by proteome and transcriptome analysis. J Bacteriol, 2002, **184** (9): 2500~2520
35 Tian Q, Stepaniants S B, Mao M, *et al*. Integrated genomic and proteomic analyses of gene expression in mammalian cells. Mol Cell Proteomics, 2004, **3** (10): 960~969
36 Oliver D J, Nikolau B, Wurtele E S. Functional genomics: high-throughput mRNA, protein, and metabolite analyses. Metab Eng, 2002, **4** (1): 98~106
37 Kitano H. Systems biology: a brief overview. Science, 2002, **295** (5560): 1662~1664

Progress in The Comparison of Transcriptome and Proteome

WU Song-Feng, ZHU Yun-Ping*, HE Fu-Chu*
(*Laboratory of Genomics and Proteomics, Beijing Institute of Radiation Medicine, Beijing* 100850, China)

Abstract Moderate correlations between transcriptome and proteome were found in most studies. According to different data types, current studies could be divided into four categories. They are comparison on single point, comparison between two differential points, comparison among multiple time-sequence points and comparison among multiple non-time-sequence points. In addition to the experimental error and the different datasets, the post-transcription restriction and regulatory would affect the correlation very much. Different results might be got for the different genes, cells, organs, organisms and even the different developmental phases. Because of the differentia and complementary between them, parallel studies of transcriptome and proteome trended to be performed, which could get the panorama screen of gene expression and discover the genes regulated in post-transcription.

Key words transcriptome, proteome, correlation, translational regulation, codon bias

*Corresponding author .
HE Fu-Chu. Tel: 86-10-66931246, E-mail: hefc@nic.bmi.ac.cn; ZHU Yun-Ping. Tel: 86-10-66932248, E-mail: zhuyp@hupo.org.cn
Received: September 2, 2004 Accepted: October 28, 2004