张正东,申　铁,周文卫,等. 基于 EST 序列的茶代谢网络的构建[J]. 江苏农业科学,2017,45(11):29 – 32.
doi:10.15889/j.issn.1002 – 1302.2017.11.008

基于 EST 序列的茶代谢网络的构建

张正东[1], 申　铁[2], 周文卫[2], 谢晓尧[2]
(1. 贵州大学计算机科学与技术学院,贵州贵阳 550001; 2. 贵州师范大学贵州省信息与计算科学重点实验室,贵州贵阳 550001)

摘要:茶树体内的生化反应所生成的各种功能性化合物是茶叶具有营养和健康功能的物质基础,也是茶叶品质的决定因素。这些生化反应由茶树基因编码的酶催化并组成复杂的代谢网络。首先通过开源工具包 jsoup 开发异步数据采集程序,从布伦瑞克酶数据库(braunschweig enzyme database,简称 BRENDA)和美国国立生物技术信息中心(NCBI)网站上获取酶序列及其催化反应、GI 号、EC 编码对应关系等相关信息,建立本地酶数据库;其次从 NCBI 上下载 FASTA 格式的茶树表达序列标签(expressed sequence tag,简称 EST)序列数据,通过 GI 号查询本地酶数据库,得到酶催化反应信息,继而基于超图思想利用 Cytoscape Web API 重构茶代谢网络;最后对 EST 序列信息进行统计分析,并从多个维度对构造的代谢网络进行拓扑特性、KEGG 路径、生物意义的深入分析,对茶树内生化反应的理解、新功能基因的挖掘、茶叶品质的提升、新茶产品的开发具有重要意义。

关键词:Cytoscape Web;EST;超图;代谢网络;茶叶
中图分类号: Q811.4　　**文献标志码**: A　　**文章编号**:1002 – 1302(2017)11 – 0029 – 04

茶是世界上一种重要的饮料[1]。茶叶品质是茶叶具有营养和健康功能的物质基础,其决定因素是茶叶中的各种功能性化合物[2]。研究表明,茶叶中蕴含的活性物质能够促进身体健康和预防多种疾病。比如,茶叶中的多酚类物质有很强的抗氧化性和生理活性,具有很好的抗衰老效果[3]。茶多酚及其氧化物能够吸收放射性物质锶 90、钴 60,具有一定的抗辐射作用[4]。此外,茶多酚(主要是儿茶素类化合物)具有预防多种器官癌症、代谢综合征、心血管疾病以及神经退行性疾病的作用[5-7]。

茶叶中的功能性化合物来源于茶树基因编码的酶[8]。酶是代谢反应的生物催化剂,其活性由基因转录和翻译的特定氨基酸序列决定[9-12]。茶叶中的酶促反应组成复杂的生化反应网络,即代谢网络[13]。代谢网络的基本功能是不停地与外界环境进行物质和能量交换,维持茶树体的生命特征[14]。此外,代谢网络对于茶叶中的物质合成至关重要,这些物质是决定茶叶品质和等级的关键要素[15-16]。研究茶叶中的酶及其催化的代谢反应,对于茶树品种的开发、品质的提升、新型茶产品的研发加工具有重要作用。

茶叶酶的特性取决于氨基酸种类和线性排列,这些氨基酸由茶树基因编码[17]。因此,本研究通过异步数据采集程序从布伦瑞克酶数据库(BRENDA)、美国国立生物技术信息中心(NCBI)网站上获取酶序列及其催化反应、GI 号、EC 编码等相关信息,建立本地酶数据库;从 NCBI 上下载茶树表达序列标签(EST)序列数据,通过查询本地酶数据库鉴别出 EST 序列对应的茶叶酶,继而构造茶代谢网络,从多个维度对构造的代谢网络进行拓扑特性和生物信息统计分析,并讨论分析结果所蕴含的生物学意义。

收稿日期:2016 – 09 – 23
基金项目:国家自然科学基金(编号:31200626);贵州省科技厅联合基金(编号:黔科合 LH[2015]7773)。
作者简介:张正东(1984—),男,安徽六安人,博士研究生,研究方向为生物信息学。E – mail:923534069@qq.com。
通信作者:谢晓尧,博士,教授,博士生导师,主要从事生物大数据分析研究。E – mail:2715236679@qq.com。

[14] Song X J, Huang W, Shi M, et al. A QTL for rice grain width and weight encodes a previously unknown RING – type E3 ubiquitin ligase[J]. Nature Genetics, 2007, 39(5): 623 – 630.
[15] Weng J, Gu S, Wan X, et al. Isolation and initial characterization of GW5, a major QTL associated with rice grain width and weight[J]. Cell Research, 2008, 18(12): 1199 – 1209.
[16] Hong Z, Ueguchi – Tanaka M, Umemura K, et al. A rice brassinosteroid – deficient mutant, ebisu dwarf (d2), is caused by a loss of function of a new member of cytochrome P450[J]. The Plant Cell, 2003, 15(12): 2900 – 2910.
[17] Mao H, Sun S, Yao J, et al. Linking differential domain functions of the GS3 protein to natural variation of grain size in rice[J]. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107(45): 19579 – 19584.
[18] Li M, Tang D, Wang K, et al. Mutations in the F – box gene LARGER PANICLE improve the panicle architecture and enhance the grain yield in rice[J]. Plant Biotechnology Journal, 2011, 9(9): 1002 – 1013.
[19] Wang S, Wu K, Yuan Q, et al. Control of grain size, shape and quality by OsSPL16 in rice[J]. Nature Genetics, 2012, 44(8): 950 – 954.
[20] Yoshida A, Sasao M, Yasuno N, et al. TAWAWA1, a regulator of rice inflorescence architecture, functions through the suppression of meristem phase transition[J]. Proceedings of the National Academy of Sciences of the United States of America, 2013, 110(2): 767 – 772.

1　材料与方法

1.1　EST 数据采集

茶树 EST 序列数据来源于 NCBI 数据库。在 NCBI 首页搜索“Camellia sinensis”，选择“protein”，共获得 38 619 条 FASTA 格式的茶树 EST 氨基酸序列数据。

1.2　酶数据库构建

酶及其催化反应信息来源于 BRENDA[18]。BRENDA 中共保存了 6 759 种酶 EC 编码、推荐命名和催化反应等信息。由于数据量较大，本研究利用开源工具包 jsoup 开发异步数据采集程序，解析 BRENDA 中所有酶及其催化反应的底物和产物等相关信息。对于没有催化反应信息的酶，如 EC 1.1.1.5，将其过滤掉，最终共获得 5 221 个酶及其催化反应数据。EST 序列的 GI 号、酶 EC 编码对应关系数据也来源于 BRENDA。由于 NCBI 中序列数据会被不断完善和修正，当 EST 序列信息被更新时，其 GI 号也将被赋予新值，而 BRENDA 中保留的仍然是旧的 GI 号，因此，将会出现 1 个 EC 编码可能对应多个 GI 号的情况。这种情况下，首先判定 EST 序列数据是否被更新，若被更新，追踪更新历史信息并找到最近的 GI 号，此过程通过异步数据采集程序自动完成，采集到的数据保存在本地酶数据库中。

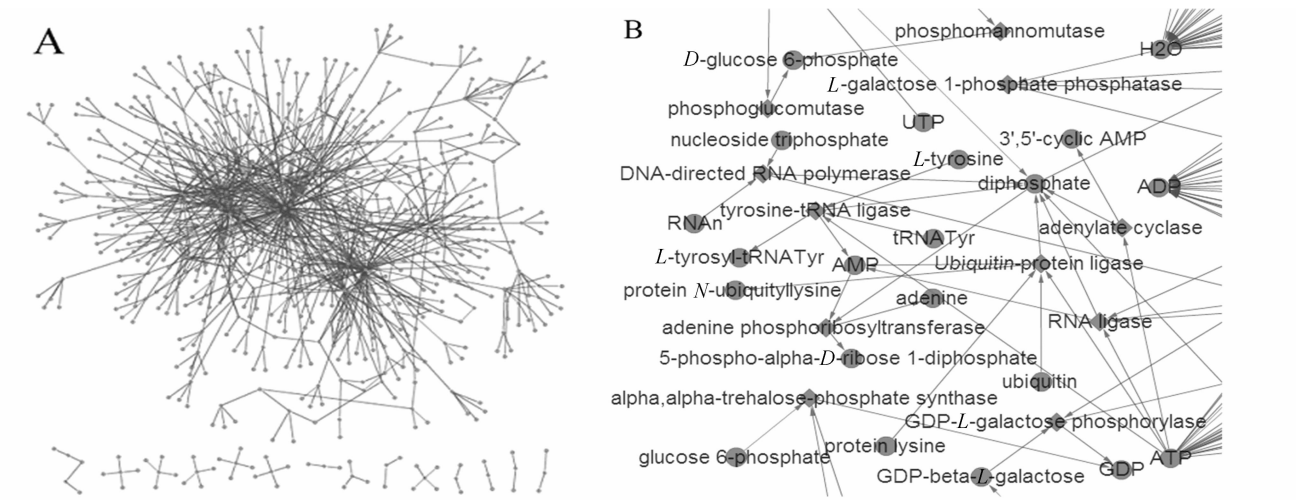
1.3　酶基因筛查

从 NCBI 上下载的 FASTA 格式文件的每个序列都有 1 个 GI 号作为唯一标识，以便于对序列进行监控和管理[19]。GI 号位于 FASTA 文件序列描述信息的第 1 行（以“>”开始）。通过 GI 号查询本地酶数据库可以获得酶的 EC 编码，进而得到酶及其催化反应信息。

1.4　代谢网络的构建和可视化

代谢网络的可视化采用 Cytoscape Web 实现。Cytoscape Web 是一款开源、交互式、高可定制的基于浏览器的网络可视化工具，采用 Flex/ActionScript 实现，支持 GraphML、XGMML、SIF 等多种交互文本格式[20]。本研究采用 GraphML 格式与 Cytoscape Web 进行数据交互。Cytoscape Web 提供非常丰富的 JavaScript API，利用这些 API 可以设置点、边的颜色、形状、权重等各种网络参数，也可实现各种回调方法与网络交互。

代谢反应可能涉及到多个底物和产物，普通图每条边最多连接 2 个顶点，因此采用普通图表示代谢网络，无论是酶还是化合物作为顶点，都要作一些额外限制，很难完整地展现代谢网络的全部信息。而超图（hypergraph）的超边可以连接多个顶点[21]，普通图可视为超边最多连接 2 个顶点的超图特例。超图可以完整地表示网络的全部信息，是代谢网络等复杂网络的最佳形式化表示方法。因此，本研究采用有向超图作为代谢网络的形式化表示方法。酶和化合物均作为超图的顶点，菱形表示酶顶点，圆形表示化合物顶点。若化合物是酶催化反应的底物，在酶和化合物之间有 1 条有向超边，方向指向酶；反之，有向超边方向则指向化合物。构造的代谢网络如图 1 所示。



A—15个独立子网络组成的代谢网络；B—282个反应组成的最大子网络的部分，其中：phosphomannomutase—磷酸甘露糖酶；*D*-glucose 6-phosphate—*D*-葡萄糖6-磷酸；*L*-galactose 1-phosphate phosphatase—*L*-半乳糖1-磷酸磷酸酶；phosphoglucomutase—葡萄糖磷酸变位酶；nucleoside triphosphate—核苷三磷酸；3′,5′-cyclic AMP—3′,5′-环腺苷一磷酸；*L*-tyrosine—*L*-酪氨酸；DNA-directed RNA polymerase—DNA定向RNA聚合酶；diphosphate—二磷酸；tyrosine-tRNA ligase—酪氨酸-tRNA连接酶；RNAn—RNA模版；tRNA Tyr—酪氨酸转移RNA；adenylate cyclase—腺苷酸环化酶；L-tyrosyl-tRNA Tyr—*L*-酪氨酰-酪氨酸转移RNA；AMP—腺苷一磷酸；Ubiquitin protein ligase—泛素蛋白连接酶；protein *N*-ubiquityllysine—*N*-泛素化蛋白质；adenine—腺嘌呤；RNA ligase—RNA连接酶；adenine phosphoribosyltransferase—腺嘌呤磷酸核糖转移酶；5-phospho-alpha-*D*-ribose 1-diphosphate—5-磷酸-*α*-*D*-核糖1-二磷酸酯；ubiquitin—泛素；alpha,alpha-trehalose-phosphate synthase—*α*，*α*-海藻糖磷酸合成酶；GDP-*L*-galactose phosphorylase—鸟苷二磷酸-*L*-半乳糖磷酸化酶；glucose 6-phosphate—葡萄糖6-磷酸；protein lysine—蛋白赖氨酸；GDP-beta-*L*-galactose—鸟苷二磷酸-*β*-*L*-半乳糖

图1　基于EST序列的茶代谢网络

2　结果与分析

2.1　EST 序列信息

EST 是来源于不同组织的 cDNA 序列，是基因转录的 mRNA 的一部分[22]。EST 可代替 mRNA 用于生物信息学软件分析以增强对基因功能的理解。本研究从 NCBI 数据库中，共搜集整理 38 619 条 EST 序列信息，氨基酸序列总长度为 17 585 719 aa，单条序列最长为 9 212 aa，最短为 4 aa，平均序列长度为 455 aa。这些序列中，非酶序列有 36 205 条，氨基酸序列总长度为 16 468 574 aa，单条序列最长为 9 212 aa，最

短为 4 aa,平均序列长度为 454 aa;筛选得到的酶序列有 2 414 条,氨基酸序列总长度为 1 117 145 aa,单条序列最长为 2 436 aa,最短为 10 aa,平均序列长度为 462 aa(表 1)。

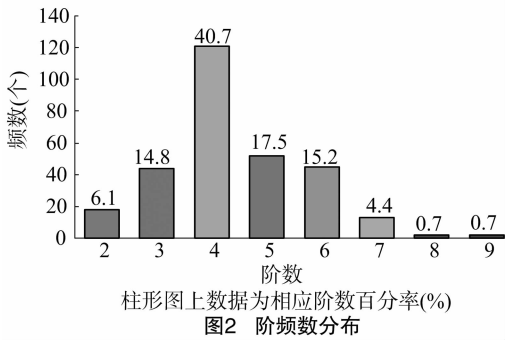
表 1　EST 序列信息

EST	序列数(条)	总长度(aa)	最大长度(aa)	最短长度(aa)	平均长度(aa)
全部 EST	38 619	17 585 719	9 212	4	455
非酶 EST	36 205	16 468 574	9 212	4	454
酶 EST	2 414	1 117 145	2 436	10	462

注:aa 表示氨基酸。

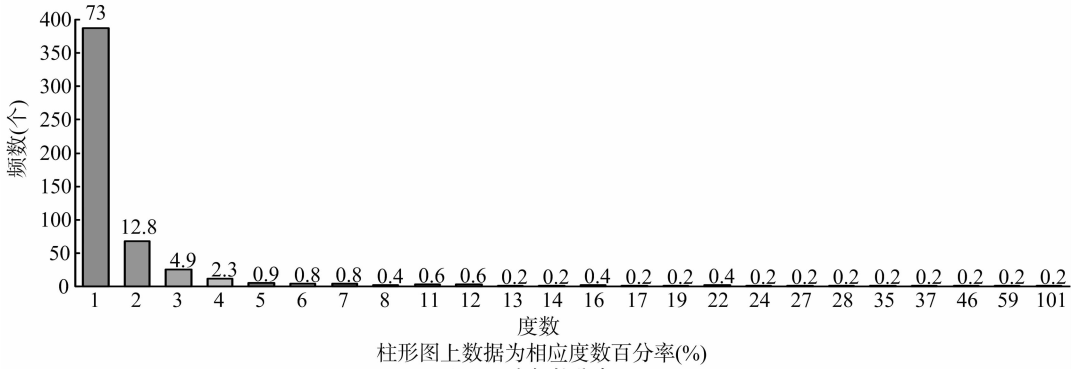
2.2　代谢网络统计

在 2 414 条酶序列重构的代谢网络中,共有 297 个酶促反应,包含 297 个酶和 530 个化合物。代谢网络最大阶为 9,最小阶为 2,平均阶为 4,阶频数分布如图 2 所示;最大度为 101,最小度为 1,平均度为 2,度频数分布如图 3 所示。阶定义为超边所连接的点的个数,即酶促反应的化合物数量;度的定义和普通图中一样,为顶点关联的超边个数,即化合物参与的代谢反应数量(表 2)。



柱形图上数据为相应阶数百分率(%)

图2　阶频数分布



柱形图上数据为相应度数百分率(%)

图3　度频数分布

表 2　基于 EST 序列的茶代谢网络信息

酶反应数(个)	化合物数(个)	最大阶	最小阶	平均阶	最大度	最小度	平均度	子网络数(个)	最大子网络反应数(个)
297	530	9	2	4	101	1	2	15	282

2.3　代谢网络 KEGG 路径分析

代谢网络的一个重要特性是代谢路径及其所涉及到的化合物,即 KEGG 路径分析,这对于理解构建的代谢网络在整个网络中的位置和作用有重要意义。因此,本研究将所有的代谢反应映射到 KEGG 路径。如图 4 所示,2 个最大的路径是次生代谢物、抗生素的生物合成,分别包含 44、16 个反应,这种情况是合理的,因为这 2 个路径位于高层次的分类,包含的反应较多;第二大路径是嘌呤,包含 11 个反应;其他较大的路径是氨酰 - tRNA、半胱氨酸和蛋氨酸、乙醛酸和二甲酸、嘧啶和丙酮酸,每个均包含 7 个反应;色氨酸、淀粉和蔗糖路径也包含 5 个以上反应,这些路径主要是碳相关网络并分布在中心碳代谢周围。所以构造的代谢网络主要分布在中心碳代谢周围,并被单体生物合成路径围绕,同时也包含其他分散的网络。

2.4　代谢网络详述

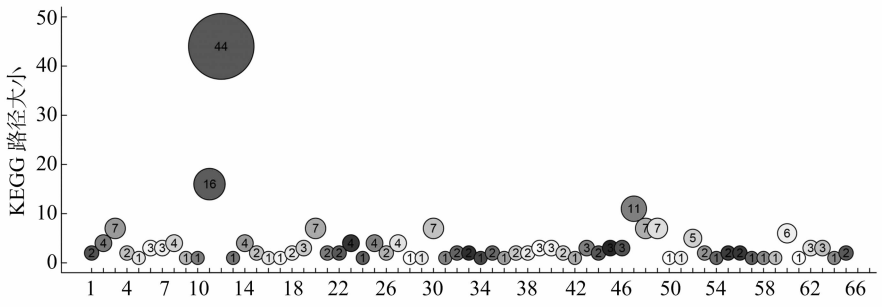
整个代谢网络由 15 个彼此间没有交集的独立子网络组成,其中最大子网络由 282 个反应构成,1 个子网络由 2 个反应构成,其余 13 个子网络均由 1 个反应构成。

最大子网络包含茶树碳中心代谢系统的主要网络,如糖酵解途径、磷酸戊糖途径、回补途径、三羧酸循环的绝大部分;此外,该网络还涵盖部分氨基酸合成代谢途径、核苷酸代谢、一碳单位代谢、糖类物质代谢、脂肪酸合成与分解代谢等重要代谢途径,同时还覆盖泛醌、NADPH、NADH、ATP、ADP、acetyl - CoA 等各类辅因子及辅酶的生成与转化途径。这些途径能够实现茶树主要物质分解、能量合成、能量转移等主要的生化活动。

另外,该网络还涉及相当数量的次生代谢网络,存在与儿茶素类物质代谢相关的黄酮醇合成酶、苯丙氨酸解氨酶、花白素还原酶等,为将儿茶素类物质代谢放到基因组规模代谢网络背景下进行研究提供便利;同时,该网络还存在没食子酸、花青素、二氢黄酮、原儿茶酸等具体物质的相关反应。

3　结论

茶叶品质的决定因素是茶树体内的生化反应所生成的各种功能性化合物,这些生化反应由茶树基因编码的酶催化并组成复杂的代谢网络。研究茶树的代谢网络对于了解茶树内的生化反应、挖掘茶树的功能基因、提升茶叶的品质、开发新的茶产品具有基础性与指导性的重要意义。本研究从 NCBI 上获得茶树的 EST 序列,通过 GI 号确定对应的酶及其催化反应,继而基于超图思想构造茶树的代谢网络,并作拓扑结构和生物意义的深入分析。后续笔者会不断地完善数据和方法,增加新的功能,如本地 BLAST 序列比对。最终,希望提供



*x*轴表示KEGG路径名称索引；*y* 轴表示KEGG路径大小，即包含的反应数，气泡半径和路径大小成比例，气泡里数字表示反应数量。KEGG路径名称和索引的对应关系：1—丙氨酸、天冬氨酸和谷氨酸代谢；2—氨基糖和核苷酸糖代谢；3—氨酰-tRNA生物合成；4—氨基苯甲酸降解；5—花生四稀酸代谢；6—精氨酸和脯氨酸代谢；7—精氨酸合成；8—维生素C代谢；9—阿特拉津降解；10—苯甲酸降解；11—抗生素的生物合成；12—次生代谢产物的生物合成；13—生物素代谢；14—丁酸代谢；15—光合生物碳固定；16—原核生物碳固定途径；17—类胡萝卜素生物合成；18—柠檬酸循环（TCA 循环）；19—氰酸代谢；20—半胱氨酸和蛋氨酸代谢；21—黄酮类化合物生物合成；22—叶酸合成；23—果糖和甘露糖代谢；24—半乳糖代谢；25—谷胱甘肽代谢；26—甘油磷脂代谢；27—甘氨酸、丝氨酸和苏氨酸代谢；28—糖酵解/糖原异生；29—糖基磷脂酰肌醇(GPI)-锚生物合成；30—乙醛酸和二甲酸代谢；31—组氨酸代谢；32—硫辛酸代谢；33—赖氨酸降解；34—单酰胺生物合成；35—氮代谢；36—叶酸一碳单位库；37—氧化磷酸化；38—泛酸和CoA生物合成；39—戊糖和葡萄糖醛酸的相互转变作用；40—戊糖磷酸途径；41—苯丙氨酸代谢；42—苯丙氨酸、酪氨酸和色氨酸生物合成；43—苯丙素合成；44—磷脂酰肌醇信号系统；45—卟啉和叶绿素代谢；46—丙酸代谢；47—嘌呤代谢；48—嘧啶代谢；49—丙酮酸代谢；50—核黄素代谢；51—鞘脂代谢；52—淀粉和蔗糖代谢；53—类固醇生物合成；54—类固醇激素生物合成；55—苯乙烯降解；56—硫代谢；57—T细胞受体信号转导通路；58—牛磺酸和亚硫磺酸代谢；59—硫胺素代谢；60—色氨酸代谢；61—酪氨酸代谢；62—缬氨酸、亮氨酸和异亮氨酸生物合成；63—缬氨酸、亮氨酸和异亮氨酸降解；64—*β*-丙氨酸代谢；65—mTOR 信号转导通路

图4　基于EST序列的茶代谢网络 KEGG 路径分析结果

一款普适工具，输入任意来源的任意序列均可轻易解析出该序列对应的酶及其催化反应信息，构建代谢网络。

参考文献：

[1] Cabrera C, Artacho R, Giménez R. Beneficial effects of green tea: a review[J]. J Am Coll Nutr, 2006, 25(2): 79 – 99.

[2] Abuajah C L, Ogbonna A C, Osuji C M. Functional components and medicinal properties of food: a review[J]. J Food Sci Technol, 2015, 52(5): 2522 – 2529.

[3] Khan N, Mukhtar H. Tea and health: studies in humans[J]. Curr Pharm Des, 2013, 19(34): 6141 – 6147.

[4] Chen H X, Zhang M, Qu Z H, et al. Antioxidant activities of different fractions of polysaccharide conjugates from green tea (*Camellia Sinensis*)[J]. Food Chem, 2008, 106(2): 559 – 563.

[5] Yang C S, Wang X, Lu G, et al. Cancer prevention by tea: animal studies, molecular mechanisms and human relevance[J]. Nat Rev Cancer, 2009, 9(6): 429 – 439.

[6] Kanwar J, Taskeen M, Mohammad I, et al. Recent advances on tea polyphenols[J]. Front Biosci, 2012(4): 111 – 131.

[7] Chen Z M, Lin Z. Tea and human health: biomedical functions of tea active components and current issues[J]. J Zhejiang Univ Sci B, 2015, 16(2): 87 – 102.

[8] Bonnely S, Davis A L, Lewis J R, et al. A model oxidation system to study oxidised phenolic compounds present in black tea[J]. Food Chem, 2003, 83(4): 485 – 492.

[9] Yun J, Kang S, Park S, et al. Characterization of a novel amylolytic enzyme encoded by a gene from a soil – derived metagenomic library[J]. Appl Environ Microbiol, 2004, 70(12): 7229 – 7235.

[10] Annaluru N, Ramalingam S, Chandrasegaran S. Rewriting the blueprint of life by synthetic genomics and genome engineering[J]. Genome Biol, 2015, 16(1): 1 – 12.

[11] Seelig B. mRNA display for the selection and evolution of enzymes from *in vitro* – translated protein libraries[J]. Nat Protoc, 2011, 6(4): 540 – 552.

[12] Karigar C S, Rao S S. Role of microbial enzymes in the bioremediation of pollutants: a review[J]. Enzyme Res, 2011(2011): 805187.

[13] Caetano – Anollés G, Yafremava L S, Gee H, et al. The origin and evolution of modern metabolism[J]. Int J Biochem Cell Biol, 2009, 41(2): 285 – 297.

[14] Wagner A, Fell D A. The small world inside large metabolic networks[J]. Proc Biol Sci, 2001, 268(1478): 1803 – 1810.

[15] Nishikawa T, Gulbahce N, Motter A E. Spontaneous reaction silencing in metabolic optimization[J]. PLoS Comput Biol, 2008, 4(12): e1000236.

[16] Janga S C, Babu M M. Network – based approaches for linking metabolism with environment[J]. Genome Biol, 2008, 9(11): 239 – 244.

[17] Griffiths A J F, Miller J H, Suzuki D T, et al. An introduction to genetic analysis: gene – protein relations[M]. 7th ed. New York: W H Freeman, 2000.

[18] Scheer M, Grote A, Chang A, et al. BRENDA, the enzyme information system in 2011[J]. Nucleic Acids Res, 2011(39): D670 – D676.

[19] McGinnis S, Madden T L. BLAST: at the core of a powerful and diverse set of sequence analysis tools[J]. Nucleic Acids Res, 2004(32): W20 – W25.

[20] Lopes C T, Franz M, Kazi F, et al. Cytoscape web: an interactive web – based network browser[J]. Bioinformatics, 2010, 26(18): 2347 – 2348.

[21] Berge C. Packing problems and hypergraph theory: a survey[J]. Ann Discrete Math, 1979(4): 3 – 37.

[22] Parkinson J, Blaxter M. Expressed sequence tags: an overview[J]. Methods Mol Biol, 2009, 533: 1 – 12.