

摘要: 新基因的功能预测可以给科研人员提供很有价值的指导信息, 在科研中起着很重要的作用。本文基于 GO (Gene Ontology) 分类系统, 结合 BLAST 的结果信息, 对来自拟南芥的十四个新基因进行了功能注释, 有效地提高了注释的可信度, 给进一步的实验提供了有价值的指导信息。

关键词: BLAST GO GoFigure

随着越来越多的基因组测序完成, 科研人员通过互联网可以得到海量的序列信息, 因此有效地利用这些序列, 对未知基因的功能进行预测以指导进一步的实验变得重要起来。现在国际一般预测新基因功能的方法与已知功能的序列数据库进行序列比对, 找到与其相似程度较高的序列, 通过已知序列的功能来推测未知序列的功能。但这种方法有着存在着手工性太强、预测准确性不高等缺点。GO 的出现给新基因的功能注释提供了一套标准, 同时提供了一种更有效更准确的预测方法。为了验证这一方法的可行性, 笔者选了 14 条拟南芥新基因, 分别用 NCBI BLAST 和 GoFigure 两种方法进行了预测, 并进行了比较分析, 以此希望对广大的科研工作者有所帮助。

一、BLAST 概论

BLAST 是目前常用的数据库搜索程序, 它是 Basic Local Alignment Search Tool 的缩写, 意为“基本局部相似性比对搜索工具”。国际著名生物信息中心都提供基于 Web 的 BLAST 服务器。BLAST 程序之所以使用广泛, 主要因为其运行速度比 FastA 等其它数据库搜索程序快, 而改进后的 BLAST 程序允许空位的插入。我们可以访问 NCBI 的网站在线进行 BLAST 和 FastA 的搜索

BLAST 软件包实际上是综合在一起的一组程序, 不仅可用于直接对蛋白质序列数据库和核酸序列数据库进行搜索, 而且可以将检测序列翻译成蛋白质或将数据库翻译成蛋白质后再进行搜索, 以提高搜索结果的灵敏度(表 3.1)。位置特异性叠代 BLAST (Position-Specific Iterated BLAST, 简称 PSI-BLAST) 则是对蛋白质序列数据库进行搜索的改进, 其主要思想是通过多次叠代找出最佳结果。具体做法是利用第一次搜索结果构建位置特异性分数矩阵, 并用于第二次的搜索, 第二次搜索结果用于第三次搜索, 依此类推, 直到找出最佳搜索结果。此外, BLAST 不仅可用于检测序列对数据库的搜索, 还可用于两个序列之间的比对。

表 1 BLAST 程序检测序列和数据库类型

程序名	检测序列	数据库类型	方法
Blastp	蛋白质	蛋白质	用检测序列蛋白质搜索蛋白质序列数据库
Blastn	核酸	核酸	用检测序列核酸搜索核酸序列数据库
Blastx	核酸	蛋白质	将核酸序列按 6 条链翻译成蛋白质序列后搜索蛋白质序列数据库
Tblastn	蛋白质	核酸	用检测序列蛋白质搜索由核酸序列数据库按 6 条链翻译成的蛋白质序列数据库

Tblastx	核酸	核酸	将核酸序列按 6 条链翻译成蛋白质序列后搜索由核酸序列数据库按 6 条链翻译成的蛋白质序列数据库
---------	----	----	--

二、GO 概论

GO(gene ontology)是基因本体论联合会(Gene Ontology Consortium)所建立的数据库，旨在建立一个适用于各种物种的，对基因和蛋白功能进行限定和描述的，并随着研究的不断深入而更新的语言词汇标准。GO 是多种生物学本体论语言中的一种，提供了三层结构的系统定义方式，用于描述基因产物的功能。

Gene Ontology (GO)项目正是为了能够使对各种数据库中基因产物功能描述相一致的努力结果。这个项目最初是由 1988 年对三个模式生物数据库的整合开始：: [FlyBase](#) (果蝇数据库 Drosophila), [Saccharomyces Genome Database](#) (酵母基因组数据库 SGD) and the [Mouse Genome Database](#) (小鼠基因组数据库 MGD)。从那开始，GO 不断发展扩大，现在已包含数十个动物、植物、微生物的数据库。

GO 的定义法则已经在多个合作的数据库中使用，这使在这些数据库中的查询具有极高的一致性。这种定义语言具有多重结构，因此在各种程度上都能进行查询。举例来说，GO 可以被用来在小鼠基因组中查询和信号转导相关的基因产物，也可以进一步找到各种生物地受体酪氨酸激酶。这种结构允许在各种水平添加对此 基因产物特性的认识。

GO 发展了具有三级结构的标准语言 (ontologies)，如表所示。根据基因产物的相关分子功能，生物学途径，细胞学组件而给予定义，无物种相关性。

本体论	内容
分子功能本体论	基因产物个体的功能，如与碳水化合物结合或 ATP 水解酶活性等
生物学途径本体论	分子功能的有序组合，达成更广的生物功能，如有丝分裂或嘌呤代谢等
细胞组件本体论	亚细胞结构、位置和大分子复合物，如核仁、端粒和识别起始的复合物等

三、基因的准备

本次测试得到了生物技术楼李林川的支持，提供了该实验室提供刚筛选的 14 条基因，有一定的实验支持，由于这些序列暂时还没发表，所以本论文公开前请与作者和李临川同学联系。

四、基因功能的预测与比较分析

以 At5g09980 为例，说明两种预测方法的优缺点。

1. BLAST

下面是 At5g09980 蛋白序列在 NCBI 网站进行 BLAST 的结果，采用默认参数，比对程序为 blastp:

Score	E	Sequences producing significant alignments:	(bits)	Value
gi 4539001 emb CAB39622.1 		putative protein [Arabidopsis th...	1134	0.0
gi 24030488 gb AAN41392.1 		unknown protein [Arabidopsis tha...	1122	0.0
gi 37534852 ref NP_921728.1 		putative pol polyprotein [Oryz...	390	e-107
gi 34903122 ref NP_912908.1 		unnamed protein product [Oryza...	385	e-105
gi 50919731 ref XP_470226.1 		Unknown protein with similarit...	360	9e-98
gi 42517136 ref NP_964000.1 		RIKEN cDNA G430022H21 gene [Mu...	244	5e-63
gi 34859963 ref XP_215698.2 		similar to KIAA1627 protein [R...	244	5e-63
gi 10047331 dbj BAB13453.1 		KIAA1627 protein [Homo sapiens]	243	1e-62
gi 24308265 ref NP_066012.1 		KIAA1627 protein [Homo sapiens...	242	2e-62
gi 48138147 ref XP_393391.1 		similar to CG7818-PA [Apis mel...	240	1e-61
gi 51513454 gb AAH80361.1 		Unknown (protein for MGC:79735) ...	239	1e-61
gi 17862044 gb AAL39499.1 		LD06016p [Drosophila melanogaste...	239	2e-61
gi 46250316 gb AAH68744.1 		MGC81236 protein [Xenopus laevis]	239	2e-61
gi 42542893 gb AAH66377.1 		Hypothetical protein zgc:77296 [...	234	7e-60
gi 31223741 ref XP_317346.1 		ENSANGP00000010457 [Anopheles ...	233	1e-59
gi 23484164 gb EAA19588.1 		Unknown-related [Plasmodium yoel...	204	6e-51
gi 50746749 ref XP_420636.1 		PREDICTED: similar to KIAA1627...	190	1e-46
gi 37360482 dbj BAC98219.1 		mKIAA1627 protein [Mus musculus]	189	3e-46
gi 23509037 ref NP_701705.1 		hypothetical protein [Plasmodi...	188	6e-46
gi 49071462 ref XP_400020.1 		hypothetical protein UM02405.1...	165	4e-39
gi 47207634 emb CAF93556.1 		unnamed protein product [Tetrao...	152	3e-35
gi 50256650 gb EAL19373.1 		hypothetical protein CNBH0670 [C...	128	7e-28
gi 50304541 ref XP_452225.1 		unnamed protein product [Kluyv...	102	4e-20
gi 10954013 gb AAG25704.1 		KAR4-like protein [Saccharomyces...	100	2e-19
gi 50285123 ref XP_444990.1 		unnamed protein product [Candi...	100	3e-19
gi 45199255 ref NP_986284.1 		AFR736Cp [Eremothecium gossypi...	99	3e-19
gi 50548917 ref XP_501929.1 		hypothetical protein [Yarrowia...	94	1e-17
gi 6319795 ref NP_009876.1 		Transcription factor required f...	91	2e-16
gi 23612659 ref NP_704220.1 		mRNA (N6-adenosine)-methyltran...	83	3e-14
gi 50251923 dbj BAD27860.1 		putative m6A methyltransferase ...	80	2e-13
gi 23486499 gb EAA20809.1 		hypothetical protein [Plasmodium...	80	2e-13
gi 7267774 emb CAB81177.1 		putative methyltransferase [Arab...	79	4e-13
gi 12844323 dbj BAB26322.1 		unnamed protein product [Mus mu...	79	4e-13
gi 50284965 ref XP_444911.1 		unnamed protein product [Candi...	78	1e-12
gi 46249484 gb AAH68672.1 		MGC81069 protein [Xenopus laevis]	77	2e-12
gi 50423415 ref XP_460290.1 		unnamed protein product [Debar...	76	3e-12
gi 9790097 ref NP_062695.1 		methyltransferase-like 3; m6a m...	75	9e-12
gi 2460037 gb AAB71850.1 		m6A methyltransferase [Homo sapiens]	74	1e-11
gi 26351497 dbj BAC39385.1 		unnamed protein product [Mus mu...	74	1e-11
gi 15214786 gb AAH12526.1 		Mettl3 protein [Mus musculus] >g...	74	1e-11
gi 30353931 gb AAH52244.1 		Methyltransferase like 3 [Homo s...	74	1e-11
gi 21361827 ref NP_062826.2 		methyltransferase like 3; puta...	74	1e-11
gi 46434953 gb EAK94346.1 		hypothetical protein Ca019.11221...	73	3e-11
gi 46434912 gb EAK94308.1 		hypothetical protein Ca019.3736 ...	73	3e-11
gi 21064123 gb AAM29291.1 		AT20169p [Drosophila melanogaste...	72	4e-11
gi 47086489 ref NP_997945.1 		Unknown (protein for MGC:77093...	72	7e-11
gi 31206127 ref XP_312015.1 		ENSANGP00000018660 [Anopheles ...	71	1e-10
gi 33301348 sp O82486 MT70_ARATH		Probable N6-adenosine-meth...	69	5e-10
gi 28193122 emb CAD62303.1 		unnamed protein product [Homo s...	67	2e-09
gi 50412773 ref XP_457162.1 		unnamed protein product [Debar...	60	2e-07
gi 6321246 ref NP_011323.1 		Probable mRNA N6-adenosine meth...	60	2e-07
gi 49072630 ref XP_400604.1 		hypothetical protein UM02989.1...	60	2e-07
gi 47227445 emb CAG04593.1 		unnamed protein product [Tetrao...	60	3e-07
gi 50310943 ref XP_455494.1 		unnamed protein product [Kluyv...	59	4e-07
gi 50545848 ref XP_500462.1 		hypothetical protein [Yarrowia...	57	2e-06
gi 45198691 ref NP_985720.1 		AFR173Wp [Eremothecium gossypi...	56	4e-06

gi 46444505 gb EAL03779.1 	hypothetical protein Ca019.1476 ...	56	4e-06
gi 50257763 gb EAL20464.1 	hypothetical protein CNBE3850 [C...	54	2e-05

序列的相似性情况

```

>gi|4539001|emb|CAB39622.1| putative protein [Arabidopsis thaliana]
gi|7267694|emb|CAB78121.1| putative protein [Arabidopsis thaliana]
gi|7487728|pir|T04002 hypothetical protein T5L19.110 - Arabidopsis thaliana
Length = 963

Score = 1134 bits (2933), Expect = 0.0
Identities = 642/761 (84%), Positives = 642/761 (84%)

Query: 15 TWYQDGEQDGGDRSEKRRMSLKXXXXXXXXXXXXXXXXXNKSVVVVEHQDRDSKRERDG 74
          TWYQDGEQDGGDRSEKRRMSLKA                                NKSVVVVEHQDRDSKRERDG
Sbjct: 15 TWYQDGEQDGGDRSEKRRMSLKASDFESSRSGGSKSKEDNKSVVVVEHQDRDSKRERDG 74

Query: 75 RERTHGXXXXXKRKRWDEAGGLVNDGDHKSSKLSDSRHDSGGERVSVSNEHGESRRDLK 134
          RERTHG          KRKRWDEAGGLVNDGDHKSSKLSDSRHDSGGERVSVSNEHGESRRDLK
Sbjct: 75 RERTHGSSSDSSKRKRWDEAGGLVNDGDHKSSKLSDSRHDSGGERVSVSNEHGESRRDLK 134

Query: 135 SDRSLKTSSRDEKSKSRGVKDDDRGSPLKKTSGKDGSEVVREVGRSNRSKTPDADYEKEK 194
          SDRSLKTSSRDEKSKSRGVKDDDRGSPLKKTSGKDGSEVVREVGRSNRSKTPDADYEKEK
Sbjct: 135 SDRSLKTSSRDEKSKSRGVKDDDRGSPLKKTSGKDGSEVVREVGRSNRSKTPDADYEKEK 194

Query: 195 YSRKDEXXXXXXXXXXXXXXXXXQEGLKDNWKRHSXXXXXXXXXXXXXXXXXRGREREFPRQX 254
          YSRKDE          QEGLKDNWKRHS          RGREREFPRQ
Sbjct: 195 YSRKDESRGRDDGWSRDRDQEGLKDNWKRHSSSSGDKDKQKGDLLYDRGREREFPRQG 254

Query: 255 XXXXXXXXXXXXXXXXXXXKDGNRGEAVKALSSGGVSNENYDVIEIQTKPHDYVRGESGNFA 314
          KDGNRGEAVKALSSGGVSNENYDVIEIQTKPHDYVRGESGNFA
Sbjct: 255 RERSEGERSHGRLGGRKDGNRGEAVKALSSGGVSNENYDVIEIQTKPHDYVRGESGNFA 314

Query: 315 RMTESGQQPPKKPSNNEEWAHNQEGRQRSETFGFGSYGEDSRDEAGEASSDYSGAKARN 374
          RMTESGQQPPKKPSNNEEWAHNQEGRQRSETFGFGSYGEDSRDEAGEASSDYSGAKARN
Sbjct: 315 RMTESGQQPPKKPSNNEEWAHNQEGRQRSETFGFGSYGEDSRDEAGEASSDYSGAKARN 374

Query: 375 QRGSTPGRTNFVQTPNRYQTPQGTRGNRPLRGGKGRPAGGRENQQAIPMPIMGSPFAN 434
          QRGSTPGRTNFVQTPNRYQTPQGTRGNRPLRGGKGRPAGGRENQQAIPMPIMGSPFAN
Sbjct: 375 QRGSTPGRTNFVQTPNRYQTPQGTRGNRPLRGGKGRPAGGRENQQAIPMPIMGSPFAN 434

Query: 435 LGMPPPSPPIHSLTPGMSPIPGTSVTPVFMPPFAPTLIWPARGVDGNMLXXXXXXXXXXXX 494
          LGMPPPSPPIHSLTPGMSPIPGTSVTPVFMPPFAPTLIWPARGVDGNML
Sbjct: 435 LGMPPPSPPIHSLTPGMSPIPGTSVTPVFMPPFAPTLIWPARGVDGNMLPVPPVLSPLPP 494

Query: 495 XXXXXRFPSIXXXXXXXXXXXXXXXXXSDRGGPPNFPGSNISQMGGRGMPDKTSGGGWVPPRG 554
          RFPSI          SDRGGPPNFPGSNISQMGGRGMPDKTSGGGWVPPRG
Sbjct: 495 GPSGPRFPSIGTPPNPMFFTPPGSDRGGPPNFPGSNISQMGGRGMPDKTSGGGWVPPRG 554

Query: 555 XXXXXXXXXXXXEQNDYSQNFVDTGMRPQNFIRELELTNVEDYPKLRELIQKDEIVSNSA 614
          EQNDYSQNFVDTGMRPQNFIRELELTNVEDYPKLRELIQKDEIVSNSA
Sbjct: 555 GGPPGKAPSRGEQNDYSQNFVDTGMRPQNFIRELELTNVEDYPKLRELIQKDEIVSNSA 614

Query: 615 SAPMYLKGDLEVELSPELFGTKFDVILVDPPEEYVHRAPGVSDSMEYWTFEDIINLKI 674
          SAPMYLKGDLEVELSPELFGTKFDVILVDPPEEYVHRAPGVSDSMEYWTFEDIINLKI
Sbjct: 615 SAPMYLKGDLEVELSPELFGTKFDVILVDPPEEYVHRAPGVSDSMEYWTFEDIINLKI 674

```

```

Query: 675 EAIADTPSFLFLWVDGVDVLEQGRQCLKKWGFRRCEDICWVKTNKSNAAPT LRHDSRTVF 734
          EAIADTPSFLFLWVDGVDVLEQGRQCLKKWGFRRCEDICWVKTNKSNAAPT LRHDSRTVF
Sbjct: 675 EAIADTPSFLFLWVDGVDVLEQGRQCLKKWGFRRCEDICWVKTNKSNAAPT LRHDSRTVF 734

Query: 735 QRSKEHCLMGIKGTVRRSTDGHIIHANIDTDVIIAEPPYG 775
          QRSKEHCLMGIKGTVRRSTDGHIIHANIDTDVIIAEPPYG
Sbjct: 735 QRSKEHCLMGIKGTVRRSTDGHIIHANIDTDVIIAEPPYG 775

.....

```

从搜索结果来看，**evaluate** 比较高的序列与 **At5g09980** 蛋白相似性较高，因此我们推断在功能上也很可能存在相关性，通观所有命中的序列，推测可能与 **KIAA1627** 蛋白功能相关。但是，由于 **At5g09980** 以前很少有人研究，其相关序列的功能注释也很少，同时上述命中序列缺乏很好的一致性，实验科学家较难理解，同时估测结果可信度也不大。

2. GoFigure

斯坦福大学开发的基于 **GO** 分类数据库和 **BLAST** 相似性搜索基因自动注释工具。**GoFigure** 底层是通过 **GODEl** 来实现的，通过 **BLAST** 搜索用 **GO** 注释过的蛋白质数据库，像 **SwissProt**, **Flybase (Drosophila)**, **the Saccharomyces Genome Database (SGD)**, **Mouse Genome Informatics (MGI)** and **Wormbase (nematode)**, 来搜索同源序列，根据 the **Expect values (E-values)** 值来预测基因的功能。通过 **GoFigure** 网站，可以提交 **DNA** 或者蛋白序列，同时也可以提供一个电子信箱地址，这样预测结果会自动发送至邮箱中，非常的方便。预测结果以表格和图形两种形式出现，清晰具体地注释了目标序列在分子功能、生物学途径、细胞组件三方面的功能。当然也可以点击 **AmiGO** 的链接浏览相关 **GO** 条目的具体信息。

下面是 **GoFigure** 的预测的表格结果和图形结果：

At4g09980	C	5634	nucleus
At4g09980	F	16422	mRNA (2'-O-methyladenosine-N6-)-methyltransferase activity
At4g09980	P	1510	RNA methylation
At4g09980	P	7126	Meiosis

表 2 预测的表格结果

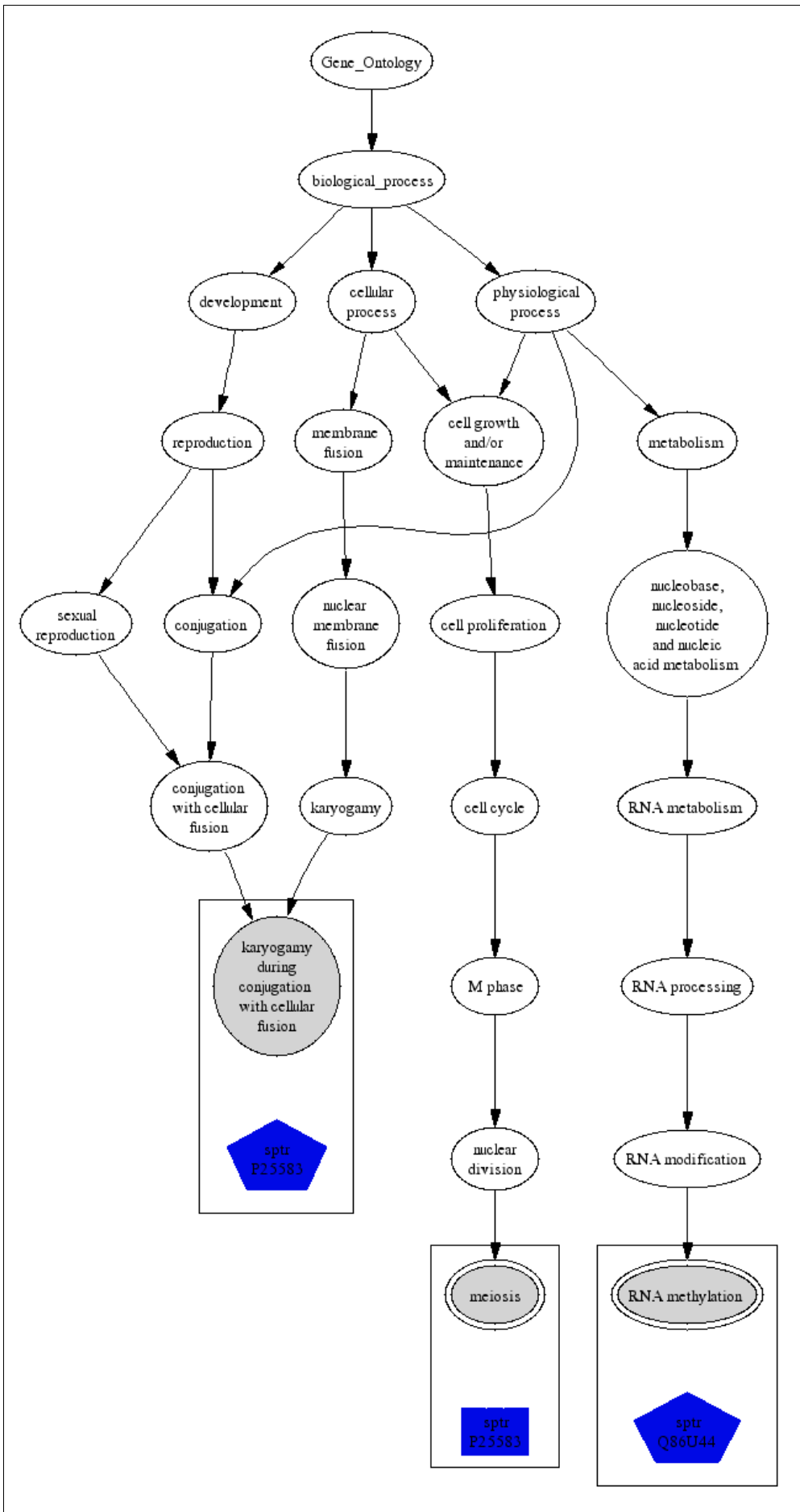


图 1 生物学途径途径

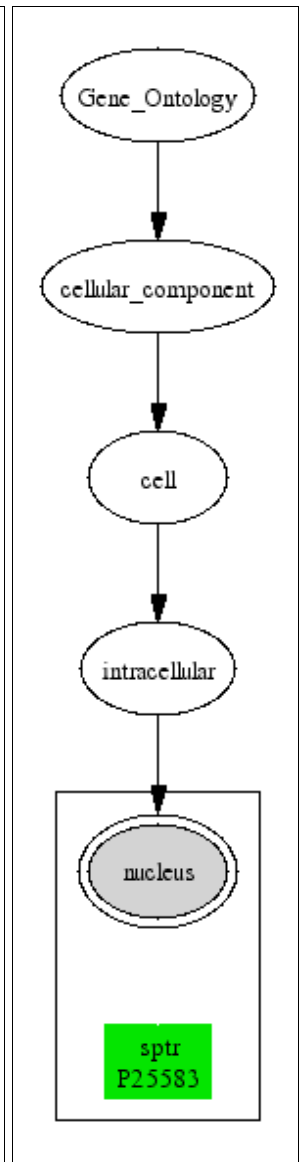


图 2 细胞组件角度

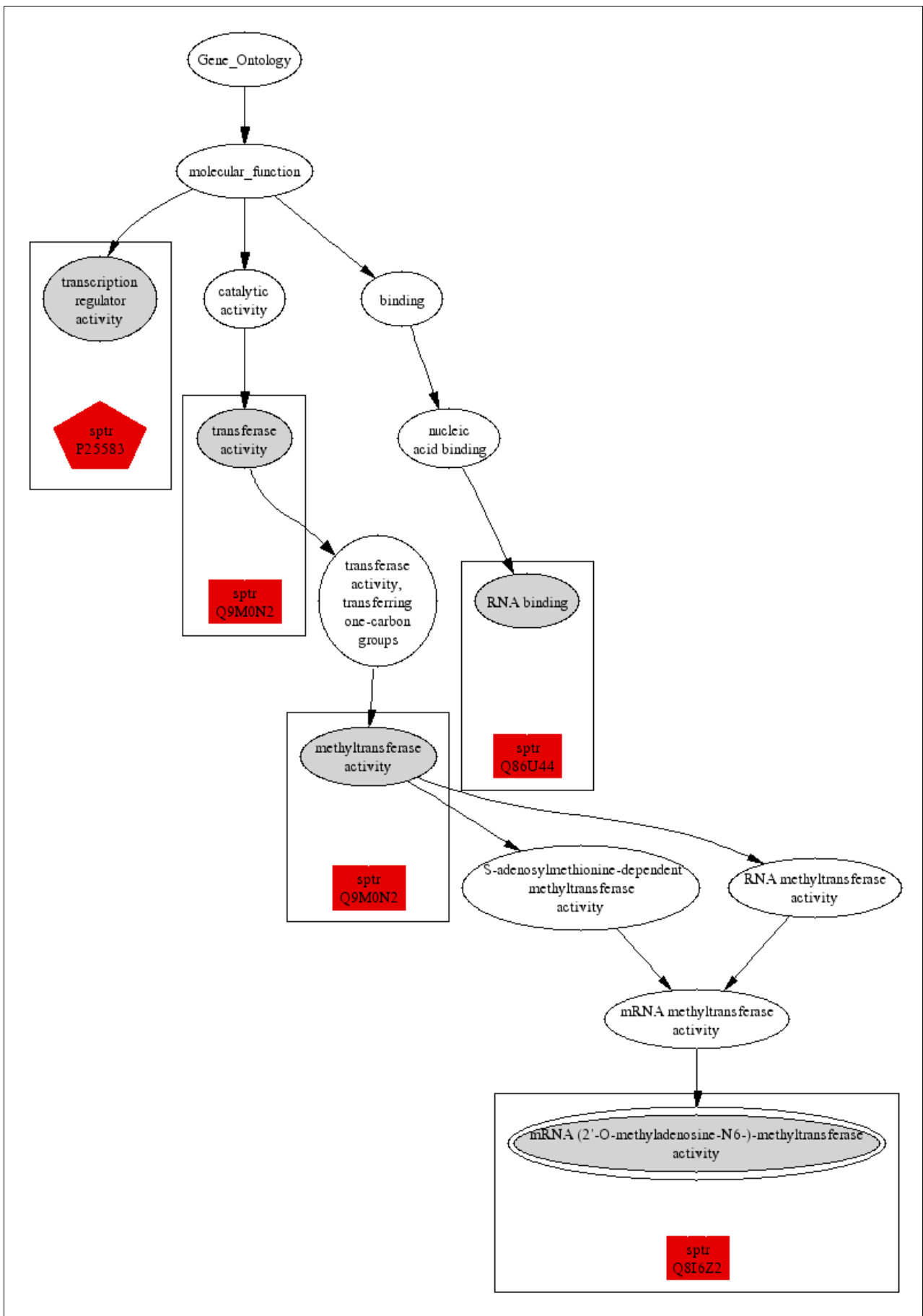


图3 分子功能角度

由于有很好的 AmiGO 分类数据库做支持，很好地整合了比较基因组方法的成果，所以

GoFigure 预测的可信度更高更准确一些。预测结果(表 2)表明 At4g09980 与细胞核内的甲基转移有关,很可能是一种甲基转移酶,同时 GoFigure 还给出了可视化的结构图表示,还可以利用相关链接浏览 AmiGO 中其他物种相关蛋白的注释信息。

从 14 条基因预测的结果来看,有 8 条可以通过 GoFigure 预测到功能,6 条不能预测。虽然不是 100%,但是考虑到 GO 尚在发展之中和相似性比较的局限性,应该是可以接受的。

五、结论

BLAST 一直是注释新基因功能的基础工具,但它也有自身难以克服的缺陷,不适合专门用来做基因功能注释,而现在基于 GO(Gene Ontology)发展的基因注释工具,充分地利用了比较基因组方法和 BLAST 相似性搜索的优点,有效地提高了自动注释的准确性,同时提供了形象的可视化信息,为科研人员提供了快速有效的指导信息。

当然 GO 现在还在发展之中,有些序列可能得不到预测结果,所以最好结合 BLAST 来分析,这样就会使预测更具目的性,也更准确。

六、参考文献

1. Ashburner, M., Ball, et al. 2000. Gene ontology: Tool for unification of biology. The Gene Ontology Consortium. Nat. Genet: 25-29.
2. Altschul, S.F., Madden, T.L., et al. 1997. Gapped BL/AST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402
3. S.B. Primrose, R.M. Twyman. 2003. Principles of Genomes Analysis and Genomics. Blackwell Publing.
4. NCBI BLAST (<http://www.ncbi.nlm.nih.gov/BLAST/>)
5. Gene Ontology Consortium (<http://www.geneontology.org/index.shtml>)
6. GoFigure (<http://udgenome.ags.udel.edu/gofigure/>)