拟南芥 SBP 转录因子基因家族分析

本章提要

本章以植物特异转录因子 SBP 基因家族为实例，讲述如何利用生物信息数据库资源和软件工具，对该家族进行系统分析。本章所用数据主要来自北京大学生物信息中心构建的植物转录因子数据库，研究对象主要为拟南芥基因组中 16 个 SBP 转录因子基因。所用生物信息学工具包括基因结构显示、双序列比对、蛋白质功能域识别、蛋白质保守域预测、序列图标构建、多序列比对、系统发育树构建、表达谱分析，以及蛋白质三维结构空间图形显示等。研究结果表明，SBP 基因家族为绿色植物所特有，在拟南芥基因组中有 16 个基因座，其基因结构差异很大，经可变剪接得到 27 个转录本，编码 16 个 SBP 转录因子，其蛋白质序列 DNA 结合结构域保守序列为 79 个氨基酸残基，含两个互不重叠的锌指结构序列模体。基因结构、系统发育和蛋白质序列保守结构域分析表明，上述 16 个 SBP 基因可分为 4 组，每组中均有一对重复基因。不同组基因结构和所编码蛋白质序列差异很大，可能与该基因家族功能多样性有关。

引言

我们知道，转录前调控、转录调控和转录后调控是真核生物基因表达调控的重要组成部分，其中转录调控通过顺式作用元件和反式作用因子相互作用实现。顺式作用元件（cis-element）泛指 DNA 序列一个片段，通常位于被调控基因的上游，主要包括启动子（promoter）、增强子（enhancer）和抑制子（suppressor）三类。反式作用因子泛指与顺式作用元件直接或间接结合并参与靶基因转录过程的调控因子，通称转录因子（transcription factor）。转录因子可以分为两大类，即通用转录因子（general transcription factor）和特异转录因子（specific transcription factor）。通用转录因子与靶基因上游约 10-35 位的 TATA 框（TATA-box）或启动子区域转录起始位点（transcription start site）DNA 序列结合，并与 II 型 RNA 聚合酶一起，形成转录起始复合物。特异转录因子种类繁多、功能复杂，它们与靶基因上游各种特定 DNA 序列片段结合，激活或抑制靶基因转录活性，以调控靶基因在不同组织、不同细胞、不同环境条件下特异表达。如无特别说明，通常所说的转录因子即指特异转录因子。

转录因子蛋白质序列除包含 DNA 结合结构域（DNA binding domain，DBD）外，一般还含有转录调控结构域（transcription regulation domain），主要用于调控靶基因转录活性，既可激活转录，也可抑制转录。转录因子中的核定位信号（nuclear localization signal, NLS），可引导转录因子在胞浆内合成后通过核膜进入细胞核。此外，有些转录因子含寡聚化结构域，可形成二聚体或多聚体复合物，具有更为复杂的调控机制。

显然，DNA 结合结构域是转录因子必须具备的功能单元，具有特定空间结构，常见的有锌指（zinc finger）结构、螺旋－转角－螺旋（helix-turn-helix）、螺旋－回环－螺旋（helix-loop-helix）、亮氨酸拉链（leucine zipper）等。通常，转录因子 DNA 结合结构域特定部位（如 alpha 螺旋）嵌入 DNA 双螺旋大沟，蛋白质序列上精氨酸、赖氨酸等带电或极性残基与 DNA 序列上碱基结合，结合位点（transcription factor binding site, TFBS）碱基序列特异性决定了能够与其结合的转录因子种类。

植物转录因子数据库

按照DNA结合结构域序列特征，可以将转录因子分为不同种类，即不同家族。2000 年，植物中第一个模式生物拟南芥（*Arabidopsis thaliana*）全基因组测序完成（AGI, 2000）。Riechmann

等通过文献检索，收集整理了已知植物转录因子家族，预测了拟南芥基因组共编码 1533 个转录因子，占拟南芥总基因数 5.9%，其中 45%是植物特异转录因子（Riechmann et al., 2000）。2003 年，北大－耶鲁合作中心等研究组对拟南芥全基因组转录因子进行了cDNA克隆，得到了 1200 多个拟南芥转录因子基因的cDNA序列（Gong et al., 2004），并进一步开展蛋白质/蛋白质、蛋白组/DNA相互作用研究。2005 年，我们通过文献调研，收集了 64 个植物转录因子家族信息，其中 48 个已经由蛋白质家族数据库PFAM收录，建立了DNA结合结构域隐马氏模型序列谱（HMM Profile），可以用来预测拟南芥基因组中该 48 个家族转录因子。另外 16 个家族则可根据文献报道，以DNA结合结构域为种子序列，用BLAST程序搜索拟南芥基因组中可能的同源基因。利用上述两种方法，预测得到 1922 个拟南芥转录因子，并进行了基于家族水平和基因水平的注释，构建了拟南芥转录因子数据库（http://datf.cbi.pku.edu.cn，DATF）。之后，又构建了水稻（*Oryza sativa*）、杨树（*Populus pinus*）、苔藓（*Physcomitrella patens*）、衣藻（*Chlamydomonas reinhardtii*）等具有全基因组序列的植物转录因子数据库，以及玉米、棉花等正在进行EST测序的植物转录因子数据库（ http://planttfdb.cbi.pku.edu.cn），并提供了统一的浏览、检索和数据下载界面（Guo et al., 2008）。

例如，拟南芥转录因子数据库 DATF 搜集的 SBP 转录因子家族共有 16 个基因， 其中大部分已有文献报道（Gardon et al., 1999），并按类 SBP 基因（SPL）的方式命名。DATF 中每个转录因子均有染色体定位编码信息，如 At1G53160，“At”为拟南芥拉丁名 *Arabidopsis thaliana* 的双字母缩写，1G 表示该转录因子基因位于 1 号染色体，53160 为编号。若该基因有不同剪接方式，则在后面缀以数字以示区别，如 At1G53160.1，At1G53160.2 等。每个基因均有详细注释，主要包括（图 1）以下内容。

- 基本信息：基因编码、名称、基因全长、CDS 编码序列长度、所编码蛋白质序列长度和理论分子量，以及该基因的简单说明。
- 基因结构：编码方向、染色体定位、编码区起始和终止位置，并用图形方式显示基因结构。
- 蛋白质结构域：结构域名称、简介、起始和终止位置，并用图示方式显示结构域信息。
- 表达信息：不同剪接方式的 EST 信息。
- 基因本体注释系统：从 InterPro 注释结果得到的该基因 GO 注释。
- 交叉链接：NCBI 数据库 GenBank 和 GenPept，拟南芥基因组数据库 TAIR、TIGR、MIPS 和 SIGnal T-DNA Express，TRANSFAC、PubMed。
- 序列信息：全长基因、蛋白编码序列 CDS、所编码蛋白质序列。
- 其它注释信息：基因复制、核定位信号、UniGene 注释，以及克隆和测序信息等。

除了对每个转录因子进行详细注释外，根据文献报道，对每个家族也进行了概要介绍，包括结合位点顺式元件信息、已知 DNA 结合结构域三维结构信息，并提供了 DNA 结合结构域多序列比对结果和系统发育树。

图 1 - 拟南芥 SBP 转录因子基因 AtSPL4 第一个剪接体 AT1G69170.1 的注释信息 A: 基本信息; B: 基因和蛋白质结构域信息; C: 数据库注释和交叉索引; D: 序列信息

拟南芥 SBP 转录因子基因家族研究背景

拟南芥转录因子数据库 DATF 中部分为植物特异转录因子家族，至今只在植物中发现，SBP 基因家族就是一个植物特异转录因子家族。1996 年，德国科隆植物育种马普研究所（Max Planck Institute for Plant Breeding Research）Huijser 领导的研究小组从金鱼草（*Antirrhinum majus*）中克隆得到两个基因，它们能与花发育相关基因 Squamosa 的启动子结合并调控其表达（Klein et al., 1996），命名为 Squamosa 启动子结合蛋白（Squamosa promoter Binding Protein, SBP），简称 SBP1 和 SBP2。之后，先后在拟南芥和苔藓中进行了 SBP 基因家族的 cDNA 克隆，发现在这两个植物中有多个 SBP 转录因子基因拷贝（Cardon et al., 1999; Riese et al., 2007）。由于多数基因尚未进行功能鉴定，Huijser 等将它们称为类 SBP（SBP-like, SPL）基因。因为这些基因都有编码 DNA 结合结构域的保守核苷酸序列，又称为 SBP 盒（SBP-box）基因，SBP 盒编码的蛋白质序列称 SBP 结构域（SBP-domain），含 79 个氨基酸残基。Yamasaki 等（2004）利用核磁共振方法测定了拟南芥中两个 SBP 结构域的三维溶液构象，均含 2 个锌指结构（Zinc finger）。其中一个（SPL3）N-端锌指结构序列模体为 Cys3His，C-端锌指结构为 Cys2HisCys；另一个（SPL7）N-端锌指结构序列模体为 Cys4，C-端锌指结构与 SPL3 相同，也是 Cys2HisCys。其与 SBP 结构域结合的顺式元件具有以 GTAC 为核心的迴文结构。SBP 结构域 C 端富含碱性氨基酸，可能为双向核定位信号（nuclear localization signal, NLS）（Birkenbihl et al., 2005）。

据文献报道，SBP 转录因子家族具有许多重要生物学功能，包括花和果实发育、孢子发育、激素应答、抗霉菌侵蚀等。拟南芥 SPL3 基因的转录受发育调控，主要集中在花序顶端分生组织、花分生组织和花器官原基，可能与控制开花时间有关（Cardon et al., 1997）。拟南芥 SPL8 基因影响孢子囊发育，进而影响植物生殖（Unte et al., 2003），并可能在赤霉素介导的发育过

程中起局部调节作用（Zhang et al., 2006）。最近报道证实，拟南芥三个 SBP 基因（SPL3、SPL4 和 SPL5）中具有 microRNA156 的调控位点（Wu, 2006; Gandikota et al., 2007）。

除拟南芥外，其它植物 SBP 转录因子研究也屡有报道。玉米 SBP 转录因子 Liguleless1（LG1）缺失突变体不能形成舌叶和叶耳(Moreno et al. 1997)。白桦 BpSPL1 基因能特异结合 BpMADS5 启动子，参与调节花发育 (Lannenpaa et al. 2004)。衣藻（*Chlamydomonas reinhardtii*）中铜应答调控子 1（Copper Response Regulator 1, CRR1）基因含典型的 SBP 结构域（Eriksson et al. 2004; Kropat et al., 2005）。最近研究表明，SBP 基因在植物果实成熟和发育中起重要作用。野生玉米（teosinte）粒有硬壳包裹，而栽培玉米粒则无外壳包裹。这种表型变化主要由玉米果实构造基因（teosinte glume architecture, tga1）控制（Wang et al., 2005）。该基因为 SBP 转录因子家族成员，野生玉米第 6 位氨基酸为赖氨酸（Lys），而栽培玉米中为天冬酰氨（Asn）。Manning 等(2006)发现，控制西红柿果实成熟的关键基因（Colorless non-ripening locus）是一个 SBP 基因（LeSPL-CNR），启动子区域甲基化修饰突变体可抑制果实成熟。

表 1 拟南芥 16 个 SBP 转录因子基因及其 27 个不同剪接体

No	Name	TAIR ID	RefSeq ID	N	GenPet ID	L	Gene ID
1	AtSPL01A	At2G47070.1	NM_180137.2	10	NP_850468.1	881	819321
2	AtSPL02A	At5G43270.1	NM_180791.1	4	NP_851122.1	419	834345
3	AtSPL02B	At5G43270.2	NM_123693.4	4	NP_199141.1	419	834345
4	AtSPL02C	At5G43270.3	NM_203146.2	4	NP_974875.1	419	834345
5	AtSPL03A	At2G33810.1	NM_128940.2	2	NP_565771.1	131	817948
6	AtSPL04A	At1G53160.1	NM_104194.3	2	NP_175723.1	174	841749
7	AtSPL04B	At1G53160.2	NM_202285.1	2	NP_974014.1	174	841749
8	AtSPL05A	At3G15270.1	NM_112390.3	2	NP_188145.1	181	820758
9	AtSPL06A	At1G69170.1	NM_105584.5	3	NP_177077.3	405	843248
10	AtSPL07A	At5G18830.1	NM_121888.2	10	NP_197384.1	801	832001
11	AtSPL07B	At5G18830.2	NM_180519.1	10	NP_850850.1	775	832001
12	AtSPL08A	At1G02065.1	NM_148426.4	3	NP_683267.1	333	839275
13	AtSPL09A	At2G42200.1	NM_129782.2	3	NP_181749.1	375	818820
14	AtSPL10A	At1G27370.1	NM_102499.3	4	NP_174057.2	396	839626
15	AtSPL10B	At1G27370.2	NM_202192.1	4	NP_973921.1	396	839626
16	AtSPL10C	At1G27370.3	NM_001084136.1	3	NP_001077605.1	396	839626
17	AtSPL10D	At1G27370.4	NM_001036019	3	NP_001031096.2	396	839626
18	AtSPL11A	At1G27360.1	NM_202191.1	4	NP_973920.1	393	839625
19	AtSPL11B	At1G27360.2	NM_001084134.1	4	NP_001077603.1	393	839625
20	AtSPL11C	At1G27360.3	NM_001084135.1	4	NP_001077604.1	393	839625
21	AtSPL11D	At1G27360.4	NM_102498.2	4	NP_564280.1	393	839625
22	AtSPL12A	At3G60030.1	NM_115866.3	10	NP_191562.1	927	825173
23	AtSPL13A	At5G50570.1	NM_180830.3	3	NP_851161.1	359	835126
24	AtSPL13B	At5G50570.2	NM_124435.2	3	NP_568731.1	359	835126
25	AtSPL14A	At1G20980.1	NM_101951.3	10	NP_173522.1	1035	838692
26	AtSPL15A	At3G57920.1	NM_115654.2	3	NP_191351.1	354	824961
27	AtSPL17A	At5G50670.1	NM_124445.2	3	NP_568740.1	359	835138

N: 外显子数; L: 蛋白质序列长度

拟南芥 SBP 转录因子基因结构

表 1 列出 DATF 数据库中收集的 16 个 SBP 转录因子基因共 27 个剪接体的基本信息，包括名称、染色体定位、核酸参考序列数据库 RefSeq 代码、外显子个数、蛋白质序列数据库 GenPept 代码、编码蛋白质序列长度和数据库 Gene 代码等，并通过检索 NCBI 基因数据库 Gene，得到最近更新的数据。为便于和已有文献及数据库对照，我们沿用文献和数据库中以类 SBP（SBP-like, SPL）的习惯命名，冠以物种名缩写 At；同一基因若有不同转录本，则分别以 A、B、C 予以区分，如 AtSPL01A、AtSPL02A、AtSPL02B。

16 个 SBP 基因在染色体上分布情况如下：1 号染色体 6 个（AtSPL04、AtSPL06、AtSPL08、AtSPL10、AtSPL11、AtSPL04），2 号染色体 3 个（AtSPL01、AtSPL03、AtSPL09），3 号染色体 3 个（AtSPL05、AtSPL12、AtSPL15），5 号染色体 4 个（AtSPL02、AtSPL07、AtSPL13、AtSPL17）。我们知道，可变剪接（alternative splicing）是真核生物重要特征。拟南芥 SBP 转录因子家族 16 个基因中 2 个基因（AtSPL 10 和 AtSPL 11）有 4 种剪接方式，各编码 4 个 SBP 转录因子，所编码蛋白质序列相同；1 个基因（AtSPL02）有 3 种剪接方式，编码 3 个长度为 419 个残基的蛋白质序列；3 个基因（AtSPL04、AtSPL 07 和 AtSPL 13）有 2 种剪接方式，各编码 2 个 SBP 转录因子，其中 AtSPL07 所编码蛋白质序列长度不同。另外 9 个基因（AtSPL 01、AtSPL 03、AtSPL 05、AtSPL 06、AtSPL 12、AtSPL 12、AtSPL 14、AtSPL 15 和 AtSPL 17）只有一种剪接方式，各编码一个 SBP 转录因子。基因 AtSPL08（At1G02065）有两种可变剪接体 At1G02065.1 和 At1G02065.2，编码两个不同长度的蛋白质，NCBI 参考序列数据库 RefSeq 代码为 NM_148426 和 NM_202009，分别编码长度为 333 和 246 个残基的蛋白质序列 NP_683267 和 NP_973738。其中 NP_973738 仅含 SBP 结构域 N 端部分序列，不是一个完整的 SBP 基因，表中不予列入。基因 At1G76580 曾被预测为 SBP 转录因子，并以 AtSPL16 命名；其 mRNA 序列（GenBank：NM_106308）编码 809 个氨基酸残基（NP_177784），不含 SBP 结构域，表中不予列入。AtSPL13（At5G50570）和 AtSPL17 （At5G50670）位于 5 号染色体相邻部位，其序列完全相同，可能是尚未分化的复制基因。

应该注意，随着拟南芥功能基因组、蛋白组、代谢组研究不断深入，新的可变剪接形式不断发现，可变剪接形式可能不断更新。位于美国 Stanford 的拟南芥信息资源网站（The Arabidopsis Information Resource, TAIR）是国际上最为权威的拟南芥基因组数据库和拟南芥基因组注释系统，具有丰富的数据资源和最新的注释信息。拟南芥转录因子数据库 DATF 的每个条目都有 TAIR 链接，可以直接查看最新更新信息。

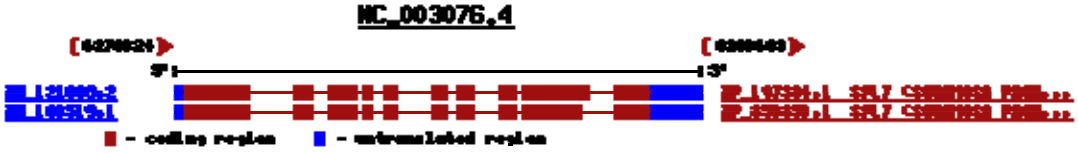


图 2 拟南芥 SBP 转录因子 AtSPL07 基因结构。本图显示拟南芥 SBP 转录因子 AtSPL07 基因结构，可用 TAIR 编码检索 NCBI 基因（Gene）数据库获得。图中显示该基因有两个可变剪接体，由正链转录，转录起始位点为 6276204，终止位点位 6280683，共 4480 个碱基。两个可变剪接体均有 10 个外显子，其中第 9 个外显子长度不同（图中箭头所示）。图中左侧为 RefSeq 数据库中 mRNA 序列编号（NM_121888.1 和 NM_180519.1），右侧为 GenPept 数据库中蛋白质序列代码（NP_197384.1 和 NP_850850.1）。图形上方 NC_003076.4 为核酸序列数据库 GenBank 编号。

对于拟南芥等已经完成全基因组测序，或某些具有包括外显子、内含子、5’和 3’非翻译区（untranslated region, UTR）等全长序列的基因，GenBank 和 EMBL 核酸序列数据库的序列特征表中一般都会给出上述序列片段的位置信息。根据它们的位置信息，可以用图形方式显示基因结构。查看 NCBI 基因数据库（Gene），直接可以用图形方式在浏览器中显示基因结构。例如，以 TAIR 编号 At5G18830 检索 NCBI 基因数据库，可以得到基因 AtSPL07 的详细信息，并用图形显示 2 个可变剪接体信息，包括外显子、内含子、非翻译区等（图 2）。北京大学生物信息中心开发的基因结构显示系统（Gene Structure Display System, GSDS），可以用来同时显示多个基因结构（ 郭安源等，2007）。从拟南芥基因转录因子数据库 DATF 得到 16 个基因 27 种剪接体的全长基因和编码区核苷酸序列，利用 GSDS（http://gsds.cbi.pku.edu.cn），可得到它们的基因结构显示图（图 3）。图中可以看出，拟南芥 SBP 转录因子家族 16 个成员基因结构差异很大，有的只有 2 个外显子，如 AtSPL03A，有的多达 10 个外显子，如 AtSPL01A。除 At5G18330 的两个剪接体 AtSPL07A 和 AtSPL07B 蛋白质编码序列长度不同外，其余 5 个基因的不同剪接体编码相同的蛋白质序列，它们的区别仅在非编码区或非翻译区。

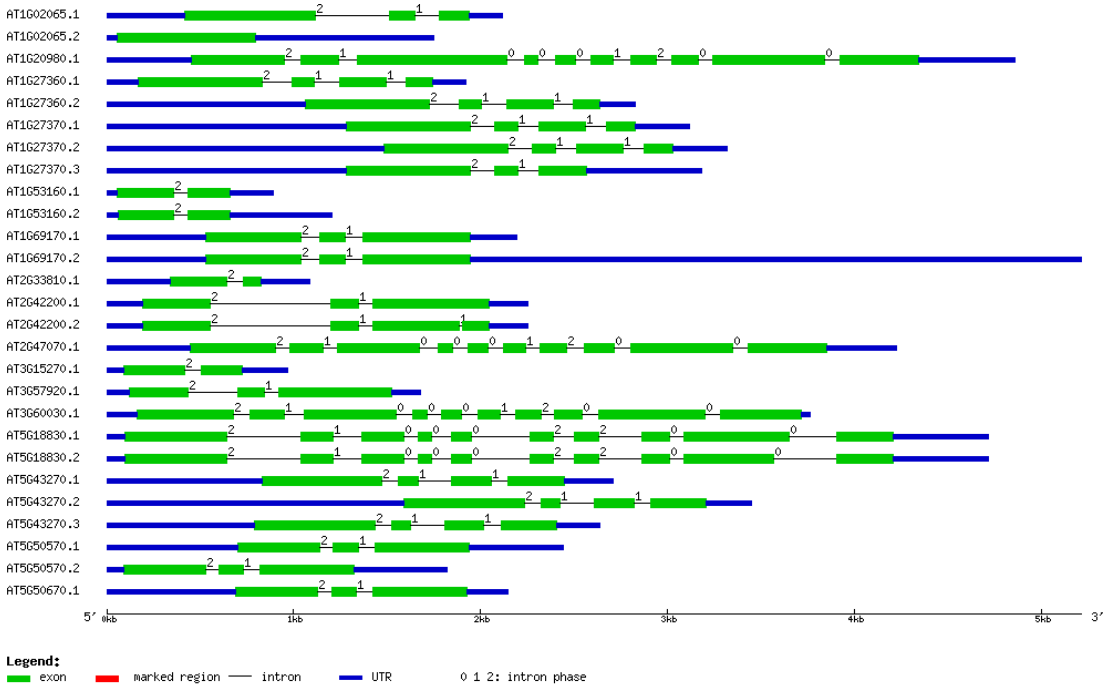


图 3 拟南芥 SBP 转录因子家族 16 个基因 27 个不同剪接体基因结构（拟更新，并用 SVG 格式编辑）

拟南芥 SBP 转录因子 AT5G18330 两个剪接体

提取 At5G18330 的两个剪接体 AtSPL07A 和 AtSPL07B 蛋白质序列，长度分别为 801 和 775 个氨基酸残基（框 1）。

```
AtSPL07A 601 RCQIKRYNRVLNYLIQNNSASILGNVLHNLETLVKKMEPDSLVHCTCDCD    650
             ||||||||||||||||||||||||||||||||||||||||||||||||||||
AtSPL07B 601 RCQIKRYNRVLNYLIQNNSASILGNVLHNLETLVKKMEPDSLVHCTCDCD    650

AtSPL07A 651 VRLLHENMDLASDIHRKHQSPIESKVNPPSSGCCCVSSQKDIPSRILNFN    700
             |||||||||||||||||||||||
AtSPL07B 651 VRLLHENMDLASDIHRKHQSPIES------------------------    674

AtSPL07A 701 KDPEAGLDCKERIQADCSPDSGGKETDPLLNKEVVMNVNDIGDWPRKSCI    750
             ||||||||||||||||||||||||||||||||||||||||||||||||||||
AtSPL07B 675 KDPEAGLDCKERIQADCSPDSGGKETDPLLNKEVVMNVNDIGDWPRKSCI    724
```

Partial output of Sequence alignment between AtSPL07A encoded by At5G18830.1 (GenPept AC: NP_197384.1, 801AA) and AtSPL07B encoded by At5G18830.2 (GenPept AC: NP_850850.1, 775AA).

框 1 – 拟南芥转录因子基因 At5G18330 两个剪接体 AtSPL07A 和 AtSPL07B 蛋白质序列（上）和序列比对结果（下）。用 EMBOSS 程序包中全局比对程序 Needle 对编码蛋白质序列比对，显示 AtSPL07B 在 675-700 位处有一个长度为 26 个氨基酸残基的缺失片段。

用 EMBOSS 程序包中 DotPath 程序进行序列比较，以点阵图方式显示比对结果（图 4）。从图中可以清晰看到，AtSPL07B（Y 轴）C-末端约 700 位处有一段缺失，与图 4 基因结构图中显示结果一致。用 EMBOSS 程序包中全局比对程序 Needle 进行序列比对，可将这段缺失序列定位于 675-700 处，长度为 26 个氨基酸残基。这一缺失片段是否会影响该基因功能，或者是一个假基因，需要通过实验验证。

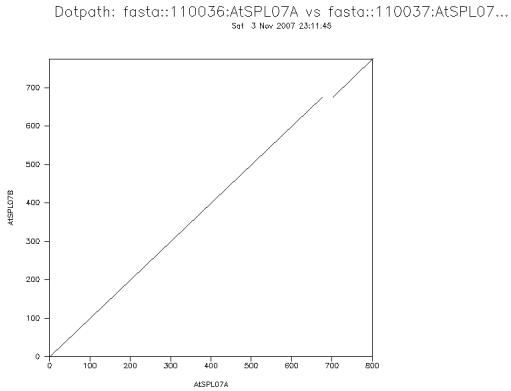


图 4 用点阵图显示拟南芥转录因子基因 At5G18330 两个剪接体 AtSPL07A 和 AtSPL07B 编码蛋白质序列差异

为进一步推测基因At5G18330 两个剪接体AtSPL07A和AtSPL07B可能的功能，将它们编码的蛋白质序列递交到蛋白质功能域识别网站SMART (http://smart.embl-heidelberg.de/)结果表明，这两个蛋白质序列N-端均包含SBP结构域（137-257），C-端有一个长度为 20 个残基的跨膜螺旋片段，分别位于AtSPL07A的 763-782 和AtSPL07B的 737-756，未能找到其它已知功能域。用EMBOSS软件包中alpha螺旋轮绘制程序可得到该跨膜螺旋残基分布（图 5）。图中可以看出，该跨膜螺旋 20 个氨基酸残基中除N-端有 1 个亲水残基苏氨酸（Thr）、C-端有 2 个亲水残基酪氨酸（Tyr）和组氨酸（His）外，核心片段序列仅有一个亲水残基苏氨酸（Thr），其它均为疏水残基，具有跨膜螺旋氨基酸序列的典型特征。

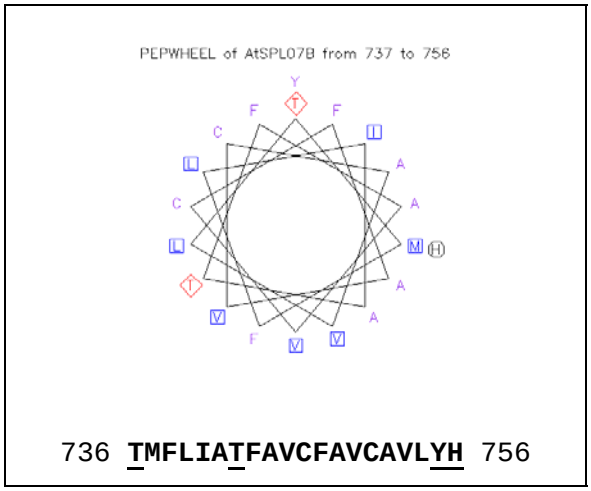


图 5 – 基因 At5G18330 剪接体 AtSPL07B 737-756 位氨基酸残基序列和 alpha 螺旋轮

上述分析结果表明，AtSPL07B 剪接体可能是错误剪接结果，蛋白质功能域识别未给出明确结果，在以后蛋白质序列分析中，我们不予考虑，仅取剪接体 AtSPL07A。此外，其它 14 个 SBP 基因不同剪接体所编码的蛋白质序列均相同，我们仅以其中的第一个剪接体为代表，如 At5G43270 有 3 个剪接体 AtSPL02A、AtSPL02B 和 AtSPL02C，均编码 419 个氨基酸残基。在以下分析拟南芥 15 个转录因子蛋白质序列特征时，我们只取第 1 个剪接体，并略去物种名 At 和区分不同剪接体的字母 A，如 At5G43270 的 3 个剪接体以 AtSPL02A 为代表，名称简化为 SPL02。

拟南芥 SBP 转录因子蛋白质序列分析

我们知道，转录因子与 DNA 序列顺式元件结合，是基因转录调控的关键步骤。除了与 DNA 结合外，蛋白质分子间的相互作用也是实现复杂调控机制的基础。上述 15 个 SBP 转录因子中，SPL04、SPL07 和 SPL12 三个成员的 DNA 结合结构域的溶液构象通过核磁共振获得（Yamasaki, 2004；Yamasaki, 2006）。SPL03、SPL08 和 SPL14 的功能已有文献报道（Birkenbihl; 2005; Unte, 2003; Birkenbihl, 2005; Stone, 2005；Zhang, 2007），其它 12 个成员的功能尚待研究。利用生物信息学工具，对 SBP 转录因子基因家族进行序列分析，以推测其功能多样性，分析该基因家族演化进程。

表 2 列出 15 个拟南芥 SBP 转录因子的基本特征，包括编码序列外显子数目、序列长度、DNA 结合结构域起始位置。从表中可以看出，它们的长度差异很大，为便于比较分析，根据外显子数和序列长度，我们将它们分成 4 组，分别为 G1、G2、G3 和 G4（表 3）。G1 组 3 个成员 SPL03、SPL04 和 SPL05 均由 2 个外显子编码、序列长度在 200 个残基以内，最短的 SPL03 仅有 131 个残基，最长的 SPL05 含 181 个残基，两者相差 50 个残基。G2 组包括 5 个成员，均含 3 个外显子；按序列长度从小到大分别为：SPL08、SPL15、SPL13、SPL09 和 SPL06，最短的 SPL08 含 333 个残基，最长的 SPL06 含 405 个残基，两者相差约 80 个残基。G3 组 3 个成员 SPL11、SPL10 和 SPL02 均含 4 个外显子，序列长度分别为 393、396 和 419 个残基，最短的 SPL11 和最长的 SPL02 仅差 26 个残基。G4 组 4 个成员均由 10 个外显子编码、序列长度均在 800 以上，最短的 SPL07 为 801 个残基，最长 SPL14 由 1035 个残基组成，两者相差 200 多个残基。总的看来，4 组内部序列长度也有一定差异，但总体差异较小，不超过 30%；而组间差异很大，G4 组最长的 SPL14 全长 1035 个残基，约为 G1 组最短的 SPL03 的 8 倍。

表 2 – 拟南芥 15 个 SBP 转录因子

Name	GenPet ID	N	G	L	S	E	References
SPL01	NP_850468.1	10	G4	881	104	182	Cardon, 1999
SPL02	NP_851122.1	4	G3	419	167	245	Cardon, 1999
SPL03	NP_565771.1	2	G2	131	52	130	Cardon, 1999; Birkenbihl; 2005; Wu, 2006
SPL04	NP_175723.1	2	G2	174	52	130	Cardon, 1999; Yamasaki, 2004; Wu, 2006
SPL05	NP_188145.1	2	G2	181	61	139	Cardon, 1999; Wu, 2006
SPL06	NP_177077.3	3	G3	405	122	200	Cardon, 1999
SPL07	NP_197384.1	10	G4	801	136	214	Cardon, 1999; Yamasaki, 2004
SPL08	NP_683267.1	3	G2	333	186	264	Cardon, 1999; Unte, 2003; Birkenbihl, 2005; Zhang, 2007
SPL09	NP_181749.1	3	G2	375	72	150	Cardon, 1999
SPL10	NP_174057.2	4	G3	396	174	252	Cardon, 1999
SPL11	NP_973920.1	4	G3	393	173	251	Cardon, 1999
SPL12	NP_191562.1	10	G4	927	125	203	Cardon, 1999; Yamasaki, 2006
SPL13	NP_851161.1	3	G2	359	99	177	
SPL14	NP_173522.1	10	G4	1035	118	196	Stone, 2005
SPL15	NP_191351.1	3	G2	354	57	135	

N: Exon number; G: Group; L: Length; S: DBD start; E: DBD end

首先，用 ClustalW 程序对上述 15 个拟南芥 SBP 转录因子进行多序列比对。结果表明，所有这些序列均有保守的 DNA 结合结构域，长度约为 80 个残基，大部分位于 N-端。仔细分析表明，由于这 15 个转录因子序列长度相差很大，将所有序列放在一起进行比对，很难得到其它有用信息。

表 3 – 拟南芥 15 个 SBP 转录因子分组

Group	Name	N	Length
G1	SPL03	2	131
G1	SPL04	2	174
G1	SPL05	2	181
G2	SPL08	3	333
G2	SPL15	3	354
G2	SPL13	3	359
G2	SPL09	3	375
G2	SPL06	3	405
G3	SPL11	4	393
G3	SPL10	4	396
G3	SPL02	4	419
G4	SPL07	10	801
G4	SPL01	10	881
G4	SPL12	10	927
G4	SPL14	10	1035

根据表 3 提供的分组信息，将 15 个序列分成 4 组分别进行比对，结果显示，G1 组 3 个序列（SPL03、SPL04 和 SPL05）均由 2 个外显子编码，序列之间相似性很高，尤其是 SPL04 和 SPL05，长度仅相差 3 个残基。用 EMBOSS 软件包中程序 Needle 对 SPL04 和 SPL05 进行全局

比对，两者一致序列达 65.1%，相似序列达 76.9%（框 2）。仔细观察这 3 个序列的氨基酸组成，发现它们具有以下特点：极性氨基酸比非极性氨基酸多、带电氨基酸比不带电氨基酸多、带正电氨基酸比带负电氨基酸多。

```
SPL03 MSMRRSKAEGKRSLR----ELSEEEEEEEETEDEDTFEEEE---ALEKKQKGKATS--SS 51
SPL04 MEGKRSQGQGYMKKK---SYLVEEDMETDTDEEEEVGRDR------VRGSRGSINRGGSL 51
SPL05 MEGQRTQRRGYLKDKATVSNLVEEEMENGMDGEEEDGGDEDKRKKVMERVRGPSTDRVPS 60
       *. :*:: .*  . :     * **: *     :*:   :.       .  :*  .   .

SPL03 GVCQVESCTADMSKAKQYHKRHKVCQFHAKAPHVRISGLHQRFCQQCSRFHALSEFDEAK 111
SPL04 RLCQVDRCTADMKEAKLYHRRHKVCEVHAKASSVFLSGLNQRFCQQCSRFHDLQEFDEAK 111
SPL05 RLCQVDRCTVNLTEAKQYYRRHRVCEVHAKASAATVAGVRQRFCQQCSRFHELPEFDEAK 120
       :***: **.::.:** *::**:**:.****. .  ::*:.********** * ******

SPL03 RSCRRRLAGHNERRRKSTTD---------------------------------------- 131
SPL04 RSCRRRLAGHNERRRKSSGESTYGEGSGRRGINGQVVMQNQERSRVEMTLPMPNSSFKRP 171
SPL05 RSCRRRLAGHNERRRKISGDS-FGEGSGRRGFSGQ-LIQTQERNRVDRKLPMTNSSFKRP 178
      **************** : :

SPL03 ---
SPL04 QIR 174
SPL05 QIR 181
```

Output of Sequence alignment among 3 AtSPL TFs, SPL03 (GenPept AC: NP_565771.1, 131AA), SPL04 (GenPept AC: NP_175723.1, 174AA) and SPL05 (GenPept AC: NP_188145.1, 181AA).

```
SPL04   1 MEGKRSQGQGYMKKK---SYLVEEDMETDTDEEEEVGRD----------RV      38
          |||:|:|.:||:|.|    |.||||:||...|.|||.|.|          ||
SPL05   1 MEGQRTQRRGYLKDKATVSNLVEEEMENGMDGEEEDGGDEDKRKKVMERV      50

SPL04  39 RGSRGSINRGGSLRLCQVDRCTADMKEAKLYHRRHKVCEVHAKASSVFLS      88
          ||.   |.:|..| |||||||||.::.|||.|:|||:|||||||||:..::
SPL05  51 RGP--STDRVPS-RLCQVDRCTVNLTEAKQYYRRHRVCEVHAKASAATVA      97

SPL04  89 GLNQRFCQQCSRFHDLQEFDEAKRSCRRRLAGHNERRRKSSGESTYGEGS    138
          |:.||||||||||:|.|||||||||||||||||||.||:| :||||
SPL05  98 GVRQRFCQQCSRFHELPEFDEAKRSCRRRLAGHNERRRKISGDS-FGEGS    146

SPL04139 GRRGINGQVVMQNQERSRVEMTLPMPNSSFKRPQIR    174
         ||||.:||:: |.|||:||:..|||.||||||||||
SPL05147 GRRGFSGQLI-QTQERNRVDRKLPMTNSSFKRPQIR    181
```

Output of Sequence alignment between SPL04 (GenPept AC: NP_175723.1, 174AA) and SPL05 (GenPept AC: NP_188145.1, 181AA) showing high similarity (Identities: 65.1%, Similarities: 76.9%).

框 2 – 3 个拟南芥转录因子 SPL03、SPL04 和 SPL05 多序列比对（上），SPL04 和 SPL05 双序列比对（下）。

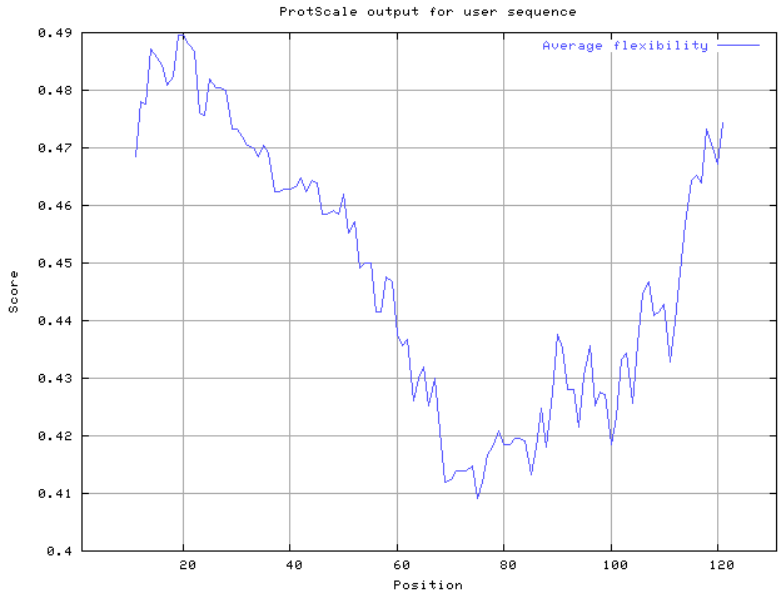
表 4 列出用 EMBOSS 软件包中 Pepstats 程序分析这三个序列氨基酸组成，以及 SwissProt 蛋白质序列数据库中 20 种氨基酸残基的平均含量。可以发现，这三个转录因子精氨酸（Arg）和谷氨酸（Glu）含量普遍高于平均水平，而异亮氨酸（ILe）和亮氨酸（Leu）含量均低于平均水平，均不含色氨酸（Trp）。其中 SPL03 的 131 个残基中极性氨基酸 89 个，占 2/3 以上；带电氨基酸 61 个，占 46.6%，几乎接近一半；带正电氨基酸 35 个，多半分布在 DNA 结合结构域，而 26 个带负电氨基酸中 21 个为谷氨酸，主要分布在序列的 N-端，其序列柔性和溶剂可及性均较高（图 6）。上述氨基酸组成特征，可能有利于与 DNA 结合。

表 4 – G1 组 3 个拟南芥转录因子氨基酸残基分布

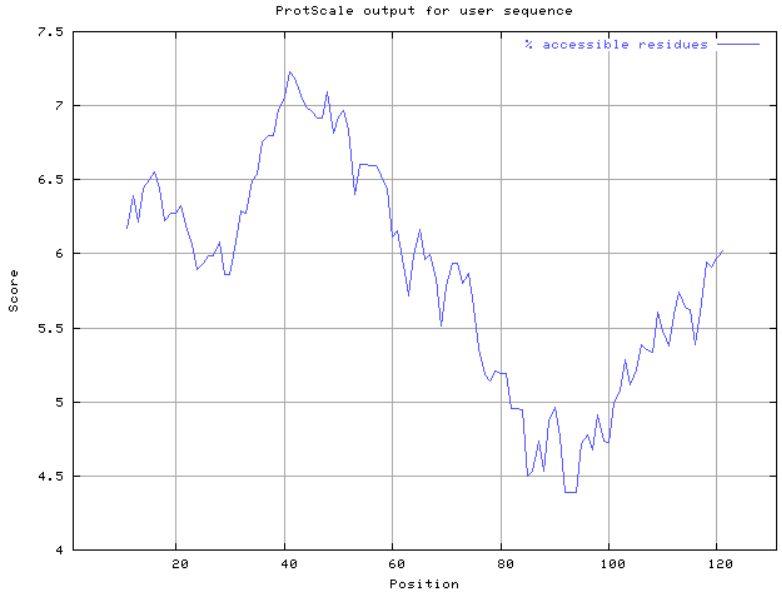
AA	SwissProt*	SPL03		SPL04		SPL05	
	%	No	%	No	%	No	%
Ala	7.9	10	7.6	6	3.4	9	5.0
Cys	1.5	6	4.6	6	3.4	6	3.3
Asp	5.3	5	3.8	8	4.6	10	5.5
Glu	**6.7**	**21**	**16.0**	**17**	**9.8**	**18**	**9.9**
Phe	4.0	5	3.8	5	2.9	6	3.3
Gly	6.9	5	3.8	16	9.2	15	8.3
His	2.3	7	5.3	5	2.9	4	2.2
Ile	5.9	1	0.8	3	1.7	3	1.7
Lys	5.9	13	9.9	11	6.3	11	6.1
Leu	9.7	6	4.6	9	5.2	8	4.4
Met	2.4	3	2.3	7	4.0	5	2.8
Asn	4.1	1	0.8	6	3.4	6	3.3
Pro	4.9	1	0.8	3	1.7	5	2.8
Gln	4.0	7	5.3	11	6.3	11	6.1
Arg	**5.4**	**15**	**11.5**	**25**	**14.4**	**29**	**16.0**
Ser	6.9	14	10.7	17	9.8	12	6.6
Thr	5.4	6	4.6	5	2.9	8	4.4
Val	6.7	4	3.1	10	5.7	12	6.6
Trp	1.1	0	0.0	0	0.0	0	0.0
Tyr	3.0	1	0.8	4	2.3	3	1.7
	99.9	131	100.0	174	100.0	181	100.0

*Percentage of 20 residues in SwissProt (Release 54.0 of 24-Jul-07)

G2 组 5 个序列情况比较特殊，尽管它们均由 3 个外显子编码，序列长度差别也不是很大，但它们之间序列相似性程度较小。利用 EMBOSS 软件包中 Polydot 程序，可以画出它们相似性关系的点阵图（图 7）。从图中可以看出，它们之间有一个保守序列片段，在 SPL08 中靠近 C-端，SPL06 中位于序列中间，而在其它三个序列中则靠近 N-端。实际上，这就是它们共有的 DNA 结合结构域。由于 ClustalW 是基于全局序列相似性的多序列比对程序，若不调节参数，上述 5 个序列多序列比对结果显示，SPL08 的 DNA 结合结构域不能正确匹配。而用基于局部序列相似性的多序列比对程序 POA（http://www.bioinformatics.ucla.edu/poa/）比对结果表示（框 3）。从图中还可看出，5 个序列中 SPL15 和 SPL09 有一定的相似性，图中显示一条不连续的对角线。



Profile of flexibility of the SPL03 sequence generated by EXPASSY online tool Protscale (http://cn.expasy.org/tools/protscale.html) with windows size 21.



Profile of accessibility of the SPL03 sequence generated by EXPASSY online tool Protscale (http://cn.expasy.org/tools/protscale.html) with windows size 21.

图 6 – 拟南芥转录因子 SPL03 序列柔性图谱（上）和溶剂可及性图谱（下）

```
SPL15   TARCQVEGCRDLSNVKAYYSRHKVCCIHSKSSKVIVSGLHQRFCQQCSRFHQLSEFDLE
SPL09   IPRCQVEGCGMDLTNAKGYYSRHRVCGVHSKTPKVTVAGIEQRFCQQCSRFHQLPEFDLE
SPL13   MPICLVDGCDSDFSNCREYHKRHKVCDVHSKTPVVTINGHKQRFCQQCSRFHALEEFDEG
SPL06   NPLCQVYGCSKDLSSSKDYHKRHRVCEAHSKTSVVIVNGLEQRFCQQCSRFHFLSEFDDG
SPL08   QQQHHLTLYGQTNSNNQFLHHHHHHHSLYGSTTTTTPYGASDPIYHPHSSAPPASLFSYD
             :       :. :  : :*:    :.:. .   *  : : :  *       *.

SPL15   KRSCRRRLACHNERRRKPQP-TTALFTSHYSRIAPSLYGNPNAAMIKSVLGDP-TAWSTA
SPL09   KRSCRRRLAGHNERRRKPQPASLSVLASRYGRIAPSLYENGDAGMNGSFLGNQEIGWPSS
SPL13   KRSCRKRLDGHNRRRRKPQP--------EHIGRPAN-FFTGFQGSKLLEFSGGS-HVFPTT
SPL06   KRSCRRRLAGHNERRRKPAFYFLPGKRHKLLRTSQDVVGNKFLENSSLVLPESFPGSLLY
SPL08   QTGPGSGSGSSYNFLIPKTEVDFTSNRIGLNLGGRTYFSAADDDFVSRLYRRSRPGESGM
        : .          .                    .              .

SPL15   R----SVMQRPGP---WQINPVRETHPHMNVLSHGSSSFTTCPEMINNNST------DSS
SPL09   RTLDTRVMRRPVSSPSWQINPMNVFS--QGSVGGGGTSFSS-PEIMDTKLESYKGIGDSN
SPL13   S----------------VLNPSWGNSLVSVAVAANGSSYGQSQSYVVGSSP-------AK
SPL06   R----------------VIDEDDHRTSRLVSFKDEPTCSMFPTNEQNSSRTYESKPAIYS
SPL08   AN---------------SLSTPRCQAEGCNADLSHAKHYHRRHKVCEFHSK------AST
                         :.                .      .                .

SPL15   CALSLLSNSYPIHQQQLQTPTN------TWRPSSGFDSMISFSDKVTMAQPPPISTHQPP
SPL09   CALSLLSNPHQPHDNNNNNNNNNNNNNTWRASSGFGPMT-----VTMAQPP-------P
SPL13   TGIMFPISSSPNSTRSIAKQFPFLQEEESSRTASLCERMT--------------------
SPL06   TEVSSIWDLHETAASRSTRALSLLSAQSQQHLSKFPNTTFS------------------
SPL08   VVAAGLSQRFCQQCSRFHLLSEFDNGKRSCRKRLADHNRR--------------------
               .                        :

SPL15   ISTHQQYLSQTWEVIAGEKSNSHYMSPVSQISEPADFQISNGTTMGGFELY-LHQQVLKQ
SPL09   APSQHQYLNPPWVFKDNDNDMSPVLN-LGRYTEPDNCQISSGTAMGEFELSDHHHQSRRQ
SPL13   -SCIHDSDCALSLLSSSSSSVPHLLQPPLSLS-----QEAVETVFYGSGLFENASAVSDG
SPL06   ITQPNQNLNHSSSTDYHQMEQPLWIDPGKTNSAGSSSCKGKGTSTVDLLQLSSHLQRIEQ
SPL08   RRKCHQSASATQDTGTGKTTPKSPNDSGVKAS-----SSPSSNAPPTISLECFRQRQFQT
            ::          .       .     :          .
```

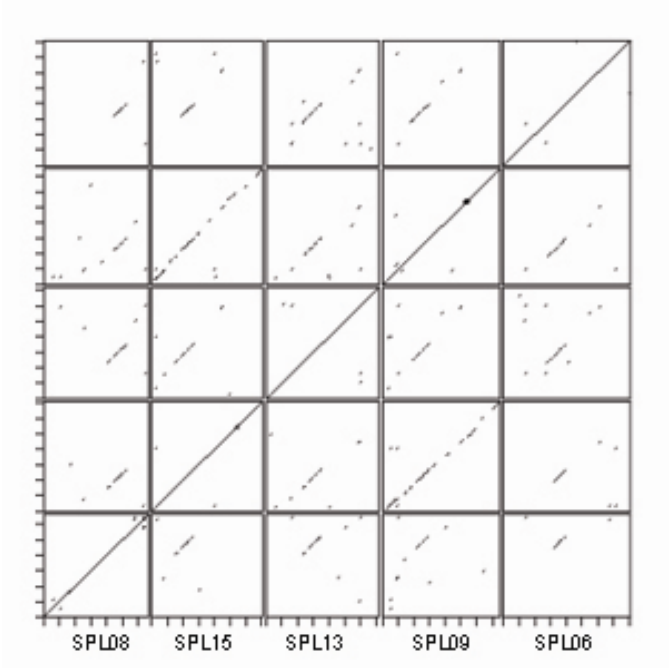
Partial output of ClustalW multiple sequence alignment for 5 SPL TFs showing mismatched of the SPL08 with other 4 sequences in the conserved sequences of the DNA binding domain.

```
SPL08   MAN-SLSTPRCQAEGCNADLSHAKHYHRRHKVCEFHSKASTVVAAGLSQ
SPL15   VRK-SSTTARCQVEGCRMDLSNVKAYYSRHKVCCIHSKSSKVIVSGLHQ
SPL13   V-G-TNQMPICLVDGCDSDFSNCREYHKRHKVCDVHSKTPVVTINGHKQ
SPL09   S-GQSGQIPRCQVEGCGMDLTNAKGYYSRHRVCGVHSKTPKVTVAGIEQ
SPL06   L---CSQNPLCQVYGCSKDLSSSKDYHKRHRVCEAHSKTSVVIVNGLEQ

SPL08   FCQQCSRFHLLSEFDNGKRSCRKRLADHNRRRRKCHQSAS-ATQDTGTG
SPL15   FCQQCSRFHQLSEFDLEKRSCRRRLACHNERRRK-PQPTT-ALFTSHYS
SPL13   FCQQCSRFHALEEFDEGKRSCRKRLDGHNRRRRK-PQP-----------
SPL09   FCQQCSRFHQLPEFDLEKRSCRRRLAGHNERRRK-PQPASLSVLASRYG
SPL06   FCQQCSRFHFLSEFDDGKRSCRRRLAGHNERRRK-P-AFY-FLPGKRHK
```

Partial output of POA multiple sequence alignment for 5 SPL TFs showing the conserved sequences of the DNA binding domain among 5 sequences

框3 — G2组5个拟南芥转录因子ClustalW比对（上）和POA比对结果（下）。



Output of Polydot for the 5 AtSPL TFs showing the conserved fragment of the DNA binding domain of among all 5 sequences with low sequence similarity. The second TF (SPL15) and the forth TF (SPL09) has a higher similarity indicated by a diagonal dot line. A pattern of low sequence complexity in SPL09 can be seen with a black area along the solid diagonal line. Word size 4 was chosen rather than 6 as the default for better sensitivity.

图 7 – 用点阵图显示 G2 组 5 个 SPB 转录因子序列相似性

G3 组 3 个序列均由 4 个外显子编码，长度分别为 393（SPL11）、396（SPL10）和 419（SPL02）个残基。三者序列相似性极高，尤其是 SPL10 和 SPL11，序列长度仅差 3 个残基，一致序列占 75.4%，相似序列占 82.1%，空位仅占 3.7%，是所有 15 个转录因子中最为相似的两个成员。

G4 组 4 个序列（SPL07、SPL01、SPL12、SPL14）均由 10 个外显子编码，序列长度均在 800 个残基以上，其中最短的 SPL07 为 801 个残基，约为其它三组中最长的 SPL02 的 2 倍；而最长的 SPL14 含 1035 个残基，是最短的 SPL03 的 8 倍。SPL01 和 SPL12 序列长度分别为 881 和 927 个残基，序列比对结果表明，这两个序列具有较高的序列相似性，两者之间一致序列占 69.3%，相似序列占 78.9%。

表 5 – 拟南芥 4 对 SBP 转录因子双序列比对结果

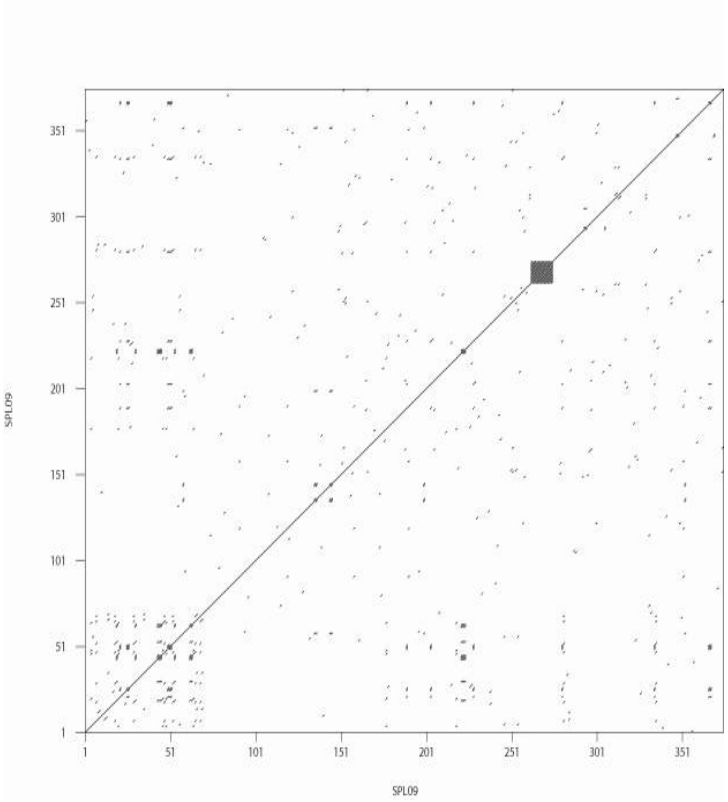
Group	Name	Length	Identities (%)	Similarities (%)	Gaps (%)
G1	SPL04/SPL05	174/181	65.1	76.9	9.1
G2	SPL15/SPL09	354/375	52.5	60.5	17.8
G3	SPL11/SPL10	393/396	75.4	82.1	3.7
G4	SPL01/SPL12	881/927	69.3	78.9	9.1

综合上述 4 组分析结果，发现每组中均有一对序列相似性较高的序列（表 5）。为什么会有这种现象呢？是偶然巧合，或是有内在的生物学背景？这是一个值得思考的问题。研究表明，拟南芥基因组至少经历了两次全基因组水平重复。在之后的长期演化过程中，由于整条或部分染色体片段丢失，以及结构域重排、转座等种种原因，大部分重复基因丢失，或演化成其它功能不同的新基因。而拟南芥基因组也演化成了今天我们看到的二倍体。然而，这种基因组水平的重复，依然可以通过比较基因组研究发现（Wang et al. 2005; Wang et al., 2007）。上述拟南芥 4 对基因中，SPL04/AtSPL5、SPL9/SPL15 和 SPL01/SPL12 很可能就是通过基因组水平重复而产生。德国慕尼黑蛋白质序列信息中心（Munich Information Center for Protein Sequence, MIPS）拟南芥基因组数据库（MIPS Arabidopsis Thaliana Dtabase，MATDB，http://mips.gsf.de/proj/thal/db/）将 SPL01 和 SPL12 注释为复制基因。据报道，SPL04 /SPL05 和 SPL9/SPL15 可能于芸苔属（Brassica）植物分化即拟南芥物种形成之前就已经存在（Blanc & Wolfe, 2004; Bowers et al., 2003）。而 SPL10/SPL11 具有同样基因结构，序列高度相似（82.1%），在染色体上彼此相邻。据分析，它们可能是通过串联重复机制产生。

复制基因经过演化，可能会产生不同功能，也可能功能相同，但表达方式不同，也就是说，复制基因中的不同成员尽管编码区序列相似性较高，所编码的蛋白质具有相同功能，但由于非编码区调控序列的不同，可能会在不同组织、不同发育阶段或不同环境条件下表达。利用瑞士苏黎世基因表达谱分析网站（https://www.genevestigator.ethz.ch/）提供的基因相关性分析工具（Gene Correlator）对上述拟南芥复制基因表达差异分析结果表明，2 对复制基因（SPL10/SPL11 和 SPL09/SPL15）既有共同表达、也有不同表达的特征，可能意味着正在不断分化。而 SPL12 的表达覆盖了 SPL1 的表达，也就是说 SPL01 表达时，SPL12 一定表达，而 SPL12 表达时，SPL01 不一定表达。同样，SPL04 的表达覆盖了 SPL05 的表达。这种表达模式是否意味着其中一个基因的表达已经不占主导地位，在今后的演化过程中，可能会逐步退居到无足轻重的地位，或者是生物界常见的冗余现象，值得进一步探讨。

基于上述分析，我们对拟南芥 15 个转录因子的序列特征有了一个初步了解，下面，我们通过蛋白质功能域预测、保守结构域识别等工具，进一步分析它们的序列特征和可能的功能。

利用德国海德堡欧洲分子生物学实验室（European Molecular Biology Laboratory, EMBL）蛋白组功能域预测网站（Simple Modular Architecture Research Tool, SMART, http://smart.embl-heidelberg.de/），对 15 个转录因子序列可能的功能域进行预测，结果表明，除 SBP 结构域外，G4 组 4 个成员 C-端均有跨膜螺旋，SPL01、SPL12 和 SPL14 各有 2 个长度为 29-31 个残基的 Ankrin 结构域，可能与蛋白质相互作用有关。预测结果表明，大部分转录因子中均有低复杂度重复序列片段，有的序列中有多个这样的片段。用点阵图方法可以清晰显示这样的重复序列片段（图 8）。这种简单的重复序列片段是否具有生物学意义，是一个值得研究的问题。荷兰 Wageningen 大学构建了蛋白质重复序列数据库，为研究蛋白质序列重复片段提供了基础。我们知道，蛋白质序列中各种氨基酸残基并非随机分布，由若干个连续残基构成的序列片段也非随机分布。近年来，郝柏林领导的研究组利用组分矢量方法，对 200 多个细菌基因组中蛋白质组分进行分析，用于比较不同基因组的系统发育，取得了很有意义的成果（Qi et al., 2005），所构建的系统发育树可以通过北京大学生物信息中心 CVTree 网站浏览和下载（http://cvtree.cbi.pku.edu.cn）。

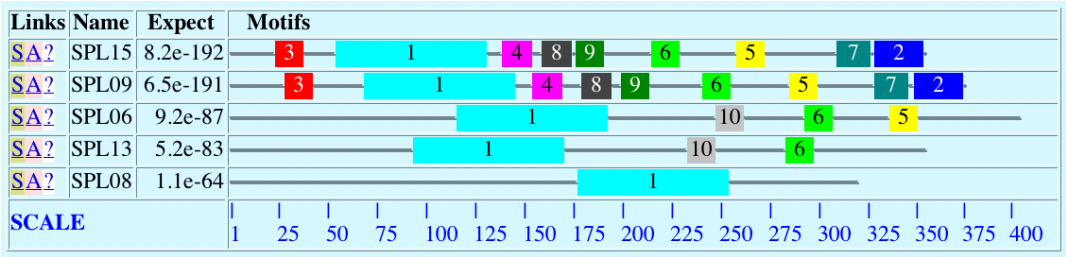


Dot matrix plot for SPL09 showing low complexity regions using the Dottup program in EMBOSS with window size 1. The bottom-left corner box indicates the N-terminal imperfect repeat fragment "MEMGSNSGPGHGPGQAESGGSSTESSSFSGGLMFGQKIYFEDGGGGSGSSSSGGRSNRRVRGGGSGQS GQ", the black box shows the single amino acid repeat "NNNNNNNNNNNNNN" in position 262-275.

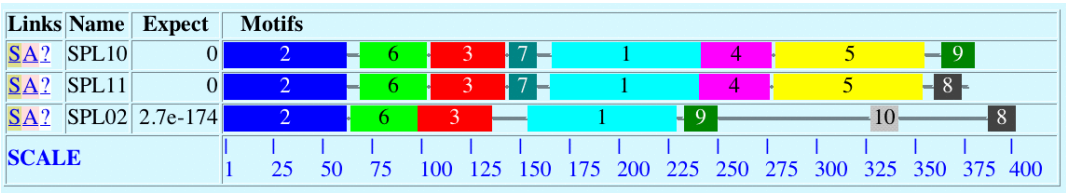
图 8 – 用点阵图显示 SPL09 中低复杂度序列片段

利用序列保守结构域识别系统网站（Multiple EM for Motif Elicitation, MEME, http://meme.sdsc.edu/），对上述拟南芥 SBP 转录因子保守结构域进行识别。首先，将 15 个 SBP 转录因子同时提交，结果表明，它们都具有长度为 79 个残基的 SBP 结构域。此外，G2、G3 和 G4 组中 7 个成员有一个共同的保守区域"ALSLLS"，均处于 C-端，即 G2 组 SPL06、SPL09、SPL13 和 SPL15 的 3 个外显子，G3 组 SPL02、SPL10 和 SPL11 的第 4 个外显子。据文献报道，这一保守位点的 mRNA 序列可能是 microRNA156 的调控靶序列（Rhoades et al., 2002）。对 G1 组 3 个转录因子 SPL03、SPL04 和 SPL05 的 CDS 序列分析表明，这一 microRNA 调控位点位于它们的 3' 非翻译区，其核苷酸序列为"GTGCTCTCTCTCTTCTGTCA"（Guo et al. unpublished data）。

为提高识别灵敏度，对 4 组长度不同的 SBP 转录因子分别进行 MEME 搜索，可以得到每组序列之间多个保守结构域，并用图形方式显示组内各成员之间保守结构域（图 9），从另一个侧面反映它们之间序列相似性关系。与 Clustalw 等多序列比对相比，结果更加可靠，输出更加直观清晰。



MEME search result for 5 AtSPLs of Group 2. Maximum width: 80, minimum width: 10, maximum number of motifs: 10.



MEME search result for 3 AtSPLs of Group 3. Maximum width: 80, minimum width: 10, maximum number of motifs: 10.

图 9 – 多重期望序列模体识别系统分析拟南芥 SBP 转录因子蛋白质序列保守结构域图形输出

拟南芥 SBP 转录因子 DNA 结合结构域序列和结构分析

多序列比对结果和序列图标分析表明（框 4），SBP 结构域可分为三个区域：N-端保守区（3-30）、中间非保守区（31-41）和 C-端保守区（42-76）（图 10）。C-端序列保守程度高于 N-端，除形成锌指结构序列模体残基半胱氨酸和组氨酸外，另有一个保守序列模式“KRSCR[RK][RK]Lx2HNxRR[RK][KR]”，主要由碱性氨基酸残基组成。据文献报道，SBP 转录因子在胞浆内合成后进入细胞核，需要核定位信号介导（Birkenbihl, et al. 2005），而这一保守序列片段就是介导该转录因子以主动运输方式通过核膜进入细胞核的识别信号。Birkenbihhl 等利用基因工程定位突变方法证明，该序列片段第 3 位保守残基丝氨酸十分关键，将它突变成门冬酰胺（Asp）后进入细胞核的能力降低。

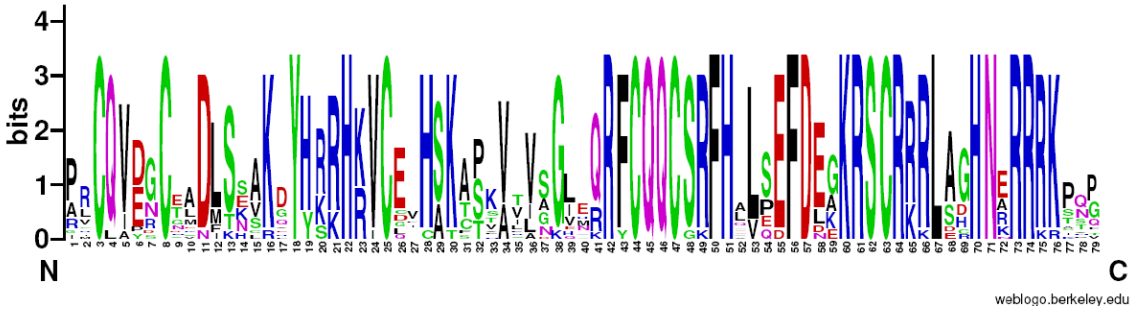


图 10 – 15 个拟南芥转录因子 DNA 结合结构域序列图标

框 4 – 拟南芥 SBP 转录因子家族 DNA 结合结构域序列和比对结果

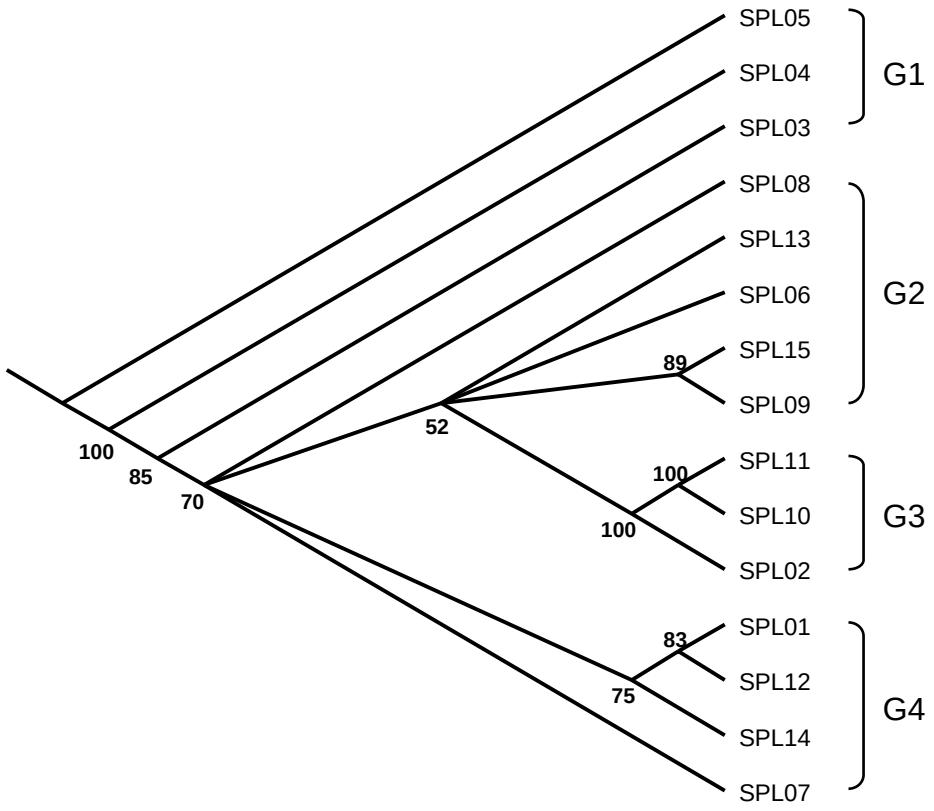
```
>AtSPL01A ( start= 104 )
AVCQVENCEADLSKVKDYHRRHKVCEMHSKATSATVGGILQRFCQQCSRFHLLQEFDEGKRSCRRRLAGHNKRRRKTNP
>AtSPL09A ( start= 72 )
PRCQVEGCGMDLTNAKGYYSRHRVCGVHSKTPKVTVAGIEQRFCQQCSRFHQLPEFDLEKRSCRRRLAGHNERRRKPQP
>AtSPL12A ( start= 125 )
ICCQVDNCGADLSKVKDYHRRHKVCEIHSKATTALVGGIMQRFCQQCSRFHVLEEFDEGKRSCRRRLAGHNKRRRKANP
>AtSPL14A ( start= 118 )
PMCQVDNCTEDLSHAKDYHRRHKVCEVHSKATKALVGKQMQRFCQQCSRFHLLSEFDEGKRSCRRRLAGHNRRRRKTTQ
>AtSPL02A ( start= 167 )
PHCQVEGCNLDLSSAKDYHRKHRICENHSKFPKVVVSGVERRFCQQCSRFHCLSEFDEKKRSCRRRLSDHNARRRKPNP
>AtSPL11A ( start= 173 )
PRCQIDGCELDLSSAKGYHRKHKVCEKHSKCPKVSVSGLERRFCQQCSRFHAVSEFDEKKRSCRKRLSHHNARRRKPQG
>AtSPL10A ( start= 174 )
PRCQIDGCELDLSSSKDYHRKHRVCETHSKCPKVVVSGLERRFCQQCSRFHAVSEFDEKKRSCRKRLSHHNARRRKPQG
>AtSPL06A ( start= 122 )
PLCQVYGCSKDLSSSKDYHKRHRVCEAHSKTSVVIVNGLEQRFCQQCSRFHFLSEFDDGKRSCRRRLAGHNERRRKPAF
>AtSPL08A ( start= 186 )
PRCQAEGCNADLSHAKHYHRRHKVCEFHSKASTVVAAGLSQRFCQQCSRFHLLSEFDNGKRSCRKRLADHNRRRRKCHQ
>AtSPL03A ( start= 52 )
GVCQVESCTADMSKAKQYHKRHKVCQFHAKAPHVRISGLHQRFCQQCSRFHALSEFDEAKRSCRRRLAGHNERRRKSTT
>AtSPL15A ( start= 57 )
ARCQVEGCRMDLSNVKAYYSRHKVCCIHSKSSKVIVSGLHQRFCQQCSRFHQLSEFDLEKRSCRRRLACHNERRRKPQP
>AtSPL04A ( start= 52 )
RLCQVDRCTADMKEAKLYHRRHKVCEVHAKASSVFLSGLNQRFCQQCSRFHDLQEFDEAKRSCRRRLAGHNERRRKSSG
>AtSPL05A ( start= 61 )
RLCQVDRCTVNLTEAKQYYRRHRVCEVHAKASAATVAGVRQRFCQQCSRFHELPEFDEAKRSCRRRLAGHNERRRKISG
>AtSPL13A ( start= 99 )
PICLVDGCDSDFSNCREYHKRHKVCDVHSKTPVVTINGHKQRFCQQCSRFHALEEFDEGKRSCRKRLDGHNRRRRKPQP
>AtSPL07A ( start= 136 )
ARCQVPDCEADISELKGYHKRHRVCLRCATASFVVLDGENKRYCQQCGKFHLLPDFDEGKRSCRRKLERHNNRRKRKPV


          10        20        30        40        50        60        70        80
         ----:----|----:----|----:----|----:----|----:----|----:----|----:----|----:----|
Consensus PxCQVDGCxxDLSNAKxYHRRHKVCEVHSKAxxVVVSGLxQRFCQQCSRFHxLSEFDExKRSCRRRLAGHNERRRKxxx
SPL04     RL....R.TA.mKE..L..........a..SS.Fl...N..........D.Q....A.................SSG
SPL05     RL....R.TVn.tE..Q.y...r.....a..SAAT.a.vR..........E.P....A.................ISG
SPL03     GV...eS.TA.m.K..Q..k.....qF.a..PH.Ri...H..........A......A.................STT
SPL01     AV...eN.EA...KV.D.........m....TSAT.G.iL..........L.Q....G............k....TNP
SPL12     IC....N.GA...KV.D.........i....TTAl.G.iM..........V.E....G............k....ANP
SPL14     .M....N.TE...h..D..............TKAl.GKQM..........L......G............R....TTQ
SPL08     .R..Ae..NA...h..H.........F....ST..Aa..S..........L.....NG.....k...D..R....CHQ
SPL06     .L...Y..SK...ss.D..k..r...A...TSV.i.n..E..........F.....dG.................PAF
SPL13     .I.L....DS.F..CrE..k.....d....TPV.Tin.HK..........A.E....G.....k..D...R....PQP
SPL09     .R...e..GM..t...G.yS..r..G....TPK.T.a.iE..........Q.P...LE.................PQP
SPL15     AR...e..RM....V.A.yS.....Ci...sSK.i....H..........Q.....LE.........C.......PQP
SPL11     .R..i...EL...s..G...k.....K...CPK.S....Er.........Av.....K.....k..sH..A....PQG
SPL10     .R..i...EL...ss.D...k.r...T...CPK......Er.........Av.....K.....k..sH..A....PQG
SPL02     .H...e..NL...s..D...k.ri..N...FPK.....vEr.........C......K........sD..A....PNP
SPL07     AR...PD.EA.i.EL.G..k..r..LRCaT.SF..lD.ENk.y....Gk..L.Pd...G......k.ER..N..krKPV
```

上述分析结果表明，15 个拟南芥转录因子可以分为 4 组，每组内部各成员序列长度差别较小，序列相似性较高。为进一步搞清它们之间的关系，我们以 SBP 结构域构建系统发育树，可以得到以下几点结论（图 11）。

- 4 对重复基因分别聚在同一分支上，并有较高支持率（SPL04/SPL05: 100，SPL10/SPL11: 100，SPL09/SPL15: 89，SPL01/SPL12: 83）；
- G1 组 3 个成员 SPL03、SPL04 和 SPL05 以较高支持率聚在一起，G2 组 5 个成员中，重复基因对 SPL09/SPL15 能够较好地聚在一起，G3 组 3 个成员 SPL02、SPL10 和 SPL11 以较高支持率聚在一起，G4 组 4 个成员中，有 3 个 SPL01、SPL12 和 SPL14 以较高支持率聚在一起；

- G2 组 4 个成员（SPL06、SPL13、SPL09 和 SPL15）与 G3 组 3 个成员以 52%的支持率聚在一起；
- G4 组 SPL07、G2 组 SPL08 和 G2 组其它 4 个成员、G3 组全部 3 个成员以 70%支持率聚在一起。



Phylogenetic tree of the DNA binding domain of 15 AtSPLs using the neighbor joining method of Phylip. The tree was drawn using MEGA. Bootstrap values greater than 50 are shown along the tree branch. Labels of four different groups (G1-G4) were added manually.

图 11 – 15 个拟南芥转录因子 SBP 结构域系统发育树

上述结果表明，拟南芥 SBP 基因家族 15 个转录因子的 SBP 结构域序列之间的异同，与基因结构、序列长度有一定关系，但并不完全一一对应。其中 G2 组 5 个成员之间 DNA 结构域序列差别较大，这一点与它们全长蛋白质多序列比对差异较大结果一致，特别是 SPL08，与组内其它成员在系统发育树上不聚在一起。同样，SPL07 与组内其它 3 个成员差别较大。多序列比对结果显示，和其它 14 个序列差异较大（框 4），79 个位点中共有 40 个差异位点，占 50.6%，明显高于平均值 30.97%（表 6）。

表 6 – 拟南芥 15 个 SBP 转录因子 DNA 结合多序列比对结果统计

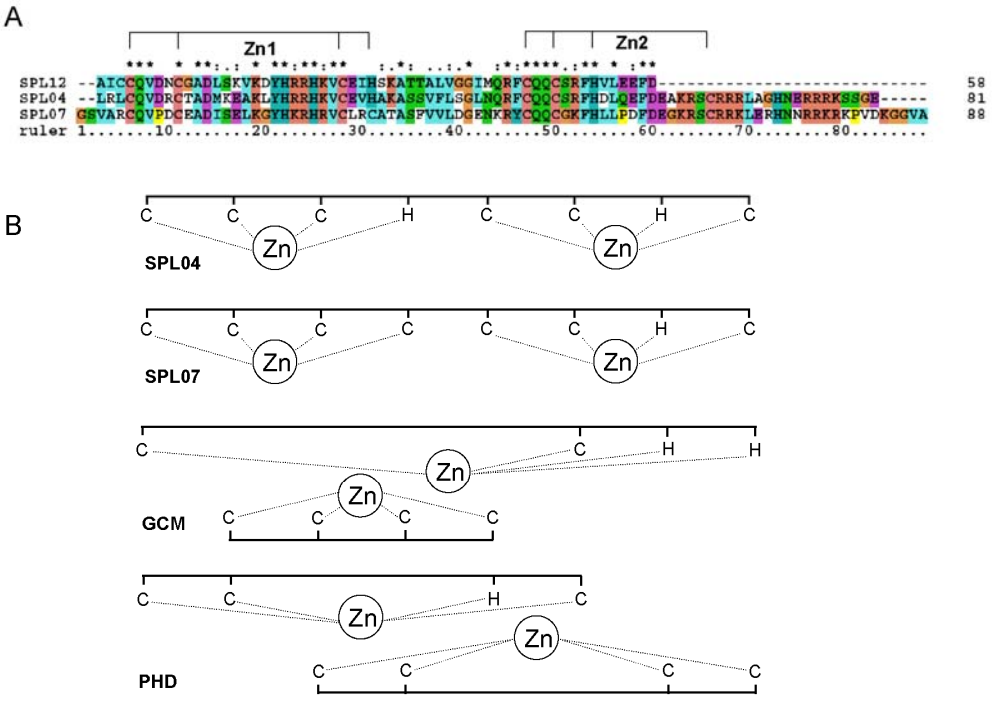
Name	Group	Identities	Similarities	Differences	Changes
SPL04	G1	58	3	18	26.58
SPL05	G1	54	7	18	31.65
SPL03	G1	56	6	17	29.11
SPL01	G4	55	4	20	30.38
SPL12	G4	56	4	19	29.11
SPL14	G4	59	2	18	25.32
SPL08	G2	57	4	18	27.85
SPL06	G3	57	7	15	27.85
SPL13	G2	52	6	21	34.18
SPL09	G2	55	6	18	30.38
SPL15	G2	56	5	18	29.11
SPL11	G2	55	7	17	30.38
SPL10	G3	54	9	16	31.65
SPL02	G3	55	8	16	30.38
SPL07	G4	39	12	28	50.63
Average		55	6	19	30.97

表 7 – 拟南芥 3 个 SBP 转录因子 DNA 结合结构域核磁共振溶液构象参数

Name	PDB	Position	L	Zinc motif	Alpha	Beta	Reference
SPL04	1UL4	51-131	81	C54C59C76H79	K67-H73	V85-L87	Yamasaki, 2004
				C95C98H102C114	C76-K81	L90-F94	
						F101-D103	
SPL07	1UL5	135-220	88	C138C143C160C163	H154-R158	Q139-V140	Yamasaki, 2004
				C179C182H186C198	C160-S167	C143-E144	
					P189-F191	F168-L171	
						E174-Y178	
						F185-L187	
SPL12	1WJ0	124-181	58	C127C132C149H152	C149-K154	I125-C126	Yamasaki, 2006
						D135-L136	
						A158-V160	
						I163-Q165	

我们知道，锌指结构是转录因子 DNA 结合结构域的典型特征，锌离子与氨基酸残基侧链形成四配位键。最早发现的锌指序列模体为 C2H2，即 2 个半胱氨酸 Cys 和 2 个组氨酸 His。典型的锌指序列模体还有 C3H、C2HC 和 C4 等。进一步分析 15 个转录因子 SBP 结构域序列特征，发现其 N-端保守区含 C-x5-C-x13-H-x2-C-x2-[CH]序列模体，即 C2HC2 或 C2HCH；C-端保守区含 Cx2C-x3-H-x11-C-x6-H 序列模体，即 C2HCH。两者各有 5 个半胱氨酸或组氨酸，共计 10 个残基。2004 年，日本筑波国家高新科技产业研究院（National Institute of Advanced Industrial Science and Technology, AIST）Yamasaki 等用核磁共振方法测定了 SPL04 和 SPL07 两个转录因子 SBP 结构域的溶液构象（Yamasaki et al., 2004）；2006 年，又测定了 SPL12 SBP 结构域 N-端 的溶液构象（Yamasaki et al., 2004）。结果表明，拟南芥转录因子 SBP 结构域含两个锌指结构（表 7），N-端序列模体为 C3H（SPL04 和 SPL12）或 C4（SPL07），C-端序列

模体均为 C2HC（图 12）。由于 SPL12 只测定了 N-端部分结构，第 2 个锌指序列模体未在图中显示。与以往已经发现的双锌指结构转录因子不同，SBP 结构域两个锌指结构互相独立，在序列上互不交叉。



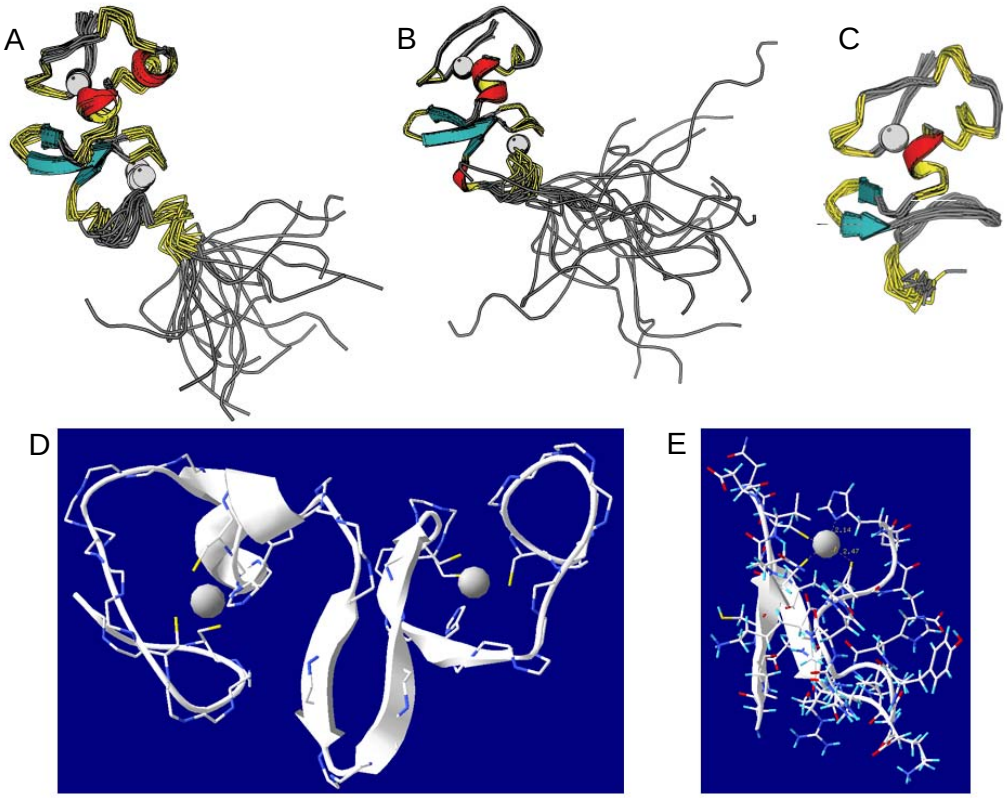
Zinc finger motifs in SBP domains and other transcription factors. A: Sequence alignment of a fragment of the DNA binding domain of SPL12 (PDB: 1WJ0), and two SBP domains SPL04 (PDB: 1UL4), SPL07 (PDB: 1UL5). The three dimensional structures were solved by NMR (Yamasaki et al., 2004; Yamasaki et al., 2006). The two zinc finger motifs are connected by lines at the top. B: Three different patterns of zinc finger connection. SBP: Arabidopsis SPL04 (PDB: 1UL4); GCM: Glia cell missing transcription factor (PDB: 1ODH); PHD: Plant homeo-domain (PDB: 1WEE).

图 12 – SBP 结构域锌指结构序列模体

这 3 个转录因子 DNA 结合结构域的溶液构象具有以下几个特征：

- 整体结构以无规卷曲为主，较少 alpha-螺旋和 beta-折叠等二级结构单元（图 13）；
- N-端锌指结构部分构象变化不大，而 C-端碱性区域构象变化很大（图 13A, B）
- 两个锌指结构各自形成独立的结构单元，通过一对 beta-折叠片连接（图 13E）
- 分子表面多为带电或亲水残基（图 13F）。

上述结构特征可能与其和 DNA 结合的特性有关。



Three-dimensional structure of SBP domains (Yamasiki, et al., 2004; Yamasaki et al., 2006). A: NMR structure of the DNA binding domain of SPL04 (PDB: 1UL4); B: NMR structure of the DNA binding domain of SPL07 (PDB: 1UL4); C: NMR structure of the N-terminal fragment of the DNA biding domain of SPL12 (PDB: 1WJ0); D: the N-terminal part of 1UL4 showing two zinc finger motifs; E: the first zinc finger region of 1WJ0 showing all side chains and two beta strands. Pictures A, B and C were retrieved from the Jena Library of Biological Molecules (http://www.fli-leibniz.de/IMAGE.html), Pictures E and F were produced with Swiss-PDB Viewer (http://swissmodel.expasy.org/spdbv/).

图 13 – 3 个 SBP 结构域核磁共振溶液构象

本章小结

本章已拟南芥 SBP 转录因子家族转录因子为例，介绍如何利用生物信息学数据库和软件工具，对 SBP 转录因子家族基因结构、蛋白质序列、蛋白质结构等进行分析，从中得到该基因家族不同成员之间的相互关系和演化历程，推断它们可能得生物学功能。必须说明，对植物特异转录因子 SBP 家族研究和全面的生物信息学分析，应参考有关文献（Cardon et al., 1999; Riese et al. 2007; Yang et al, 2007; Guo et al. unpublished data），本章主要分析拟南芥中 16 个 SBP 基因和它们的编码蛋白质，目的在于是读者对相关生物信息数据库和分析工具有一个比较系统的了解，并用于分析自己研究中遇到的与此类似的实际问题。

习题

1. 分析拟南芥基因 At1G02065 两种剪接体 AtSPL08A 和 AtSPL08B 基因结构和序列差异。
2. 找出 G1 组三个成员 CDS 序列中 microRNA 调控位点。
3. 找出拟南芥 SPL03 在其它物种中的直系同源基因，构建系统发育树。
4. 分析 1UL4 和 1UL5 结构异同。

References

1. Birkenbihl RP, Jach G, Saedler H, Huijser P. Functional dissection of the plant-specific SBP-domain: overlap of the DNA-binding and nuclear localization domains. J Mol Biol. 2005 Sep 23;352(3):585-96. [PMID: 16095614]
2. Cardon G, Hohmann S, Klein J, Nettesheim K, Saedler H, Huijser P. Molecular characterisation of the Arabidopsis SBP-box genes. Gene. 1999 Sep 3;237(1):91-104. [PMID: 10524240]
3. Cardon G, Hohmann S, Nettesheim K, Saedler H, Huijser P. Functional analysis of the *Arabidopsis thaliana* SBP-box gene SPL3: a novel gene involved in the floral transition. Plant J. 1997 Aug;12(2):367-77. [PMID: 9301089]
4. Gandikota M, Birkenbihl RP, Hohmann S, Cardon GH, Saedler H, Huijser P. The miRNA156/157 recognition element in the 3' UTR of the Arabidopsis SBP box gene SPL3 prevents early flowering by translational inhibition in seedlings. Plant J. 2007 Feb;49(4):683-93.[PMID: 17217458]
5. Gao G, Zhong Y, Guo A, Zhu Q, Tang W, Zheng W, Gu X, Wei L, Luo J. DRTF: a database of rice transcription factors.Bioinformatics. 2006 May 15;22(10):1286-7. Epub 2006 Mar 21.[PMID: 16551659]
6. Gong, W., Y.P. Shen, L.G. Ma, Y. Pan, Y.L. Du, D.H. Wang, J.Y. Yang, L.D. Hu, X.F. Liu, C.X. Dong, L. Ma, Y.H. Chen, X.Y. Yang, Y. Gao, D. Zhu, X. Tan, J.Y. Mu, D.B. Zhang, Y.L. Liu, S.P. Dinesh-Kumar, Y. Li, X.P. Wang, H.Y. Gu, L.J. Qu, S.N. Bai, Y.T. Lu, J.Y. Li, J.D. Zhao, J. Zuo, H. Huang, X.W. Deng, and Y.X. Zhu. Genome-wide ORFeome cloning and analysis of Arabidopsis transcription factor genes. Plant Physiol. 2004. **135:** 773-782.
7. Guo AY, Chen X, Gao G, Zhang H, Zhu QH, Liu XC, Zhong YF, Gu X, He K, Luo J. PlantTFDB: a comprehensive plant transcription factor database. Nucleic Acids Res. 2007 Oct 12; [PMID: 17933783]
8. Guo AY, Zhu QH, Chen X, Luo JC. [GSDS: a gene structure display server] Yi Chuan. 2007 Aug;29(8):1023-6. Chinese.[PMID: 17681935] （郭安源, 朱其慧, 陈新, 罗静初 GSDS：基因结构显示系统. 遗传）
9. Huijser P, Klein J, Lönnig WE, Meijer H, Saedler H, Sommer H. Bracteomania, an inflorescence anomaly, is caused by the loss of function of the MADS-box gene squamosa in Antirrhinum majus.EMBO J. 1992 Apr;11(4):1239-49.[PMID: 1563342]
10. Klein J, Saedler H, Huijser P. A new family of DNA binding proteins includes putative transcriptional regulatorsof the Antirrhinum majus floral meristem identity gene SQUAMOSA. Mol Gen Genet. 1996 Jan 15;250(1):7-16.[PMID: 8569690]
11. Kropat J, Tottey S, Birkenbihl RP, Depege N, Huijser P, Merchant S. A regulator of nutritional copper signaling in Chlamydomonas is an SBP domain protein that recognizes the GTAC core of copper response element. Proc Natl Acad Sci U S A. 2005 Dec 20;102(51):18730-5. [PMID: 16352720]

12. Lee J, Park JJ, Kim SL, Yim J, An G. Mutations in the rice liguleless gene result in a complete loss of the auricle, ligule, and laminar joint. Plant Mol Biol. 2007 Nov;65(4):487-99. Epub 2007 Jun 27. [PMID: 17594063]
13. Manning K, Tor M, Poole M, Hong Y, Thompson AJ, King GJ, Giovannoni JJ, Seymour GB. A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. Nat Genet. 2006 Aug;38(8):948-52. [PMID: 16832354]
14. Qi J, Wang B, Hao BI. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J Mol Evol. 2004 Jan;58(1):1-11. [PMID: 14743310]
15. Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B. and Bartel, D.P., 2002. Prediction of plant microRNA targets. Cell 110, 513-520.
16. Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, Yu G. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science. 2000 Dec 15;290(5499):2105-10. [PMID: 11118137]
17. Riese M, Hohmann S, Saedler H, Munster T, Huijser P. Comparative analysis of the SBP-box gene families in P. patens and seed plants. Gene. 2007 Oct 15;401(1-2):28-37. Epub 2007 Jul 10. [PMID: 17689888]
18. Stone JM, Liang X, Nekl ER, Stiers JJ. Arabidopsis AtSPL14, a plant-specific SBP-domain transcription factor, participates in plant development and sensitivity to fumonisin B1. Plant J. 2005 Mar;41(5):744-54. [PMID: 15703061]
19. Unte US, Sorensen AM, Pesaresi P, Gandikota M, Leister D, Saedler H, Huijser P. SPL8, an SBP-box gene that affects pollen sac development in Arabidopsis. Plant Cell. 2003 Apr;15(4):1009-19.[PMID: 12671094]
20. Wang X, Shi X, Hao B, Ge S, Luo J. Duplication and DNA segmental loss in the rice genome: implications for diploidization. New Phytol. 2005 Mar;165(3):937-46. [PMID: 15720704]
21. Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, Faller M, Bomblies K, Lukens L, Doebley JF. The origin of the naked grains of maize. Nature. 2005 Aug 4;436(7051):714-9. [PMID: 16079849]
22. Wu G, Poethig RS. Temporal regulation of shoot development in *Arabidopsis thaliana* by miR156 and its target SPL3. Development. 2006 Sep;133(18):3539-47. [PMID: 16914499]
23. Yamasaki K, Kigawa T, Inoue M, Tateno M, Yamasaki T, Yabuki T, Aoki M, Seki E,Matsuda T, Nunokawa E, Ishizuka Y, Terada T, Shirouzu M, Osanai T, Tanaka A, Seki M, Shinozaki K, Yokoyama S.A novel zinc-binding motif revealed by solution structures of DNA-binding domains of Arabidopsis SBP-family transcription factors. J Mol. Biol. 337, 49-63. [PMID: 15001351]
24. Yamasaki K, Kigawa T, Inoue M, Yamasaki T, Yabuki T, Aoki M, Seki E, Matsuda T, Tomo Y, Terada T, Shirouzu M, Tanaka A, Seki M, Shinozaki K, Yokoyama S. An Arabidopsis SBP-domain fragment with a disrupted C-terminal zinc-binding site retains its tertiary structure. FEBS Lett. 2006 Apr 3;580(8):2109-16. [PMID: 16554053]
25. Yang Z, Wang X, Gu S, Hu Z, Xu H, Xu C. Comparative study of SBP-box gene family in Arabidopsis and rice. Gene. 2007 Apr 18; PMID: 17629421
26. Zhang Y, Schwarz S, Saedler H, Huijser P. SPL8, a local regulator in a subset of gibberellin-mediated developmental processes in Arabidopsis. Plant Mol Biol. 2007 Feb;63(3):429-39. [PMID: 17093870]

27. Zhu QH, Guo AY, Gao G, Zhong YF, Xu M, Huang M, Luo J. DPTF: a database of poplar transcription factors. Bioinformatics. 2007 May 15;23(10):1307-8. Epub 2007 Mar 28.[PMID: 17392330]