Botanical Research 植物学研究, 2021, 10(5), 724-733
Published Online September 2021 in Hans. http://www.hanspub.org/journal/br
https://doi.org/10.12677/br.2021.105091

Hans汉斯

拟南芥IDZ基因的系统发育、分子进化分析

王嘉琪

湖南农业大学生物科学技术学院，湖南 长沙
Email: corawang07@163.com

收稿日期：2021年8月6日；录用日期：2021年9月18日；发布日期：2021年9月28日

摘　要

Cys2His2 (C2H2)型锌指蛋白是广泛存在于真核生物转录因子中的DNA结合模体。在大多数情况下，含有C2H2型锌指结构的蛋白是基因表达调控中的重要的转录调节因子，在细胞发育、分化和抑制恶性细胞转化(抑瘤)等过程中起着重要作用。拟南芥是第一个完成全基因组测序的植物，并作为模式生物被广泛运用于植物生物学各领域的研究中。拟南芥IDZ基因家族以C2H2型锌指蛋白结构为特征。本文运用生物信息学对拟南芥IDZ基因家族中的11个蛋白编码基因进行理化性质、系统发育分析、蛋白保守基序、染色体定位、互作蛋白网络、蛋白质三维结构分析。结果提示同组成员包含相似的保守基序和序列结构，IDZ基因在染色体上分布均匀，不存在串联重复片段。本研究为进一步研究IDZ家族的生物学功能和价值提供了科学依据。

关键词

拟南芥，IDZ基因家族，C2H2锌指蛋白，系统发育分析

Phylogenetic and Molecular Evolution Analysis of IDZ Gene in *Arabidopsis thaliana*

Jiaqi Wang

College of Bioscience and Biotechnology, Hunan Agricultural University, Changsha Hunan
Email: corawang07@163.com

Received: Aug. 6th, 2021; accepted: Sep. 18th, 2021; published: Sep. 28th, 2021

Abstract

The Cys2His2 (C2H2) type zinc finger protein is a DNA binding motif that exists widely in eukaryotic transcription factors. In most cases, proteins containing C2H2 type zinc finger structures are

文章引用: 王嘉琪. 拟南芥 IDZ 基因的系统发育、分子进化分析[J]. 植物学研究, 2021, 10(5): 724-733.
DOI: 10.12677/br.2021.105091

important transcription regulators in regulation of gene expression and play an important role in cell development, differentiation and inhibition of malignant cell transformation (tumor suppression). *Arabidopsis* was the first plant to complete the whole genome sequencing and is widely used as a model organism in various areas of plant biology. The IDZ gene family in *Arabidopsis* is characterized by the structure of C2H2 type zinc finger protein. In this study, 11 protein coding genes in IDZ gene family of *Arabidopsis thaliana* were analyzed by using bioinformatics, including physical and chemical properties, phylogenetic analysis, protein conserved motif, chromosome location, interacting protein network, and three-dimensional structure analysis. The results showed that the IDZ gene was distributed evenly on the chromosome and there were no tandem repeats in the IDZ gene. This study provides a scientific basis for further study on the biological function and value of the IDZ family.

Keywords

***Arabidopsis thaliana*, IDZ Gene Family, C2H2 Zinc Finger Protein, Phylogenetic Analysis**

1. 引言

拟南芥(*Arabidopsis thaliana*)是十字花科(Cruciferae)芥菜科的一种小型开花植物。它分布在世界各地，最早在 16 世纪由约翰内斯·塔尔(Johannes Thal)报道。五十多年来，它已被广泛用于研究植物突变和经典的遗传分析。拟南芥现在已成为植物生物学不同研究领域的模式生物。

拟南芥是一种具有 2n = 10 条染色体的二倍体植物。拟南芥成为第一个完成全基因组测序植物是基于以下几点优势：1) 基因组约为 120 Mb，且结构简单，重复序列少；2) 从种子萌发到结实期为 6 周，世代时间短；3) 产生大量种子。测序工作由国际合作组织拟南芥基因组计划(*Arabidopsis* Genome Initiative, AGI)完成。拟南芥作为一种重要的农业作物来说，是一种无价的资源。

Cys2His2 (C2H2)型锌指蛋白是广泛存在于真核转录因子中的 DNA 结合模体。锌指结构是由两个或三个 β 折叠和一个 α-螺旋组成的短蛋白质序列。位于特定位置的两个半胱氨酸和两个组氨酸残基与锌离子结合以稳定结构。另外四个氨基酸残基定位在 α-螺旋的 N 端，通过与 DNA 主槽中的氢供体和受体相互作用参与 DNA 结合。一个蛋白质中锌指的数量可以在很大范围内变化，从而使目标 DNA 序列具有可变性。锌指蛋白除具有 DNA 结合作用外，还具有蛋白质与蛋白质、RNA 与蛋白质的相互作用。在大多数情况下，含有 C2H2 型锌指蛋白的蛋白质是基因表达的转运调节因子，在细胞发育、分化和抑制恶性细胞转化(抑瘤)等过程中起着重要作用[1]。

本实验通过对拟南芥 IDZ 基因家族进行生物信息学分析，揭示拟南芥 IDZ 的结构及进化特点，为后期功能基因挖掘及深入功能分析提供基础数据。

2. 材料与方法

2.1. 拟南芥 IDZ 基因家族的选取及理化分析

拟南芥 IDZ 基因家族信息来源于 TAIR 数据库(https://www.*arabidopsis*.org/index.jsp)，并通过 NCBI 获取 IDZ 基因家族蛋白质序列信息，并通过 ExPaSy [2] [3] [4] (https://web.expasy.org/compute_pi/)在线工具进行理化性质分析

2.2. 拟南芥 IDZ 基因家族的多序列比对

运用 T-coffee [5]在线多序列比对工具(http://tcoffee.crg.cat/apps/tcoffee)对拟南芥 IDZ 基因家族序列做多序列比对，将结果导入 ESPript 3.0 [6] (https://espript.ibcp.fr/ESPript/cgi-bin/ESPript.cgi)，Secondary structure depiction 的 Parameters 参数设置为 Sec.structure labels: α1，β2，α2，β2，…，Sequence similarities depiction parameters 设置为%Equivalent，结果经过检验验证。

2.3. 拟南芥 IDZ 基因家族系统进化树构建

使用 ClustalW 对拟南芥 IDZ 基因家族序列进行多序列比对，参数设置为 Pairwise Alignment 的 Gap Opening Penalty: 10，Gap Extension Penalty: 0.1；Multiple Alignment 的 Gap Opening Penalty: 10，Gap Extension Penalty: 0.2；Protein Weight Matrix 设置为 Identity，Residue-specific Penalties 设置为 ON，Hydrophilic Penalties 设置为 ON，Gap Separation Distance: 4，End Gap Separation 设置为 off，Use Negative Matrix 设置为 OFF，Delay Divergent Cutoff(%)：30。将结果用 MEGA 7 [7]构建系统进化树，构建采用临近算法(Neighbor-Joining, NJ)，Bootstrap 检验 1000 次，Model/Method 设置为 p-distance，Rates among Sites 设置为 Uniform rates, Gaps/Missing Data Treatment 设置为 Partial deletion，Site Coverage Cutoff(%)：50，得出的系统进化树设置 Hide valueslowerthan 60%。

2.4. 拟南芥 IDZ 基因家族的结构分析

使用 MEME [8] (https://meme-suite.org/meme/tools/meme)在线工具对拟南芥 IDZ 基因家族进行分析，参数设置 motif 为 10。将结果与进化树导入 TBtools [9]将蛋白质结构域可视化，使用 Gene Structure View (Advanced)功能，勾选 Fill in Gradient Mode 和 Motif Num，将 Width 设置为 1500，Hight 设置为 300。

使用 Pfam [10] (http://pfam.xfam.org/)数据库进行数据分析，可得蛋白结构域信息。

2.5. 拟南芥 IDZ 基因家族的染色体定位分析

从 TAIR 数据库上拉取拟南芥 IDZ 基因家族 Genomic Locus 相关信息制表，使用 MG2C 2.0 [11] (http://mg2c.iask.in/mg2c_v2.0/)导入相关信息绘制染色体定位图。

2.6. 蛋白互相作用网络构建

将蛋白质名称输入 STRING v11.5 [12] (https://string-db.org/)数据库分析可构建出蛋白互相作用网络，利用分析功能可得网络统计数据，以及 GO 基因富集等相关分析结果。

2.7. 蛋白三维结构分析

将蛋白质名称输入 Uniprot [13] (http://beta.uniprot.org/)数据库可得蛋白质三维结构，根据 AlphaFold [14]产生的置信评分(plDDT)不同，进行注释分析。

3. 结果与分析

3.1. 拟南芥 IDZ 基因家族的选取及理化分析

从 TAIR 数据库获取拟南芥 IDZ 家族成员共 11 个[15] [16]，利用 ExPaSy 在线工具理化性质分析显示(表 1)，该蛋白家族分子量大小为 64463.70 (AtIDZ2)至 44513.05 (AtIDZ8)，等电点为 9.63 (AtIDZ3)至 8.36 (AtIDZ5)。

Table 1. Physicochemical properties of IDZ gene family in *Arabidopsis thaliana*
表 1. 拟南芥 IDZ 基因家族理化性质

蛋白质 Protein	基因 ID Gene ID	分子量 Mw (kD)	等电点 PI	染色体 Chromosome	染色体定位 Chromosome location
AtIDZ1	AT3G45260	49,940.73	9.41	Chr3	16596358:16598811
AtIDZ2	AT2G02070	64,463.70	9.18	Chr2	505375:510781
AtIDZ3	AT5G03150	55,200.86	9.63	Chr5	745421:749028
AtIDZ4	AT1G03840	55,826.16	8.64	Chr1	967470:970334
AtIDZ5	AT5G44160	51,189.08	8.36	Chr5	17772856:17775749
AtIDZ6	AT3G50700	47,943.19	9.44	Chr3	18840411:18843053
AtIDZ7	AT5G66730	52,626.36	8.93	Chr5	26641586:26644738
AtIDZ8	AT4G02670	44,513.05	9.23	Chr4	1176093:1178489
AtIDZ9	AT3G13810	56,726.80	8.70	Chr3	4544364:4547513
AtIDZ10	AT1G55110	50,844.58	8.64	Chr1	20560258:20563269
AtIDZ11	AT5G60470	50,965.24	9.27	Chr5	24320595:24322790

3.2. 拟南芥 IDZ 基因家族的多序列比对

利用 T-coffee 在线多序列比对工具(http://tcoffee.crg.cat/apps/tcoffee)对拟南芥 IDZ 基因家族序列做多序列比对，将结果导入 ESPript 3.0 (https://espript.ibcp.fr/ESPript/ESPript/)可得多序列比对可视图(图 1)。由结果可知在序列 40-200、350-370、410 处存在高度保守区段。

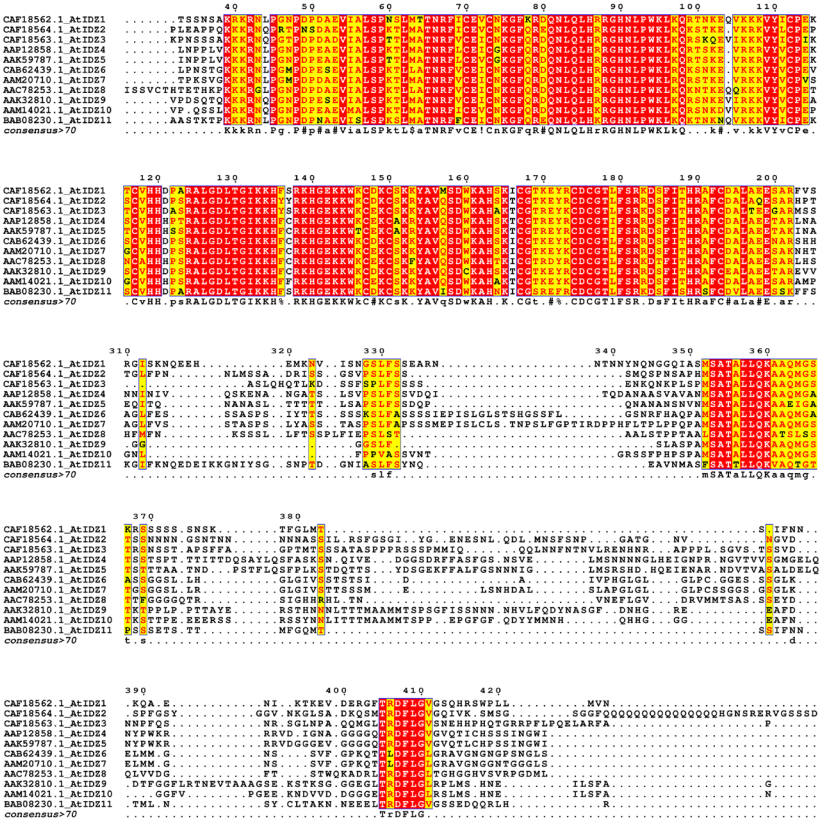

Figure 1. Sequence alignment of IDZ gene family in *Arabidopsis thaliana*
图 1. 拟南芥 IDZ 基因家族序列比对结果

3.3. 拟南芥 IDZ 基因家族的进化树分析

拟南芥 IDZ 基因家族系统进化树上都为旁系同源，是原始基因复制分离而产生的同源基因，由图(图 2)可知原始的基因因复制产生了分支，随后又各自复制产生多分支，在进化过程中现在 IDZ 家族成员里 AtIDZ4 的基因与原始基因的差异最小，而 AtIDZ2、AtIDZ11 的基因与原始基因差异最大。

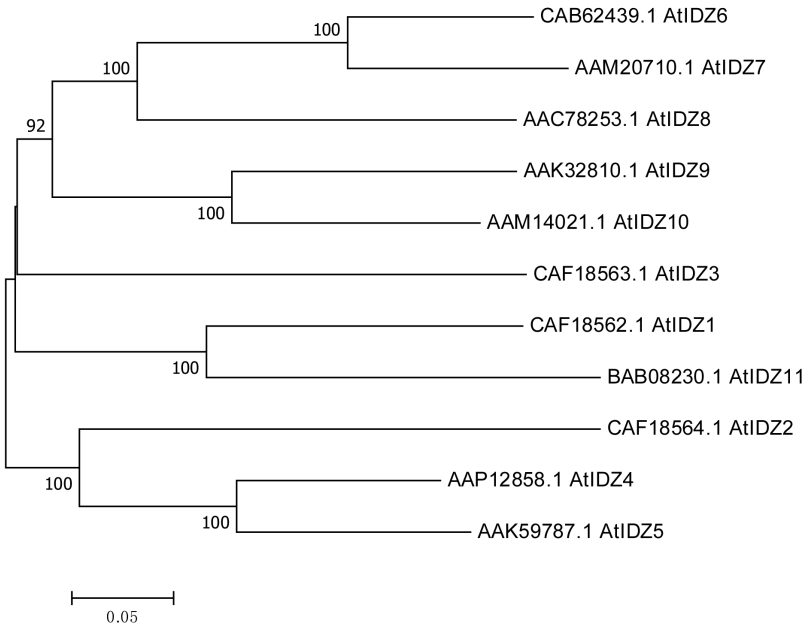

Figure 2. Phylogenetic tree of IDZ gene in *Arabidopsis thaliana* (NJ)
图 2. 拟南芥 IDZ 基因的系统进化树(Neighbor-Joining Method, NJ)

3.4. 拟南芥 IDZ 基因家族的结构分析

根据 TBtools 结果图(图 3)可知拟南芥 IDZ 基因家族都有 motif1-7、motif10，motif9 仅存在于 AtIDZ9、AtIDZ10 头部和 AtIDZ2 尾部，motif8 仅存在于 AtIDZ4、AtIDZ5 尾部，因此这五个蛋白质也许有其特殊功能。

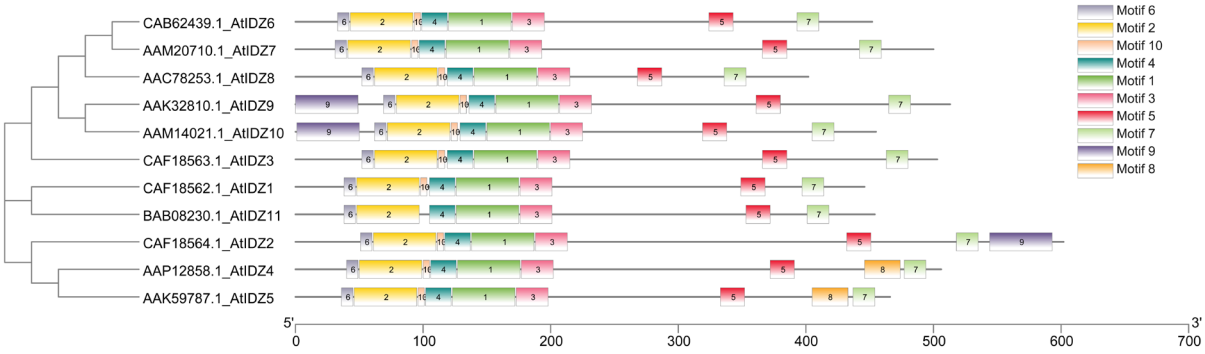

Figure 3. IDZ protein motif in *Arabidopsis*
图 3. 拟南芥 IDZ 蛋白质基序

根据 Pfam 数据库搜寻结果可得(表 2)，由结果可知拟南芥 IDZ 基因家族蛋白含有保守 IDZ 结构域，该结构域包含一个 zf-C2H2_jaz，一个 zf_C2H2。zf-C2H2_jaz 结构域家族发现于古生菌和真核生物中，jaz

含有 4 个 C2H2 型锌指基序，它们由长(28~38)氨基酸连接序列连接。jaz 在所有被检测的组织中都有表达，并定位于细胞核，主要是核仁。jaz 优先结合双链(ds) RNA 或 RNA/DNA 杂交而不是 DNA。单个锌指结构域的突变表明，锌指结构域不仅是 dsRNA 结合的必要条件，也是其核仁定位的必要条件，这表明了依赖于蛋白质的核酸结合能力的复杂贩运机制。此外，jaz 可能属于一类以双链 RNA 结合为特征的锌指蛋白，并可能通过其独特的双链 RNA 结合特性来调节细胞生长[17]。

Table 2. IDZ gene family domains in *Arabidopsis*
表 2. 拟南芥 IDZ 基因家族结构域

蛋白名称 Protein name	结构域 1 Domain1	个数 Number	结构域 2 Domain2	个数 Number	结构域 3 Domain3	个数 Number	结构域 4 Domain4	个数 Number
AtIDZ1	zf-C2H2_jaz	1	zf-C2H2	1	zf-H2C2_2	1		
AtIDZ2	zf-C2H2_jaz	1	zf-C2H2	2	zf-Sec23_Sec24	2		
AtIDZ3	zf-C2H2_jaz	1	zf-C2H2	2	zf-Sec23_Sec24	2		
AtIDZ4	zf-C2H2_jaz	1	zf-C2H2	1	zf-H2C2_2	1	NRIP1_repr_2	1
AtIDZ5	zf-C2H2_jaz	1	zf-C2H2	1	zf-H2C2_2	1		
AtIDZ6	zf-C2H2_jaz	1	zf-C2H2	2				
AtIDZ7	zf-C2H2_jaz	1	zf-C2H2	2	zf_met	1		
AtIDZ8	zf-C2H2_jaz	1	zf-C2H2	2	zf-MYST	1		
AtIDZ9	zf-C2H2_jaz	1	zf-C2H2	2	zf-Sec23_Sec24	2		
AtIDZ10	zf-C2H2_jaz	1	zf-C2H2	2	zf-Sec23_Sec24	2	C1_4	2
AtIDZ11	zf-C2H2_jaz	1	zf-C2H2	1	zf-H2C2_2	1		

3.5. 拟南芥 IDZ 基因家族的染色体定位

从 TAIR 数据库上拉取拟南芥 IDZ 基因家族 Genomic Locus 相关信息制表(表 3)，使用 MG2C 2.0 (http://mg2c.iask.in/mg2c_v2.0/)导入相关信息绘制可得染色体定位图(图 4)。

Table 3. Chromosome distribution information table of IDZ gene family in *Arabidopsis*
表 3. 拟南芥 IDZ 基因家族染色体分布信息表

蛋白质名称 protein name	基因 gene_id	基因起始 gene_start	基因终止 gene_end	染色体 chrom_id
AtIDZ1	AT3G45260	16596358	16598811	Chr3
AtIDZ2	AT2G02070	505375	510781	Chr2
AtIDZ3	AT5G03150	745421	749028	Chr5
AtIDZ4	AT1G03840	967470	970334	Chr1
AtIDZ5	AT5G44160	17772856	17775749	Chr5
AtIDZ6	AT3G50700	18840411	18843053	Chr3
AtIDZ7	AT5G66730	26641586	26644738	Chr5
AtIDZ8	AT4G02670	1176093	1178489	Chr4
AtIDZ9	AT3G13810	4544364	4547513	Chr3
AtIDZ10	AT1G55110	20560258	20563269	Chr1
AtIDZ11	AT5G60470	24320595	24322790	Chr5

IDZ 家族染色体定位结果(图 4)显示，AtIDZ4、AtIDZ10 的基因位于 1 号染色体，AtIDZ2 位于染色体 2，AtIDZ1、AtIDZ6、AtIDZ9 位于 3，AtIDZ8 位于 4 号染色体，AtIDZ3、AtIDZ5、AtIDZ7、AtIDZ11 位于 5 号染色体。基因复制分析结果表明 IDZ 基因家族没有串联重复基因。

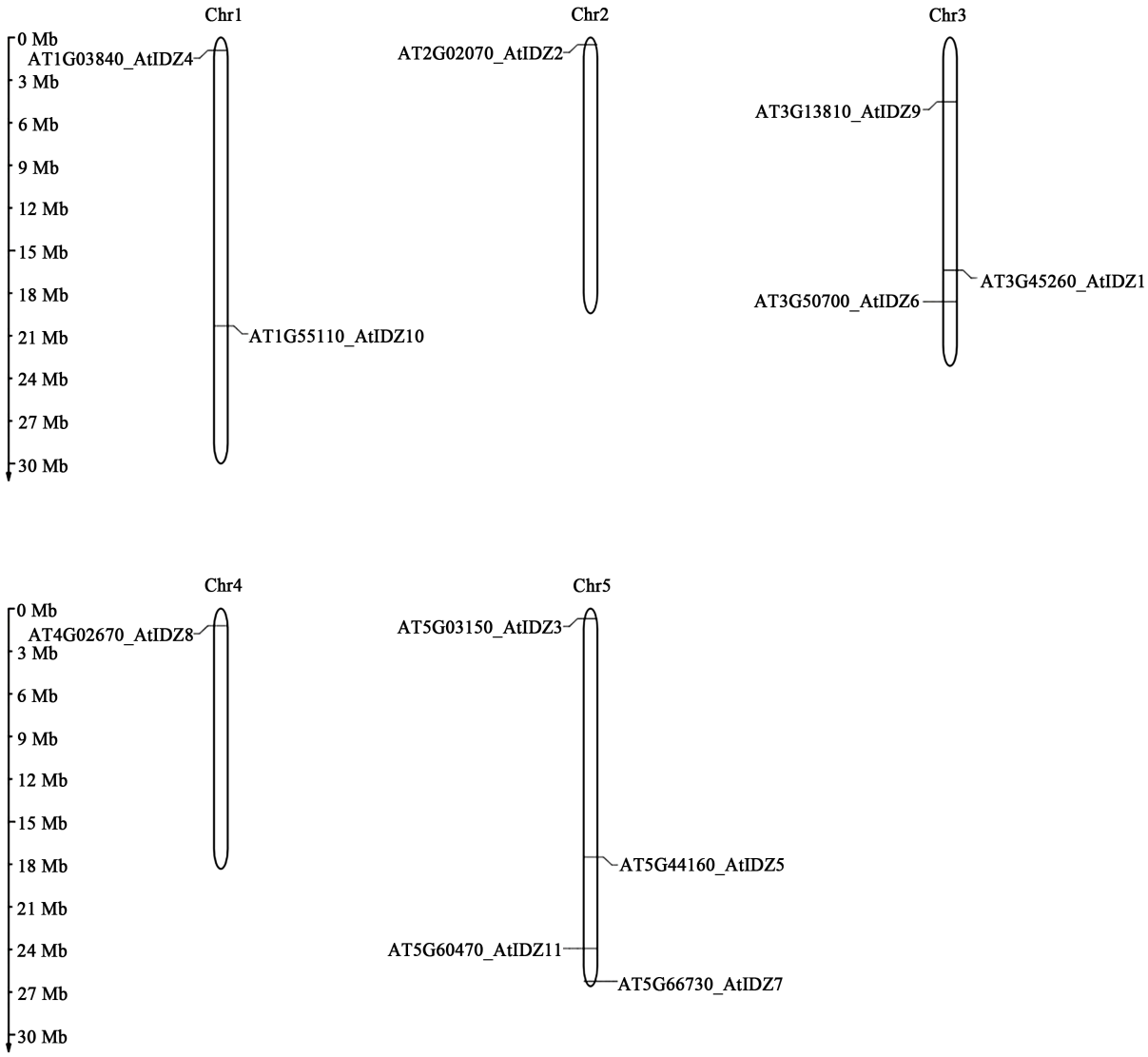

Figure 4. Chromosome mapping of IDZ gene family in *Arabidopsis thaliana*
图 4. 拟南芥 IDZ 基因家族染色体定位图

3.6. AtIDZ2 的蛋白互相作用网络

通过 STRING 数据库分析 AtIDZ2 蛋白-蛋白互相作用网络，结果显示节点数为 11，如图(图 5)所示。其中以 AtIDZ2 为中心存在多个与之相互作用的蛋白分子，且互作基因的 GO 基因富集分析结果显示其蛋白涉及的生物学过程多与细胞过程、细胞代谢的调节相关，分子功能多与 DNA 结合的转录因子活性相关。且这些蛋白多是高迁移率族蛋白(high mobility group protein, HMG 蛋白)，HMG 蛋白在染色质结构与功能及基因表达调控过程中均发挥重要作用，提示 AtIDZ 家族成员广泛参与生物体生长发育过程中的基因表达调控过程。

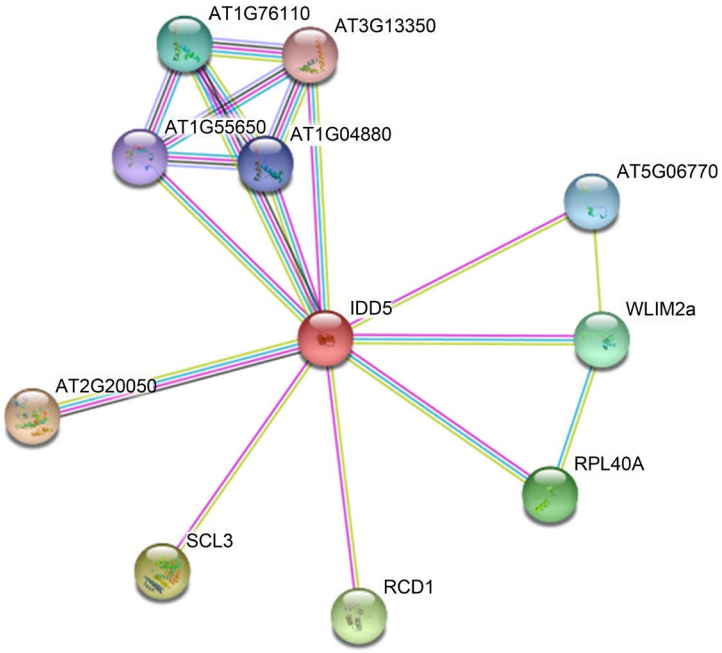

Figure 5. AtIDZ2 (IDD5) interacting protein network
图 5. AtIDZ2 (IDD5)互作蛋白网络图

3.7. AtIDZ2 三维结构分析

AlphaFold 产生一个 0 到 100 之间的每残留置信评分(pLDDT)。一些 pLDDT 含量低的区域可能是孤立无结构的。AtIDZ2 蛋白三维结构模型见图 6。

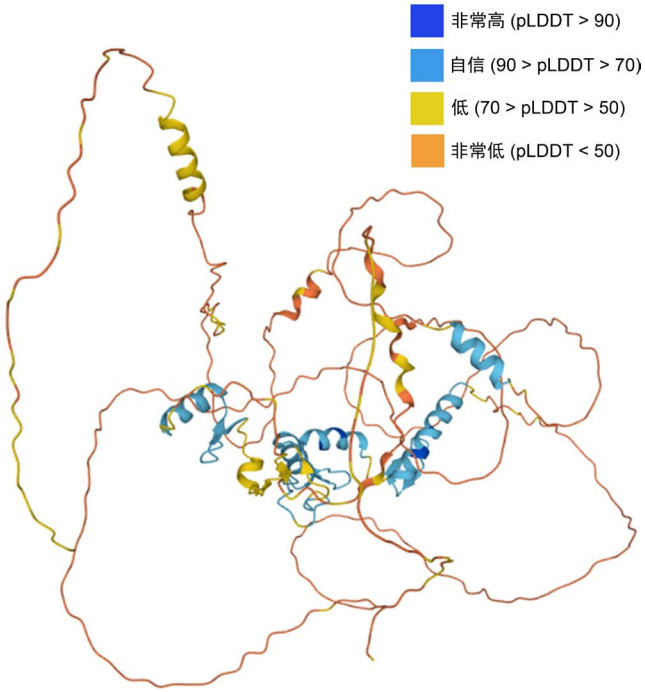

Figure 6. AtIDZ2 protein three-dimensional structure model
图 6. AtIDZ2 蛋白三维结构模型

4. 讨论与结论

模式生物拟南芥被广泛用于植物生物学各领域研究，因而它的相关生物信息学分析对于植物生物学领域来说是一项非常重要的工作。C2H2 型锌指结构蛋白质在植物细胞发育、分化和抑制恶性细胞转化等过程中起着重要作用。本研究通过分析拟南芥 IDZ 基因家族发现 11 个家族成员均有保守的 IDZ 结构域，该结构域包含一个 zf-c2h2_jaz，一个 zf_C2H2。根据蛋白基序分析可知家族成员蛋白 AtIDZ9、AtIDZ10 头部和 AtIDZ2 尾部有一种不保守基序，AtIDZ4、AtIDZ5 尾部有另一种不保守基序。由进化树分析可知 IDZ 家族成员里 AtIDZ4 的基因与原始基因的差异最小，而 AtIDZ2、AtIDZ11 的基因与原始基因差异最大。基因家族复制分析结果表明 IDZ 基因家族没有串联重复基因，且它们多存在于不同染色体上，有发生染色体同片段复制事件的可能性。后续研究选择了 IDZ 基因家族里的 AtIDZ2 进行分析，分析结果表明该蛋白与细胞过程、细胞代谢和调节相关，这印证了 IDZ 结构域含有 C2H2 型锌指结构域的结果可靠性。本研究为拟南芥 IDZ 基因家族深入的功能研究奠定了基础。此外，AtIDZ2、AtIDZ4、AtIDZ5、AtIDZ9、AtIDZ10 的基序特殊性提示它们可能在进化上具有趋异的功能，有潜在的继续研究价值，相关工作将有利于揭示生物生长发育调控过程的重要机制。

参考文献

[1] Razin, S.V., Borunova, V.V., Maksimenko, O.G., *et al.* (2012) Cys2His2 Zinc Finger Protein Family: Classification, Functions, and Major Members. *Biochemistry* (*Moscow*), **77**, 217-226. https://doi.org/10.1134/S0006297912030017

[2] Bjellqvist, B., Hughes, G.J., Pasquali, Ch., Paquet, N., Ravier, F., Sanchez, J.-Ch., Frutiger, S. and Hochstrasser, D.F. (1993) The Focusing Positions of Polypeptides in Immobilized pH Gradients Can Be Predicted from Their Amino Acid Sequences. *Electrophoresis*, **14**, 1023-1031. https://doi.org/10.1002/elps.11501401163

[3] Bjellqvist, B., Basse, B., Olsen, E. and Celis, J.E. (1994) Reference Points for Comparisons of Two-Dimensional Maps of Proteins from Different Human Cell Types Defined in a pH Scale Where Isoelectric Points Correlate with Polypeptide Compositions. *Electrophoresis*, **15**, 529-539. https://doi.org/10.1002/elps.1150150171

[4] Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D. and Bairoch, A. (2005) Protein Identification and Analysis Tools on the ExPASy Server. In: Walker, J.M., Ed., *The Proteomics Protocols Handbook*, Humana Press, Totowa, 571-607. https://doi.org/10.1385/1-59259-890-0:571

[5] Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. *Journal of Molecular Biology*, **302**, 205-217. https://doi.org/10.1006/jmbi.2000.4042

[6] Robert, X. and Gouet, P. (2014) Deciphering Key Features in Protein Structures with the New ENDscript Server. *Nucleic Acids Research*, **42**, W320-W324. https://doi.org/10.1093/nar/gku316

[7] Kumar, S., Stecher, G. and Tamura, K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, **33**, 1870-1874. https://doi.org/10.1093/molbev/msw054

[8] Bailey, T.L. and Elkan, C. (1994) Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers. In: *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, 28-36.

[9] Chen, C., *et al.* (2020) TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Molecular Plant*, **13**, 1194-1202. https://doi.org/10.1016/j.molp.2020.06.009

[10] Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D. and Bateman, A. (2021) Pfam: The Protein Families Database in 2021. *Nucleic Acids Research*, **49**, D412-D419.

[11] 晁江涛, 孔英珍, 王倩, 孙玉合, 龚达平, 吕婧, 刘贯山. MapGene2Chrom 基于 Perl 和 SVG 语言绘制基因物理图谱[J]. 遗传, 2015, 35(1): 91-97.

[12] Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., Jensen, L.J. and von Mering, C. (2019) STRING v11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Research*, **47**, D607-D613. https://doi.org/10.1093/nar/gky1131

[13] The UniProt Consortium (2021) UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Research*, **49**, D480-D489.

[14] Senior, A.W., Evans, R., Jumper, J., *et al*. (2020) Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature*, **577**, 706-710. https://doi.org/10.1038/s41586-019-1923-7

[15] Haasen, D., Kohler, C., Neuhaus, G. and Merkle, T. (1999) Nuclear Export of Proteins in Plants: AtXPO1 Is the Export Receptor for Leucine-Rich Nuclear Export Signals in *Arabidopsis thaliana*. *Plant Journal*, **20**, 695-705. https://doi.org/10.1046/j.1365-313X.1999.00644.x

[16] Merkle, T. (2001) Nuclear Import and Export of Proteins in Plants: A Tool for the Regulation of Signalling. *Planta*, **213**, 499-517. https://doi.org/10.1007/s004250100621

[17] Yang, M., May, W.S. and Ito, T. (1999) JAZ Requires the Double-Stranded RNA-Binding Zinc Finger Motifs for Nuclear Localization. *Journal of Biological Chemistry*, **274**, 27399-27406. https://doi.org/10.1074/jbc.274.39.27399