

Doi:10.11840/j.issn.1001-6392.2023.04.004

基于特定序列的长牡蛎视黄酸反应元件预测

李言柯¹, 徐晓莹², 魏磊¹, 黄宝玉¹, 张美激¹,
韩怡静¹, 刘雅琼¹, 王晓梅³, 王晓通¹

(1. 鲁东大学 农学院, 山东 烟台 264025; 2. 烟台市海洋经济研究院, 山东 烟台 264034;
3. 中国水产科学研究院长岛增殖实验站, 山东 烟台 265800)

摘 要: 视黄酸受体和核受体超家族中的大部分成员对细胞整个分化、增殖过程都具有调控功能。视黄酸受体结合配体后激活, 通过结合靶基因启动子区特定的核苷酸序列调控靶基因表达。视黄酸受体结合序列是由核心序列[A/G]G[T/G]TCA 间隔不同碱基构成的重复序列, 称为视黄酸反应元件。为了实现对长牡蛎基因组中含有的视黄酸反应元件的快速筛选预测, 本研究利用 Perl 工具编写了一个可以批量筛选视黄酸反应元件的脚本, 并对长牡蛎基因组中启动子区域序列进行筛选预测, 共筛选到 412 个启动子区含有视黄酸反应元件的基因。随后, 将这些基因在各种数据库中比对分析, 预测其参与的生物学过程及可能的生物学功能。结果显示, 大部分基因与蛋白质结合、核苷酸结合、水解酶活性、蛋白激酶活性等功能有关。

关键词: 长牡蛎; 视黄酸受体; 视黄酸反应元件; 基因组; Perl

中图分类号: P714+.5; Q811.4; TP313 **文献标识码:** A **文章编号:** 1001-6932(2023)04-0398-09

Prediction of retinoic acid response elements based on specific sequences in the pacific oyster *Crassostrea gigas*

LI Yanke¹, XU Xiaoying², WEI Lei¹, HUANG Baoyu¹, ZHANG Meiwei¹, HAN Yijing¹, LIU Yaqiong¹,
WANG Xiaomei³, WANG Xiaotong¹

(1. School of Agriculture, Ludong University, Yantai 264025, China;

2. Yantai Institute of Marine Economy, Yantai 264034, China;

3. Changdao Enhancement and Experiment Station, Chinese Academy of Fishery Sciences, Yantai 265800, China)

Abstract: Most members of the retinoic acid receptor and nuclear receptor superfamily regulate the process of cell differentiation and proliferation. RA receptors are activated by binding ligands and regulate the expression of target genes by binding specific nucleotide sequences in the promoter region of target genes. The retinoic acid receptor binding sequence is composed of [A/G]G[T/G]TCA, which is called retinoic acid response element. In order to achieve rapid screening and prediction of genes containing retinoic acid reaction elements on *Crassostrea gigas* genome, this study used Perl programming to predict the genes with retinoic acid response elements in the promoter region of genes on *Crassostrea gigas* genome. A total of 412 gene containing RAREs in the promoter regions were screened. Then, these genes were compared and analyzed in various databases to predict their biological processes and possible biological functions. The results showed that most of these genes were related to protein binding, nucleotide binding, hydrolase activity and protein kinase activity.

Keywords: *Crassostrea gigas*; retinoic acid receptor; retinoic acid response element; genome; Perl

收稿日期: 2022-06-04; 修订日期: 2022-07-19

基金项目: 国家自然科学基金 (41876193; 41906088; 42076088); 国家重点研发计划 (2018YFD0901400); 山东省泰山学者专项基金 (tsqn201812094); 山东省现代农业产业技术体系(SDAIT-14-03); 山东省高等学校“青创科技计划”(2019KJF004)

作者简介: 李言柯, 硕士研究生, 主要从事海洋生物基因组学研究, 电子邮箱: leeray001@163.com

通信作者: 刘雅琼, 博士, 讲师, 主要从事分子生物学研究, 电子邮箱: lyqpka@163.com

王晓梅, 博士, 副研究员, 主要从事海洋生物学研究, 电子邮箱: wangxiaom@cafs.ac.cn

视黄酸 (Retinoic Acid, RA) 是维生素 A 的一种衍生物, 在整个生物体细胞的分化、增殖以及程序性凋亡等过程中都发挥着调控功能^[1]。视黄酸在细胞内的受体蛋白具有抑制致癌因子发挥致癌作用的效应, 并在白血病^[2]、肝癌^[3]、乳腺癌^[4]等肿瘤的生长中发挥作用, 这也引起了研究者们对视黄酸受体的关注。实际上, 人们对视黄酸受体怎样控制细胞功能的研究大多是直接根据核受体超家族中的甲状腺激素及固醇类激素等的作用机制来理解的^[5]。视黄酸受体属于核受体, 结合配体后被激活, 进入细胞核中, 通过结合位于靶基因调控区域的特定 DNA 序列调节基因表达。1987 年, 人们通过预测核受体超家族中除类固醇、维生素 D 等之外的大量未曾发现的且与调节蛋白密切相关的基因, 从而确定了视黄酸受体的存在。随后发现了 RAR (RETINOIC ACID RECEPTOR) 和 RXR (RETINOID X RECEPTOR) 两大类视黄酸受体蛋白^[6], 见表 1。

表 1 视黄酸受体类型

视黄酸受体	亚类	基因功能
RAR α	$\alpha 1, \alpha 2$	反式激活
RAR β	$\beta 1, \beta 2$	抑制
RAR γ	$\gamma 1, \gamma 2$	配子, 细胞启动子活性
RXR α	-	-
RXR β	-	-
RXR γ	-	-

视黄酸受体与核受体超家族的基本结构相同, 都是由 A/B、C、D 及 E 四个对应不同作用的结构域组成。其中, 高度保守的 C 区含有 DNA 结合结构域 (DNA Binding Domain, DBD), 具有特异性的 DNA 识别序列^[7]。核受体的 DNA 结合结构域包括两个“锌指 (zinc fingers)”, 每个“锌指”表现出不同的功能: 由三个氨基酸组成一个半体, 且近 N 端的成为“P-box”; 由五个氨基酸组成且近 C 端的称为“D-box”^[8]。以“P-box”为氨基酸基础序列的核受体可以被分为两组, 第一组是以糖皮质激素、盐皮质激素、孕激素和雄激素受体为代表, 第二组包括其他所有配体激活的核受体, 包括维生素 D、全反式视黄酸、9-顺视黄酸受体和大部分具有与已确认的受体相似的结构、但尚未确定其配体和功能的孤儿受体。E 区是第二保守的区域, 含有配体结合结构域 (Ligand-Binding

Domain, LBD), 其功能相对复杂, 包含配体结合口袋, 是主要二聚化结合区域, 并具有配体依赖的反激活功能^[7]。

核受体通过直接结合其调控的靶基因启动子区特异的 DNA 序列从而发挥功能, 其结合的特异 DNA 序列又称为激素反应元件 (Hormone Response Elements, HREs)^[9]。视黄酸受体结合的靶基因调控区特定序列被称为视黄酸反应元件 (Retinoic Acid Response Elements, RAREs)。研究表明, 主要的 RAREs 的核心序列为 [A/G]G[T/G]TCA, 此外还有一些近似序列也可以构成 RAREs^[10-11]。视黄酸受体识别位点由核心序列组成, 两个核心序列之间间隔不同数量的核苷酸组成不同的同向重复序列 (Direct Repeats, DRs), 例如, 间隔 1 个核苷酸组成的序列为 DR1, 间隔 2 个核苷酸组成的序列称为 DR2。第一个发现的 RA 反应元件为两个拷贝的 [A/G]GGTCA 核心序列组成的同向重复序列^[12]。目前为止发现的大多数 RAREs 是由 [A/G]GGTCA 间隔 5 个碱基组成的同向重复序列 DR5, 是全反式视黄酸受体识别的片段。此外, [A/G]GGTCA 间隔两个核苷酸组成的同向重复序列 DR2 同样也作为 RAREs 发挥作用。另外还鉴定到少量的 9-顺式视黄酸反应元件, 由同向重复的两个核心序列间隔 1 个碱基构成的 DR1 组成。研究表明, 脊椎动物中 RAR 可以结合 DR1、DR2 以及 DR5, 而 RXR 识别序列为 DR1^[12]。

除了同向重复外, 视黄酸反应元件同样识别以 8 个碱基为间隔且两个核心序列以外翻重复 (Everted Repeats, ER) 构成的 ER8^[13], 以及以零个碱基和一个碱基为间隔的且两个核心序列以回文序列 (palindrome) 构成的反向重复序列 (Inverted Repeat, IR) IR0 和 IR1。另外 RARs 可以结合同向重复、外翻重复和反向重复序列及可变长度的间隔碱基的复合结构。不过, 视黄酸反应元件的相对效力和特异性取决于重复序列的构型以及间隔碱基组成的核苷酸序列。研究发现以 DR5 组成的视黄酸反应元件通常在反式激活测定中比 DR2 和 ER8 有效, 而且比间隔多个碱基的复合构型更有效。已知的视黄酸反应元件由 Giguère 等^[5]整理 (见表 2)。

在生物信息学 (Bioinformatics) 中, Perl (Practical Extraction and Report Language) 有着得天独厚的优势。首先, Perl 作为一个脚本语言, 省去编译

的周期过程，有着编译语言的功能，又有 Shell 脚本的快捷。其次，Perl 不会因为数据太大而崩溃，在生物测序领域动辄上百 G 乃至 T 数量级的数据需要处理，Perl 很好地契合这一点。最后，Perl 中有着无比强大的正则表达式可以快速处理字符串，非常适合处理生物信息中大数据量的序列。

表 2 已知的视黄酸反应元件^[5]

类型	基因	序列
→ →		
DR1	HBV	CGGGGTAAAGGTTTCAGG
	mHHCl	TCAGGTCAGGGGTGGGG
→ →		
DR2	mCRBP-I	GTAGGTCAAAAGGTCAGA
	hApoA1	AGGGGTCAAGGGTTCAGT
→ →		
DR5	hRARβ2	AGGGTTCACCGAAAGTTCACT
	mCP-H	GCAGGTCACGACAGGGCATA
← →		
ER8	m7F-crystallin	AGTGACCCTTTTAACCAGGTCAGT
	hMCAD	ATTGACCTTTCTCTCCGGGTAAAG
→ ←		
IR	TREpal	TCAGGTCATGACCTGA
	hOST	CTAGGTGACTCACCGG

蛋白质中酪氨酸磷酸化修饰过程是调节信号转导途径和关键细胞功能的主要机制之一。酪氨酸残基的可逆磷酸化修饰受到酪氨酸激酶和酪氨酸磷酸酶的共同调节^[14]。酪氨酸激酶和酪氨酸磷酸酶在控制细胞生长和许多其他功能中发挥关键作用，酪氨酸磷酸化对许多细胞信号通路至关重要^[15]。此外，有研究表明，酪氨酸磷酸化也影响一些代谢酶的活性^[16]。另外，酪氨酸可用于蛋白质、生物胺和黑色素的生物合成，或通过五种酶促反应分解为能量，产生乙酰乙酸等并进一步分解^[17]。

长牡蛎 (*Crassostrea gigas*) 属于软体动物门，双壳纲，牡蛎目，牡蛎科，生长在全世界范围的温热带海域中^[18-19]。在世界范围内，长牡蛎具有很高的经济价值和生态意义。对长牡蛎视黄酸受体 CgRAR 和 CgRXR 已有一定的研究^[20-22]，但是其所调控的下游基因及下游基因的功能的相关报道较少。本文通过 Perl 语言编写一个可以预测视黄酸反应元件的脚本，在长牡蛎基因组中挖掘具有视黄酸反应元件的基因，并通过生物信息学方法

对这些基因的功能进行预测。通过本文的研究，有助于整体了解长牡蛎中视黄酸受体调控的基因及其可能参与的生物学过程，为长牡蛎视黄酸受体调控的信号途径及下游基因功能的分子生物学研究提供一定的基础。

1 材料与方法

1.1 牡蛎全基因组数据筛选获取

2012 年，中国科学家完成世界上第一张牡蛎的全基因组序列图谱，牡蛎实际基因组大小为 800 M，这次组装基因组最终版本为 558 M。组装率达到 70%^[23]。本文所用到的牡蛎基因数据文件 (FASTA) 和基因注释文件 (GFF) 下载于 NCBI (ftp://ftp.ncbi.nlm.nih.gov/)，基因版本号为：GCA_000297895.1。此版本数据信息新，基因 ID 标明确，主流数据库都是以此来进行功能注释，方便后面进行功能注释。

1.2 在 Linux、Mac、Windows 下配置运行环境

在牡蛎全基因组中预测含有视黄酸反应元件的基因，首先从牡蛎全基因组文件中获取所有可能的序列，并获取序列所在 Scaffold 上的位点。在基因描述文件里，获取序列下游调控的基因。并获取基因的起始位点、终止位点、正负链。然后利用得出的基因 ID 对数据库 (GO, KEGG, Swissport, Trembl 等) 进行对比获取 KEGG 富集图等数据。

软件基于 Perl 语言编程在多个操作系统都可以运行，在 windows 中，需要提前安装 Perl 编程语言 (https://www.perl.org/)，打开 cmd 执行命令。在 macos 与 Liunx 中打开终端执行命令即可。

示例代码：

```
perl PredictRAR.pl -fa GCF_902806645.1_cgigas_uk_roslin_v1_genomic.fna -gff GCF_902806645.1_cgigas_uk_roslin_v1_genomic.gff -upstreamOut up.txt -intervalOut inter.txt
```

1.2.1 配置视黄酸反应元件模式匹配

从前文中分析出，视黄酸受体结合位点由两个核心序列间隔不同的数量的碱基构成，首先我们列表分析视黄酸反应元件的主要类型。本文选择文献报道的主要核心序列 [A/G]G[T/G]TCA，在表 3 中列出 [A/G]G[T/G]TCA 构成的重复序列，作为 RAREs 的预测依据，表中类型皆用正则表达式

表3 视黄酸反应元件模式

类型	正链	负链
DR	[AG]G[TG]TCA.{1,2}[AG]G[TG]TCA	TGA[AC]C[TC].{1,2}TGA[AC]C[TC]
	[AG]G[TG]TCA.{5}[AG]G[TG]TCA	TGA[AC]C[TC].{5}TGA[AC]C[TC]
ER	TGACC[TC].{8}[AG]GGTCA	TGACC[TC].{8}[AG]GGTCA
IR	[AG]GGTCA.{0,1}TGACC[TC]	[AG]GGTCA.{0,1}TGACC[TC]

表达。

通过模式化之后序列符合的视黄酸反应元件的所有结构。接下来就是通过 Perl 处理牡蛎基因组和基因描述文件。

1.2.2 处理牡蛎基因组文件

在生物信息学中，FASTA 文件是一种最常见的表示核苷酸序列或氨基酸序列的文本格式。每一个独立的染色体、基因或蛋白质都是由一个“>”开头，后面跟基因描述文字为第一行，第二行开始为核苷酸序列或氨基酸序列，核苷酸 ATCGN 等代码表示。

通过 Perl 存一个哈希变量%seq，把第一行的 ID 定义为 Key，之后的序列直到下一个“>”定义为 Value。如此就可以把数据全部捕获在内存中，等待调用。

利用 Perl 的正则表达式用每一个模式对每一个基因进行匹配，若匹配成功，则把匹配的哈希对应的键值重新赋值给一个哈希变量。并获取匹配上的序列和所在的位置，储存于一个二维哈希。基因组文件处理完毕。

1.2.3 处理基因描述文件

现版本的基因描述文件版本是由 Sanger 研究所研发定义的第三代文本格式文件，可以用来描述基因、mRNA、CDS、exon 等序列所在的染色体或 Scaffold、基因位点、正负链、起始位点、终止位点及其基因 ID、类型描述。

利用被 Shell 处理过后的基因描述文件，在 Perl 中省了多余的处理步骤，但是由于基因描述文件的信息过多，仍然需要通过筛选过滤出有用的基因信息。

首先利用 Split 命令对基因描述文件进行分隔，并把 Scaffold、正负链、基因起始和终止位点、基因 ID 分别存入变量中，其余的定义为 undef。基因描述文件处理完成。

1.2.4 提取预测视黄酸反应元件

哈希%seq 作为一个全局变量并储存每一个

Scaffold 和基因序列可以随时被调用，利用之前已经正则表达式模式化的视黄酸反应元件对每一个哈希值进行模式匹配 m//。捕获的序列储存到一个数组中，并把匹配上的序列和 Scaffold 另存于一个新哈希。

由于基因组装只组装一条链，而两条链上都有基因，这导致了基因在描述文件中的正负链之分。组装在同一条链上正负链基因呈反向互补结构。因此，正链上基因调控区是在上游，负链上的基因调控区是下游并反向互补。为确保预测的准确性和完整性，本文还捕获了基因中的预测序列并保存于不同文件中。

通过捕获的序列所在 Scaffold 的位置与基因描述文件中的正链基因起始文件对比，如果捕获序列在基因上游 2Kb 内，则把基因描述文件保存的变量存入一个上游基因合集文件。负链的基因与终止位点进行对比，在基因下游 2Kb 内同样把保存的变量存入上游基因合集文件。如果在起始位点和终止位点中，不论正负链都将保存的变量存于一个基因内合集文件。至此预测基因序列已经完成，接下来进行基因数据库功能注释。

1.3 DNA 水平的基因功能注释

通过 Blast 将预测的视黄酸反应元件所调控的基因于 NR、SwissPort、GO、GOG、KEGG 等数据库进行基因注释。

Blast (Basic Local Alignment Search Tool) 是美国国家生物技术信息中心开发的一种通过对比待测基因与蛋白质数据库或者基因数据库发现生物序列之间相似区域的工具。Blast 程序能不限制待测基因和参考数据库快速进行对比并利用统计学来计算序列同源性。本文将与预测基因相似性最大的基因的功能注释表示为预测基因的功能注释。

NR 数据库 (Non-Redundant Database) 是美国国家生物技术信息中心非冗余蛋白质数据库，记录了 GenBank、EMBL、DDBJ 和 PDB 数据库中

的非冗余蛋白序列。NR 是以核酸序列为基础的交叉索引，可以将核酸序列数据与蛋白序列数据联系起来。

SwissPort 数据库是由欧洲生物信息学研究所和其他组织共同维护的数据库。数据库中每一条蛋白质序列都包含着蛋白质的基因名、引用文献信息、物种来源、分类学注释等。基本信息之后，会注释说明蛋白质的功能、翻译后修饰及 3D 结构等等。SwissPort 与 PIR、GPCR 等 30 多个数据库建立交叉引用，扩展了数据库和注释信息。

GO 数据库（Gene Ontology）是由基因本体联合会创立的国际标准化的基因功能数据库。GO 数据库中不含序列信息，注重注释基因与蛋白质的功能，提供了三级结构的标准注释，包括分子功能、细胞组件和生物过程。本文中利用 GO 数据库进行了富集分析。

COG 数据库（Cluster of Orthologous Groups of proteins）直译为蛋白直系同源聚类，COG 数据库中每一个蛋白质都被假定为来自同一个祖先蛋白，为直系同源或者旁系同源。直系同源是指不同物种蛋白是由共同祖先基因进化而来，可能有相似的功能。旁系同源是指同一物种内基因由于自我复制而分离的同源基因。COG 数据库又分两个数据库，COG 与 KOG。本文所用数据库是 KOG，并进行注释图解。

KEGG 数据库（Kyoto Encyclopedia of Genes and Genomes）是由日本京都大学生物信息学中心创立维护整合了目前生物化学中关于基因组、化学和系统功能信息的数据库。相较于其他数据库，KEGG 最大的特色就是系统分析批量基因产物在细

胞中的代谢通路并进行聚类。本文将预测出的含有视黄酸反应元件的基因进行 KEGG 通路图分析。

2 结果

2.1 视黄酸反应元件预测结果

2.1.1 基因上游视黄酸反应元件

通过 Perl 编程对基因组文件和基因组描述文件处理，预测出的含有视黄酸反应元件的基因一共 412 个，其中一条 Scaffold 上最多有 7 个视黄酸反应元件。正链基因上总共有 215 个，负链基因上总共有 197 个。其中以同向重复（DR）的元件为 328 个，反向重复为 15 个，回文结构为 70 个。部分预测上游视黄酸反应元件及其基因位置见表 4，由于产出的结果过多，只展示部分信息。

2.1.2 基因内视黄酸反应元件

为保证预测的视黄酸反应元件的数据全面性和完整性，本文还预测了基因内的视黄酸反应元件。基因内的视黄酸反应元件总数为 3 072 个，其中一个 Scaffold 上最多有 15 个视黄酸反应元件。正链基因上有 1 575 个，负链基因上有 1 497 个。其中以同向重复（DR）的元件 2 506 个，反向重复为 130 个，回文结构为 436 个。

2.2 GO 数据库注释

GO 数据库在分子功能、生物学过程及细胞组件三个方面对提供的基因进行富集分析。通过分子功能注释发现大部分基因与蛋白质结合、核苷酸结合、水解酶活性以及蛋白激酶活性等相关。其中与酪氨酸残基磷酸化修饰相关的基因有 4 个，具体见表 5。

表 4 部分预测基因上游视黄酸反应元件及其基因位置

染色体	基因 ID	正负链	类型	视黄酸反应元件	位置	起始位点	终止位点
JH815662.1	CGI_10003111	-	DR	tgacccttgacct	53475	46856	51766
JH815691.1	CGI_10001586	+	DR	AGGTCACGGGGTCA	4756	5754	9932
JH815726.1	CGI_10005369	-	IR	AGGTCATGACCC	7731	6734	7225
JH815739.1	CGI_10013078	+	DR	GGTTCATGAGGTCA	183964	185679	186080
JH815740.1	CGI_10009032	-	ER	TGACCetaagtccaaggtca	113348	101388	111532
JH815769.1	CGI_10011943	+	DR	AGTTCACAGTTCA	104027	104554	112431
JH815769.1	CGI_10011952	+	DR	AGTTCAGGAGTTCA	290741	291335	292999
JH815778.1	CGI_10000522	-	DR	TGACCTGTGAACT	5651	2565	5479
JH815782.1	CGI_10013037	+	DR	agGTCAAAGGTCA	110747	112223	169446
...	...	...	...	...	...	...	...

通过生物过程注释得出大部分基因与单细胞或单体过程、多细胞系统发育及胚胎发育等功能相关。其中酪氨酸残基磷酸化修饰的生物过程相关的基因有5个，具体见表6。

通过细胞组件注释得出和生物膜系统相关联的基因有21个，和细胞质相关联的有14个，与细胞核相关联的基因有10个。另外基因大多和细胞内细胞器、质膜相关（表7）。

2.3 KOG 注释

由于牡蛎是真核生物，本文选择KOG库进行注释。注释结果中发现有9个基因为酪氨酸蛋白激酶或酪氨酸磷酸酶，具体见表8。

2.4 KEGG 富集通路图

将预测的含有具有 RAREs 序列的基因在KEGG数据库中进行富集分析。结果发现大部分基因与生物系统、遗传信息处理、环境信息处理、

细胞过程和多种代谢通路有关。另外，从结果中发现上游调控的基因中有一个基因（CGI_10021536）和酪氨酸代谢有关。酪氨酸代谢通路见图1。

3 讨论

视黄酸在细胞分化、增殖、凋亡和生物生长发育中发挥着至关重要的调控作用，广泛参与生物的发育调控及免疫反应^[24]。视黄酸受体识别并结合其配体视黄酸后，通过结合靶基因启动子区特定序列调控靶基因表达，其靶基因上的特定序列即视黄酸反应元件。视黄酸反应元件具有一个核心序列[A/G]G[T/G]TCA，通过不同的排列方式和间隔构成DR1、DR2、DR5、ER8、IR0和IR1等多种重复序列组合。近年来人们对哺乳类视黄

表5 分子功能注释

分子功能	基因个数	占总数的百分比	基因 ID
protein tyrosine kinase activity	2	0.4%	CGI_10011454;CGI_10016631
protein tyrosine phosphatase activity	1	0.2%	CGI_10022937
protein tyrosine/serine/threonine phosphatase activity	1	0.2%	CGI_10022937
transmembrane receptor protein tyrosine kinase activity	1	0.2%	CGI_10026689

表6 生物过程注释

生物过程	基因个数	占总数的百分比	基因 ID
transmembrane receptor protein tyrosine kinase signaling pathway	2	0.4%	CGI_10026689;CGI_10013750
peptidyl-tyrosine dephosphorylation	1	0.2%	CGI_10022937
peptidyl-tyrosine phosphorylation	2	0.4%	CGI_10016631;CGI_10011454

表7 细胞定位注释

细胞定位	基因个数	占总数的百分比	基因 ID
membrane	21	5.1%	CGI_10002418;CGI_10003993;CGI_10001538;CGI_10021444;CGI_10001503;CGI_10015752;CGI_10025404;CGI_10017638;CGI_10025685;CGI_10026689;CGI_10013420;CGI_10019245;CGI_10014064;CGI_10028688;CGI_10009032;CGI_10017324;CGI_10001190;CGI_10004935;CGI_10010799;CGI_10023672;CGI_10020425
cytoplasm	14	3.4%	CGI_10019054;CGI_10013750;CGI_10001538;CGI_10005550;CGI_10022033;CGI_10013420;CGI_10009641;CGI_10010799;CGI_10015138;CGI_10001625;CGI_10015055;CGI_10004383;CGI_10006853;CGI_10007733
nucleus	10	2.4%	CGI_10027388;CGI_10025673;CGI_10015055;CGI_10002814;CGI_10012696;CGI_10026910;CGI_10022956;CGI_10007796;CGI_10019964;CGI_10014448
cell part	8	1.9%	CGI_10003993;CGI_10001190;CGI_10008150;CGI_10017324;CGI_10005377;CGI_10018855;CGI_10000675;CGI_10017638
intracellular part	7	1.7%	CGI_10024156;CGI_10002469;CGI_10001586;CGI_10024696;CGI_10018157;CGI_10028675;CGI_10007577
plasma membrane	5	1.2%	CGI_10002831;CGI_10026116;CGI_10011454;CGI_10026098;CGI_10028688
intracellular organelle	5	1.2%	CGI_10028595;CGI_10000979;CGI_10022818;CGI_10006853;CGI_10001625



注：图中通路标注为4.11.28，基因ID为CGI_10021536 (KEGG: Kyoto Encyclopedia of Genes and Genomes)

表8 KOG 注释

基因 ID	KOG 注释	功能类型定义
CGI_10011454	Protein tyrosine kinase	Signal transduction mechanisms
CGI_10013750	Protein tyrosine phosphatase, contains fn3 domain	Signal transduction mechanisms
CGI_10014308	Tyrosine kinase specific for activated (GTP-bound) p21cdc42Hs	Signal transduction mechanisms
CGI_10016631	Fibroblast/platelet-derived growth factor receptor and related receptor tyrosine kinases	Signal transduction mechanisms
CGI_10017027	Tyrosine kinase specific for activated (GTP-bound) p21cdc42Hs	Signal transduction mechanisms
CGI_10022656	Protein tyrosine phosphatase	Signal transduction mechanisms
CGI_10022937	Protein tyrosine phosphatase CDC14	Defense mechanisms
CGI_10026689	Nerve growth factor receptor TRKA and related tyrosine kinases	-
CGI_10002469	Protein tyrosine phosphatase	Signal transduction mechanisms

酸及视黄酸受体功能和作用的研究取得了一定的进展，但在其他非模式物种中，特别是无脊椎动物中的研究还相对匮乏。比如，长牡蛎中虽然对视黄酸受体有了初步的功能分析^[20,22]，但是其在识别并结合视黄酸后调控的下游基因还少有报道。

Perl 语言是一种功能丰富的计算机编程语言，具有简单易学、兼容性好以及文本处理能力强大等优点。鉴于这些优点，Perl 语言非常适合用于以序列文本为基础的生物信息学分析，可以针对个性化的生物问题编写脚本，帮助我们解决很多生物学问题。

本研究采用 Perl 语言编写了一个脚本，通过这个脚本可以对含有视黄酸反应元件的基因进行预测。本文以长牡蛎基因组为对象，使用该脚本预测到长牡蛎基因组中有 412 个基因启动子区含有视黄酸反应元件，暗示了长牡蛎视黄酸受体可能可以通过结合这些基因启动子区的 RAREs 调控这些基因的表达。对预测到的含有 RAREs 的基因进行生物信息学分析，结果发现，牡蛎中具有视黄酸反应元件调控的基因大多与单细胞或单体过程、多细胞系统发育、胚胎发育等功能相关。从结果中可以得出，长牡蛎中具有视黄酸反应元件调控的基因和其他生物中的功能类似，主要与生物生长发育和细胞分化增殖有关。在细胞组件类别中，大多分布在生物膜系统，这也表明了这些基因对外来因子进入细胞内具有调控作用，从 KEGG 通路图中发现，预测基因和小细胞肺癌通路相关，这也进一步验证了视黄酸对癌症因子的调控作用。分子功能注释和 KOG 注释结果发现，多个具有 RAREs 的基因编码长牡蛎酪氨酸激酶或酪氨酸磷酸酶，可能与一些关键蛋白酪氨酸残基

的磷酸化和去磷酸化修饰有关。此外，本研究预测的视黄酸反应元件所调控的基因中有一部分是和酪氨酸通路代谢相关的，曾有学者报道视黄酸可以促进酪氨酸酶基因的表达^[25]，本结果也进一步证明了视黄酸和酪氨酸的紧密关系。

综上所述，本研究编写了一个可以高通量预测基因启动子区视黄酸反应元件的 Perl 脚本，并且通过这个脚本在长牡蛎基因组中预测到 412 个启动子区具有视黄酸反应元件的基因，这为今后视黄酸调控长牡蛎的生理与发育分子机制的研究提供了理论基础。此外，这个脚本同样也可以用于其他已完成基因组测序的物种中，预测具有视黄酸反应元件的基因。

参 考 文 献

[1] LOHNES D, MARK M, MENDELSON C, et al. Function of the retinoic acid receptors (RARs) during development (I). Craniofacial and skeletal abnormalities in RAR double mutants[J]. Development, 1994, 120 (10): 2723-2748.

[2] 林建树,余英豪.全反视黄酸治疗急性早幼粒白血病:初步结果[J].国际输血及血液学杂志,1993(3):44-47.

[3] 夏明明,陈惠黎,蒋秉坤.异视黄酸对人肝癌细胞的诱导分化作用[J].细胞生物学杂志,1991,13 (3):127-130.

[4] 郎海滨,糜漫天,朱俊东.赫赛汀及 9-顺式视黄酸对乳腺癌联合抑制作用[J].中国公共卫生,2008,24 (3):311-312.

[5] GIGUÈRE V. Retinoic acid receptors and cellular retinoid binding proteins: complex interplay in retinoid signaling[J]. Endocrine Reviews, 1994, 15 (1): 61-79.

[6] TINI M, OTULAKOWSKI G, BREITMAN M L, et al. An everted repeat mediates retinoic acid induction of the gamma F-crystallin gene: evidence of a direct role for retinoids in lens development[J]. Genes Development, 1993, 7 (2): 295-307.

- [7] BASTIEN J, ROCHETTE-EGLY C. Nuclear retinoid receptors and the transcription of retinoid-target genes[J]. *Gene*, 2004, 328: 1–16.
- [8] WU W, NILES E G, LOVERDE P T. Thyroid hormone receptor orthologues from invertebrate species with emphasis on *Schistosoma mansoni*[J]. *BMC Evolutionary Biology*, 2007, 7: 150.
- [9] FORMAN B M, EVANS R M. Nuclear hormone receptors activate direct, inverted, and everted repeats[J]. *Annals of the New York Academy of Sciences*, 1995, 761: 29–37.
- [10] BALMER J E, BLOMHOFF R. A robust characterization of retinoic acid response elements based on a comparison of sites in three species[J]. *Journal of Steroid Biochemistry and Molecular Biology*, 2005, 96(5): 347–354.
- [11] JULIANA G M, KUMAR N E, DANIELA L, et al. A mollusk retinoic acid receptor (RAR) ortholog sheds light on the evolution of ligand binding[J]. *Endocrinology*, 2014(11): 4275.
- [12] NÄÄR A M, BOUTIN J M, LIPKIN S M, et al. The orientation and spacing of core DNA-binding motifs dictate selective transcriptional responses to three nuclear receptors[J]. *Cell*, 1991, 65(7): 1267–1279.
- [13] RAISHER B D, GULICK T, ZHANG Z, et al. Identification of a novel retinoid-responsive element in the promoter region of the medium chain acyl-coenzyme A dehydrogenase gene[J]. *Journal of Biological Chemistry*, 1992, 267(28): 20264–20269.
- [14] VERIN A D. Tyrosine phosphorylation and endothelial cell barrier regulation[J]. *American Journal of Pathology*, 2005, 166(4): 955–957.
- [15] YU J S. From discovery of tyrosine phosphorylation to targeted cancer therapies: The 2018 Tang Prize in Biopharmaceutical Science[J]. *Biomedical Journal*, 2019, 42(2): 80–83.
- [16] TADDEI M L, PARDELLA E, PRANZINI E, et al. Role of tyrosine phosphorylation in modulating cancer cell metabolism[J]. *Biochim Biophys Acta Reviews Cancer*, 2020, 1874(2): 188442.
- [17] STERKEL M, OLIVEIRA P L. Developmental roles of tyrosine metabolism enzymes in the blood-sucking insect *Rhodnius prolixus*[J]. *Proceeding of the Royal Society B Biological Sciences*, 2017, 284(1854): 20162607.
- [18] 姜秋云, 李玲玲, 王晓娜, 等. 双酚A暴露对不同黑色素含量长牡蛎的影响研究[J]. *海洋通报*, 2019, 38(3): 319–326.
- [19] 魏钰恒, 潘英, 刘圣, 等. 乳山长牡蛎(*Crassostrea gigas*)的抗性基因表达和生存环境的季节差异[J]. *海洋通报*, 2020, 39(5): 601–608.
- [20] HUANG W, WU Q, XU F, et al. Functional characterization of retinoid X receptor with an emphasis on the mediation of organotin poisoning in the Pacific oyster (*Crassostrea gigas*)[J]. *Gene*, 2020, 753: 144780.
- [21] ANDRÉ A, RUIVO R, FONSECA E, et al. The Retinoic Acid Receptor (RAR) in molluscs: function, evolution and endocrine disruption insights[J]. *Aquatic Toxicology*, 2019, 208: 80–89.
- [22] JIN K, JIN Q, CAI Z, et al. Molecular characterization of retinoic acid receptor CgRAR in Pacific oyster (*Crassostrea gigas*)[J]. *Front Physiol*, 2021, 12: 666842.
- [23] ZHANG G, FANG X, GUO X, et al. The oyster genome reveals stress adaptation and complexity of shell formation[J]. *Nature*, 2012, 490(7418): 49–54.
- [24] DAS B C, THAPA P, KARKI R, et al. Retinoic acid signaling pathways in development and diseases[J]. *Bioorganic Medicinal Chemistry*, 2014, 22(2): 673–683.
- [25] LI H, ZHU W. Effects of acidum vitamin A on tyrosinase activity[J]. *Chinese Medical Journal*, 2001, 114(4): 415–417.

(本文编辑: 袁泽轶)