

海豚基因组中硒蛋白基因的生物信息学预测

陈华^{①†}, 姜亮^{②†}, 倪嘉缙^{①*}, 刘琮^①, 张继红^③

① 深圳大学生命科学学院, 深圳 518060;

② 深圳大学光电学院, 深圳 518060;

③ 中国水产科学研究院黄海水产研究所, 青岛 266071

† 同等贡献

* 联系人, E-mail: jzni@szu.edu.cn

2011-04-03 收稿, 2011-10-28 接受

国家自然科学基金(31070731)和广东省自然科学基金(10151806001000023)资助

摘要 硒蛋白中的硒以硒代半胱氨酸(Sec)的形式存在. Sec 由传统的终止密码子 TGA 编码. 由于在可读框中难以区分 TGA 密码子在硒蛋白中的功能, 硒蛋白基因的预测非常困难. 利用本实验室建立的真核生物硒蛋白基因预测系统, 从海豚基因组中识别了包括含 2 个 Sec 的 2 型碘甲腺氨酸脱碘酶(DI2)、具有可变剪切编码区的 1 型碘甲腺氨酸脱碘酶(DI1)及谷胱甘肽过氧化物酶(GPX)、硒蛋白 P(SeIP)等 16 个硒蛋白, 从基因信息的角度探索了海豚中硒蛋白与其特殊生存环境及进化阶段的关系等问题.

关键词

硒
硒蛋白
海豚
基因组
基因识别

硒是一种人体必需的微量元素^[1], 缺硒与多种疾病密切相关^[2,3]. 硒在人体内主要是以硒蛋白形式存在并发挥功能. 硒蛋白家族的核心特征是含有被称为第 21 种氨基酸的硒代半胱氨酸(selenocysteine, Sec), 由传统终止码 TGA 编码. 同时, 在硒蛋白基因的 3'-非编码区(UTR)存在硒代半胱氨酸插入序列(SECIS), 指导编码区内的 TGA 解码为 Sec^[4]. 近年生物信息学、计算生物学快速发展, 但由于硒蛋白基因的特殊性, 常规基因分析方法无法预测硒蛋白, 因此必须重新分析预测已公布了基因组序列和基因注释的物种中的硒蛋白, 为研究硒蛋白的进化及生物功能提供数据和理论基础.

近 10 年来, 有关硒蛋白基因的生物信息学预测方法不断得到发展, 并先后应用于哺乳动物、鱼类、原生生物、古细菌和细菌等物种的硒蛋白识别^[5-9]. 通过生物信息学预测和随后的实验验证, 人类硒蛋白家族的数量从 14 种扩展到 25 种^[10], 在全球海洋样品超基因组中更是识别了 58 个硒蛋白家族^[11]. 上述硒蛋白预测方法的关键点是基于对 SECIS 结构的分

析预测, 即采用网上在线程序 SECISearch 进行 SECIS 分析. 然而该程序预测的 SECIS 结构类型均是从已有硒蛋白的 SECIS 结构中归纳而得到, 无法检测含有新的 SECIS 结构类型的硒蛋白基因. 针对上述问题, 我们通过对已有软件系统进行拓展、修改和再编辑, 建立了一种新的计算方法^[12], 适用于识别不同真核生物基因组中含有编码 Sec 的 TGA 的外显子, 并在此基础上组装、鉴定硒蛋白基因, 已成功预测了玻璃海鞘基因组中的硒蛋白组. 本文进一步采用该方法对海洋哺乳动物海豚的基因组进行了分析与计算.

生活在海洋中的海豚(*Tursiops truncatus*)在进化过程中保留着陆生哺乳动物胎生、哺乳和用肺呼吸等基本特征. 有报道称, 生物的生存环境, 如水生或陆生, 对物种硒蛋白的种类数量有显著影响^[13]. 多种陆生哺乳动物基因组中硒蛋白的研究工作已经开展并相继得以报道, 但国际上还没有关于海生哺乳动物硒蛋白种类和数量的报道. 本实验针对海豚生存环境和进化阶段的特殊性, 以生物信息学方法探索其硒蛋白数量和种类, 比较海洋脊椎动物与陆地脊椎动物硒蛋白

英文版见: Chen H, Jiang L, Ni J Z, et al. Bioinformatic prediction of selenoprotein genes in the dolphin genome. Chin Sci Bull, 2012, 57, doi: 10.1007/s11434-011-4970-5

的不同,有助于认识硒蛋白的发生发展,也可为海豚为代表的鲸目进化谜题的解答提供信息.

1 材料和方法

(i) 数据来源. 海豚的基因组序列从 Ensembl 生物信息学数据库(<http://www.ensembl.org>)获得. 海豚的基因组数据量为 2.40 G, 包括 111212 个基因组序列片段(scaffold). 但其基因组数据来源于概观式测序(survey sequencing), 是一种低序列覆盖率的测序方法(2.59 倍的序列覆盖率). 与高覆盖率测序(如人类基因组测序的覆盖率为 8.3 倍、小鼠的为 6.5 倍)相比, 缺乏长程连续性(多为短片段), 且往往存在缺失的部分基因组信息. 因此, 对进行硒蛋白预测的精度造成影响, 有待今后更为精确的测序数据发表后再进行校对.

(ii) 硒蛋白基因的分析预测. 海豚硒蛋白基因的识别工作在深超 21-C 超级计算机(深圳大学超级计算中心)上完成. 深超-21C 由 128 个计算结点组成, 总内存超过 128 GB, 总磁盘容量超过 10 TB. 在该硬件支持下, 我们利用图 1 所示的计算流程开展海豚硒蛋白基因识别工作.

预测步骤简述如下:

(1) 从基因组中获得所有 TGA 码, 对所有含 TGA 码的序列进行分析, 并构建含 TGA 的特殊外显子.

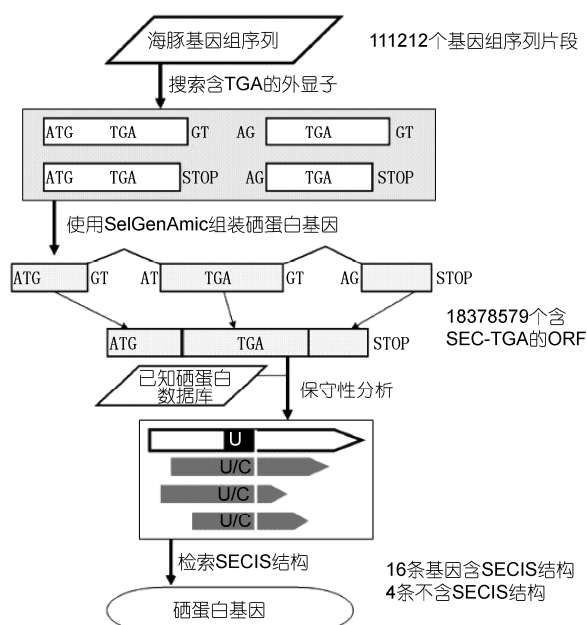


图 1 海豚基因组中硒蛋白基因识别流程图

外显子的边界是基因剪接位点, 如起始码 ATG、终止码 TGA 或 TAG 或 TAA、内含子剪切位点 AG/GT 等.

(2) 以含 TGA 码的外显子为中心, 用本实验室建立的 SelGenAmic 算法及分析系统组装外显子, 获得含 TGA 码的基因编码区.

(3) 将含 TGA 码的基因编码区翻译为氨基酸序列, 使用 BLASTp 程序, 与本实验室建立的已知硒蛋白库进行比对, 检索预测基因的 Sec 附近区域的局部保守性, 观察 Sec 残基附近序列的保守性以及该序列与硒蛋白相似序列之间是否存在 Sec/Sec 匹配或 Sec/Cys 匹配.

(4) 从海豚基因中提取预测基因编码区下游 DNA 序列, 采用 SECISearch 程序(<http://genome.unl.edu/SECISearch.html>)检索预测基因 3'-UTR 中是否存在 SECIS 结构, 验证所预测的硒蛋白基因的可信度. 对所预测的硒蛋白基因, 采用 SECISearch 程序绘制各硒蛋白基因的 SECIS 二级结构图, 其 COVE 值分布在 5.36~41.01 之间.

2 结果与讨论

2.1 预测方法的改进

本研究采用 SelGenAmic 算法及分析系统对海豚硒蛋白基因进行预测. 与我们以前用于疟蚊基因组中硒蛋白基因预测的方法相比, 该方法已有较大改进, 方法更为简单, 敏感性更高^[14]. 疟蚊硒蛋白基因预测方法中所使用的 TGA, 来源于原基因组注释中以 TGA 终止的基因片段, 因此, 上述方法在一定程度上受制于原注释基因的质量. 而本文使用的 TGA 码来源于对全基因组进行检索所获得的全部 TGA, 本方法是无遗漏地检索所有 TGA 作为编码 Sec 的密码子(简称为 Sec-TGA)的可能性, 这就极大地增加了方法的敏感度和独立性. 本方法还降低了硒蛋白基因预测中对 SECIS 结构检索结果的依赖性. 目前国际上其他实验室对硒蛋白基因的预测均将 SECIS 结构检索作为最关键的一个步骤, SECIS 结构无法被检测到的序列, 均被舍弃. 这样做存在一个严重缺陷: 对于那些无法被 SECISearch 程序检索到的特殊 SECIS 结构, 其所对应的硒蛋白基因将会在预测过程中被遗漏, 由此降低了发现新硒蛋白基因的可能性. 本文所采用的方法主要依赖于 SelGenAmic 算法对含 TGA 外显子的分析预测和对含 Sec-TGA 编码区

的组装,再通过寻找具有 Sec/Sec 或 Sec/Cys 匹配的同源序列,确定该基因是否为硒蛋白基因. SECIS 结构的检索是最后作为辅助步骤,进一步验证该预测基因是否为硒蛋白. 此方法最大限度地减少了对 SECIS 结构的依赖性,有助于发现新硒蛋白.

本文采用此方法在海豚基因组中发现了 4 条用现有软件在 3'-UTR 中未检出 SECIS 结构、但却极其类似硒蛋白的基因,如果这些基因能够在今后被实验证实为硒蛋白基因,则对拓宽硒蛋白的概念具有重要意义.

2.2 海豚硒蛋白的预测及分析

采用本实验室建立的硒蛋白基因分析预测方法,共组装了 18378579 个含 TGA 码的海豚基因编码区,最终从海豚基因组中共识别了 16 种硒蛋白.

表 1 列出了各硒蛋白基因所处的基因组 scaffold 的编号、Sec-TGA 在该片段中所处的位置、预测编码区的位置、SECIS 结构的位置和得分等信息(TGA 和 SECIS 的位置皆为首核苷酸的位置, SECIS 序列见图 2 中的二级结构). Sec-TGA 位置栏中的正负号表明该硒蛋白基因处在相应 scaffold 的正义链或负义链上.

表 1 从海豚基因组中预测的硒蛋白基因

硒蛋白名称	Scaffold 编号	Sec - TGA 位置	外显子位置	SECIS 起始位置	SECIS 得分
Iodothyron-ine deiodi-nase type 1a (DI 1a)	GeneScaf-fold_170	(+)257802	249629~249893, 257764~257804, 257805~257907, 260872~261074, 294213~294260	263509	23.21
Iodi-nase type 1b1 (DI 1c)	GeneScaf-fold_170	(+)257782	249629~249893, 257254~257273, 257764~257784, 257785~257856, 259484~259523, 260872~261074, 294213~294260	263509	23.21
Iodothyron-ine deiodi-nase type 2 (DI 2)	GeneScaf-fold_80	(-)1839818 (-)1840217	1839804~1839815, 1839816~1840214, 1840215~1840391, 1849096~1849317	1835077	27.55
Glutathione peroxidase 1 (Gpx1)	GeneScaf-fold_2343	(-)865611	864868~865259, 865479~865608, 865609~865801, 868161~868216, 869628~869758, 873662~873782, 878309~878464	864811	24.31
Glutathione peroxidase 4 (Gpx4)	GeneScaf-fold_2517	(+)5907	278~371, 5582~5792, 5870~5909, 5910~6014, 6121~6272, 6386~6410, 6570~6629, 6702~6734	6780	35.19
Selenophos-phate syn- thetase 2 (SPS2)	GeneScaf-fold_2971	(-)4858	3689~4855, 4856~5005	3137	23.53
Selenopro-teiin H (SelH)	GeneScaf-fold_1892	(+)383394	383388~383536, 403130~403301	382217	25.22
Selenopro-teiin I (SelI)	GeneScaf-fold_380	(+)306314	268182~268437, 289528~289602, 294012~294279, 301197~301245, 306251~306316, 306317~306345, 312423~312477	304961	22.43
Selenopro-teiin K (SelK)	GeneScaf-fold_161	(-)34588	21579~21759, 22734~22891, 23307~23514, 25640~25772, 25990~26083, 26600~26773, 27056~27188, 27732~27895, 29293~29463, 32382~32663, 33274~33334, 34581~34585, 34586~34667, 45096~45154	34214	29.68
Selenopro-teiin M (SelM)	GeneScaf-fold_2507	(-)327956	326653~326811, 327933~327953, 327954~327968, 329160~329307, 339805~339848	318719	5.36
Selenopro-teiin O (SelO)	GeneScaf-fold_261	(+)106123	74007~74128, 105417~106125, 106126~106134	106195	27.41
Selenopro-teiin P (SelP)	GeneScaf-fold_3416	(-)65001	43901~44136, 51934~52004, 64023~64235, 64973~64998, 64999~65188, 71390~71580	58691	16.12
Selenopro-teiin S (SelS)	Scaf-fold_88896	(+)12171	4445~4529, 5222~5356, 7318~7424, 10307~10385, 12094~12173, 12174~12179	12508	41.01
Selenopro-teiin T (SelT)	Scaf-fold_116247	(-)16653	16210~16650, 16651~16771, 23147~23304	16161	36.43
Selenopro-teiin X (SelX)	GeneScaf-fold_3522	(-)306582	293617~293700, 300237~300553, 306546~306579, 306580~306660, 307006~307154, 308943~308997	305121	23.98
Thioredoxin reductase 1b (Tr1b)	GeneScaf-fold_3383	(+)122146	113603~113719, 122086~122148, 122149~122154	122357	23.79
Thioredoxin reductase 2 (Tr2)	GeneScaf-fold_484	(-)389105	389097~389102, 389103~389226, 390073~390170, 390252~390323, 391654~391746, 391983~392078, 394351~394487, 401817~401874	388071	24.84

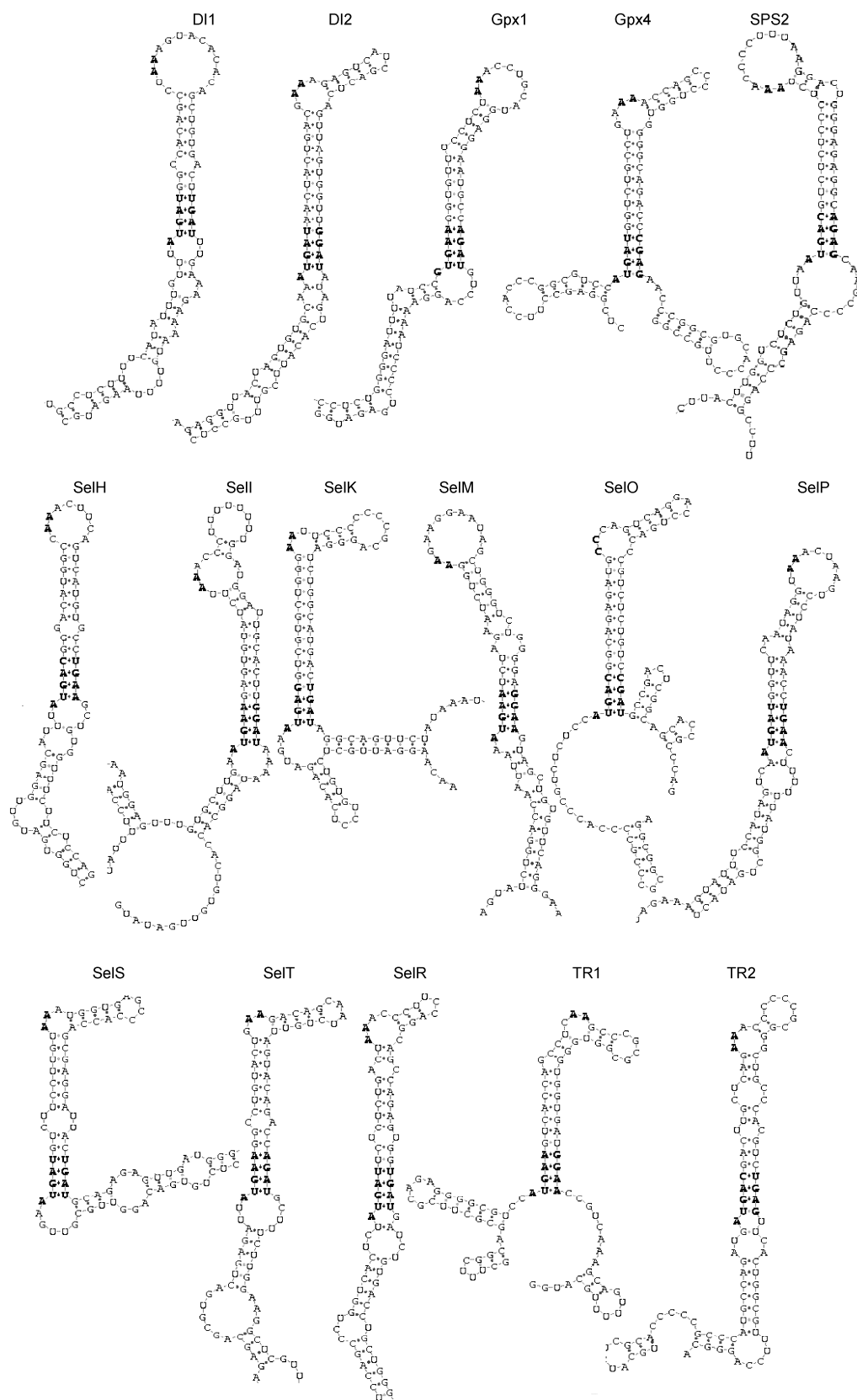


图 2 海豚硒蛋白基因的 SECIS 元件二级结构图

所预测到的 16 种硒蛋白均能在 NCBI 非冗余蛋白质库(non-redundant database, 简称 NR 库)中找到各自的相似性序列, 说明 Sec-TGA 码附近序列具有保守性和功能性. 在这 16 种硒蛋白基因编码区的下游序列中均检测出 SECIS 结构, 这些 SECIS 结构所处位置及二级结构见表 1 和图 2, 其中 14 个基因的 SECIS 结构为经典的 AUGA-AA-GA 型. 另外, Gpx1 的 SECIS 结构是特殊的 GUGA-AA-GA 型, 与线虫的硫氧还蛋白还原酶(TR)的 SECIS 相似^[15]; 硒蛋白 O (SeIO)的 SECIS 结构也是特殊的 AUGACC-GA, 与人类和小鼠的硒蛋白 M(SeIM)的 SECIS 相似^[7].

本实验发现的海豚硒蛋白 1 型碘甲腺氨酸脱碘酶 DI1a 和 DI1c 是 *DI1* 基因的 2 个可变剪切体, 图 3 所示的多序列比对说明该蛋白的 2 种剪切体 Sec (在图 3(a)和(b)中 U 代表 Sec)附近的局部序列均表现出与其他物种 DI1 不同亚型的相似性. 通过海豚 DI1a 和 DI1c 的基因结构比较, 说明由于不同剪切位点的使用, 导致基因组同一区段(图 3(c)和(d)中基因组坐标上的核酸位数标识)出现不同结构的基因亚型(DI1a 有 4 个外显子, TGA 出现在第 2 个外显子上; 而 DI1c 有 6 个

外显子, 其第 1 个和后 2 个外显子与 DI1a 相同, 而 TGA 出现在第 3 个外显子上). 该可变剪切基因的发现, 也体现了本实验硒蛋白预测方法的优点: 可发现同一条基因的不同可变剪切, 而国际上其他硒蛋白生物信息学识别方法未见报道有相似功能和结果.

此外, 如图 4 的基因结构图所示, 本实验发现的海豚硒蛋白 DI2 有 2 个 Sec-TGA, 都处在该基因第 2 外显子中, 分别位于该 Scaffold 的 1839818 和 1840217 核苷酸位置. DI2 的多序列比对分析表明, 海豚 DI2 的 2 个 Sec(U)附近局部序列均有高度保守性(图 5), 因此这 2 个 TGA 皆被翻译作 Sec. 已知多 Sec 的硒蛋白(含有 1 或 1 个以上 Sec 的硒蛋白)种类非常稀少(已报道有 SeIP 和 SeIL 等^[16,17]), 而多 Sec 意味着该 DI2 可能具有更强的氧化还原等功能活性, 启发我们在以后的工作中以该类蛋白为对象, 进一步研究硒蛋白功能活性与 Sec 数量之间的关系.

除上述 16 种硒蛋白, 本实验还预测到了 4 个特殊蛋白(表 2), 这些蛋白的编码区内均含有编码 Sec 的 TGA, 而且其 Sec 附近的氨基酸序列具有高保守性(图 6), 依据相似性分别归属于 4 种硒蛋白家族:

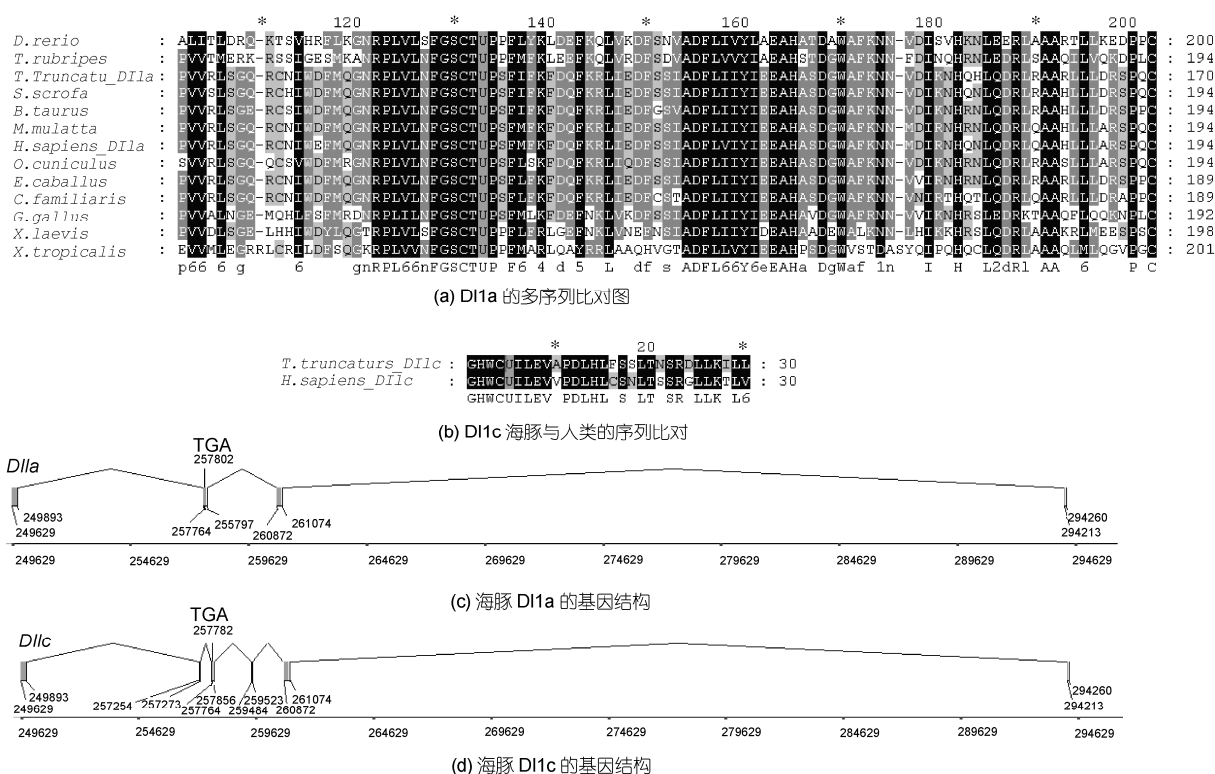


图 3 海豚 DI1a 和 DI1c 的多序列比对和基因结构

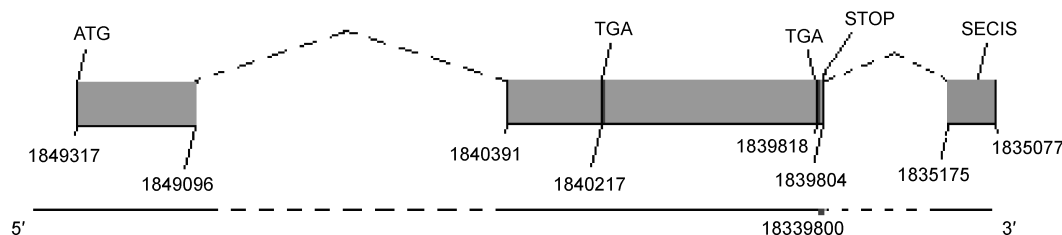


图 4 海豚 DI2 的基因结构图

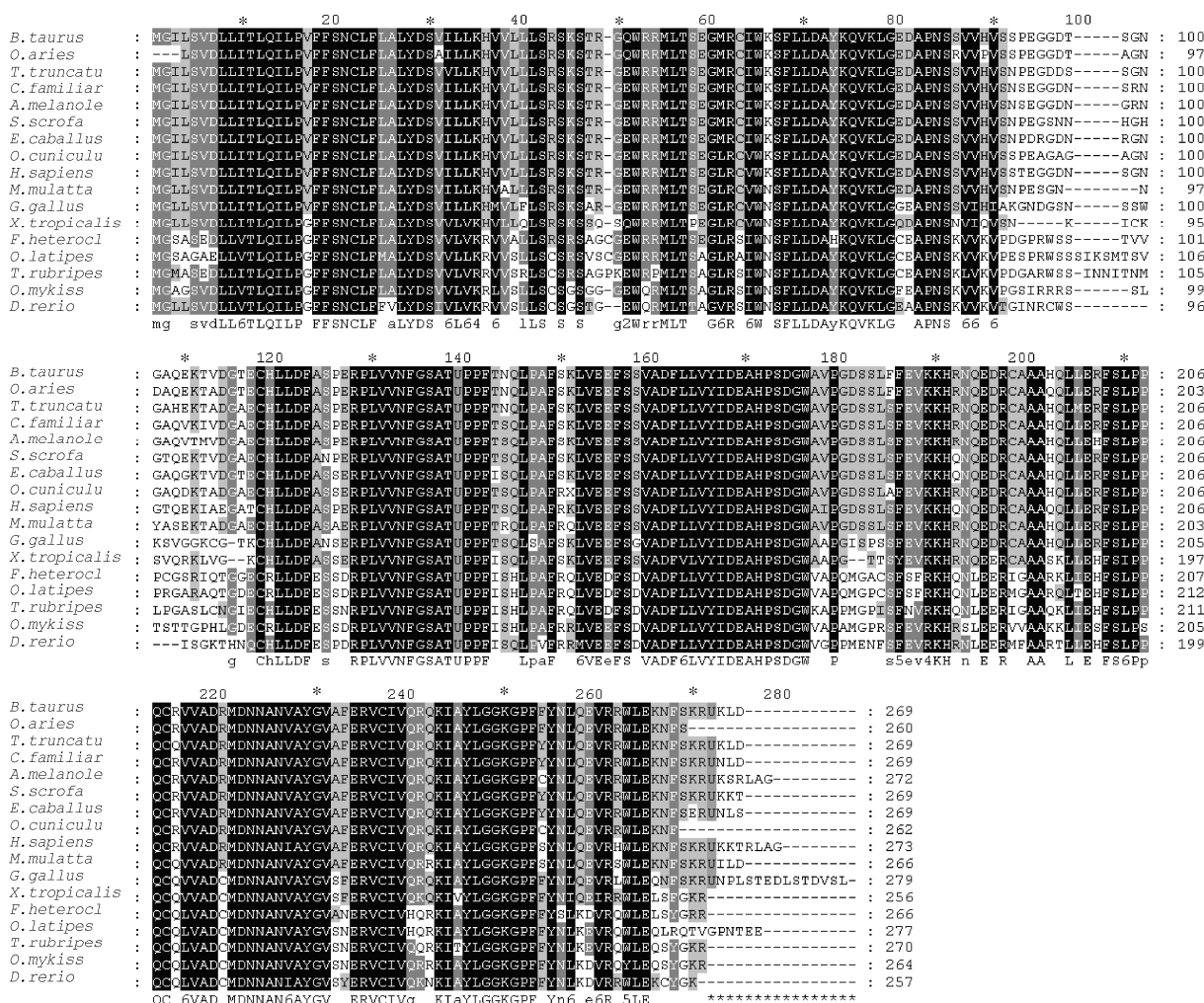


图 5 DI2 蛋白多序列比对图

表 2 海豚中预测到的 4 个潜在硒蛋白

蛋白名称	基因组测序片段	Sec-TGA 位置
15 kD selenoprotein (Sel 15)	GeneScaffold_466	(-)127180
Iodothyronine deiodinase type 3 (DI 3)	Scaffold_112862	(+)157377
Glutathione peroxidase 6 (Gpx6)	GeneScaffold_3345	(-)117122
Thioredoxin reductase 1a (TR1a)	GeneScaffold_238	(-)108067

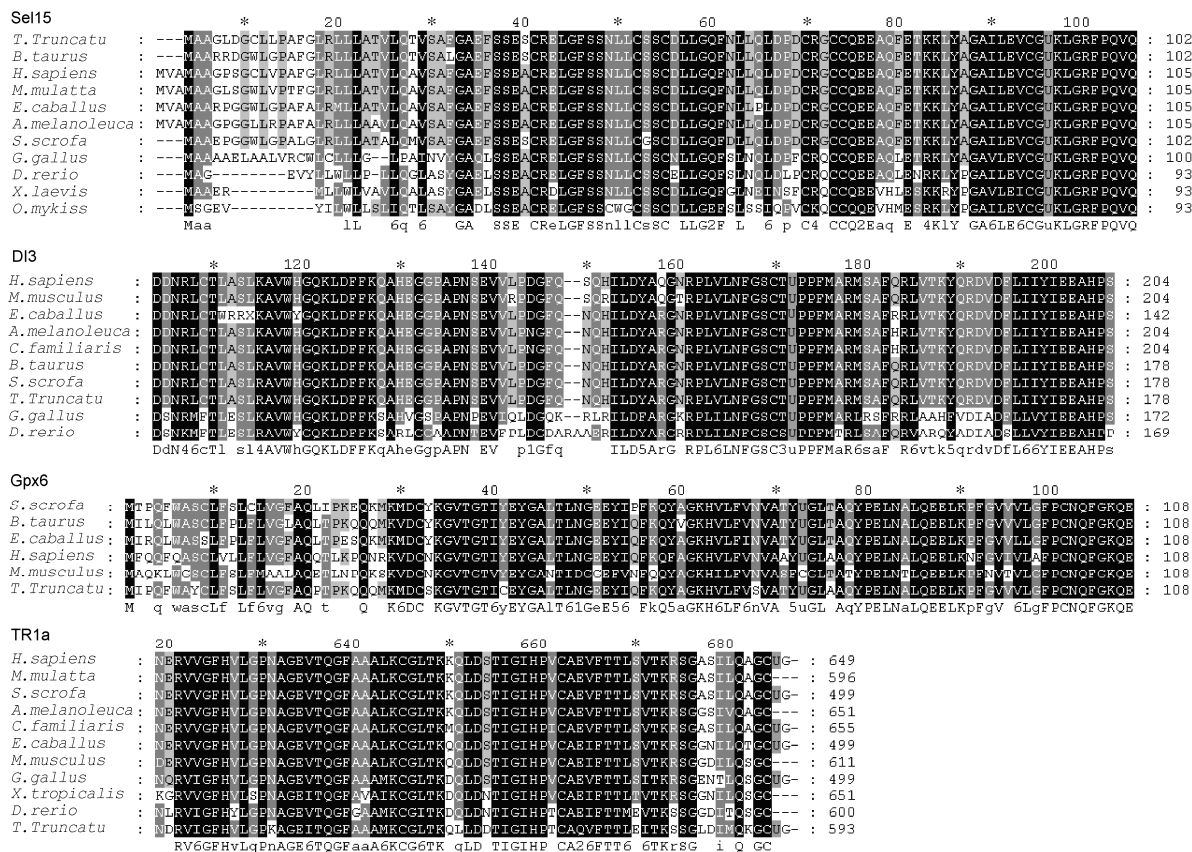


图6 无 SECIS 结构的 4 个潜在硒蛋白相似性比对图

15 kD 硒蛋白(Sel 15)、甲状腺脱碘酶 3 (DI 3)、谷胱甘肽过氧化物酶 6 (Gpx6)和硫氧还蛋白还原酶 1a (TR1a). 然而这 4 个蛋白的 3'-UTR 经 SECISearch 程序检索之后, 却未发现存在 SECIS 结构. 尽管如此, 这 4 种蛋白仍可能有潜在的 SECIS 结构, 因为: (i) 该蛋白的 SECIS 结构可能较为特殊, 现有的 SECIS 检索程序无法检测出; (ii) 海豚基因组测序的覆盖度或者准确性不足, 导致相应 SECIS 结构所在序列的缺漏或错误. 因为本实验将能否检测到 SECIS 结构, 作为判断预测结果是否为硒蛋白的一个必要指标, 故暂将这 4 条基因看作潜在的硒蛋白基因.

2.3 硒蛋白 DI2 的进化分析

由于本实验所获得的海豚硒蛋白 DI2 具有特殊的基因结构(2Sec), 而且 NCBI 的 NR 库中, 有较多的 DI2 同家族硒蛋白可用于多序列比对分析, DI2 被选为研究海豚中硒蛋白进化问题的对象. 图 5 所示的含海豚 DI2 的多序列蛋白比对图表明, 已记录于 NR 库的 DI2 硒蛋白主要出现在哺乳动物、两栖动物、鸟类

和鱼类中. 生活在水中的鱼类以及大量时间在水中生存的两栖动物的 DI2 只有 1 个 Sec, 没有序列末端的 Sec, 而绝大部分陆生的哺乳动物和鸟类都含有 2 个 Sec. 曾有报道称, 不同物种硒蛋白种类和数量受其生存环境是水生或陆生的影响较大^[13], 但本实验识别的海豚 DI2 给出了一个反例, 海豚并没有因为水生而如鱼类或两栖类般失去第 2 个 Sec, 而是更接近其他陆生动物或鸟类. 因此, 我们认为生存环境的改变对硒蛋白中 Sec 数量的影响有限.

根据图 5 中 DI2 的多序列比对结果, 采用 MEGA 3.1 软件, 构建了海豚硒蛋白 DI2 的进化树(图 7). 由图 7 可见, 海豚 DI2 与陆生偶蹄类动物具有最近的亲缘关系. 以海豚为代表的鲸目(Cetaceans)的起源问题一直是生物学家感兴趣而由难以解答的谜题. 依据化石和解剖学证据, 鲸长期被认为由一种已灭绝的外貌似狼的蹄类肉食动物中爪兽(Mesonychids)进化而来^[18]. 而 2005 年有关鲸类的分子进化研究报道指出, 鲸与陆生偶蹄类, 特别是河马具有最近的亲缘关系^[19]. 虽然缺乏河马的 DI2 序列信息, 但我们的计算

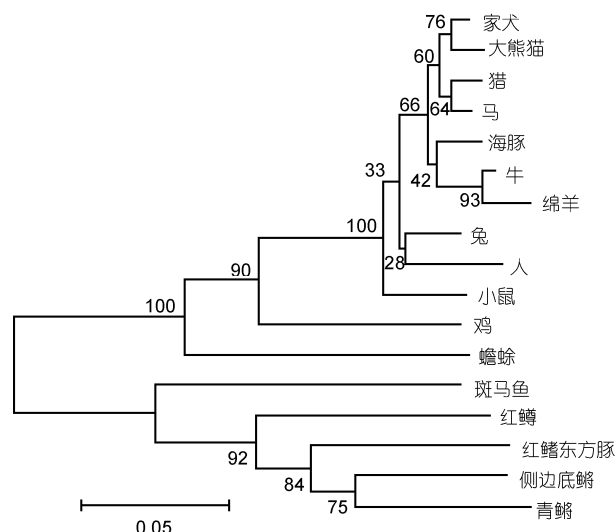


图7 海豚中硒蛋白DI2的进化树

结果同样支持海豚与偶蹄类动物，如牛、绵羊和马等具有较近的亲缘关系的观点。

3 结论

本文采用真核基因组预测系统，在海豚基因组中预测获得16个含SECIS结构的硒蛋白和4个未检测到SECIS结构的潜在硒蛋白。硒蛋白种类和数量都与人类等哺乳动物相似。本文发现了具有多剪切结构的海豚DI1基因的a亚型和c亚型结构，证明本方法具有能发现可变剪切硒蛋白基因的优点。而双Sec结构的DI2的发现，及相应多序列比对和系统发生学的分析结果，支持了海豚等鲸类动物与陆生偶蹄类有较近的亲缘关系这一观点。本文的研究方法和结果为研究海洋脊椎动物硒蛋白的分布与进化和Sec残基在硒蛋白中的演化提供了理论依据。相应的序列信息，如核酸氨基酸序列和SECIS结构位置序列等为具有重要生理功能硒蛋白(DI, GPX或TR等)的克隆表达及功能活性研究提供了必要的数据来源，有助于推动与硒蛋白有关的药物或人类健康的研究发展。

参考文献

- Low S C, Berry M J. Knowing when not to stop: Selenocysteine incorporation in eukaryotes. *Trend Biochem Sci*, 1996, 21: 203–208
- Stadman T C. Selenocysteine. *Annu Rev Biochem*, 1996, 65: 83–100
- 刘琼, 姜亮, 田静, 等. 硒蛋白的分子生物学及与疾病的关系. *化学进展*, 2009, 21: 819–830
- Kryukov G V, Kryukov V M, Gladyshev V N. New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. *J Biol Chem*, 1999, 274: 33888–33897
- Castellano S, Morozova N, Morey M, et al. Reconsidering the evolution of eukaryotic selenoproteins: A novel nonmammalian family with scattered phylogenetic distribution. *Embo Rep*, 2004, 5: 71–77
- Novoselov S V, Hua D, Lobanov A V, et al. Identification and characterization of Fep15, a new selenocysteine-containing member of the Sep15 protein family. *Biochem J*, 2006, 394: 575–579
- Korotkov K V, Novoselov S V, Hatfield D L, et al. Mammalian selenoprotein in which selenocysteine (Sec) in-corporation is supported by a new form of sec insertion sequence element. *Mol Cell Biol*, 2002, 22: 1402–1411
- Kryukov G V, Chapple C, Gladyshev V N. Selenium metabolism in zebrafish: Multiplicity of selenoprotein genes and expression of a protein containing 17 selenocysteine residues. *Genes Cells*, 2000, 5: 1049–1060
- Zhang Y, Gladyshev V N. An algorithm for identification of bacterial selenocysteine insertion sequence elements and selenoprotein genes. *Bioinformatics*, 2005, 21: 2580–2589
- Kryukov G V, Castellano S, Novoselov S V, et al. Characterization of mammalian selenoproteomes. *Science*, 2003, 300: 1439–1443
- Zhang Y, Gladyshev V N. Trends in selenium utilization in marine microbial world revealed through the analysis of the Global Ocean Sampling (GOS) Project. *PLoS Genet*, 2008, 4: e1000095
- Jiang L, Liu Q, Ni J. *In silico* identification of the sea squirt selenoproteome. *BMC Genom*, 2010, 11: 1471–2164–11–289
- Lobanov A V, Fomenko D E, Zhang Y, et al. Evolutionary dynamics of eukaryotic selenoproteomes: Large selenoproteomes may associate with aquatic life and small with terrestrial life. *Genome Biol*, 2007, 8: R198
- 刘琼, 姜亮, 陆平, 等. 蚊虫基因组中新硒蛋白的计算机识别. *中国科学 C 辑: 生命科学*, 2006, 36: 451–458
- Taskov K, Chapple C, Kryukov G V, et al. Nematode selenoproteome: The use of the selenocysteine insertion system to decode one codon in an animal genome? *Nucleic Acids Res*, 2005, 33: 2227–2238
- Tujebajeva R M, Harney J W, Berry M J. Selenoprotein P expression, purification, and immunochemical characterization. *J Biol Chem*, 2000, 275: 6288–6294
- Shchedrina V A, Novoselov S V, Malinouski M Y, et al. Identification and characterization of a selenoprotein family containing a diselenide bond in a redox motif. *Proc Natl Acad Sci USA*, 2007, 104: 13919–13924
- Erfurt J, Averianov A. Enigmatic ungulate-like mammals from the eocene of central Asia. *Naturwissen-Schafien*, 2005, 92: 182–187
- Price S A, Bininda-Emonds O R, Gittleman J L. A complete phylogeny of the whales, dolphins and even-toed hoofed mammals (Cetartiodactyla). *Biol Rev*, 2005, 80: 445–473