Article
Chinese Science Bulletin
Bioinformatics
May 2012 Vol.57 No.13: 1533–1541
doi: 10.1007/s11434-011-4970-5

Bioinformatic prediction of selenoprotein genes in the dolphin genome

CHEN Hua[1†], JIANG Liang[2†], NI JiaZuan[1*], LIU Qiong[1] & ZHANG JiHong[3]

[1] *College of Life Sciences, Shenzhen University, Shenzhen 518060, China;*
[2] *College of Optoelectronic Engineering, Shenzhen University, Shenzhen 518060, China;*
[3] *Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao 266071, China*

Received April 3, 2011; accepted October 28, 2011; published online February 23, 2012

Selenium (Se), an essential trace element *in vivo*, is present mainly as selenocystein (Sec) in various selenoproteins. The Sec residue is translated from an in-frame TGA codon, which traditionally functions as a stop codon. Prediction of selenoprotein genes is difficult due to the lack of an effective method for distinguishing the dual function of the TGA codon in the open reading frame of a selenoprotein gene. In this article a eukaryotic bioinformatic prediction system that we have developed was used to predict selenoprotein genes from the genome of the common bottlenose dolphin, *Tursiops truncatus*. Sixteen selenoprotein genes were predicted, including selenoprotein P and glutathione peroxidase. In particular, a type II iodothyronine deiodinase was found to have two Sec residues, while the type I iodothyronine deiodinase gene has two alternative splice forms. These results provide important information for the investigation of the relationship between a variety of selenoproteins and the evolution of the marine-living dolphin.

selenium, selenoprotein, selenocystein, dophin, genome, gene prediction

Citation: Chen H, Jiang L, Ni J Z, et al. Bioinformatic prediction of selenoprotein genes in the dolphin genome. Chin Sci Bull, 2012, 57: 1533–1541, doi: 10.1007/s11434-011-4970-5

Selenium (Se) is an essential trace element [1] and Se-deficiency is related to several diseases [2,3]. Selenium *in vivo* is primarily present in various selenoproteins, which generally function as antioxidants to maintain the balance of the redox state. The active site of selenoproteins contains selenocysteine (Sec), the 21st amino acid, which is encoded by a TGA codon, traditionally a termination codon. Meanwhile, a stem-loop structure designated as the Sec insertion sequence (SECIS) element is necessary for introducing Sec into selenoproteins. The SECIS element is located in the 3'-untranslated region (UTR) of selenoprotein genes in eukaryotes and archaea, or located immediately downstream of the in-frame TGA in bacteria [4]. In recent years, the rapid development of bioinformatics and computational biology has made gene identification from genome sequences faster and easier; however, the special structure of selenoprotein genes leads to difficulties in their identification with bioinformatics methods. Therefore, the genome sequences and other annotated bio-information in databases, such as GenBank from NCBI need to be re-analyzed, to investigate the evolution and function of selenoproteins.

Over the past decade, several bioinformatic methods for the prediction of selenoprotein genes have been developed, and have been used for the identification of selenoproteins from many species, including human, fish, protozoa, archaea and bacteria [5–9]. The number of human selenoproteome members increased from 14 to 25, via bioinformatic predictions and subsequent experimental verification [10]. In addition, up to 58 selenoprotein families have been identified recently in the Global Ocean Sampling (GOS) Project [11]. The online SECIS search analysis is the key procedure of the methods described above. However, the model used by this online SECIS search program was generalized from

†These authors contributed equally to the work.
*Corresponding author (email: jzni@szu.edu.cn)

all known SECIS elements. Therefore, these methods cannot be used to find selenoproteins which have new types of SECIS structure. To solve this problem, we developed a new method based on expanded, modified and re-edited versions of reported programs, which is suitable for the identification of exons containing TGA codons coding for Sec, and subsequent identification of selenoprotein genes. This method was successfully used to identify the selenoproteome from the sea squirt genome [12]. With this improved method, the genome of dolphin, a marine mammal, was searched and analyzed.

After millions of years of evolution, the bottlenose dolphin (*Tursiops truncatus*), a marine mammal, has retained several characteristics of terrestrial mammals, such as being viviparous, breast-feeding and breathing with lungs. The environment in which an animal lives, such as aquatic or terrestrial, has a key impact on the number of members and the variety of selenoproteins of an organism [13]. A number of studies of selenoproteins in terrestrial mammals have been reported in recent years, but no study on marine mammal selenoproteins has previously been reported. In this article, focusing on the special marine environment and evolutionary position of dolphins, a bioinformatic method was used to investigate the variety and numbers of selenoproteins from the dolphin genome. The comparison between selenoproteins in marine and terrestrial vertebrates, will provide information for exploring the evolution of selenoproteins, and also for answering the evolutionary puzzle of Cetaceans.

1 Materials and methods

1.1 Sequence databases and resources

The genome sequences of dolphin were obtained from the Ensembl Project Genome Databases (http://www.ensembl.org). The text file size of the dolphin genome data is about 2.40 Gb, containing approximately 111212 scaffolds. These data were obtained by a survey sequencing approach of which the assembly coverage is low (2.59X). Compared with high-coverage assembly sequencing data (human with 8.3X assembly and mouse 6.5X), more short-segment sequences were present, leading to higher probability of partial missing annotation of the genome sequences. Therefore, the low-coverage data had a negative impact on our study, and additional searches will be performed when higher coverage assembly data are released.

1.2 Identification procedures

The identification of dolphin selenoproteins was performed on the Deep-Super 21-C of the super computer center of Shenzhen University. Deep-Super 21-C is composed of 128 grids with 128 Gb memories and more than 10 Tb storage volume. With this powerful hardware platform, we accomplished the identification procedures of dolphin selenoproteins shown in Figure 1.

(i) The whole genome sequence of dolphin was scanned to find all TGA codons. Special exons with TGA codons and other normal gene splice signals such as the start codon ATG, stop codons, TGA\TAG\TAA, and splice sites AG\GT were identified.

(ii) Using the SelGenAmic selenoprotein gene assembly algorithm reported in our earlier work, open reading frames, including in-frame TGA codons, were identified on the basis of special exons with TGA codons found in the previous step.

(iii) All coding sequences, including in-frame TGA codons, were translated into amino acid sequences. Local sequences flanking the Sec residues were extracted for detecting similarity in the selenoprotein database built in our laboratory by the BlastP program. Those sequences with conservation in the local regions flanking the Sec residue were screened out, and alignments containing Sec/Cys pairing (simplified as U/C pair), i.e., the Sec-containing local sequence, must have its homologous sequences containing Cys residues in the position of Sec in multiple alignments.

(iv) The downstream sequences of these predicted genes were extracted from the genome sequence, and searched by theSECISearch program (http://genome.unl.edu/SECISearch.html, 2.19). The secondary RNA structure and Cove score of each predicted selenoprotein gene were obtained. The Cove scores of dolphin SECIS elements were distributed between 5.36 and 41.01.

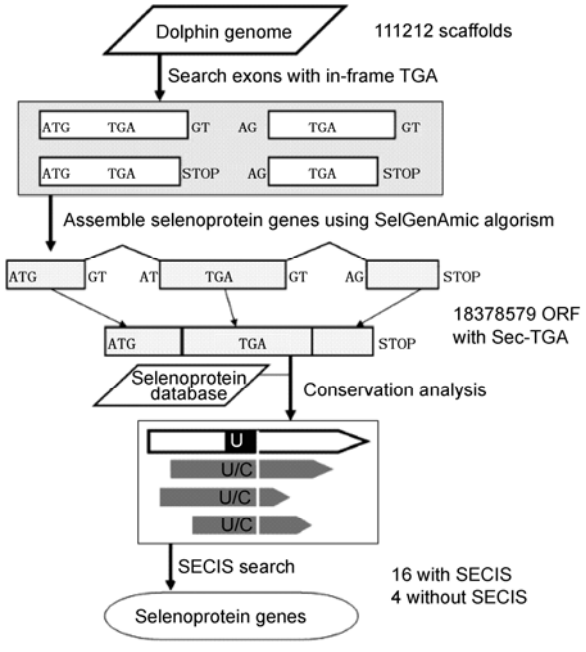

Figure 1 Flow chart describing general procedures for the identification of selenoprotein genes from the dolphin genome.

2 Results and discussion

2.1 Relative advantage of prediction method

A method based on the SelGenAmic algorithm was used to predict selenoprotein genes from the dolphin genome. Compared with the method used for the identification of selenoproteins from the anopheles genome [14], the method used in this paper has several advantages, leading to a simpler and more accurate method. In the anopheles study, all TGA codons obtained from reported data were misannotated as STOP codons. Therefore, this method is highly influenced by the quality of the released anopheles data. Using the method in this paper, all TGA codons in dolphin genome sequences were scanned and tested without any skipping. This method improved the sensitivity of detecting selenoproteins from a genome due to all TGAs in the genome being investigated for the possibility of coding a Sec residue.

This method reduced the dependency of SECIS element search results for the prediction of selenoprotein genes. Most selenoprotein prediction methods consider the SECIS element search procedure as an indispensable step; all sequences for which no SECIS element was detected would be discarded. A disadvantage occurs under this strategy; special selenoproteins with non-canonical structures that cannot be detected by the reported SECIS search programs would be lost. Our method has the merit of independently finding possible ORFs for all TGA codons without the help of SECIS information. This is because the SelGenAmic algorithm could enumerate all ORFs for all TGA codons in a genome. Therefore, even with special SECIS structures, these selenoproteins will not be lost in the analysis using the SelGenAmic-based method. They can be found during protein sequence conservation analysis. In this study, four dolphin genes, for which no SECIS element was found in their 3′-untranslated regions, were identified. But the conservation analysis shows that all of these protein sequences are very similar with known selenoproteins. In other words, if these proteins are shown experimentally to be real selenoproteins, it could be a significant discovery expanding our concept of selenoprotein genes.

2.2 Prediction and analysis of dolphin selenoproteins

Using the method we developed according to the SelGenAmic algorithm, 18378579 open reading frames, including in-frame TGAs were assembled, and 16 selenoproteins were found in the dolphin genome. Important information is shown in Table 1.

This information includes the genome sequence (scaffold) from which the selenoprotein gene was identified, the position of the Sec-coding TGA codon, SECIS element and ORF, and the COVE score of each SECIS element. The minus or positive sign in the brackets before Sec-TGA position numbers indicate the gene located on the negative or positive sense strand. Similar protein sequences can be found for all of these 16 selenoproteins in the non-redundant (NR) database from NCBI, indicating the conserved and functional regions flanking the Sec residue of these predicted selenoproteins. In addition, SECIS elements were found in the downstream sequences of these proteins, the positions and secondary structures of which are shown in Table 1 and Figure 2.

Among them in Figure 2, 14 are classic AUGA-AA-GA type SECIS elements, the other 2 are special types. The SECIS element of dolphin Gpx1 is GUGA-AA-GA, which is similar to that of nematode thioredoxin reductase (TR) [15], and the SECIS element of dolphin selenoprotein O (Sel O) is AUGA-CC-GA, which is similar to that of human and mouse SelO [7].

Subtype a and subtype c of iodothyronine deiodinase type 1 (DI1) are two alternative splicing forms of DI1. As shown in Figure 3, the local sequence flanking the Sec residue (indicated by U) of different subtypes of dolphin DI1 show similarity with corresponding splicing forms of other species. As shown in Figure 3(c) and (d), different splicing forms of selenoprotein genes with various gene structures could be present in the same region of genomic sequence. There are four exons in DI1a, the TGA codon is in the second exon. There are six exons in DI1c, the first and the last two exons are the same as DI1a, the TGA codon is in the third exon. The discovery of DI1a and DI1c, reflected another merit of our method, is able to find alternatively splicing selenoproteins genes. Because earlier methods, like Geneid_SP, only considered one of these two TGAs as the best Sec codon, they could not detect the alternative splicing form of DI2 containing another Sec-TGA codon.

As shown in the gene structures of Figure 4, two Sec residues were also found in dolphin DI2, both of which were located in the second exon (with positions 1838718 and 1840217). Multiple alignments of DI2 show that both of them have local sequence conservation (Figure 5). There are very rare selenoproteins (SelP and SelL) that have 2 or more than 2 Sec residues [16,17]. More Sec residues mean higher redox activity, and further work could be done based on this protein to investigate the relationship between function and numbers of Sec residues in selenoproteins.

Four other special proteins were predicted (Table 2). As shown in Figure 6, all of these 4 proteins present homology with known selenoproteins in their amino acid sequence regions flanking Sec residues. According to the multiple alignment results, these proteins were classified as the 15 kD selenoprotein (Sep15), iodothyronine deiodinase type 3 (DI3), glutathione peroxidase 6 (Gpx6) and thioredoxin reductase 1a (TR1a). However, no SECIS element was detected by the available programs, such as SECISearch, in their 3′ untranslated regions. Nevertheless, genes of these four proteins may still have potential SECIS elements, because of the following reasons. Firstly, the structures of their potential SECIS elements could be non-canonical,

Table 1 Helminth species (strains) with complete mitochondrial genomes characterized to date

Selenoprotein	Scaffold	Sec-TGA position	Exon position	SECIS position	SECIS COVE score
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-tein H (SelH)	GeneScaf-fold_1892	(+)383394	383388–383536, 403130–403301	382217	25.22
Selenopro-tein I (SelI)	GeneScaf-fold_380	(+)306314	268182–268437, 289528–289602, 294012–294279, 301197–301245, 306251–306316, 306317–306345, 312423–312477	304961	22.43
Selenopro-tein 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-tein M (SelM)	GeneScaf-fold_2507	(–)327956	326653–326811, 327933–327953, 327954–327968, 329160–329307, 339805–339848	318719	5.36
Selenopro-tein O (SelO)	GeneScaf-fold_261	(+)106123	74007–74128, 105417–106125, 106126–106134	106195	27.41
Selenopro-tein P (SelP)	GeneScaf-fold_3416	(–)65001	43901–44136, 51934–52004, 64023–64235, 64973–64998, 64999–65188, 71390–71580	58691	16.12
Selenopro-tein S (SelS)	Scaf-fold_88896	(+)12171	4445–4529, 5222–5356, 7318–7424, 10307–10385, 12094–12173, 12174–12179	12508	41.01
Selenopro-tein T (SelT)	Scaf-fold_116247	(–)16653	16210–16650, 16651–16771, 23147–23304	16161	36.43
Selenopro-tein 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

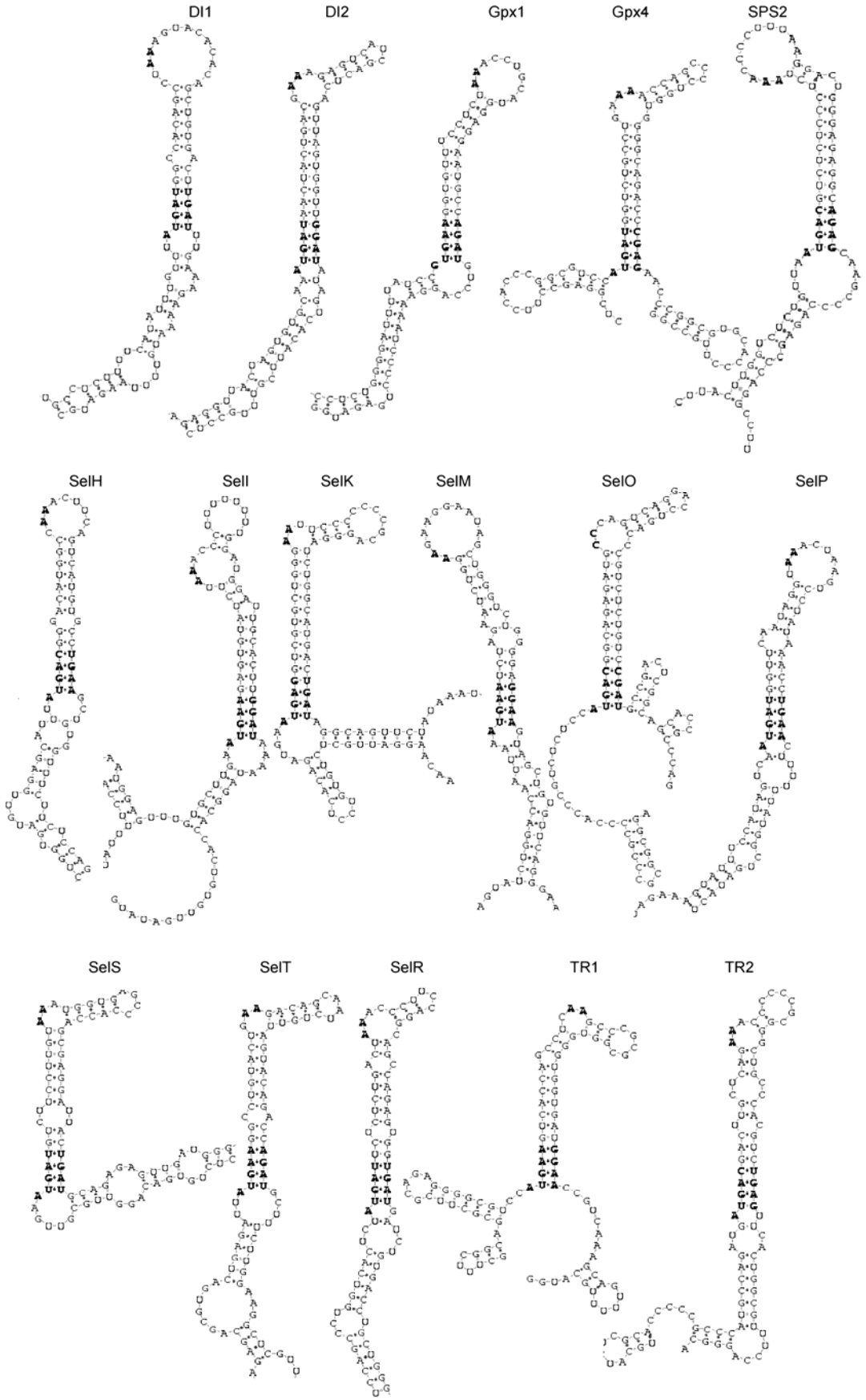

Figure 2 Secondary structure of the SECIS elements of dolphin selenoprotein genes.

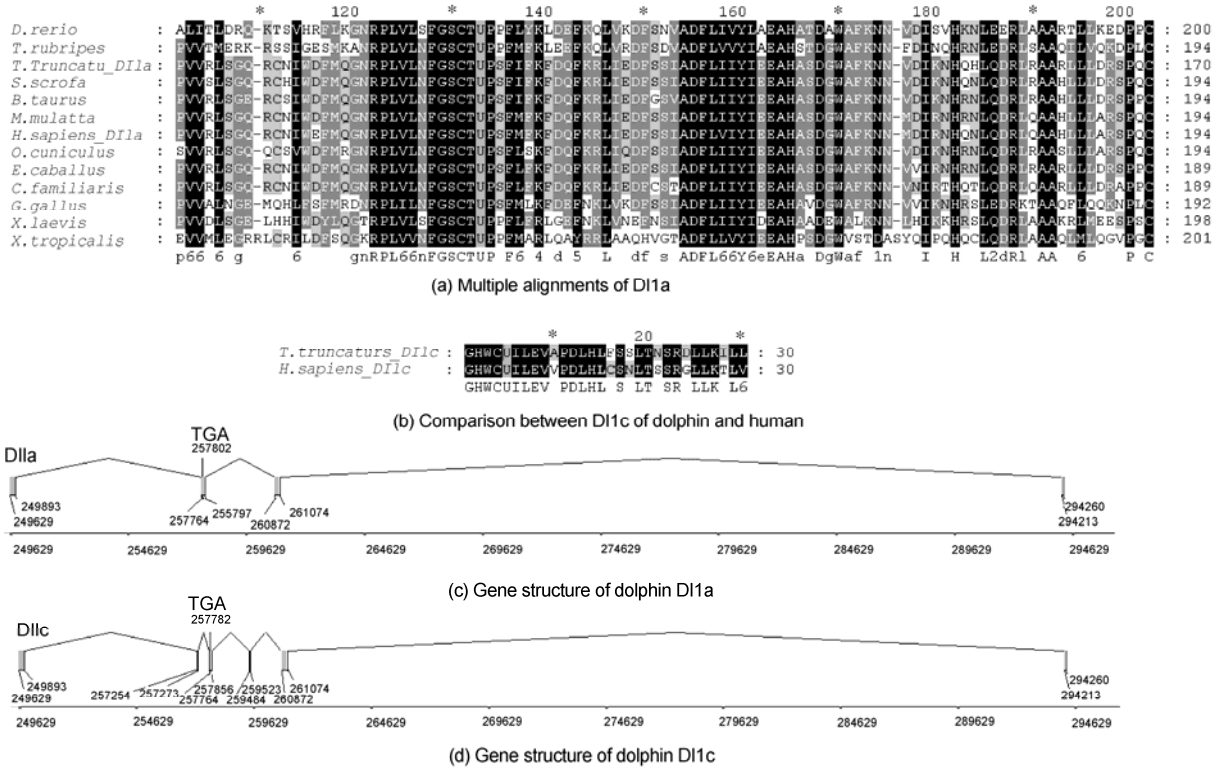

Figure 3 Multiple alignments and gene structures of Dolphin DI1a and DI1c.

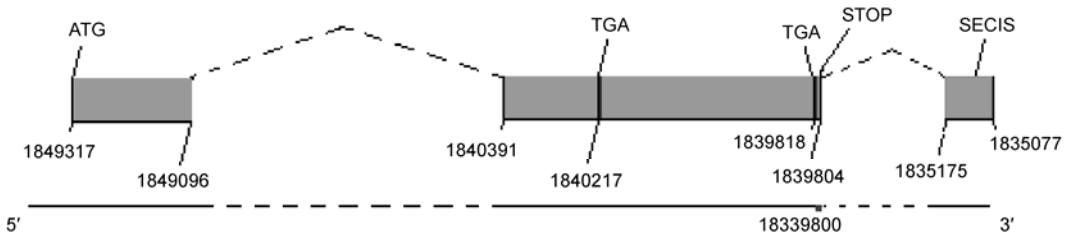

Figure 4 Gene structure of dolphin DI2.

which cannot be detected by the known SECIS model. Secondly, the sequences including the SECIS elements could be mis-annotated or totally discarded because of the low-coverage assembly of the dolphin genome. Because no SECIS elements were detected, all of these 4 proteins are temporarily classified as potential selenoproteins.

2.3 Evolutionary analysis of DI2

Considering the special gene structure (2 Sec residues) and the abundance of DI2 from other organisms in the NR database, the DI2 of dolphin was chosen for evolutionary analysis. As the multiple alignment shows (Figure 5), all reported DI2 genes were mainly present in mammals, amphibians, birds, and fishes. The DI2 of aquatic living fishes and amphibians, which spend most of their time under water, have only one Sec, while 2 Sec residues are found in most terrestrial mammals and birds. As we discussed in the introduction, the living environment, such as aquatic or terrestrial, is a key impact element for the number of members and the variety of selenoproteins of an organism [13]. However, the situation occurring for dolphin DI2 shows a counter-example for this hypothesis; the marine environment did not cause the loss of the second Sec of dolphin DI2, as is the case for fishes and amphibians. Therefore, aquatic or terrestrial environments have limited influence on the Sec quantity in various species.

According to the multiple alignments shown in Figure 6, we constructed a phylogenetic tree of DI2 using the program MEGA 3.1. As shown in Figure 7, DI2 of dolphin is most closely related to D12 of terrestrial cloven-hoofed mammals. The evolutionary origin of cetaceans has been an interesting puzzle and has been difficult to answer for a long time. Based on fossil and anatomical evidence, whales have long been considered to be evolved from extinct wolf-like hoofed animals named mesonychids [18]. A report on the

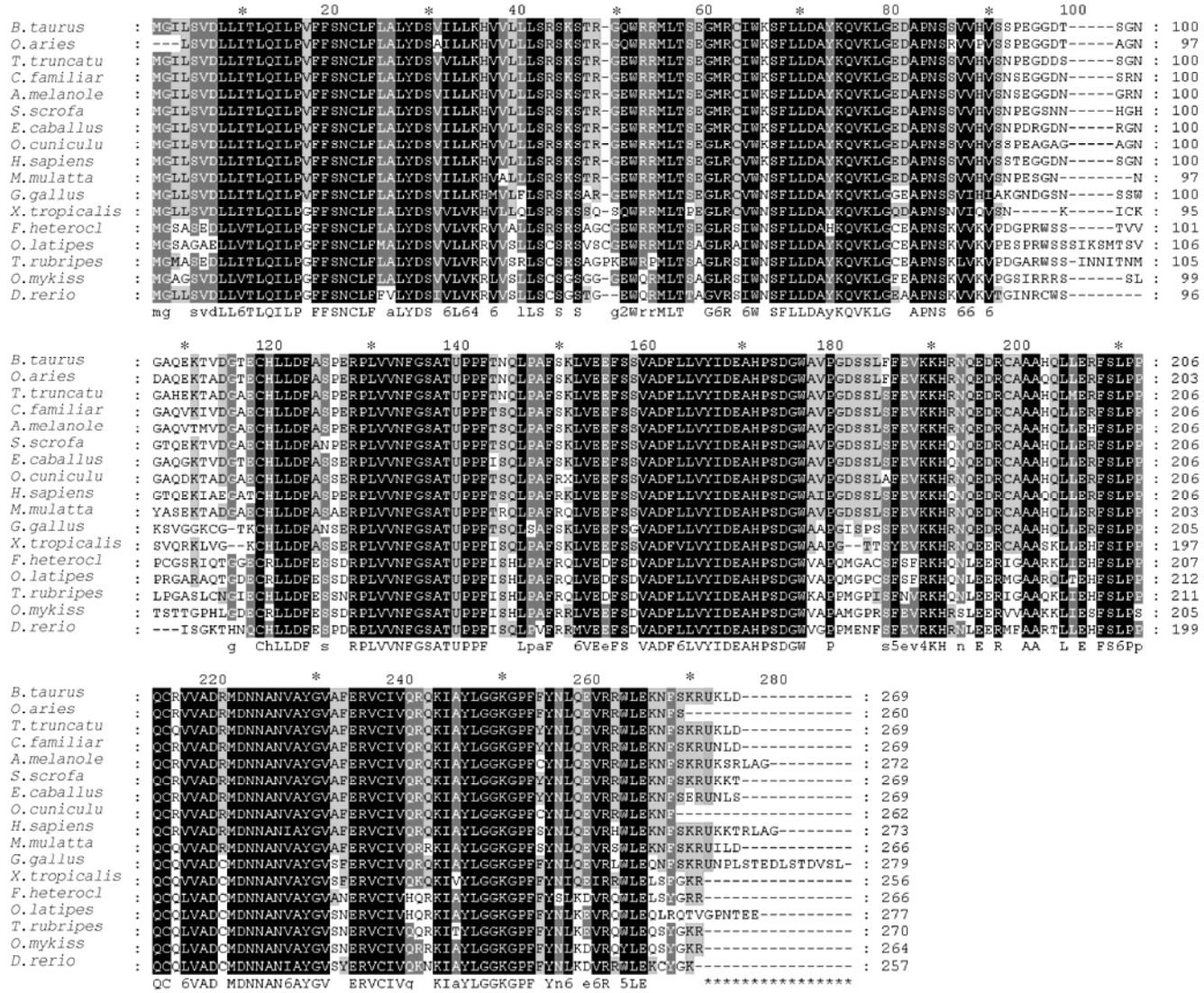

Figure 5 Multiple alignment of DI2.

Table 2 Four potential selenoproteins predicted from the dolphin genome

Protein	Scaffold	Sec-TGA position
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

molecular evolution of whales published in 2005 pointed out that whales are most closely related to terrestrial cloven-hoofed mammals,, especially the hippopotamus [19]. Although no sequence information is available for the hippopotamus, the phylogenetic analysis done in this paper shows the closest relationship is between dolphin and other hoofed animals, such as cattle, sheep and horse, which is additional evidence to support the hypothesis reported in 2005 [19].

3 Conclusions

Using the eukaryotic selenoprotein gene prediction method, 16 selenoprotein genes with SECIS elements and 4 potential selenoproteins without detectable SECIS elements were identified in the dolphin genome. The variety and numbers of selenoproteins found in dolphin are similar to those in human and other mammals. The different subtypes of DI1 found in dolphin proved that our method has the ability to

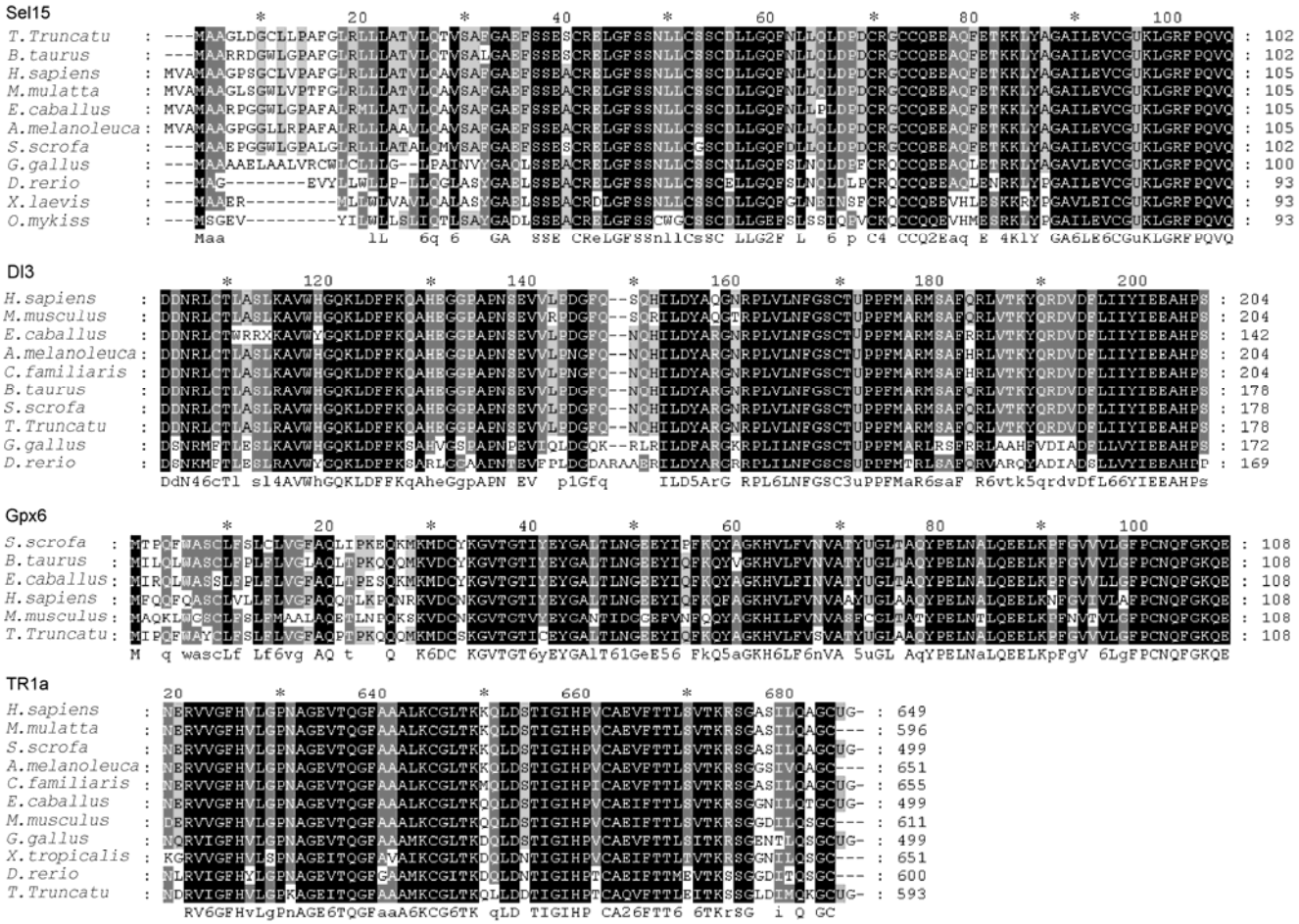

Figure 6 Multiple alignment of four potential dolphin selenoproteins.

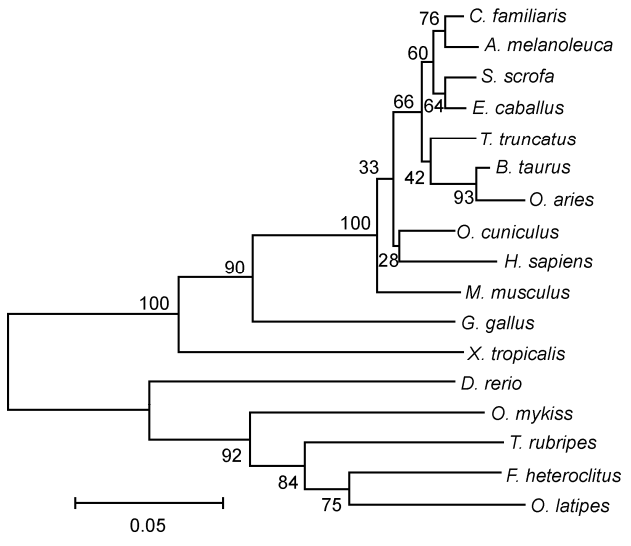

Figure 7 Phylogenetic tree of DI2.

find selenoproteins with alternative splice forms. The discovery of DI2 with 2 Sec residues, and the subsequent evolutionary analysis of this protein family supported the view that whales are most closely related with terrestrial cloven-hoofed mammals. The method and results of this paper, provides an important theoretical basis for the exploration of the evolution and distribution of selenoproteins in marine vertebrates. The biological information, such as amino acid sequences and SECIS structures, reported in this paper are important data for the functional research of human diseases related to selenoproteins, such as DI, Gpx and TR, and to selenium related drug development.

This work was supported by the National Natural Science Foundation of China (*31070731*) *and the Natural Science Foundation of Guangdong Province* (*10151806001000023*).

1 Low S C, Berry M J. Knowing when not to stop: Selenocysteine incorporation in eukaryotes. Trends Biochem Sci, 1996, 21: 203–208
2 Stadman T C. Selenocysteine. Annu Rev Biochem, 1996, 65: 83–100
3 Liu Q, Jiang L, Tian J, et al. The molecular biology of selenoproteins and their effects on diseases. Prog Chem, 2009, 21: 819–830
4 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
5 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

6 Novoselov S V, Hua D, Lobanov A V, et al. Identification and characterization of Fep15, a new selenocys-teine-containing member of the Sep15 protein family. Biochem J, 2006, 394: 575–579
7 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
8 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
9 Zhang Y, Gladyshev V N. An algorithm for identification of bacterial selenocysteine insertion sequence elements and selenoprotein genes. Bioinformatics, 2005, 21: 2580–2589
10 Kryukov G V, Castellano S, Novoselov S V, et al. Characterization of mammalian selenoproteomes. Science, 2003, 300: 1439–1443
11 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
12 Jiang L, Liu Q, Ni J. *In silico* identification of the sea squirt selenoproteome. BMC Genomics, 2010, 11: 289
13 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
14 Jiang L, Liu Q, Chen P, et al. New selenoproteins identified *in silico* from the genome of *Anopheles gambiae*. Sci China Ser C Life Sci, 2007, 50: 251–257
15 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
16 Tujebajeva R M, Harney J W, Berry M J. Selenoprotein P expression, purification, and immunochemical characterization. J Biol Chem, 2000, 275: 6288–6294
17 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
18 Erfurt J, Averianov A. Enigmatic ungulate-like mammals from the eocene of central Asia. Naturwissen-Schafien, 2005, 92: 182–187
19 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