BMC Genomics

RESEARCH ARTICLE Open Access

In silico identification of the sea squirt selenoproteome

Liang Jiang[1,2,3], Qiong Liu[2] and Jiazuan Ni*[1,2,3]

Abstract

Background: Computational methods for identifying selenoproteins have been developed rapidly in recent years. However, it is still difficult to identify the open reading frame (ORF) of eukaryotic selenoprotein gene, because the TGA codon for a selenocysteine (Sec) residue in the active centre of selenoprotein is traditionally a terminal signal of protein translation. Although the identification of selenoproteins from genomes through bioinformatics methods has been conducted in bacteria, unicellular eukaryotes, insects and several vertebrates, only a few results have been reported on the ancient chordate selenoproteins.

Results: A gene assembly algorithm SelGenAmic has been constructed and presented in this study for identifying selenoprotein genes from eukaryotic genomes. A method based on this algorithm was developed to build an optimal TGA-containing-ORF for each TGA in a genome, followed by protein similarity analysis through conserved sequence alignments to screen out selenoprotein genes form these ORFs. This method improved the sensitivity of detecting selenoproteins from a genome due to the design that all TGAs in the genome were investigated for its possibility of decoding as a Sec residue. Using this method, eighteen selenoprotein genes were identified from the genome of *Ciona intestinalis*, leading to its member of selenoproteome up to 19. Among them a selenoprotein W gene was found to have two SECIS elements in the 3'-untranslated region. Additionally, the disulfide bond formation protein A (DsbA) was firstly identified as a selenoprotein in the ancient chordates of *Ciona intestinalis, Ciona savignyi* and *Branchiostoma floridae*, while selenoprotein DsbAs had only been found in bacteria and green algae before.

Conclusion: The method based on SelGenAmic algorithm is capable of identifying eukaryotic selenoprotein genes from their genomes. Application of this method to *Ciona intestinalis* proves its successes in finding Sec-decoding TGA from large-scale eukaryotic genome sequences, which fills the gap in our knowledge on the ancient chordate selenoproteins.

Background

Selenium (Se), an essential trace element *in vivo*, is closely linked to Keshan disease, Kaschin-Beck disease, cancer and virus propagation. It also plays important roles in cell growth, proliferation and aging. Selenium *in vivo* is primarily present in various selenoproteins, which generally function as antioxidants to maintain the balance of redox state. The active site of selenoprotein is selenocysteine (Sec or U), the 21st amino acid encoded by a TGA codon in the open reading frame (ORF) of the gene [1]. Traditionally, TGA codon only signals the termination of protein synthesis; however, it can also be translated into a Sec residue when a specific stem-loop structure, designated as the Sec insertion sequence (SECIS) element, is located in the 3'-untranslated region (UTR) of a selenoprotein gene in eukaryotes and archaea, or located immediately downstream of the Sec-decoding TGA (designated as Sec-TGA) in bacteria [2-5]. The amino acid sequences flanking the active Sec residue are more conserved than other regions less functionally or structurally important in selenoproteins. These conserved regions play key roles in redox balance, metal combination, Sec-Sec/Sec-S bond formation and protein folding *in vivo*. Additionally, the Sec residue is highly analogous to the Cys in biochemical properties, which accounts for the fact that in most homologs of a selenoprotein the active Sec is replaced by Cys residues.

* Correspondence: jzni@szu.edu.cn
[1] Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, PR China
Full list of author information is available at the end of the article

Several methods for *in silico* prediction of selenoprotein genes have been developed over the past decade, based on the structural characteristics of selenoprotein genes. These methods have been used separately or together for the identification of selenoproteins from the genomes or expressed sequence tag (EST) libraries in many species, such as human, fish, insects, green algae, nematodes and protozoa [6-11]. Recent application on the analyses of environmental metagenomic sequences have also been succeeded in finding prokaryotic selenoprotein genes [12,13]. To identify selenoprotein genes from the genome, several computational methods have to be combined for use, including the methods of RNA secondary-structure prediction, eukaryotic or prokaryotic ORF prediction, sequence alignment analysis and phylogenetic analysis. With the development and application of those methods, the size of selenoprotein family is growing, for example, the members of human selenoproteome increased from 14 to 25. In addition, up to 58 selenoprotein families have been identified recently in the Global Ocean Sampling (GOS) Project, which shed light on the evolution of selenoproteins according to their distribution in different species and environments.

For the prediction of selenoprotein genes, it is indispensable to construct complete or partial ORFs containing the Sec-TGA codons. In a non-intron DNA sequence like prokaryotic genome, it is relatively easy to build a theoretical ORF containing a TGA codon termed "interrupted ORF" (i-ORF) as shown in Figure 1. However, in eukaryotic genome, the intron-exon structure of gene makes it difficult to build an i-ORF. Most of earlier studies on eukaryotic selenoprotein identification were performed by the following scheme. Firstly, RNA prediction algorithms were used to predict SECIS elements. Secondly, the SECIS elements were used to inform gene prediction algorithms to predict i-ORFs, of which a suitable SECIS element must be downstream [14]. The disadvantage of this scheme is that if any special-structure SECIS elements, which has not been discovered so far and included into the known SECIS models, existed in a newly sequenced genome, then no SECIS information can be used to inform the gene prediction algorithms to find the upstream ORFs. Naturally thinking, to identify the selenoproteins with special-structure SECIS, it must be able to predict i-ORFs without the help of SECIS information.

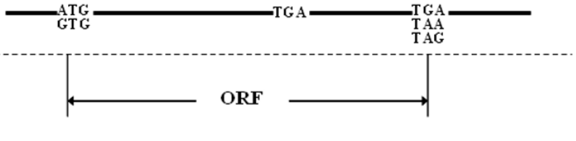


Figure 1 A buildup of theoretical i-ORF containing a TGA codon in prokaryotic genomes.

For eukaryotic genome, a SECIS-independent gene prediction approach was previously introduced in 2004 [11]. A modified gene prediction program named geneid was developed to identify 20 human selenoprotein genes. The gene assembly algorithm GenAmic used by geneid only builds the optimal gene structures with the TGA-containing exons having higher coding potential scores. Selenoproteins with lower score TGA-containing exons, such as human selenoproteins K, S and T, were not identified by the GenAmic based method. Although the sensitivity of this SECIS-independent gene prediction approach is high (20/25), it is not so good as the SECIS-dependent approach that discovered all 25 human selenoproteins [14]. In this paper, a new gene assembly algorithm named SelGenAmic was constructed to develop a similar SECIS-independent gene prediction method for the identification of selenoproteins. Compared with the GenAmic algorithm used by geneid, the SelGenAmic is more sensitive because its target is to build an optimal gene structure for each TGA. Thus no TGA codon is neglected for building i-ORFs. Finally, amino acid conservation assessment is used to find the real selenoproteins from these i-ORFs.

While much research has been done on the identification of selenoproteins of lower bacteria, unicellular eukaryotes, insects and higher vertebrates, only a few results have been reported for the sea squirts, one of the closest relatives in between invertebrates and vertebrates. Sea squirts are widely used for biological research in post-genomic era [15]. Abundant sequence data from the sea squirt genome project enable us to investigate selenoprotein distribution in these species, which provides valuable insights into selenium utilization and selenoprotein gene evolution [16,17]. Additionally, the sea squirt's worldwide distribution, short life cycle, and even its transparent body make it a potentially appropriate model organism for studying eukaryotic selenoproteins. Up to the present, only a few selenoproteins were found in the *Ciona intestinalis*, such as selenoprotein L (SelL), which was identified from EST sequences [18].

In this study, eighteen selenoproteins of *Ciona intestinalis* were identified from the genome by the method we developed according to the SelGenAmic algorithm. Combined with the SelL previously found from EST sequences, the members of selenoproteome in *Ciona intestinalis* increased to 19. Among those selenoproteins, disulfide bond formation protein A (DsbA) was the first time to be found as a selenoprotein in multicellular organisms. Its homologous DsbA in amphioxus and *Ciona savignyi*, two other ancient animals, were also identified to be selenoproteins.

Results and Discussion

Basic idea

TGA coding for Sec is one of the key characters of selenoprotein genes. If we scan all TGA codons from a genome, all Sec codons will be included. It is relatively easy to build a theoretical i-ORF from a prokaryotic genome as shown in Figure 1. This task can be carried out by finding a start codon and a stop codon in the nucleotide sequences flanking any TGA codon. All of i-ORFs can be translated into amino acid sequences and compared with known proteins to find potential selenoprotein genes. This method has been reported to be used for the discovery of rare amino acids, selenocysteine and pyrrolysine, in prokaryotes [19].

However building eukaryotic i-ORFs is much more difficult than that of prokaryotes. Many potential gene prediction signals like start codon ATG, stop codons TGA/TAA/TAG, and splice sites AG/GT can be found in the sequences flanking any TGA in a eukaryotic genome. It is rather complicated to choose suitable signals to build exons, and consequently suitable exons to build i-ORFs during selenoprotein identification.

To address this issue, a method was presented in this study. Firstly, all TGA codons were found from a genome, and supposed to be signals of Sec. All other signals such as start codon, stop codon and splice sites are also predicted. Secondly, common exons (c-exons) were built with common signals as shown in Figure 2A and interrupt exons (i-exons) containing TGA were built by concatenating common signals and TGA as shown in Figure 2B. Thirdly, the gene assembly algorithm SelGenAmic was used to build the best ORF for each i-exon. Figure 3 shows the process of building a best ORF for an i-exon. The best ORF which has the maximal coding potential is composed of this i-exon and other c-exons.

If all i-exons and i-ORFs were enumerated from a genome to build a set, theoretically it should include all potential selenoprotein genes with Sec-TGA codons. However, the vast majority of these i-genes (genes containing Sec-TGA) will be biologically meaningless. To filter out such meaningless i-genes, the conservation of amino acid sequence in the local regions flanking the Sec residue (shown in Figure 4) was used to screen out i-genes that are more likely to be selenoprotein genes.

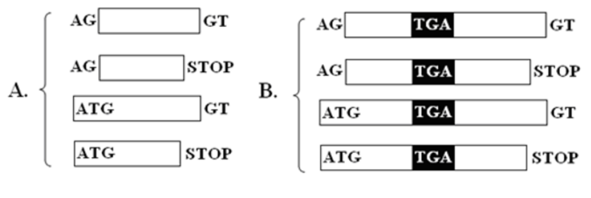


Figure 2 Different types of exons used to assemble selenoprotein genes. A. c-exons in which no in-frame TGA codon is allowed. B. i-exons in which an in-frame TGA codon is present and codes for a Sec residue. The signals such as splice sites AG and GT, start codon and stop codons are indicated in the figure.

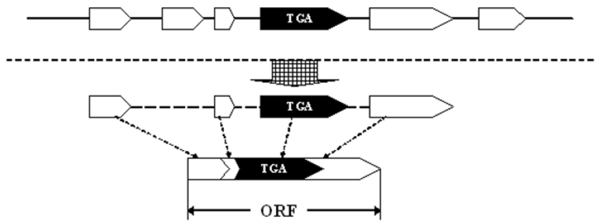


Figure 3 Building an ORF containing a TGA codon. The i-exon is indicated by a black arrow, and the c-exons by white arrows. An ORF containing an i-exon is formed by concatenating suitable c-exons and the i-exon.

General identification procedure

General procedures of our method are shown in Figure 5 and described as follows.

(1) Obtaining i-exons and c-exons from the genome. The whole genome sequence of *Ciona intestinalis* containing 4390 scaffolds was scanned to find all TGA codons and other signals including ATG, TAA\TAG, and AG\GT. All i-exons and c-exons were built from these signals. The coding potential of any exon was calculated as the sum of the scores of the signals, plus the log-likelihood ratio of a Markov model for coding DNA.

(2) Assembling i-genes from i-exons and c-exons. For each i-exon, a best ORF which has the maximal coding potential score was built with our gene assembly algorithm SelGenAmic. 2275466 i-genes were built from *Ciona intestinalis*.

(3) Searching for Sec/Cys pairing and the conservation of its flanking regions. All i-genes were translated into amino acid sequences. Local sequences flanking the Sec residue were extracted for detecting similarity in the NCBI non-redundant (nr) protein database by the BLASTp program in order to obtain multiple sequence alignments. Those sequences were screened out with conservation in the local regions flanking the Sec residue, and alignments containing Sec/Cys pairing (simplified as U/C pair), i.e., the Sec-containing local sequence must

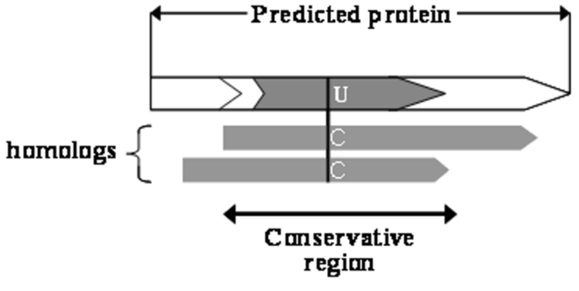


Figure 4 A predicted protein which has conservative amino acid sequences flanking the Sec residue is more likely to be a selenoprotein. In its homologs, the Sec (U) residue is expected to be replaced by Cys (C) residues.

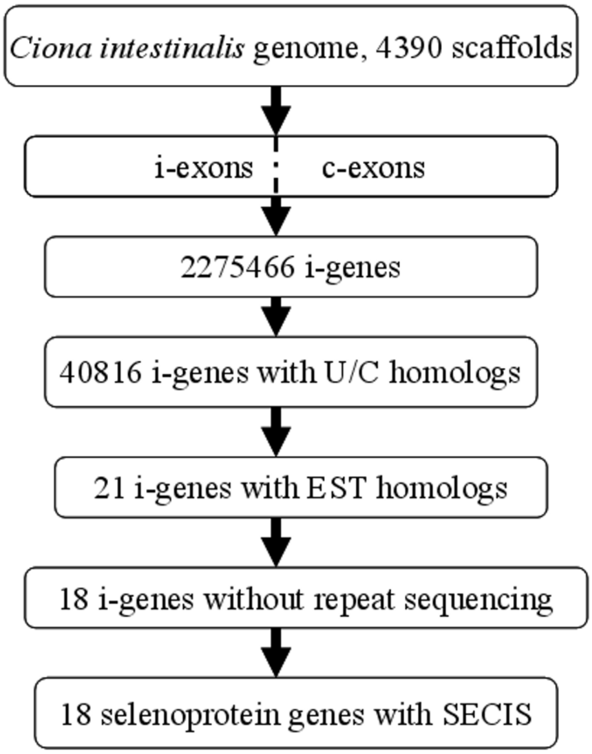

Ciona intestinalis genome, 4390 scaffolds
i-exons
c-exons
2275466 i-genes
40816 i-genes with U/C homologs
21 i-genes with EST homologs
18 i-genes without repeat sequencing
18 selenoprotein genes with SECIS

Figure 5 The flow chart describing general procedures for the identification of selenoprotein genes from the *Ciona intestinalis* genome.

have its homologous sequences contain Cys residues in the position of Sec in multiple alignments. 40816 i-genes were screened out by this step.

(4) Searching against EST databases and splicing the ESTs. Similarity analysis was performed against EST databases to obtain spliced ESTs for the 40816 i-genes. The local DNA sequences flanking the TGA of each i-gene were searched by BLASTn against the EST database of *Ciona intestinalis*. 21 i-genes were screened out after discarding the i-genes in which TGAs were analyzed to be the products of sequencing errors. Among those 21 i-genes, three of them are almost the same as another i-gene when comparing them with each other. Their corresponding genome segments are also very similar. Therefore, we concluded that the three analogous i-genes were caused by repeated sequencing and should be discarded. Finally, 18 i-genes were found to have complete ORFs and UTRs.

(5) Checking for SECIS elements. All 18 i-genes were found to have downstream SECIS elements in their UTRs, which further confirmed them as selenoprotein genes.

Selenoproteins identified

The selenoprotein genes identified in this article are shown in Table 1. All of them are members of known selenoprotein families. The gene structures of selenoproteins in *Ciona intestinalis* are shown in the Supplemental Figure S1, S2 in the Additional file 1, along with the posi tions of exons, introns, UTRs and Sec-TGA codon. The secondary structures of SECIS elements and multiple sequence alignments of the conserved amino acid sequences flanking the Sec residue are shown in Supplemental Figure S3 and S4 in the Additional file 1

Eight selenoproteins were found from the genome which was misannotated previously in the NCBI database. Brief information about these misannotated proteins is listed in Table 2, and comparisons between the gene structures of misannotated and newly identified genes are shown in Supplemental Figure S1 in the Additional file 1. No information was found in the NCBI database for the other newly identified selenoprotein genes in this article.

It should be mentioned that *Ciona intestinalis SelL*, a selenoprotein gene found previously by Gladyshev, was not identified by our method. *SelL* gene was constructed from EST sequences, and proved to be a real selenoprotein of sea squirt before [18]. Comparing the sequences of *SelL* with the genome of *Ciona intestinalis* by Sim4 and

Table 1: Selenoproteins identified from the genome of *Ciona intestinalis*

Selenoprotein gene	Position
selenoprotein N (SelN)	2q
Gpx like protein a (Gpx a)	12p
Gpx like protein b (Gpx b)	1q
Gpx like protein c (Gpx c)	3q
Gpx like protein c (Gpx d)	14q
Gpx like protein c (Gpx e)	scaffold_161
selenophosphate synthetase (SPS)	8q
selenoprotein W, 1 (SelW1)	scaffold_63
selenoprotein W, 2 (SelW2)	1p
15 kDa selenoprotein (Sel 15)	scaffold_127
iodothyronine deiodinase type 3 (DI 3)	9p
selenoprotein H (SelH)	4q
selenoprotein S (SelS)	4q
selenoprotein K (SelK)	7q
thioredoxin reductase (TR)	8q
selenoprotein O (SelO)	6q
selenoprotein T (SelT)	3q
DSBA like protein	3q

* The genomic chromosome or scaffold from which the selenoprotein gene was identified. The abbreviated name of each selenoprotein is shown in parentheses.

Table 2: The misannotated selenoprotein genes of *Ciona intestinalis*

Selenoprotein gene	Brief information in the NCBI database.
SPS	XM_002129490 PREDICTED: similar to selenophosphate synthetase 1 (LOC100182599)
SelH	XM_002124882 PREDICTED: similar to Selenoprotein H (SelH) (LOC100184326)
SelO	XM_002124815 PREDICTED: similar to predicted protein (LOC100179326)
SelS	XM_002131362 PREDICTED: similar to selenoprotein S (LOC100186840)
DsbA	XM_002126620 PREDICTED: hypothetical protein LOC100182954 (LOC100182954)
SelK	XM_002128679 PREDICTED: similar to selenoprotein K (LOC100184686)
TR3	XM_002131964 PREDICTED: similar to thioredoxin reductase 3 (LOC100178436)
Gpx c	XM_002121522 PREDICTED: Ciona intestinalis similar to C11E4.1 (LOC100182197)

BLASTn, no significantly similar regions of *SelL* were found in the genome data. The results implied that incomplete sequencing of the genome may cause the omission of genomic regions containing the *SelL* gene. The method developed in this paper is used to predict selenoproteins from genomes, thus it is impossible to find the *Ciona intestinalis SelL* gene from the genome data that do not include this gene.

Unique SelW1 with two SECIS elements

Two SECIS elements were detected by SECISearch 2.19 in the 3'-UTR of the *SelW1* gene of *Ciona intestinalis*. The gene structure is shown in Figure 6. All gene structures are schematically shown in this article as the 5'-end on the left. The complete ORF and two SECIS elements are located in the first exon. Two SECIS elements were named SECIS 1 and SECIS 2.

The primary sequences and secondary structures of the two SECIS elements are shown in Figure 7. Both of them belong to the form II SECIS element, which have an additional minihelix in the apical loop [2]. The two SECIS elements were both detected by SECISearch 2.19 under the "default" pattern to limit the conservation and energy cutoff. In addition, SECIS 2 could also be detected under the "strict" pattern. The COVE scores were 24.91 for SECIS 1 and 23.70 for SECIS 2, which are much higher than 15, the recommended COVE score cutoff for SECISearch2.19 [14]. The COVE score was calculated by matching the query sequence with the known SECIS secondary model. If the score exceeds 15, the query sequence is considered as a true SECIS according to the recommendation of SECISearch2.19. Therefore, both SECIS elements could be functionally active. Up to the present, two potential SECIS elements have only been found in *SelP* and human *Sep 15* genes [6,20]. The *SelP* gene contains more than one Sec-TGAs, and experimental evidence has shown that its two SECIS elements are necessary for the efficient incorporation of multiple Sec residues into SelP. Whereas the *Sep 15* gene contains only one Sec residue, and its upstream SECIS has been proved nonfunctional [21]. Interestingly, the predicted *Ciona SelW1* gene also contains one Sec residue and two SECIS elements. Only SECIS 2 can be detected by the "strict" pattern of SECISearch 2.19. Those suggested that SECIS 2 has a higher possibility of being functional. From the characteristics of *Ciona SelW1* gene structure, it seems that this gene is more analogous to human *Sep15* that its SECIS 1 may be nonfunctional while SECIS 2 is used for Sec incorporation. However, conclusion can only be drawn after experimental results on functional analyses of these two SECIS elements.

Selenoprotein DsbA found in multicellular organisms

Selenoprotein DsbA has only been found in multiple prokaryote species, microbial marine communities, symbiotic bacterium of a gutless worm, and *picoeukaryote Micromonas* [11,12,19,22-24]. In this paper, DsbA was firstly reported as selenoprotein in a multicellular organism, *Ciona intestinalis*. The structural information for selenoprotein DsbA gene of *Ciona intestinalis* is shown in Figure 8A indicated as newly annotated, while the originally released gene is indicated as misannotated, where the upstream part of the first exon, including the Sec-TGA, was misannotated as the UTR of the gene.

Interestingly, a selenoprotein *DsbA* gene was also identified in *Branchiostoma floridae* which was misannotated as hypothetical protein BRAFLDRAFT_120127 (gi|219424997|ref|XP_002210295.1|) in the present database [25]. The upstream sequence of the coding region in this amphioxus gene was extracted from its mRNA, and translated into an amino acid sequence where the original start codon, ATG, was decoded as a methionine (Met, or M) and a TGA decoded as a Sec. The translated sequence was analysed by BLASTp against the NCBI nr database. The complete ORF of the selenoprotein *DsbA* gene was constructed by extending its aligned ESTs of *Branchiostoma floridae*. However, not enough ESTs were found to be extended to the UTR for SECIS search. Thus, the ORF downstream sequence was extracted from the genome to search for SECIS element. A positive hit was found by SECISearch 2.19. The gene structure of this newly identi-

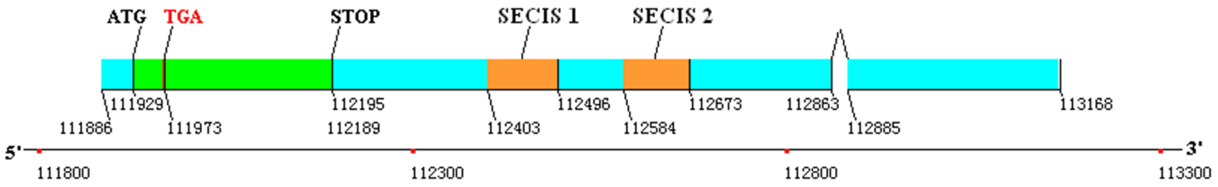
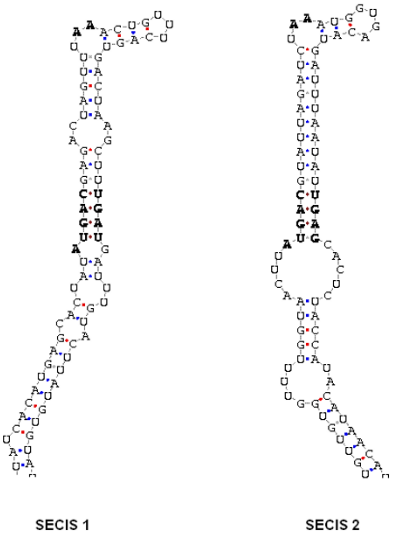


Figure 6 Gene structure of *Ciona intestinalisSelW1*. The coding region is indicated by a green rectangle, the untranslated regions by blue rectangles, and the SECIS elements by orange rectangles. The intron is indicated by lines connecting the exons. The position of each site in the sequence of chromosome or scaffold is shown by numbers and the bottom coordinates.

fied selenoprotein *DsbA* is shown in Figure 8B indicated as newly annotated, while the originally annotated coding region in the NCBI database is indicated as misannotated. The first exon of the new gene containing the Sec-TGA codon was misannotated as the 5'-UTR in the original version, and the last exon of the new gene including the stop codon and 3'-UTR was missing originally. Two long exons and a long intron were found at the end of the original gene. No similar EST sequence or amino acid sequence has been released for these two exons.

The primary sequences and secondary structures of the SECIS elements in the newly identified *DsbA* selenoprotein genes from three species are shown in Figure 9. Homologous analysis by multiple alignments of SECIS sequences in Figure 9 shows that the two SECIS elements from the *Ciona DsbA* selenoprotein genes are more similar than the one from *Branchiostoma floridae*. In addition, both *Ciona* SECIS elements belong to the form II SECIS that has an additional minihelix in the apical loop. The COVE scores of the SECIS elements of *Ciona intestinalis*, *Ciona savigyni* and *Branchiostoma floridae* were 22.40, 17.89 and 18.83, respectively. An unpaired CC motif is found in the apical loop of *Ciona savigyni* SECIS element as a conserved site, while the AA motif was found in *Ciona intestinalis* and *Branchiostoma floridae* SECIS elements. The AA motif of a SECIS element has been reported in most selenoprotein genes of different organisms, while the rare CC motif has only been reported in human, rat and mouse selenoprotein M [26].

The thiol/disulfide interchange protein DsbA, a member of the thioredoxin family of disulfide oxidoreductases, catalyzes disulfide bond formation by donating its disulfide to newly translocated proteins in many prokaryotes [27]. Eukaryotic protein disulfide isomerase (PDI) has a similar function, and both proteins have similar redox active levels [28]. DsbA contains a



```
                   *        20         *          40         *          60          *          80          *      100        *
SECIS 1:AGACTCCATT  TCATATC  ACATGAGCACTAT  ATGAC  GAGACTAGTTT  AA  ACTGTTTCAGT  GACTAAGCTT   TGAT  GATTT  GTACTTA  TGTGTATGAA : 95
SECIS 2:TTACAAATGT  TGTGGTT  TTGGTAACTT     ATGAC  GTATTAGATCT  AA  ATGGTGACA    TGATTTAATAT  TGAG  CAC    TCTACCA  TACATAACAT : 89
                                           ATGA                AA                             GA
```

Figure 7 Primary sequences and secondary structures of the two SECIS elements in the *Ciona intestinalis SelW1* gene. Core conserved functional sites are indicated by bold letters in the secondary structures and white letters on a black background in the primary sequences.

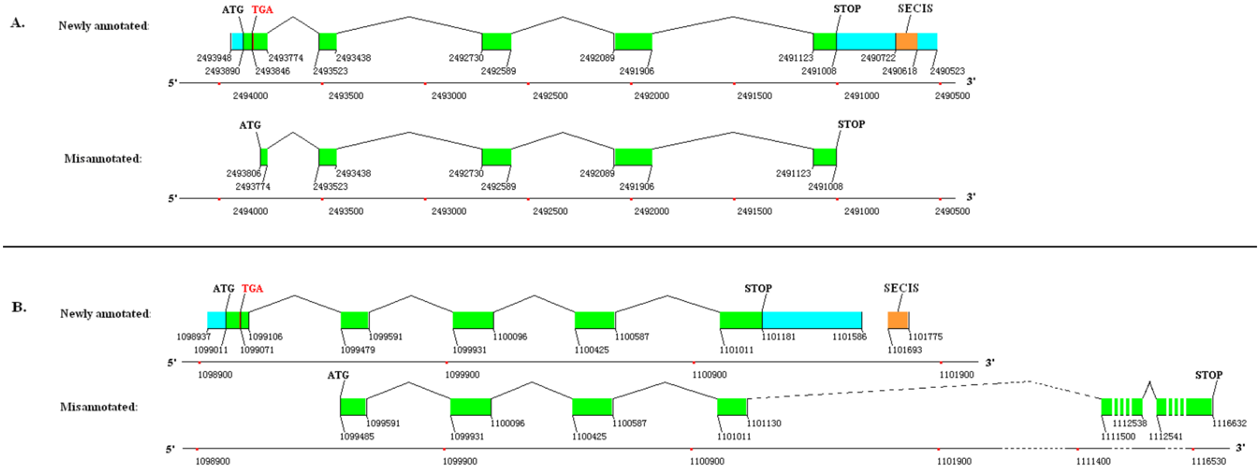
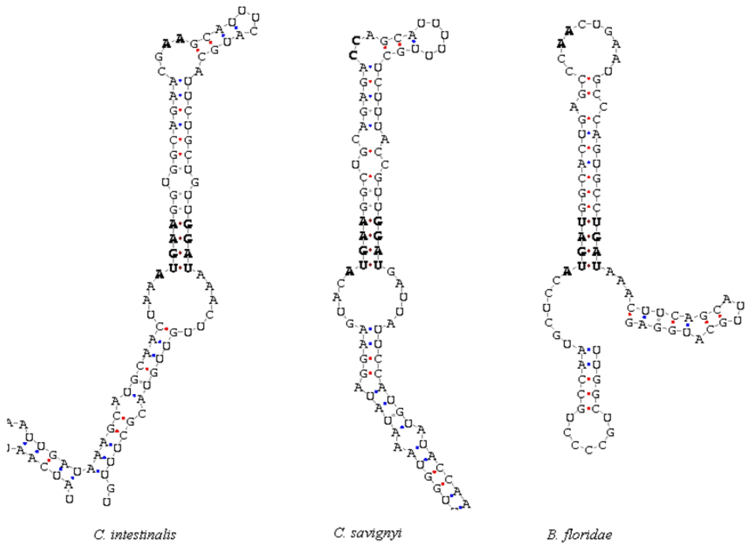


Figure 8 Comparison between the gene structures of newly identified *DsbA* selenoprotein genes and the coding regions of originally misannotated genes. **A**. *Ciona intestinalis DsbA*; **B**. *Branchiostoma floridae DsbA*. In the misannotated version, the downstream 2 exons and 1 intron are very long that they are indicated by broken line and interrupted rectangles.

redox-active CXXC motif imbedded in a TRX fold. The active CXXC motifs can be found in many other selenoproteins. Research on the structure and function of the CXXC motif in *Escherichia coli* DsbA has shown that the upstream Cys is exposed on the surface of protein, and the downstream Cys is embedded in the 3-D structure [29]. The upstream Cys has a very low pK_a value (≈3.5), and is completely ionized under physiological pH conditions [30]. Mutational analyses have revealed that the upstream Cys residue can catalyze thiol/disulfide interchange reaction without the presence of the downstream Cys [31]. These studies suggest that the upstream Cys is more important than the downstream Cys of DsbA. Interestingly, the upstream Cys is replaced with Sec in our newly identified three DsbA selenoproteins. Due to the higher redox activity of Sec compared with that of Cys, it is reasonable for us to deduce that selenoprotein



```
                           *       20          *    ▼▼▼▼ 40          *   ▼▼     60          *       80   ▼▼      *       100          *
C. intestinalis : TATCAATCAA TTGATAA AGCATGCAACTAA ATGAA GGTGGCAGAACG AA GCATTTCATG-- CATTCTGCTGTT GGAT AAA- CTTGTTG TACGCTTTGT : 95
C. savignyi     : AACTATTGTT TTGGTAA ATATAGGAAGTAC ATGAA GGCTGCAGAGA- CC AGCATTTTTTGC TCTTTACCGTT- GGAT GAT- TATTCCA TGTATACCAA : 95
B. floridae     : TTGGCTGCCC CTGCCAA TGCTCC------- ATGAT GGCACTGAGCCC AA CTGAA------- TGCCCAGTGCC- TGAT AAAC TTCAGCA TTGCATGGAG : 83
                  ta  attc  tTG tAA agct g aa ta ATGAa GGc gcaga c aa    att  t   t ttcagcg t  gGAT aAa  ttt  ca T    t  a
```

Figure 9 Primary sequences and secondary structures of the SECIS elements of *Ciona intestinalis, Ciona savigyni* and *Branchiostoma floridae DsbA* selenoprotein genes. Conserved functional sites are indicated by ?.

DsbAs with Sec in place of Cys in the reactive centres are more functionally active than non-selenoprotein DsbAs.

Evolutionary analyses of DsbA

The 100 most similar protein sequences to DsbA were obtained from the NCBI nr database by BLASTp program. Sequences from the same species were deleted, leaving 67 homologs of DsbA, including the newly identified DsbA selenoproteins. Those DsbA homologs were aligned and phylogenetically analysed. Most of the DsbAs (47 of 67) are bacterial proteins, and the remaining 20 are eukaryotic. A phylogenetic tree of these 20 eukaryotic DsbA proteins is shown in Figure 10. The multiple sequence alignments of these 20 proteins, along with 8 prokaryotic DsbA selenoproteins found previously in the Sargasso Sea microbial selenoproteome, are shown in Figure 11. The putative CXXC active sites and Sec residues are highlighted.

Phylogenetic analysis shows that eukaryotic DsbA proteins can be clearly classified into three groups: the fungi, viridiplantae and metazoa. No selenoprotein DsbA was found in the fungi group. Only one selenoprotein DsbA was found in the viridiplantae group, which belongs to *Micromonas pusilla*, a unicellular plant. The three newly predicted DsbA selenoproteins belong to the metazoa group, cionidae subgroup. They are most closely related to each other. According to the phylogenetic analysis, DsbA is widely distributed in prokaryotes. The utilization of selenium to form Sec in DsbA diverged during the evolution from prokaryotes to eukaryotes. Since both selenoprotein DsbA and non-selenoprotein DsbA (the Cys-containing form) could be found in prokaryotes, metazoa, and viridiplantae (see Figure 10), a hypothesis was deduced that the evolution of selenoprotein DsbA occurred prior to the separation of animal and plant kingdoms. Additionally, the Sec residue was lost or replaced by Cys in land plants [8]. The hypothesis is supported by the evidence that only an early-diverging unicellular plant in the viridiplantae has a selenoprotein DsbA, while other land plants have the Cys-containing proteins.

Sea squirts and amphioxus are important transitional species, bridging the gap between vertebrates and invertebrates. Amphioxus is even considered as the invertebrate that is most evolutionarily close to vertebrates. The placozoan *Trichoplax adhaerens*, a close relative to the three cionidae DsbA selenoproteins in the phylogenetic tree, is arguably the simplest free-living animal, and therefore may represent a primitive metazoan form [32]. Genome project research on these ancient species has provided important information for studying the evolution of vertebrates, and even metazoans. Interestingly, selenoprotein DsbAs were found in these primitive chordates, while only Cys-containing form DsbAs were found in the primitive metazoan. No DsbAs were found in vertebrates through homologous analysis. The reason could be that DsbA is possibly replaced during evolution by another enzyme with similar function, such as DPI or other proteins in the thioredoxin family. In higher vertebrates, the DsbA family was completely replaced by other proteins with similar function, leading to the disappearance of DsbA selenoproteins.

Relative merit of the SelGenAmic-based method

The selenoprotein identification scheme presented herein (an 'i-ORF-assembly-first, Sec-homology-second, SECIS-search-third' approach) differs from the reported methods in some ways. The methods based on SECIS search required the finding of SECIS elements to inform their gene-prediction-algorithms the proper sequences for predicting i-genes [7]. However, it is possible for some new selenoprotein genes to have certain special SECIS structures that have not been discovered up to the present, thus can't be detected by available known programs. In this case, the SECIS-search based approach is impossible to find those special selenoprotein genes. Although the method based on read-through similarity analysis (RSA) had made up this shortage [33], it could only be applied to identify prokaryotic selenoprotein genes but not the eukaryotic due to their complicated intro-exon structures. In this sense, our method has the merit of independently finding possible i-ORFs for all TGA codons in eukaryotic genomes without the help of SECIS information. This is because the SelGenAmic algorithm could enumerate all i-ORFs for all TGA codons in a eukaryotic genome. Therefore, even with special SECIS structures, those selenoproteins will not be lost in the analysis using the SelGenAmic-based method. They can be found during protein sequence conservation analysis.

In fact, the SECIS-independent-prediction method was firstly introduced to find human and fugu selenoproteins in 2003 [11,14]. However, its ORF prediction step is different from ours. The SelGenAmic algorithm we developed in this paper is to assemble i-ORF for each TGA in genomic sequences, while the GenAmic algorithm reported previously is to assemble selenoprotein genes and standard genes at the same time, using stronger coding bias for the sequence downstream Sec-TGA codon to discriminate real Sec-TGA codons and others. Although it was performed with rather high sensitivity (80%), some selenoproteins, such as SelK, SelT and SelS, were still missing during human genome analysis. To overcome this shortage, the SelGenAmic algorithm was developed to maximize the sensitivity in such way that an optimal i-ORF was built for each TGA codon, without skipping any TGA codon even with low coding bias for the downstream sequence. The SelGenAmic-based method was also applied to human genome to detect its sensitivity, and it successfully identified the whole 25 human seleno-

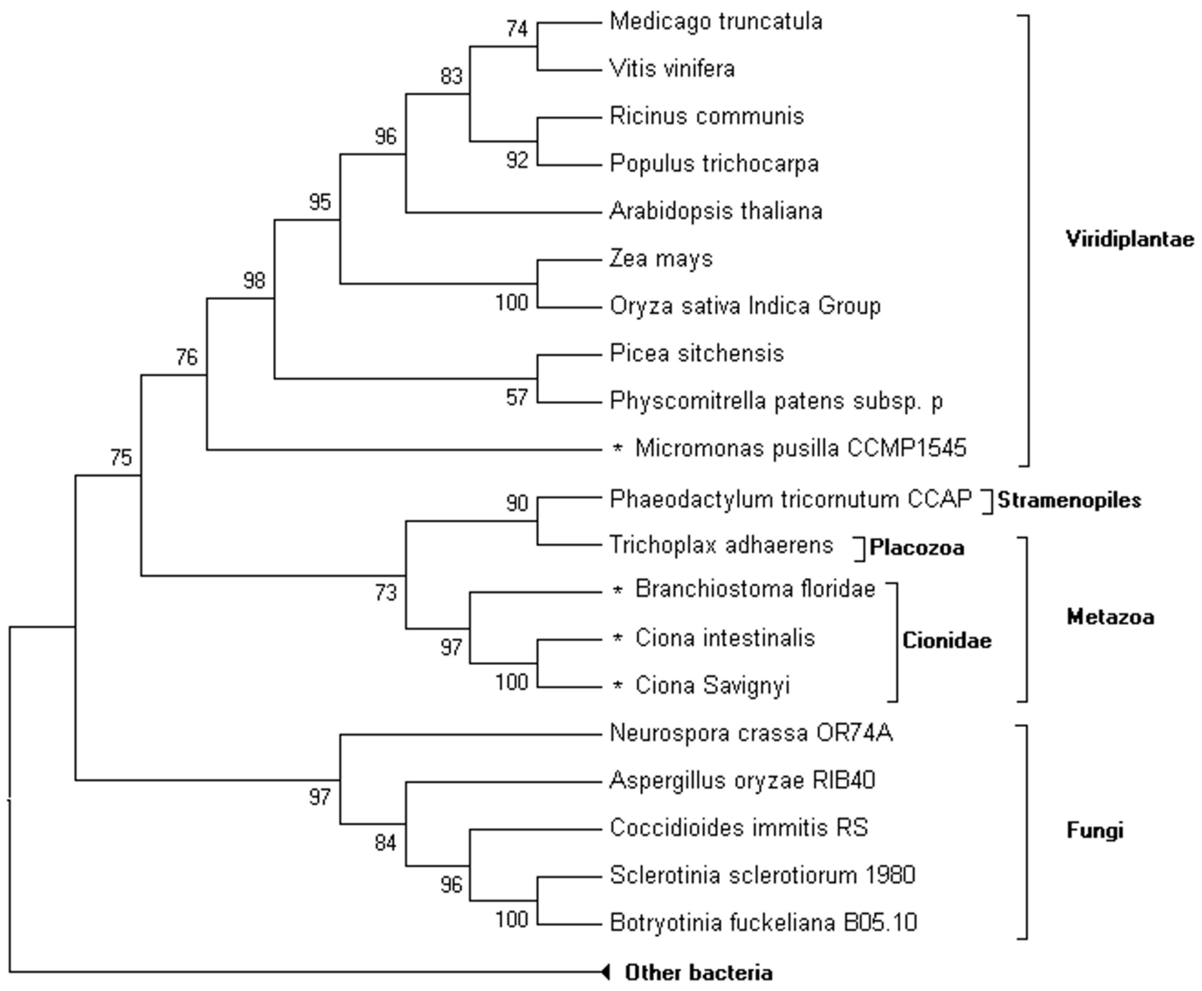


Figure 10 Phylogenetic tree of eukaryotic DsbA. Selenoproteins are marked by an asterisk. Bootstrap values are shown at each branchpoint to indicate the reliability of this tree.

proteins as shown as Supplemental Figure S5 in the Additional file 1. However, the improvement was achieved at the expense of dealing with a huge amount of data (more than 2 million biological meaningless i-genes in *Ciona*). These false positive predictions could be removed through protein conservation alignments.

Another merit of the SelGenAmic-based method is that it can find the alternative splicing genes of selenoproteins. Some selenoprotein genes, such as human DI2, have more than one alternative splicing forms. The Sec residues of two human DI2 forms were coded by different TGA codons in the gene. 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. Using our method, both forms of human DI2 were identified (data not shown), so were the two forms of DI2 in the armadillo genome (shown as Supplemental Figures S6, S7 and S8 in Additional file 1).

Conclusion

A eukaryotic selenoprotein identification method based on a gene assembly algorithm SelGenAmic was presented in this paper. It focuses on the prediction of ORF containing a Sec-TGA codon in the eukaryotic genomes. With the aid of this method, 18 selenoprotein genes were identified from the genome of *Ciona intestinalis*, leading to the member of its selenoproteome up to 19. Among them, a unique selenoprotein gene of *Ciona intestinalis SelW1* contained two SECIS elements while others had only one SECIS element in their 3'-UTRs. In addition, DsbA was firstly found to be a selenoprotein in multicellular organisms, like *Ciona intestinalis, Ciona savignyi*, and *Branchiostoma floridae.* The existence of DsbA selenoproteins in multicelluar organisms provides important

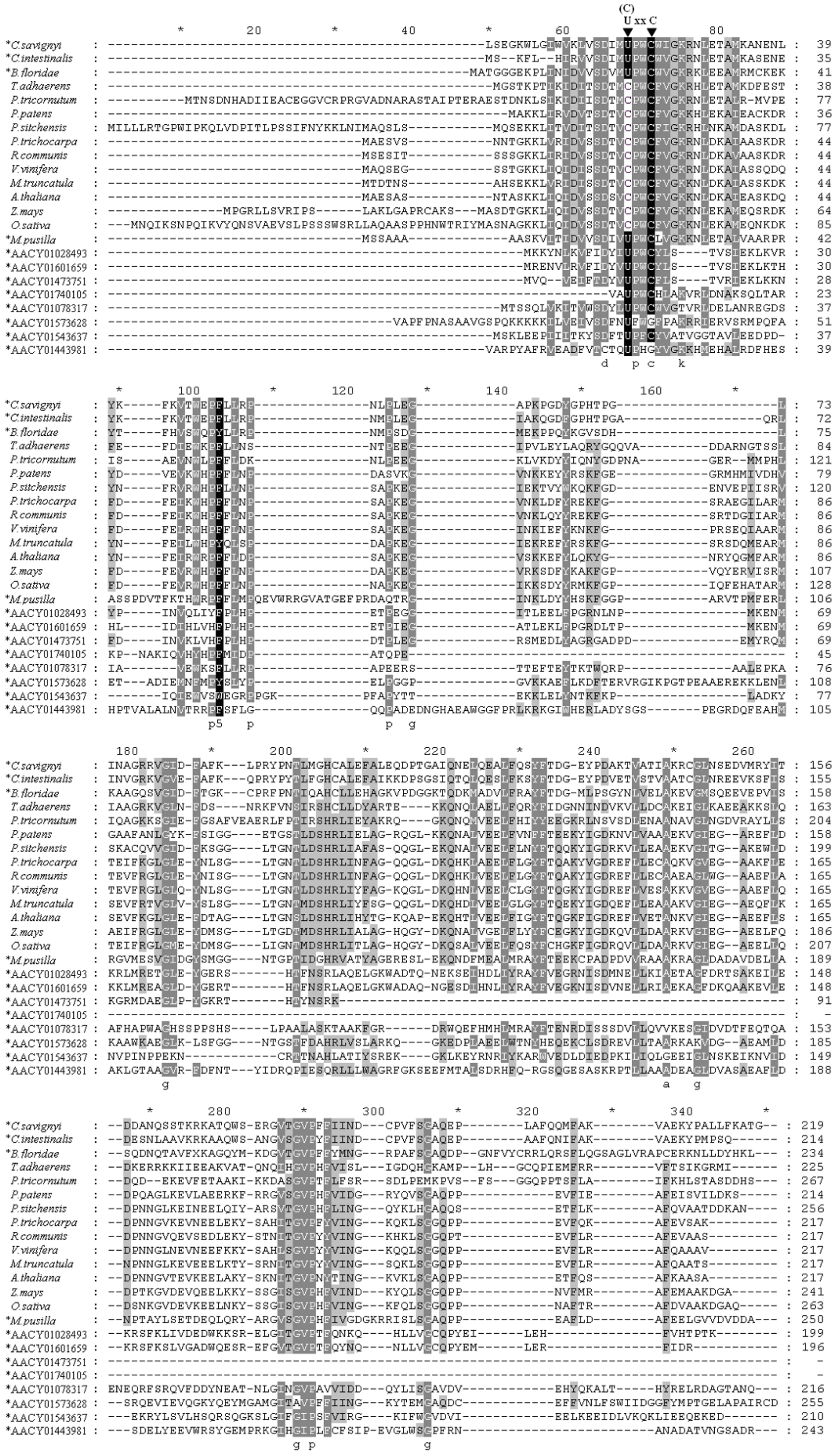


Figure 11 Multiple alignments of eukaryotic DsbA-like proteins and several prokaryotic DsbA-like selenoproteins from the microbial selenoproteome of the Sargasso Sea. Sec and Cys residues in the active site U(or C)XXC are marked on top. Species names of eukaryotes and sequence names from the Sargasso Sea are listed on the left. Selenoproteins are marked by an asterisk.

information for the exploration of the evolution of selenium utilization in invertebrates and vertebrates.

Methods

Data resources

The genome sequences were downloaded from the Ensembl Project Genome Databases http://www.ensembl.org. The release number of the genome data is JGI2.48 for *Ciona intestinalis*. The text file size of the genome data of *Ciona intestinalis* is about 172 Mb, containing approximately 176,000,000 bp, 4,390 scaffolds. All available *Ciona intestinalis* EST sequences (1,205,981 at the time of our analyses) were extracted from the NCBI database.

Construction of ORFs containing Sec-TGAs

The program Geneid (version 1.2a) [34] was used to obtain all common gene signals, such as start codon, splice sites, stop codons, and common potential exons, from genomic sequences. A series of PERL programs were edited to obtain all TGA codons from a genome, and build i-exons from common signals and TGA codons. The coding potential was obtained by Geneid with specific parameter file of ...*Ciona intestinalis* (downloaded from http://genome.crg.es/software/geneid/index.html). A PERL program was edited based on the i-gene assembly algorithm, SelGenAmic, to construct all i-genes from common exons and i-exons.

Assembly algorithm SelGenAmic

The algorithm SelGenAmic is developed from GenAmic to solve the problem of finding an optimal ORF for each i-exon. The word optimal here means that the coding potential score of such ORF is bigger than any other ORFs composed of this i-exon and other suitable c-exons.

The input data of SelGenAmic are all i-exons and c-exons along with their information such as coding potential, position and coding frame. Let $E = \{e_1, e_2, ..., e_k, (k \geq 0$, k is an integer) be the set of exons. The coding potential of these exons are shown as $P(e)$ for each e E in this paper.

Let $C = \{c_1, ..., c_m, (m \geq 0, m$ is an integer) and $T = \{t_1, ..., t_n, (n \geq 0, n$ is an integer) be the sets of c-exons and i-exons. Obviously, $E = C\ \ T$.

The principle to constrain the algorithm to choose suitable exons for concatenation is described as a function M.

$$M(e_a, e_b) = \begin{cases} 1, (a < b, e_a \text{ and } e_b \text{ can be concatenated legally}) \\ 0, (e_a \text{ and } e_b \text{ cannot be concatenated legally}) \end{cases}$$

The function M describes the relation of two exons e_a and $e_b(e_a$ E, e_b $E)$, which can be concatenated legally in numerical order. The word legal here means e_a and e_b are frames compatible, non-overlapping, and adjoining splice sites matched.

Firstly we recall the concepts of gene assembly. A gene assembly g is a sequence consisting of exons $e_{1'}$, ... $e_{q'}$ from E ($e_{i'}$ E). Thus a legal gene assembly can be described as $g = <e_{1'}, ...\ e_{q'}>$, where for all $e_{i'}$ in g, $M(e_{i'}, e_{(i+1)'}) = 1$.

The coding potential of a gene assembly is the sum of scores of assembled exons: $P(g) = P(e_{1'})+...+P(e_{q'})$. The problem to find an optimal g could thus be interpreted as to search for the gene assembly g with maximum $P(g)$, i.e., for all other genes g' constructed from E, $P(g)> = P(g')$

Thus the target of SelGenAmic is that, for each t_i T, finding an optimal gene assembly $g(t_i) = <c_{1'}, ..., c_{k'}, t_i, c_{1''}, ..., c_{k''}>$, where $c_{i'}$ C, $c_{i''}$ C, $M(c_{k'}, t_i) = 1$, $M(t_i, c_{1''}) = 1$, $M(c_{i'}, c_{(i+1)'}) = 1$, $M(c_{i''}, c_{(i+1)''}) = 1$, i.e., for the set $T = \{t_1, ..., t_n$, build a set of optimal gene assemblies $\{g(t_1), ..., g(t_n)$.

This problem is equivalent to find the best upstream assembly $g_u = <c_{1'}, ...\ c_{k'}, t_i>$ and the best downstream assembly $g_d = <t_i, c_{1''}, ...\ c_{k''}>$ for each t_i.

To solve such problem, we introduced six concepts. (1) the best upstream assembly(BUA); (2) the coding potential of BUA (CpBUA); (3) the best upstream adjoining exon (BUE); (4) the best downstream assembly (BDA); (5) the coding potential of BDA (CpBDA) and (6) the best downstream adjoining exon (BDE).

The concepts (1), (2) and (3), were used in the GenAmic algorithm of Geneid program to calculate the optimal assembly from common exons [35]. The concept (1) BUA was described as that for each exon c_i from C it is possible to find a best assembly ended with c_i, *i.e.*, $g_u(c_i) = <c_{1'}, ...\ c_i>$, where for all other assembly g' ended with c_i, $P(g_u(c_i))> = P(g')$. The coding potential $P(g_u(c_i))$ is the CpBUA (concept (3)) of c_i.

The concept (2) BUE was described as that for each c_i from C, it is possible to find a best upstream adjoining exon c_j, if and only if

1. $M(c_j, c_i) = 1$, and

2. for all other c' C, if $M(c', c_i) = 1$, $P(g_u(c_j)) > = P(g_u(c'))$.

Here we use function G to describe the relationship between a c-exon c_i and its BUE c_j: $G(c_i) = c_j$.

Obviously, the BUA $g_u(c_i)$ of c_i, can be obtained by concatenating its BUE c_i and the BUA $g(c_j)$ of c_i:

$$g_u(c_i) =< c_1, ...c_j, c_i >=< g_u(c_j), c_i >=< g_u(G(c_i)), c_i > .$$

So that, it can be easily concluded that if all the BUE for each c C were known, the BUA can be easily calculated as follow:

$g_u(c_i) = <c_0, ..., G\ (G(c_i)), G(c_i), c_i>$, ($c_0$ is used to describe the first exon of this assembly).

In Geneid, the author of GenAmic algorithm used dynamic programming to obtain all the BUE $G(c)$ for every $c \quad C$ [36]. Using the GenAmic, we can easily build 3 sets for $C = \{c_1, c_2, ..., c_k$:

the set of BUE $G(c_1), G(c_2), ..., G(c_k)\}$,

the set of BUA $\{ g_u(c_1), g_u(c_2), ..., g_u(c_k)\}$, and

the set of CpBUA $P(g_u(c_1)), P(g_u(c_2)), ..., P(g_u(c_k))$.

Knowing how to calculate the BUA $g_u(c_i)$ for each c-exon $c_i \quad C$, we can use the same method to build the BUA $g_u(t_i)$ for each i-exon $t_i \quad T$:

$g_u(t_i) = <c_{1'}, ...\ c_{k'}, t_i> = <g_u\ (G(t_i)), t_i> = <g_u\ (c_u), t_i>$, where c_u is the BUE of t_i.

The BUE $c_u \quad C$ is a c-exon, and the set of BUA $\{ g_u(c_1), g_u(c_2), ..., g_u(c_k)\}$ for each $c \quad C$ can be obtained with GenAmic algorithm in linear time. Thus, if all the BUE c_u for each i-exon $t_i \quad T$ can be obtained, its BUA can also be produced.

The BUE $\mathrm{G}(t_i) = c_u$ can be found in the following way. We find all exons c' satisfying $M(c', t_i) = 1$, and their CpBUA from $\{P(g_u(c_1)), P(g_u(c_2)), ..., P(g_u(c_k))$. By comparing their coding potentials, the c_u with maximal $P(g_u(c))$ can be found.

As shown in Figure 12, all c-exons $c'' \quad C$ located upstream of t_i and satisfying $M(c'', t_i)$ will be searched for the c_u of t_i. An i-exon t_{i-k} can be found to divide the genome into 2 regions. The t_{i-k} is a exon for which all c-exons c'' satisfying $M(c'', t_{i-k}) = 1$ are also satisfying $M(c'', t_i) = 1$. Obviously, the c-exon with maximal CpBUA in region α is the BUE of t_{i-k}. Then only c' in region β satisfying $M(c', t_i) = 1$ will be searched for the c-exon with maximal CpBUA. Let A and B be the sets of c-exons in the region α and β, respectively, then to find the c_u for t_i, can be described as follow:

$$P(g_u(\mathrm{G}(t_i))) = \max \left\{ \begin{array}{l} P(g_u(\mathrm{G}(t_{i-k}))) \\ \max \{ P(g_u(c)) : c \in ' \ \} \end{array} \right\}$$

With this equation, all BUE(s) of $t_i \quad T$ can be calculated recursively with one scan of the genome in linear time, due to the reason that only the region β for each t_i is needed to be searched.

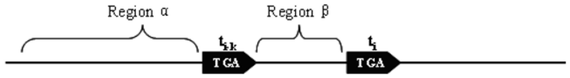


Figure 12 The regions in which c-exons are used to search for the best upstream adjoining exon of t_i.

Then the set of BUE for T is obtained as $G(t_1), G(t_2), ..., G(t_k)\}$ and

the set of BUA for T is obtained as $g_u(G(t_1)), g_u(G(t_2)), ..., g_u(G(t_k))$.

The BDA, BDE and CpBDA for all $c \quad C$ can be obtained in similar way. Let $g_d(c_i)$ be the BDA of c_i, and let function G^r describe the relation of c_i and its BDE c_j:

$G^r(c_i) = c_j$, if and only if

1. $M(c_i, c_j) = 1$ and

2. for all other $c' \quad C$, if $M(c_i, c') = 1$, then $P(g_d(c_j)) >= P(g_d(c'))$.

Thus the set of BDA, BDE and CpBDA for all $c \quad C$ can be produced similarly as follows:

$G^r(c_1), G^r(c_2), ..., G^r(c_k)\}$,

$\{ g_d(c_1), g_d(c_2), ..., g_d(c_k)\}$, and

$P(g_d(c_1)), P(g_d(c_2)), ..., P(g_d(c_k))\}$.

Then the set of BDA for T is obtained as $g_d(G^r(t_1)), g_d(G^r(t_2)), ..., g_d(G^r(t_k))$.

By concatenating BUA and BDA, the best assembly for t_i is constructed as $g = <g_u(G(t_i)), t_i, g_d(G^r(t_i))>$.

Thus, for each i-exon an optimal assembly can be constructed in linear time.

Homology analysis

The NCBI nr protein database was downloaded from the NCBI ftp server in June of 2008, containing 6,598,440 protein sequences, 2,257,741,895 total letters. BLAST programs (version 2.2.18) [36] were also obtained from the NCBI ftp server at ftp://ftp.ncbi.nih.gov/blast/db/. All i-genes were searched by the program BLASTp with an E-value cutoff at 1. All similar sequences detected were used to create multiple sequence alignments with ClustalW (version 1.83) [37]. The conservative motif containing the Sec residue of an i-gene was analysed by the program using a motif search algorithm like MAME.

Gene structure analysis

EST sequences were downloaded and compared with all predicted selenoprotein genes using the program BLASTn. Highly similar EST sequences were spliced by SeqMan program from the DNASTAR package http://www.dnastar.com/ and analysed for the selenoprotein gene structure. The constructed genes were homologously compared to genomic sequences with the program Sim4 [38] to find the locations of exons and introns in the genome, shown as position numbers in Figure 3, Figure 5 and Supplemental Figure S1, Figure S2 in the Additional file 1.

Search for SECIS elements

RNAfold (version 1.7.2) [39] and PatScan [40] were automatically used by a PERL program to detect SECIS-like structures from genome sequences. The SECIS patterns used in the present paper are the same as that in the search of human SECIS [14]. These patterns together with the SECISearch program in PERL language were kindly provided online by Doctor Charles (Karolos) Chapple http://genome.imim.es/~cchapple/. The COVE score of SECIS-like structures were evaluated by the online program SECISearch (version 2.19) [14].

Phylogenetic analysis

Multiple alignments of amino acid sequences were generated using the ClustalX program (version 1.83) [41]. The unrooted phylogenetic tree with unscaled distance branches were generated using the program MEGA3.1 http://meme.sdsc.edu/meme4_1/intro.html with the Neighbor-Joining method. Tests of the phylogenic analyses were done by 1000 replications of the Bootstrap algorithm.

Additional material

Additional file 1 Additional information of newly predicted selenoproteins. The following additional data are included within the additional file1. Gene structures of the newly identified selenoprotein genes of *Ciona intestinalis*, and comparison between the newly identified version and mis-annotated version of these genes are shown in Supplemental Figure S1 and S2. The secondary structures of the SECIS elements of *Ciona intestinalis* selenoprotein genes are shown in Supplemental Figure S3. Multiple alignments of all newly identified selenoproteins and their homologous sequences are shown in Supplemental Figure S4. Multiple alignments of all human selenoproteins predicted by the SelGenAmic-based method and their homologous sequences are shown in Supplemental Figure S5. Information on chromosome number and ORF position of these genes were also shown in Supplemental Figure S5. Gene structures of the two alternative splicing forms of DI2 predicted from the armadillo genome were shown in Supplemental Figure S6, along with their multiple alignments (Supplemental Figure S7) and secondary structure of SECIS elements (Supplemental Figure S8).

Authors' contributions

LJ carried out the whole research work including program edit, algorithm design, selenoprotein identification from the genomes, and draft making. QL was responsible for the project design, key-issue discussion and manuscript writing. JN was responsible for the project design, progress and coordination. All authors read and approved the final manuscript.

Acknowledgements

This work was funded by the National Natural Science Foundation of China (Grant No. 30570420, 20637010) and by the Shenzhen Municipal Science and Technology Project (2007-2010). We thank the Supercomputer Centre of the Shenzhen University for the use of Deep-Super 21 supercomputer.

Author Details

[1]Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, PR China, [2]College of Life Sciences, Shenzhen University, Shenzhen, 518060, PR China and [3]Graduate University of the Chinese Academy of Sciences, Chinese Academy of Sciences, Changchun 130022, PR China

Received: 13 December 2009 Accepted: 10 May 2010
Published: 10 May 2010

References

1. Hatfield DL: **Selenium: Its Molecular Biology and Role in Human Health: Springer.** Springer; 2001.
2. Kryukov GV, Kryukov VM, Gladyshev VN: **New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements.** *J Biol Chem* 1999, **274(48):**33888-33897.
3. Atkins JF, Gesteland RF: **The twenty-first amino acid.** *Nature* 2000, **407(6803):**. 463, 465
4. Bock A: **Biosynthesis of selenoproteins--an overview.** *Biofactors* 2000, **11(1-2):**77-78.
5. Hatfield DL, Gladyshev VN: **How selenium has altered our understanding of the genetic code.** *Mol Cell Biol* 2002, **22(11):**3565-3576.
6. Kryukov GV, Gladyshev VN: **Selenium metabolism in zebrafish: multiplicity of selenoprotein genes and expression of a protein containing 17 selenocysteine residues.** *Genes Cells* 2000, **5(12):**1049-1060.
7. Castellano S, Morozova N, Morey M, Berry MJ, Serras F, Corominas M, Guigo R: **In silico identification of novel selenoproteins in the Drosophila melanogaster genome.** *Embo Rep* 2001, **2(8):**697-702.
8. Novoselov SV, Rao M, Onoshko NV, Zhi H, Kryukov GV, Xiang Y, Weeks DP, Hatfield DL, Gladyshev VN: **Selenoproteins and selenocysteine insertion system in the model plant cell system, Chlamydomonas reinhardtii.** *Embo J* 2002, **21(14):**3681-3693.
9. Taskov K, Chapple C, Kryukov GV, Castellano S, Lobanov AV, Korotkov KV, Guigo R, Gladyshev VN: **Nematode selenoproteome: the use of the selenocysteine insertion system to decode one codon in an animal genome?** *Nucleic Acids Res* 2005, **33(7):**2227-2238.
10. Novoselov SV, Hua D, Lobanov AV, Gladyshev VN: **Identification and characterization of Fep15, a new selenocysteine-containing member of the Sep15 protein family.** *Biochem J* 2006, **394(Pt 3):**575-579.
11. Castellano S, Novoselov SV, Kryukov GV, Lescure A, Blanco E, Krol A, Gladyshev VN, Guigo R: **Reconsidering the evolution of eukaryotic selenoproteins: a novel nonmammalian family with scattered phylogenetic distribution.** *Embo Rep* 2004, **5(1):**71-77.
12. Zhang Y, Fomenko DE, Gladyshev VN: **The microbial selenoproteome of the Sargasso Sea.** *Genome Biol* 2005, **6(4):**R37.
13. Zhang Y, Gladyshev VN: **Trends in selenium utilization in marine microbial world revealed through the analysis of the global ocean sampling (GOS) project.** *PLoS Genet* 2008, **4(6):**e1000095.
14. Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehtab O, Guigo R, Gladyshev VN: **Characterization of mammalian selenoproteomes.** *Science* 2003, **300(5624):**1439-1443.
15. Pennisi E: **Evolution and development. Comparative biology joins the molecular age.** *Science* 2002, **296(5574):**1792-1795.
16. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, *et al.*: **The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins.** *Science* 2002, **298(5601):**2157-2167.
17. Sordino P, Belluzzi L, De Santis R, Smith WC: **Developmental genetics in primitive chordates.** *Philos Trans R Soc Lond B Biol Sci* 2001, **356(1414):**1573-1582.
18. Shchedrina VA, Novoselov SV, Malinouski MY, Gladyshev VN: **Identification and characterization of a selenoprotein family containing a diselenide bond in a redox motif.** *PNAS* 2007, **104(35):**13919-13924.
19. Fujita M, Mihara H, Goto S, Esaki N, Kanehisa M: **Mining prokaryotic genomes for unknown amino acids: a stop-codon-based approach.** *BMC Bioinformatics* 2007, **8:**225.
20. Tujebajeva RM, Harney JW, Berry MJ: **Selenoprotein P expression, purification, and immunochemical characterization.** *J Biol Chem* 2000, **275(9):**6288-6294.
21. Kumaraswamy E, Malykh A, Korotkov KV, Kozyavkin S, Hu Y, Kwon SY, Moustafa ME, Carlson BA, Berry MJ, Lee BJ, *et al.*: **Structure-expression relationships of the 15-kDa selenoprotein gene. Possible role of the protein in cancer etiology.** *J Biol Chem* 2000, **275(45):**35540-35547.
22. Zhang Y, Romero H, Salinas G, Gladyshev VN: **Dynamic evolution of selenocysteine utilization in bacteria: a balance between selenoprotein loss and evolution of selenocysteine from redox active cysteine residues.** *Genome Biol* 2006, **7(10):**R94.
23. Zhang Y, Gladyshev VN: **High content of proteins containing 21st and 22nd amino acids, selenocysteine and pyrrolysine, in a symbiotic**

deltaproteobacterium of gutless worm Olavius algarvensis. *Nucleic Acids Res* 2007, **35(15):**4952-4963.

24. Worden AZ, Lee JH, Mock T, Rouze P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV, *et al.*: **Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas.** *Science* 2009, **324(5924):**268-272.
25. Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK, *et al.*: **The amphioxus genome and the evolution of the chordate karyotype.** *Nature* 2008, **453(7198):**1064-1071.
26. Korotkov KV, Novoselov SV, Hatfield DL, Gladyshev VN: **Mammalian selenoprotein in which selenocysteine (Sec) incorporation is supported by a new form of Sec insertion sequence element.** *Mol Cell Biol* 2002, **22(5):**1402-1411.
27. Heras B, Shouldice SR, Totsika M, Scanlon MJ, Schembri MA, Martin JL: **DSB proteins and bacterial pathogenicity.** *Nat Rev Microbiol* 2009, **7(3):**215-225.
28. Wunderlich M, Jaenicke R, Glockshuber R: **The redox properties of protein disulfide isomerase (DsbA) of Escherichia coli result from a tense conformation of its oxidized form.** *J Mol Biol* 1993, **233(4):**559-566.
29. Zapun A, Cooper L, Creighton TE: **Replacement of the active-site cysteine residues of DsbA, a protein required for disulfide bond formation in vivo.** *Biochemistry-Us* 1994, **33(7):**1907-1914.
30. Nelson JW, Creighton TE: **Reactivity and ionization of the active site cysteine residues of DsbA, a protein required for disulfide bond formation in vivo.** *Biochemistry-Us* 1994, **33(19):**5974-5983.
31. Wunderlich M, Otto A, Maskos K, Mucke M, Seckler R, Glockshuber R: **Efficient catalysis of disulfide formation during protein folding with a single active-site cysteine.** *J Mol Biol* 1995, **247(1):**28-33.
32. Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Carpenter ML, *et al.*: **The Trichoplax genome and the nature of placozoans.** *Nature* 2008, **454(7207):**955-960.
33. Chaudhuri BN, Yeates TO: **A computational method to predict genetically encoded rare amino acids in proteins.** *Genome Biol* 2005, **6(9):**R94.
34. Parra G, Blanco E, Guigo R: **GeneID in Drosophila.** *Genome Res* 2000, **10(4):**511-515.
35. Guigo R: **Assembling genes from predicted exons in linear time with dynamic programming.** *J Comput Biol* 1998, **5(4):**681-702.
36. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17):**3389-3402.
37. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22(22):**4673-4680.
38. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W: **A computer program for aligning a cDNA sequence with a genomic DNA sequence.** *Genome Res* 1998, **8(9):**967-974.
39. Hofacker IL, Stadler PF: **Memory efficient folding algorithms for circular RNA secondary structures.** *Bioinformatics* 2006, **22(10):**1172-1176.
40. Dsouza M, Larsen N, Overbeek R: **Searching for patterns in genomic data.** *Trends Genet* 1997, **13(12):**497-498.
41. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25(24):**4876-4882.

doi: 10.1186/1471-2164-11-289
Cite this article as: Jiang *et al.*, In silico identification of the sea squirt selenoproteome *BMC Genomics* 2010, **11**:289