How comparative genomic helps finding selenoproteins

roderic guigó i serra
IMIM/UPF/CRG, Barcelona

Comparative genomics. IMA 2003
what are selenoproteins?

Selenoproteins are proteins that incorporate selenocysteine, the 21st aminoacid

• Mostly redox enzymes
• Distributed in the three domains of life
• About 25 known selenoproteins in mammals, but the number varies for different taxa
the selenocysteine codon?

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>Phe</td>
<td>Ser</td>
<td>Tyr</td>
<td>Cys</td>
</tr>
<tr>
<td>C</td>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
<td>Arg</td>
</tr>
<tr>
<td>A</td>
<td>Met</td>
<td>Thr</td>
<td>Asn</td>
<td>Ser</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
<td>Gly</td>
</tr>
</tbody>
</table>

Legend: STOP for stop codons
the selenocysteine codon: UGA
recoding of UGA
the dual function of UGA compounds the identification of selenoproteins

Diagram:

- Standard gene
  - 5' UTR
  - Exon 1: ATG GT
  - Exon 2: AG STOP
  - 3' UTR

- Selenoprotein
  - 5' UTR
  - Exon 1: ATG GT
  - Exon 2: AG STOP
  - SECIS: TGA
Selenoprotein search: SECIS search

SECIS came in a variety of shapes
SECIS search: PatScan

Pairing rules
r1={at, ta, gc, cg, tg, gt}

Helix I and II

5'
| p1=5...15 | p2=1...7 | A TGA N | p3=9...12 |

Apical loop

p4=0...2 | AA | p5=6...17

Helix II

3'
| r1~p3[2,1,1] | N GAN | p6=3...9 | r1~p1[2,1,1] |

1- Matches 5' chain p3
2- Follows r1 pairing rules
3- 2 mismatches allowed
4- 2 bulges allowed

1- Matches 5' chain p1
2- Follows r1 pairing rules
3- 2 mismatches allowed
4- 2 bulges allowed
SECIS search in the Drosophila genome

- 35,876 potential SECIS elements
- 1,220 thermodynamically stable
selenoprotein search: codon bias
selenoprotein search: codon bias

Protein coding codon bias
No codon bias

selenoprotein
TGA
STOP

Non selenoprotein
TGA
STOP
selenoprotein search: SECIS + codon bias

1. Predict SECIS with PatScan
2. Gene prediction with geneid

Geneid uses dynamic programming to chain input exons into gene structures maximizing a log-likelihood function. SECIS predictions and TGA-interrupted exons are now among the input exons. Chaining rules state that SECIS elements can only be chained if they terminate genes containing TGA exons, and that genes containing TGA exon can only be terminated by SECIS predictions.
selenoprotein search in Drosophila

(Castellano et al. EMBO Reports 2:697-702, 2001)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SECIS predicted</td>
<td>35876</td>
</tr>
<tr>
<td>SECIS thermo assessment</td>
<td>1220</td>
</tr>
<tr>
<td>Genes predicted</td>
<td>12194</td>
</tr>
<tr>
<td>Predicted Selenoproteins</td>
<td>(4)</td>
</tr>
<tr>
<td>Real Selenoproteins</td>
<td>3</td>
</tr>
</tbody>
</table>
dSelG

A

dSelG

5'UTR
1080698 1080820 1080986 1081046 1081274 1081380 1081454

CG1840

5'UTR
1081763 1081917 1082071 1082131

B

TGA

3'UTR

C

dSelG

1 MVVLDHNGFVWEKRWDWRRIVELFGIWFAKQIFLTPLAPFTCGNMQANPRRGNGWG

CG1840

1 MAYVDHNGFVWEKRWDWRRLRVLTFVGIFWALKQLASTISPTGNDSNDNNRRGNWG

dSelG

60 ....GGGGWGGGGGGGGGGGGGGCPGSGS ....GGLRPNRRIGRIFP...MSCNMPAGGGCIG

CG1840

61 SSSWGGGCGGGGGGGGGGGGGGPGGGSCYRGGLRPNRRIGRIPPP...SCN...AGCGG

IMA 2003,
Comparative Genomics
dSelG and dSelM are ubiquitous selenoproteins
dSelG has selenoprotein homologues in vertebrates
selenoprotein search in mammalian genomes

• Larger genome. Much more room for false positive SECIS predictions

• Poorer gene predictions.
conserved SECIS between human and mouse
characterization of mammalian selenoproteins
(Kryukov et al., Science 300:1439-1443, 2003)
selenoprotein search in other vertebrate genomes
human vs. fugu
SelU: a novel selenoprotein family
(Castellano et al., EMBO reports, revised)
SelU: scattered phylogenetic distribution
the eukaryotic selenoproteome

<table>
<thead>
<tr>
<th>Selenoprotein families</th>
<th>Sec gene</th>
<th>Cys gene</th>
<th>Sec or Cys gene</th>
<th>Gene not present</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. discoideum</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>C. reinhardii</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>A. thalana</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>S. cerevisae</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>C. elegans</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>X. laevis</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>M. musculus</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>D. rerio</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>T. rubripes</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

10/24/03 IMA 2003, Comparative Genomics
IMIM, Barcelona
Sergi Castellano
Charles Chapple

Universitat de Barcelona
Marta Morey
Montserrat Corominas
Florenici Serras

Harvard University, Boston
Nadia Morozova
Marla Berry

University of Nebraska
Gregory V. Kryukov
Sergey V. Novoselov
Vadim N. Gladyshev

IBMC, Strasbourg
Alain Lescure
Alain Krol