Phylogenetic shadowing:
A Novel Tool To Identify Primate-Specific Regulatory Elements

Approach to the Identification of Primate-Specific Regulatory Elements
Pairwise species comparison

human/mouse

human/chimp

human/baboon

Approach to the Identification of Primate-Specific Regulatory Elements
Primate phylogeny

- Lemurs
- Lorises
- New-World Monkeys
- Cercopithecinae (baboons, etc)
- Colobinae (Colobus, Langurs)
- Old-World Monkeys
- Prosimians
- Human
- Chimp, Bonobo
- Gorilla
- Orang-utan
- Gibbons

Phylogenetic shadowing:

- human
- chimp
- baboon
- rhesus monkey
- green monkey
- colobos
- dusky titi
- spider monkey

Sequence elements conserved across all species:
- candidate regulatory element

: nucleotide difference with at least one species
1. PROOF-OF-PRINCIPLE:
Identification of intron-exon boundaries of 4 genes with known boundaries

2. TEST-CASE:
Identification of regulatory elements of a primate-specific gene

Amplification of target sequence from genomic DNA of ~20 primates (new- and old-world monkeys, hominoids)

Direct sequence of the PCR product

Multiple sequence alignment with ClustalW:

<table>
<thead>
<tr>
<th></th>
<th>Allen</th>
<th>Green</th>
<th>Human</th>
<th>Chimp</th>
<th>Orangutan</th>
<th>Colobus</th>
<th>Douc</th>
<th>Francois</th>
<th>Drill</th>
<th>Mangabey</th>
<th>Owl</th>
<th>Squirrel</th>
<th>Tamarin</th>
<th>Titi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CACCCGAATTTTTTGTATTTTTTAGAGACACAGGGTTTCACTACGTTGGCCAGGC</td>
<td>CACCGTGTAATTTTTTGTATTTTTAGTAGAGGACACAGGGTTTCACTACGTTGGCCAGGC</td>
<td>CACCCGGCTAATTTTTGTATTTTTAGTAGAGGACACAGGGTTTCACTACGTTGGCCAGGC</td>
<td>CACCGGGCTAATTTTTGTATTTTTAGTAGAGGACACAGGGTTTCACTACGTTGGCCAGGC</td>
<td>CACCGGGCTAATTTTTGTATTTTTAGTAGAGGACACAGGGTTTCACTACGTTGGCCAGGC</td>
<td>CACCCGGCTAATTTTTGTATTTTTAGTAGAGGACACAGGGTTTCACTACGTTGGCCAGGC</td>
<td>CACCGGGCTAATTTTTGTATTTTTAGTAGAGGACACAGGGTTTCACTACGTTGGCCAGGC</td>
<td>CACCCGGCTAATTTTTGTATTTTTAGTAGAGGACACAGGGTTTCACTACGTTGGCCAGGC</td>
<td>CACCGGGCTAATTTTTGTATTTTTAGTAGAGGACACAGGGTTTCACTACGTTGGCCAGGC</td>
<td>CACCCGGCTAATTTTTGTATTTTTAGTAGAGGACACAGGGTTTCACTACGTTGGCCAGGC</td>
<td>CACCCGGCTAATTTTTGTATTTTTAGTAGAGGACACAGGGTTTCACTACGTTGGCCAGGC</td>
<td>CACCGGGCTAATTTTTGTATTTTTAGTAGAGGACACAGGGTTTCACTACGTTGGCCAGGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>264</td>
<td>193</td>
<td>398</td>
<td>387</td>
<td>310</td>
<td>331</td>
<td>322</td>
<td>194</td>
<td>303</td>
<td>192</td>
<td>311</td>
<td>196</td>
<td>331</td>
<td>322</td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>*</td>
<td>***</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>**</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>**</td>
<td>**</td>
<td>**********</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
</tbody>
</table>

Consensus

* *  *** ************** * **  ** ******* **********
Identification of plasminogen exon 6

Exon prediction with a two-state Hidden Markov Model:
1) site averagely conserved
2) site very well conserved

Identification of LXR-α exon 3

Exon prediction with a two-state Hidden Markov Model
Identification of CETP exon 8

Exon prediction with a two-state Hidden Markov Model

Identification of ApoB exon 19

Exon prediction with a two-state Hidden Markov Model
MAXIMUM LIKELIHOOD PHYLOGENETIC SHADOWING:

Learn mutation rates for ‘conserved’ and ‘non-conserved’ regions on a training set

Calculate log-likelihood ratios of rates for each column of the multiple alignment using the phylogenetic trees.

Plasminogen exon 6
Amplification of evolutionary distance through multiple species comparison successfully identifies known functional elements

Identification of regulatory elements of a primate-specific gene: apo(a)
apolipoprotein(a)

- Lacks a mouse ortholog:
- People with high levels are at risk of developing atherosclerosis

apo (a) is a plasma protein

apo B

Lp(a)

apo (a)

LDL
apo(a) plasma levels are highly variable

* apo(a) vary 1000-fold

* standard protein ~10-fold

~ 50% of apo(a) variability in levels is attributable to the regulation of the transcription of the gene

Phylogenetic shadowing of the apo(a) promoter
Testing the function of conserved elements

Cell culture transcription assay:

Phylogenetic shadowing of the apo(a) promoter

- Apo(a) promoter region (1.6kb)
- Luciferase

<table>
<thead>
<tr>
<th>Position (bp)</th>
<th>log[lik(fast)/lik(slow)]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>450</td>
<td></td>
</tr>
<tr>
<td>900</td>
<td></td>
</tr>
<tr>
<td>1350</td>
<td></td>
</tr>
<tr>
<td>1600</td>
<td></td>
</tr>
</tbody>
</table>

- HepG2

- Luciferase/b-gal (%full construct):
  - 100%
  - 75%
  - 50%
  - 0%
Average of expression data for conserved and non-conserved elements

![Graph showing expression data for conserved and non-conserved elements.]

Phylogenetic shadowing of the apo(a) promoter

![Graph showing phylogenetic shadowing of the apo(a) promoter.]

- **P<0.002**
Gel-shift assay to assess DNA-protein interactions

Gel-shift analysis of conserved elements in the apo(a) promoter
Quantitative analysis of the gel-shift pattern

genomic sequence from multiple primates can be used to identify functional sequence shared by this group of species
apo(a) is a recently evolved gene but...

Strategy to unravel the mechanism of evolution of the apo(a) gene

Sequence the genomic region containing apo(a) and its orthologous region in species without apo(a):

Species with apo(a):
• human: publicly available
• baboon: sequence
• hedgehog: generate BAC library and sequence

Species without apo(a):
• mouse: sequence
• lemur: generate BAC library and sequence
human apo(a) is a duplication of plasminogen

human plasminogen

hedgehog plasminogen

kringle domains

FIBRIN-BINDING
apo(a) is a unique example of convergent evolution where the same gene arose twice through the independent duplication and remodeling of the same paralogous gene.