

Ultra-Conserved Elements

Nicholas Eriksson

Department of Mathematics
University of California, Berkeley

14th December 2005

Primera Escuela Argentina de Matemática y Biología

Joint work with Mathias Drton and Garmay Leung

Outline

Ultra-Conserved
Elements

Nicholas Eriksson

Ultra-conserved elements

Ultra-conserved
elements

Datasets

Vertebrate

Drosophila

Datasets

Vertebrate

Drosophila

Biology of
ultra-conservation

Vertebrate

Drosophila

Probability of
ultra-conservation

Biology of ultra-conservation

Vertebrate

Drosophila

Summary

Probability of ultra-conservation

Ultra-conserved elements

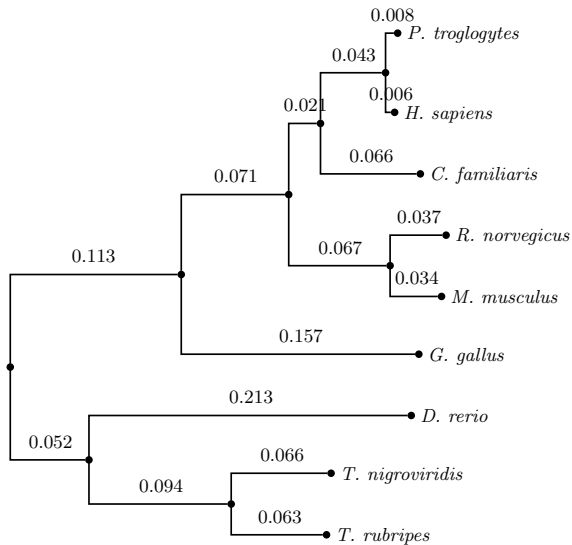
Consider 3 toy genomes in a multiple alignment of length 24:

```
G--ACCCAATAGCACCTGTTGCGG
CGCTCTCCA---CACCTGTTCCGG
CATTCT-----CTGTTTTGG
      *           ****   **
```

Nine-vertebrate alignment

Name	Scientific	Genome size
zebra fish	<i>Danio rerio</i>	1.7 Gbp
fugu fish	<i>Takifugu rubripes</i>	365 Mbp
puffer fish	<i>Tetraodon nigroviridis</i>	350 Mbp
dog	<i>Canis familiaris</i>	2.4 Gbp
human	<i>Homo sapiens</i>	2.8 Gbp
chimp	<i>Pan troglodytes</i>	~3 Gbp
mouse	<i>Mus musculus</i>	2.5 Gbp
rat	<i>Rattus norvegicus</i>	2.8 Gbp
chicken	<i>Gallus gallus</i>	1.1 Gbp

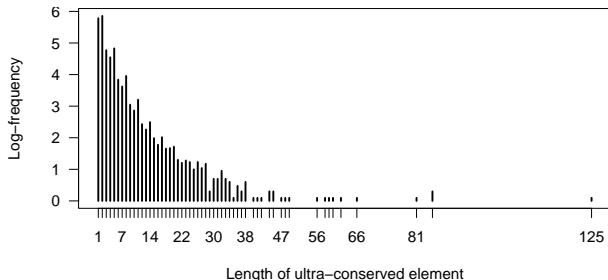
Nine-vertebrate phylogeny



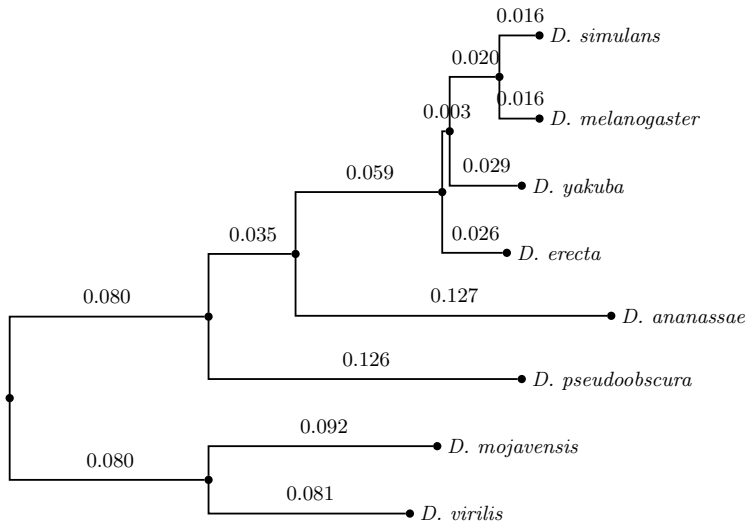
Ultra-conserved elements

In 9 vertebrate alignment

- ▶ 1,513,176 elements.
- ▶ longest element: 125 bp.
- ▶ Mean length 1.918, median length 2.
- ▶ 237 elements of length at least 20 bp.
- ▶ GC-ratio 35.8% (versus 41.0% genome wide)



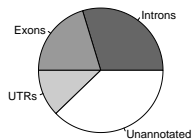
Drosophila phylogeny



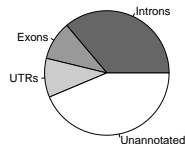
Biology of ultra-conserved elements

Nine-vertebrate alignment

Base coverage of ultra conserved elements by type:



209 elements \geq 20 bp



59 elements \geq 30 bp

Examples

Nine ultra-conserved elements cover a total of 306 bp in introns of *DPOA*, the alpha catalytic subunit of **DNA polymerase**.

Many of the flanking genes seem to be **transcription factors** such as *IRX3*, *HOXD13*, *DMRT1*, *FOXD3*.

How can we tell if the genes associated with ultra-conserved elements are “special”?

The Gene Ontology project

Ultra-Conserved
Elements

Nicholas Eriksson

Ultra-conserved
elements

Datasets

Vertebrate

Drosophila

Biology of
ultra-conservation

Vertebrate

Drosophila

Probability of
ultra-conservation

Summary

Categorizes gene products according to

- ▶ molecular function
(e.g., catalytic activity, transcription factor activity)
- ▶ biological process
(e.g., neurogenesis, pyrimidine metabolism)
- ▶ cellular component
(e.g., ribosome, rough endoplasmic reticulum)

http://www.geneontology.org/

the Gene Ontology

Search go!
gene or protein name

Gene Ontology Home

The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism. [Read more...](#)

Popular Links

Search the Gene Ontology Database

ACTG GO!

gene or protein name GO term or ID

This search uses the browser [AmiGO](#), [Browse](#) the Gene Ontology using AmiGO.

GO website

- [GO downloads](#): including [ontology files](#), [annotations](#) and the [GO database](#)
- [Tools](#) for using GO
- Request new terms or ontology changes via the [SourceForge tracker system](#); [help with new term submission](#) is available.
- [Documentation](#) on all aspects of the GO project and [the FAQ](#)
- [Gene Ontology mailing lists](#) and [contact details](#)

[Back to top](#)

News

AmiGO

Search GO
ACTG
Exact Match
Terms
Gene Symbol/Name
Submit Query
Advanced Query
Query By Sequence

Gene Product Filters
Species
All
A. japonica
A. tiger
Datasource
All
FlyBase
SGD
Evidence Code
All Curator Approved
IC
IMP
Set Filters

Query Summary
Your Query
ACTG
Exact Match
no
Target
Gene Products
Fields
Gene Symbol
Datasources
All
Evidence Codes
All
Results

http://www.godatabase.org/cgi-bin/amigo/go.cgi?action=query&view=query&query

Actg1, actin, gamma, cytoplasmic 1
gene from *Mus musculus*, data from MGI (MGI:87906)

Term	Ontology	Evidence	Reference
sarcomere organization	P	IDA	PMID:15194427
actin cytoskeleton	C	IDA	PMID:15194427
costamere	C	TAS	PMID:15194427
myofibril	C	IDA	PMID:15194427
structural constituent of cytoskeleton	F	IDA	PMID:15194427

Actg2, actin, gamma 2, smooth muscle, enteric
gene from *Mus musculus*, data from MGI (MGI:104589)

Term	Ontology	Evidence	Reference
muscle development	P	RCA	PMID:12466851
actin cytoskeleton	C	RCA	PMID:12466851
actin filament	C	RCA	PMID:12466851
motor activity	F	RCA	PMID:12466851
structural constituent of muscle	F	RCA	PMID:12466851

Actg2, actin, gamma 2
gene from *Rattus norvegicus*, data from RGD (RGD:2027)

Term	Ontology	Evidence	Reference
actin filament	C	ISS	RGD:631723

Check/Uncheck All | Get Detailed View | Submit Query

GO analysis

9-vertebrate

Ultra-Conserved
Elements

Nicholas Eriksson

Ultra-conserved
elements

Datasets

Vertebrate
Drosophila

Biology of
ultra-conservation

Vertebrate
Drosophila

Probability of
ultra-conservation

Summary

GO Annotation	<i>p</i> -value
Exons	
protein serine/threonine kinase activity	4.545×10^{-3}
transferase activity	1.494×10^{-2}
neurogenesis	1.654×10^{-2}
protein amino acid phosphorylation	2.210×10^{-2}
Introns	
regulation of transcription, DNA-dependent	8.755×10^{-4}
transcription factor activity	2.110×10^{-3}
protein tyrosine kinase activity	4.785×10^{-3}
protein amino acid phosphorylation	1.584×10^{-2}
protein serine/threonine kinase activity	2.806×10^{-2}
UTRs	
regulation of transcription, DNA-dependent	1.403×10^{-4}
transcription factor activity	3.971×10^{-3}
Flanking	
transcription factor activity	3.255×10^{-11}
regulation of transcription, DNA-dependent	2.021×10^{-8}
development	5.566×10^{-3}

A repeated ultra-conserved element

GACATGGAGA AGATCTGGCA CCACACCTTC TACAA

Appears

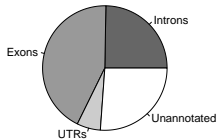
- ▶ 19 times in the human genome: Twice in exons of actin genes ACTC and ACTG and 17 times in predicted retroposed pseudogenes with actin parent genes.
- ▶ 13 times in chimp, 10 times in mouse, 5 times in both rat and dog, 4 times in tetraodon, 3 times in zebra fish, and twice in both fugu and chicken.
- ▶ Also in *C. intestinalis*, *Drosophila*, and *A. gambiae* in actin genes with 1 or 2 mutations.

Drosophila alignment

Base coverage of ultra conserved elements by type:



255 elements \geq 75 bp



59 elements \geq 100 bp

GO analysis (*Drosophila*)

GO Annotation	<i>p</i> -value
Exons, Introns, and UTRs	
synaptic transmission	3.290×10^{-9}
specification of organ identity	1.044×10^{-6}
ventral cord development	3.674×10^{-6}
RNA polymerase II transcription factor activity	4.720×10^{-6}
muscle contraction	8.714×10^{-6}
voltage-gated calcium channel activity	3.548×10^{-5}
RNA binding	7.650×10^{-5}
Flanking	
regulation of transcription	8.844×10^{-7}
neurogenesis	5.339×10^{-6}
ectoderm formation	8.285×10^{-6}
endoderm formation	2.125×10^{-5}
salivary gland morphogenesis	5.870×10^{-5}
Notch signaling pathway	1.591×10^{-4}
leg joint morphogenesis	1.788×10^{-4}
RNA polymerase II transcription factor activity	2.381×10^{-4}

Repeated ultras

None of the ultra-conserved elements correspond to annotated rRNA, regulatory regions, transposable elements or pseudogenes.

Ten of the 255 ultra-conserved elements are repeated elsewhere in the *D. melanogaster* genome. These repeats correspond to annotated tRNA or snRNA.

Probability of ultra-conserved regions

Ultra-Conserved
Elements

Nicholas Eriksson

Ultra-conserved
elements

Datasets

Vertebrate

Drosophila

Biology of
ultra-conservation

Vertebrate

Drosophila

Probability of
ultra-conservation

Summary

How long of an ultra-conserved element could occur simply by chance?

First assume that the nucleotides in the different positions in the alignment are mutually independent and apply a phylogenetic tree model.

Building the tree

- ▶ Used PAML to compute maximum likelihood parameters for JC and HKY models.
- ▶ Fixed the true tree topology.
- ▶ Used the **entire** alignment as input.

Probability

Nine-vertebrate (human)

	JC	HKY
p_{cons}	0.0456	0.0147
10	0.0001	$1.3 \cdot 10^{-9}$
20	$4.1 \cdot 10^{-18}$	$6.2 \cdot 10^{-28}$
125	$6.0 \cdot 10^{-159}$	$2.4 \cdot 10^{-220}$

Drosophila (*D. melanogaster*)

	JC	HKY
p_{cons}	0.1071	0.05969
15	$7.8 \cdot 10^{-6}$	$1.2 \cdot 10^{-9}$
75	$4.6 \cdot 10^{-64}$	$4.3 \cdot 10^{-83}$
209	$4.3 \cdot 10^{-194}$	$4.1 \cdot 10^{-247}$

A simple non-independence model

Collapse the alignment to a sequence of binary indicators of ultra-conserved positions:

NNNNN UUUNNNNNNNNN UUUUUUUNNNNNNNNN

Model this sequence with a Markov chain.

We can estimate the transition probability for $U \rightarrow U$ as $\theta = 0.4785$ for the nine-vertebrate alignment.

Probability that at least one of U ultra-conserved elements would be of length at least ℓ is

$$P(\ell, U) = 1 - (1 - \theta^{\ell-1})^U \approx U \cdot \theta^{\ell-1}$$

$$P(25, 1513176) = 0.03$$

$$P(125, 1513176) = 3 \cdot 10^{-34}$$

Summary

- ▶ What is an appropriate statistical model for ultra-conserved elements? Perhaps phylogenetic hidden Markov models?
- ▶ What length ultra-conserved element is significant for a given evolutionary distance under these models?
- ▶ Do ultra-conserved elements arise from negative selection or reduced mutation rates?
- ▶ Are there biological structures that are so sensitive to DNA sequences that they depend on the exact same sequence over hundreds of millions of years of evolution?
- ▶ Is the analysis of ultra-conserved elements a good method for comparing alignments?