Evolutionarily conserved elements

Age

JClub, October 31st, 2005

How to find functional sequences?

Look for sequences that are conserved across species.

Orthologous sequences that are significantly more similar than expected are likely to have critical functional roles.

Based on analyses of human and rodent genomes:

About 5% or more of bases in mammalian genomes are under purifying selection

Protein-coding genes account for only about 1.5% of bases

3.5% conserved (functional) noncoding sequences

Pairwise alignments and simple percent-identity based methods

VISTA, PipMaker, zPicture

‘-’: do not use phylogeny

use sliding window of fixed size

PhastCons – to identify conserved elements in multiply aligned sequences.

Based on phylogenetic hidden-Markov model and considers

1) the process by which nucleotide substitutions occur at each site in a genome and

2) how this process changes from one site to next

'+': do not require a sliding window of fixed size;

allow nearly all parameters to be estimated from the data by maximum likelihood;

efficient on large-scale datasets

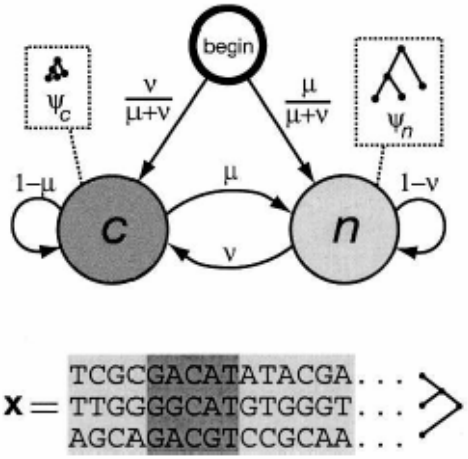


Figure 1. State-transition diagram for the phylo-HMM used by phastCons, which consists of a state for conserved regions (*c*) and a state for nonconserved regions (*n*). Each state is associated with a phylogenetic model (ψ_c and ψ_n); these models are identical except for a scaling parameter ρ ($0 \leq \rho \leq 1$), which is applied to the branch lengths of ψ_c and represents the average rate of substitution in conserved regions as a fraction of the average rate in nonconserved regions (see Methods). Two parameters, μ and ν ($0 \leq \mu, \nu \leq 1$), define all state-transition probabilities, as illustrated. The probability of visiting each state first (indicated by arcs from the node labeled "begin") is simply set equal to the probability of that state at equilibrium (stationarity). The model can be thought of as a probabilistic machine that "generates" a multiple alignment, consisting of alternating sequences of conserved (dark gray) and nonconserved (light gray) alignment columns (see example at *bottom*).

Siepel et al 2005

Four separate genome-wide multiple alignments (MULTIZ program):

- 4 vertebrates (human, mouse, rat, chicken) – reference genome human
- 4 insects – reference genome *D.melanogaster*
- 2 worms – reference genome *C.elegans*
- 7 yeasts – reference genome *S.cerevisiae*

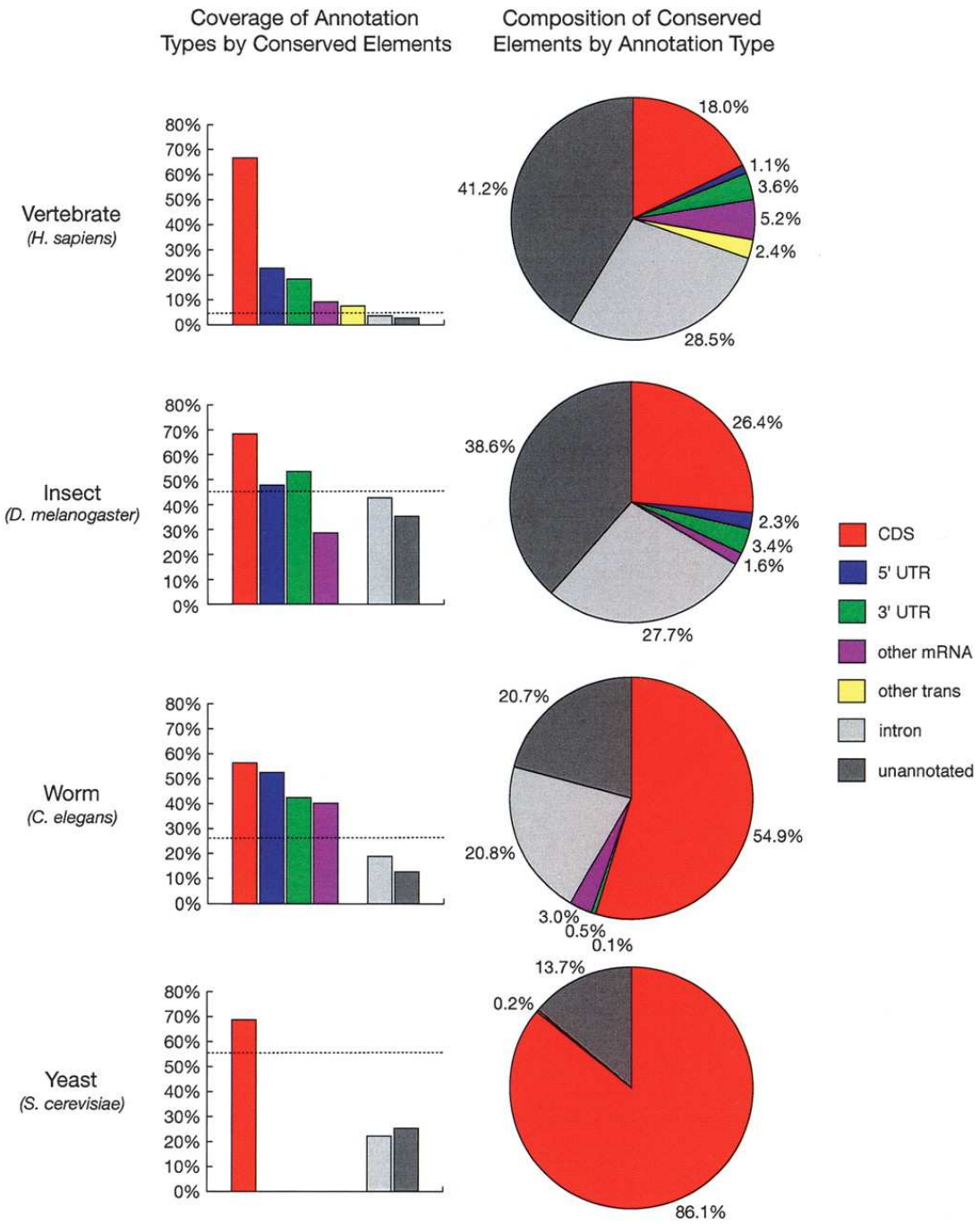
Predicted elements covered:

4.3% of the human genome

44.5 of the *D.melanogaster* genome

26.4% of *C.elegans* genome

55.6% of *S.cerevisiae* genome



Most conserved bases in vertebrates and insects do not code for proteins → the importance of gene regulation in complex eukaryotes

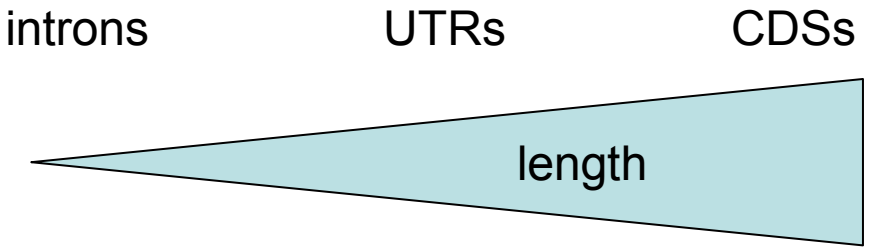
The lengths of predicted elements:

100-120 bp for vertebrate, insects and yeast groups

270 bp for worm group

(5bp – thousands of bp)

In vertebrates:



Human chr4:158,612,000-158,616,000 - UCSC Genome Browser v119 - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://genome.ucsc.edu/cgi-bin/hgTracks

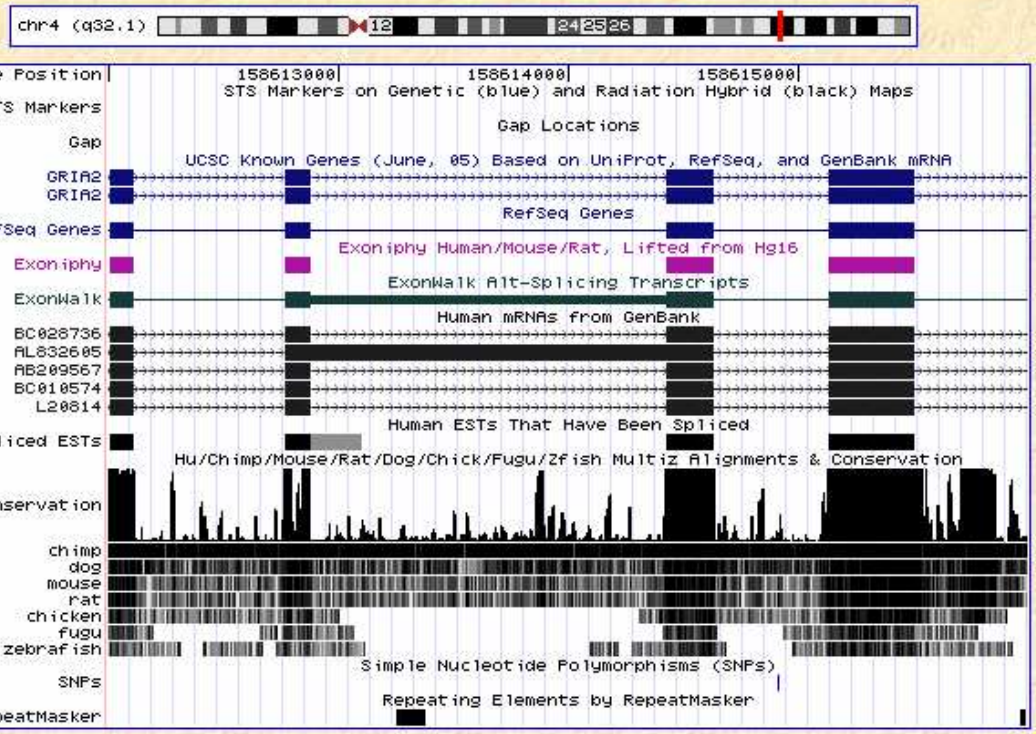
Postimees DELFI Tartu Ülikooli õppein... Department of Bioinf... Entrez PubMed meteo.physic.ut.ee

Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Help

UCSC Genome Browser on Human May 2004 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr4:158,612,000-158,616,000 jump clear size 4,001 bp. configure



move start < 2.0 >

Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for track-specific options.

move end < 2.0 >

default tracks hide all configure refresh

Done

start BIOINF - ... E:\loengu... JC311005... 3 Firefox Adobe Acr... FESTART ... ET 8:56

Human chr4:158,612,773-158,612,802 - UCSC Genome Browser v119 - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://genome.ucsc.edu/cgi-bin/hgTracks?hgsid=62718088&hgt.right1=+%3E+&position=chr4%3A158612773-158612802

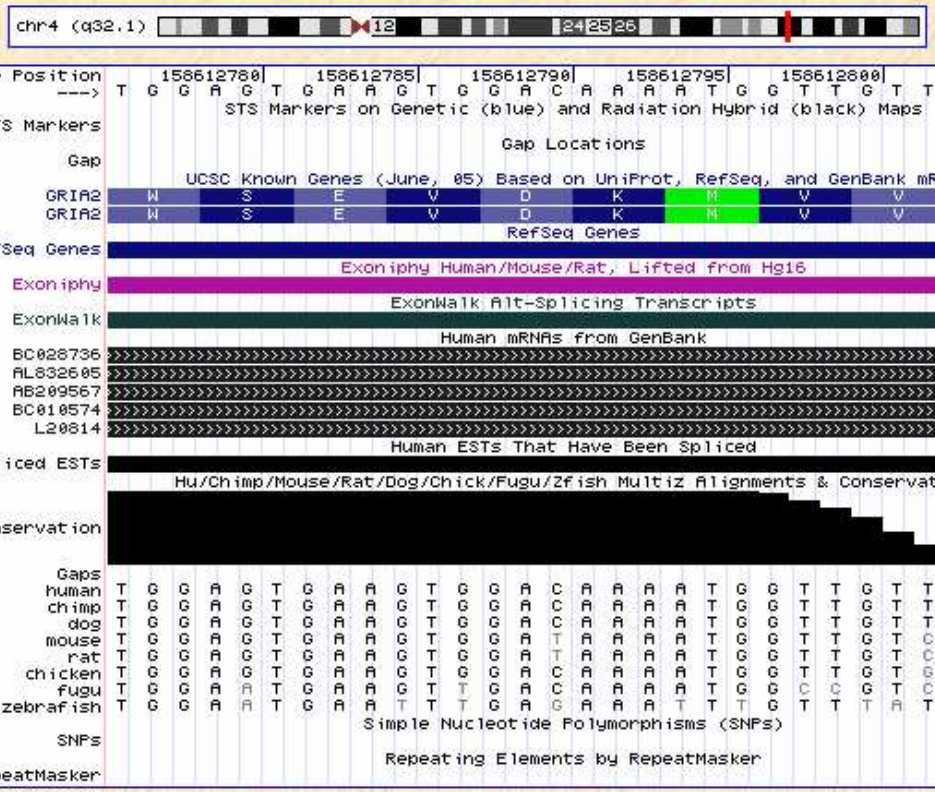
Postimees DELFI Tartu Ülikooli õppein... Department of Bioinf... Entrez PubMed meteo.physic.ut.ee

Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Help

UCSC Genome Browser on Human May 2004 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr4:158,612,776-158,612,805 jump clear size 30 bp. configure



move start < 2.0 >

Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for track-specific options

move end < 2.0 >

Done

start BIOINF - ... E:\loengu... JC311005... 3 Firefox Adobe Acr... FESTART ... ET 9:00

Highly conserved elements (HCE)

- Longer than UCEs (ultraconserved elements) (in vertebrates average 780 bp)
- Less extreme sequence conservation than in UCEs (due to the length dependency of log-odds scores)
- Based on different set of species
- Set of vertebrate HCEs is 10-fold larger than the set of UCEs
- Vertebrate HCEs include 80% of human/rodent UCEs
- More strongly asscociated with genes as genome sizes become smaller and gene densities increase
- 14.3 % top 100 vertebrate HCE's overlap 3'UTR's (in all conserved elements 5.6%)

- Significant enrichment for local secondary structures in 3'UTR, 5' UTR, introns and intergenic regions HCE's. *Manuscript in preparation (Pedersen, Bejerano and Haussler)*

- In vertebrates, intergenic HCEs are strongly enriched in stable gene deserts, suggesting that many of them may act as distal *cis*-regulatory elements

Highlights of the study:

- Larger genome and more complex organism – more conserved bases outside of known or suspected exons of protein-coding genes
- Some of the most extreme conservation is in 3'UTRs of vertebrates genes which regulate other genes
- HCEs in vertebrate 3'UTRs, introns and intergenic regions are enriched with local RNA secondary structures
- Intergenic HCEs in vertebrates are strongly enriched in stable gene deserts

- Siepel *et al* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res., 15, 1034-1051
- Bejerano *et al* (2004) Ultraconserved elements in human genome. Science, 304, 1321-1325.