Comparison of Human Sequence with the DNA of other species

Conserved elements identified by sequence alignments

- coding sequences
- regulatory elements

Human high-density oligonucleotide arrays

Rapid method of comparing the DNA of different individuals to identify haplotypes.

A feasible method for large-scale comparisons of human sequences with those of other mammals?
Comparative analysis of human chromosome 21

Chromosome 21 ~ 33 Mb

Syntenic mouse and dog BACs

~22.5 Mb of non-repetitive sequence

A series of high-density arrays consisting of ~ 180 million oligonucleotides (25-mers)

Syntenic mouse and dog sequences hybridized to arrays

BACs isolated using cross-species probes

360 mouse BACs representing 74% of human 21q

61 dog BACs representing 12% of human 21q
Syntenic mouse and dog sequences hybridized to arrays

BACs isolated using cross-species probes

360 mouse BACs representing 74% of human 21q

Human/Mouse (16.5 Mb)

61 dog BACs representing 12% of human 21q

Human/Dog (2.7 Mb)

Conserved Sequences

% Conformance
(Measure of similarity)

Fluorescent intensity
Complementary Probe greater than
Non-complementary Probes

Dog DNA hybridized

TTGAAATGGCTAGAAGCATTAGTG
AACTTACCAGATCATTGAATCAC
AACATTACCAGATCGTAGAATCAC
AACATTACCAGATCGTAGAATCAC
**Conserved Sequences**

Array hybridized with syntenic dog BACs

Conserved sequence = 30 bp in length

≥ (18/30) 60% conformance

(TGTCGGAATTCTTTGAAACCGGCTCCACATA)

(29/30) 97% conformance

(GGAAGCCACTACACGAAAACACTGAG)

(18/30) 60% conformance

**Conserved Elements**

Conserved element = 1 or more conserved sequences separated by

≤ 100 bp of non-conserved sequence
Specificity and Sensitivity of Array Data

False Positive Rate = ~ 1-2% of identified conserved elements

**Analyzed 600-kb of tiled 21q sequence**
- hybridized with random mouse DNA - 3 elements 190 bp
- hybridized with syntenic mouse DNA - 203 elements 16,010 bp

False Negative Rate = ~ 25% of conserved sequences are not detected

**22 human chromosome 21 genes previously sequenced in mouse**
- 67 exons with $E \geq 10^{-20}$ 46% not identified on array
- 123 exons with $E \leq 10^{-20}$ 15% not identified on array

Human high-density arrays provide a feasible method for large-scale comparisons of human sequences with those of other mammals.
How many conserved human-mouse and human-dog elements were identified by array analysis?

Is the distribution of conserved elements different in gene-rich and gene-poor regions?

What is the best method for distinguishing between conserved sequences due to active conservation and those due to evolutionary background?

Conserved Elements Identified by Array Analysis

**Human-mouse analysis**
- 1.6% of non-repetitive base pairs conserved
  (260,226 of 16,596,010 bp queried)
- identified 3398 conserved elements

**Human-dog analysis**
- 3.9% of non-repetitive base pairs conserved
  (101,491 of 2,595,732 bp queried)
- identified as 1292 conserved elements
What is the distribution of conserved elements in genic and non-genic intervals?

- **Nongenic**
- **Genic**

<table>
<thead>
<tr>
<th>Gene A</th>
<th>Gene B</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-kb</td>
<td>10-kb</td>
</tr>
</tbody>
</table>

- **In known exons**
- **Not in known exons**

**Approximately 38% of conserved bases are in non-genic intervals.**

- Nongenic - ~ 1.0% are conserved (86,450 of 8,533,733 bp queried)
- Genic - ~ 1.9% are conserved (140,626 of 7,510,102 bp queried)
How can we distinguish conserved sequences due to active conservation from conserved sequences due to lack of divergence?

Two mechanisms resulting in conserved sequences

1. Insufficient divergence time  
2. Active conservation

Threshold of significance

We cannot answer this question based only on length thresholds.
Use of High-Density Arrays in the Analysis of Human Chromosome 21 Sequences

SIM2 - 100-kb region upstream

Percentage of human-mouse conserved elements also conserved in the dog

Multiple-species comparisons identify actively conserved elements better than length thresholds alone.
Percent of human-dog conserved elements also conserved in the mouse.

- Conserved elements in known exons:
  - Dog: 87
  - Dog/Mouse: 132
    - ~ 60%
  - Mouse: 28
    - ~ 82%

- Conserved elements not in known exons:
  - Dog: 842
  - Dog/Mouse: 114
    - ~ 12%
  - Mouse: 120
    - ~ 50%

Why multiple species instead of just three species?

Functional DNA sequences evolve at different rates between species as well as within a single species.
Comparison of human sequence with the DNA of multiple species will be important for generating a comprehensive list of potential functional elements.

Large-scale methods for identifying actively conserved human sequences

A. 3X Shotgun

+ ~ 18,000,000 sequence reads per mammalian genome

B. Comparative hybridization

300 wafers entire human genome

+ ~ 20,000 BACs per mammalian genome
Use of High-Density Arrays in the Analysis of Human Chromosome 21 Sequences

Perlegen Sciences

Nila Patil
Renee Stokowski
John Sheehan
Cindy Chen
David Cox