Current Topics in Genome Analysis  
Fall 2006  

Week 5 (Part 2): Detection and Characterization of Non-Coding Functional Elements  
Elliott H. Margulies, Ph.D.

Sequencing Complete


Next Phase: Interpretation

Characterizing the Human Genome

- ~3 billion bases
- 20,000-25,000 protein-coding genes
**What are Genomic Functional Elements?**

- DNA sequences that either encode for some functioning unit (i.e. RNA) or that bind to proteins that perform some function

**Non-coding Functional Elements**

- Critical for gene regulation
- Maintain/Modify chromatin structure
- Candidate regions for human disease mutations
- Better understanding of human biology
- Changes in gene regulation rather than gene structure might be more influential in evolution (King & Wilson, 1975)

Identifying Functional Elements

- We understand the “language” of coding sequences (i.e., protein-coding genes)
  - Exons and introns
  - Triplet code
  - Complementary datasets (i.e., ESTs, cDNAs)

- The language of non-coding functional elements is poorly understood
  - We don’t know what to look for
  - Signal:Noise problem with short degenerate motifs

Multi-Disciplinary Approaches are Needed

- Find sequences that are likely functional
  *without prior knowledge of the function*

- Then characterize functions

Experimental Wet-lab Research

Computational Analyses
Comparative Genomics to Decode the Genome

Charles Darwin

- Served as naturalist on a British science expedition around the world (1831 -- 1836)

- The Origin of Species (1859)
  - All species evolved from a single life form
  - “Variation” within a species occurs randomly
  - Natural selection
  - Evolutionary change is gradual
Other Intellectual Foundations

- **Darwin (1859)**
  Theories of Evolution
- **Mendel (1866) (rediscovered in 1900)**
  Genes are units of heredity
- **Avery, McCarty & MacLeod (1944)**
  DNA as the “transforming principle”
- **Watson & Crick (1953)**
  Structure of DNA
- **Sanger (1977)**
  Methods of sequencing DNA

Rationale Behind Comparative Genomics

- DNA represents a “blueprint” for the structure and physiology of all living things
- All species use DNA
- Mutations occur randomly throughout the genome
  - Neutral theory of evolution (M. Kimura, 1983)
- Mutations in *functional* DNA are less likely to be tolerated

Fewer Mutations are Found in Functional DNA

- Functional sequences will be “more similar” when compared between different species

Comparative Genomics

- Find sequences that have diverged less than we expect
  
  *These sequences are likely to have a functional role*

- Our expectation is related to the time since the last common ancestor
Comparative Sequence Analysis

- Generate comparative sequence datasets
  - Targeted approaches
    - NISC Comparative Sequencing Program
      http://www.nisc.nih.gov
  - Genome-wide
    - “Finished” genomes
    - Draft whole-genome shotgun
    - Low-redundancy sequencing

- Generate multi-sequence alignments
- Downstream analysis efforts

Sequence Alignments

100% Identical
Species 1 CATGGGCAAATTGGCCCATTGGCCATGGGGGCCCACCGTA
Species 2 CATGGGCAAATTGGCCCATTGGCCATGGGGGCCCACCGTA

80% Identical
Species 1 CATGGGCAAATTGGCCCATTGGCCATGGGGGCCCACCGTA
Species 2 CA C G G C T A T C C G C C A A T T G G C T A T G G G G - C C A G

30% Identical
Species 1 CATGGGCAAATTGGCCCATTGGCCATGGGGGCCCACCGTA
Tools for Aligning Genomic Sequences (Targeted Regions)

Resource:

PipMaker—A Web Server for Aligning Two Genomic DNA Sequences
Scott Schwartz,1 Zheng Zhang,1 Kelly A. Frazer,2 Ariane Smit,1 Cathy Bieger,1 John Bouch,3 Richard Gibbs,4 Ross Hardison,3 and Webb Miller1,4

Departments of1,2 Computer Science and Engineering and 3,4 Microbiology and Molecular Biology and Center for Gene Regulation, The Pennsylvania State University, University Park, Pennsylvania USA (2000); 5Center Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, California USA (2004). 6Taco Pharmaceuticals, La Jolla, California USA (2007); Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas USA (1999).

Bioinformatics APPLICATIONS NOTE

VISTA: visualizing global DNA sequence alignments of arbitrary length
Chris Meyer1, Michael Brutger2, Jody R. Schwartz3, Alexander Polakow2, Edward M. Rubin2, Kelly A. Frazer2, Lior S. Porat1,5,6,1 and Inna Dubchak1,6,7

1National Energy Research Scientific Computing Center, 2Lawrence Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; 3Center for Bioinformatics, University of California at Berkeley, Berkeley, CA 94720, USA; 4Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania, USA; 5Department of Biophysics, The Pennsylvania State University, University Park, Pennsylvania, USA; 6Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania, USA; 7Department of Computer Science and Engineering, The University of Texas, Austin, Texas, USA.

Resources for Targeted Sequence Analysis

Resource:

zPicture: Dynamic Alignment and Visualization Tool for Analyzing Conservation Profiles
Ivan Ovcharenko1,2, Gabriela G. Loots3, Ross C. Hardison3, Webb Miller4,5, and Lisa Stubbs2,6

1,2 Energy, Environment, Biology and Institutional Computing, Lawrence Livermore National Laboratory, Livermore, California 94550, USA; 3,4,5 Genome Biology Division, Lawrence Livermore National Laboratory, Livermore, California 94550, USA; 6Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; 7Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA.


http://www.dcode.org/
Genome-wide Multi-sequence Alignments

- This is not a “solved problem”

- Significant challenges:
  - Finding the correct sequences to align
  - Not all sequences should align
  - Dealing with insertions/deletions
  - Handling duplications and rearrangements
  - Missing data challenges (i.e., sequencing gaps)

Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner

Mathieu Blanchette, W. James Kent, Cathy Riemer, Laura Elnitski, Aron F.A. Smitt, Krishna M. Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clavien, Eric D. Green, David Haussler, and Webb Miller


LAGAN and Multi-LAGAN: Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA

Michael Brudno, Chuang B. Do, Gregory M. Cooper, Michael F. Kim, Eugene Davydov, NISC Comparative Sequencing Program, Eric D. Green, Arend Sidow, and Serafim Batzoglou


MAVID: Constrained Ancestral Alignment of Multiple Sequences

Nicolas Bray and Lot Pachter

Genome Browsers

- UCSC Genome Bioinformatics
  - http://genome.ucsc.edu
- Ensembl
  - http://www.ensembl.org
- NCBi Map Viewer

Multi-sequence Alignments at UCSC

Click here for track details page
Chaining Alignments

- Chaining bridges the gulf between large syntenic blocks and base-by-base alignments.

The Challenge:

- Local alignments tend to break at transposon insertions, inversions, duplications, etc.
- Global alignments tend to force non-homologous bases to align.

The Solution:

- Chaining is a rigorous way of joining together local alignments into larger structures.

Chains join together related local alignments

Protease Regulatory Subunit 3
Net Alignments: Focus on Orthology

- Frequently, there are numerous mouse alignments for any given human region, particularly for coding regions.
- Net finds best mouse match for each human region.

*Click here for a more complicated example*
Summary of Alignments

- Not a solved problem
- Accuracy of alignment significantly affects downstream analyses
- Choosing the correct orthologous sequences to align is a major challenge

Constrained Sequences

- Highly conserved sequences
- Sequences under purifying selection
- **ECOR** – Evolutionary **COnserved Region**
  - Variant: ECR
- **CNS** – Conserved Non-coding **Sequence**
- **CNGs** – Conserved Non-**Genic** sequence
- **MCS** – Multi-species Conserved **Sequence**
- **SCAMs** – Sequence Conserved Across Multiple species
Finding Constrained Sequences

85% Identical
Species 1 CATGGGCAAATTGGCCCATGGGGCCACCGTA
Species 2 CACGGGCTAATTGGCGccATTGGCTATGGGG-CCCAGCGTA

Neutral Evolution

- No selective pressure/advantage to keep or change the DNA sequence
- Amount of observed variation correlates with:
  - Rate of mutation
  - Length of breeding cycle
  - Amount of time since the last common ancestor
- The neutral rate can vary across the genome
Types of Neutrally Evolving DNA

- **4-Fold Degenerate Sites**
  Third position of codons which can be any base and code for the same amino acid

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

- **Ancestral Repeats**
  Ancient Relics of Transposons Inserted Prior to the Eutherian Radiation

Adapted from Hedges & Kumar, *Science* 297:1283-5
Insights from Human-Rodent Sequence Comparisons

- Sequence Conservation
  - ~40% in Alignments
  - ~5% Under “Selection”
    - ~1.5% Protein Coding
    - ~3.5% Non-Coding

Determining the Fraction of Sequence Under Purifying Selection

Neutral + Functional = Genome-Wide

Genome-Wide – Neutral = Functional

Adapted From Figure 28, Nature 420:553
All 44 ENCODE Regions
29,998,016 Bases

Constrained Sequence

~60%

UTRs

8%

Coding

32%

Other

4.9%

01/06/2006 MSA-Compiled Dataset

Measures of Sequence Conservation

**Binomial-based Method**

*binCons*

Identification and Characterization of Multi-Species Conserved Sequences
Elliott H. Margulies, Mathieu Blanchette, David Haussler, and Eric D. Green

**Genomic Evolutionary Rate Profiling**

*GERP*

Distribution and intensity of constraint in mammalian genomic sequence
Gregory M. Cooper, Eric A. Stone, George Asmussen, NISC Comparative Sequencing Program, Eric D. Green, Serife Batzoglou, and Aren D. Sidow
*Genome Research* (2005) 15:901-913

**PHylogenetic Analysis with Space/Time models**

*phastCons*

Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes
Adam Siepel, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Mikhail Hattori, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W. Hillier, Stephen Richards, George M. Weinstock, Richard K. Wilson, Richard A. Gibbs, W. James Kent, Webb Miller, and David Haussler
*Genome Research* (2005) 15:1034-1050
Constrained Sequences Available from UCSC
The ENCODE Project

- **ENCODE:**
  ENCyclopedia Of DNA Elements

- **Goal:** Compile a *comprehensive encyclopedia* of all functional elements in the human genome

- **Initial pilot project:** 1% of human genome

- Apply multiple approaches to study and analyze that 1% in an international consortium

Which 1% was Selected for Analysis?

- **Manually picked**
  - Prior interest or data
  - 14 regions
  - 500 kb – 1.9 Mb

- **Randomly Selected**
  - Non-coding conservation between Human & Mouse
  - Gene Density
  - Three or four from each strata

Legend:
- Conservation
  - low
  - medium
  - high

- Gene Density
  - low
  - medium
  - high
Integration of ENCODE Data

Gene Annotation

Comparative Sequence Analysis

Promoter identification

DNA-Protein Interactions

RNA Expression

All 44 ENCODE Regions
29,998,016 Bases

Constrained Sequence

~40%

~60%

20%

8%

32%

Other ENCODE Functional Elements

UTRs

Coding

01/06/2006 MSA-Compiled Dataset
Assessing the Overlap between Constrained Sequences and Experimental Annotations

**Overlap between Constrained Sequences and Experimental Annotations**

<table>
<thead>
<tr>
<th>Overlap</th>
<th>Bases</th>
<th>20%</th>
<th>70%</th>
<th>33%</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Regions</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>75% (3 out of 4)</td>
</tr>
</tbody>
</table>

**Overlap between Constrained Sequences and Experimental Annotations**

- **Bases**
- **Regions**
  - CDS
  - 5' UTR
  - 3' UTR
  - T5F
  - TAFs
  - DSB
  - FAIRE-sites
  - Seq-Specific Factors
  - General Factors
  - All Motifs
  - ARs
Why not a Complete Correlation Between Sequence Constraint and Sequence Function?

- Likely not due to false positive experimental annotations
- Did not ascertain all functions at all time-points
- Annotation is larger than the functioning unit
- Fail to detect constraint that is not reflected in the primary sequence
- Reproducible biochemical events with no biological consequence to the organism
- Not constrained throughout all mammals
  
  *Lineage-specific constraint beyond this 5%*

Comparative Genomics can Help Identify Sequences that are Likely Functional