Gene Annotation

Yuki Juan
2003.5.12
Outline

- Genome annotation
  - EST Sequencing
  - Ab initio gene discovery
  - Non-protein coding genes
  - Structural features of genome sequences
Outline

○ Functional annotation and gene family clusters

○ Clustering of genes by sequencing similarity

○ Clusters of orthologous genes

○ Phylogenetic classification of genes

○ Gene ontology
Outline

- Genome annotation
  - EST Sequencing
  - Ab initio gene discovery
  - Non-protein coding genes
  - Structural features of genome sequences
Terms Definition

**ESTs**: Expressed sequence tags

- Single sequencing traces derived from a random set of reverse-transcribed mRNAs
- Short (500-1000 bp) sequence mRNA
- Single read
- Unannotated
- Complementary DNAs (cDNAs)
cDNA library

- poly-dT
- internal sequences
dbEST Database

Expressed Sequence Tags database

EST search method switched from IRX to Entrez. Use search box above instead of old search page.

What is EST?


Other ways to access dbEST

Other ways to access dbEST

How to submit data

How to submit data to dbEST

Information on the current release

Number of ESTs - dbEST summary by organism

Library browser

Relationship Between Gene Structure, cDNA and EST Sequences

![Diagram showing gene structure, complete cDNAs, and ESTs](image)

**Figure 2.15 Relationship between gene structure, cDNA, and EST sequences.** (A) Alternative splicing and the use of alternate 5' start sites or alternate polyA signals results in (B) the generation of multiple cDNA transcripts from individual genes. (C) EST sequences can be derived from either the 5' or 3' end of a cDNA clone and are generally incomplete, resulting in the representation of different exons in different clones from the same gene.
Primary Polymerase II Transcription of a Eukaryotic Gene

- 2~500 introns in a gene; 50~20000 bp in a intron
- < 1000 bp in a exon
Detection of Intervening Sequences by Electron Microscopy

(A) DNA
Duplex DNA
mRNA
Displaced strand of DNA

(B) Intervening sequence (intron)
Displaced strand of DNA
mRNA
Duplex DNA
Transcription and Processing of the beta-globin Gene
Alternative Splicing
Alternative Splicing
Outline

- Genome annotation
  - EST Sequencing
  - Ab initio gene discovery
  - Non-protein coding genes
  - Structural features of genome sequences
Term Definition

- *ab initio gene discovery*
  - can be used to identify protein coding sequences within a whole genome sequence.
Consensus Sequence for the Splicing of mRNA Precursors
Ab Initio Gene Discovery In Higher Eukaryotes

- Difficult

- A new generation of program using HMMs, including
  - Genie
  - Genscan
  - HMMgene
  - FGENES
HMMs

- *Hidden Markov Models*
- Invaluable in application
  - Gene finding
  - Motif identification
  - Prediction of tRNA
Schematic of The Hidden States Included in a HMM

FIGURE B  Schematic of the hidden states included in a HMM. Boxes denote signal sensors for regulatory elements, coding region start sites, intron donor and acceptor sites, and translation stop sites; arrows indicate content sensors for intergenic regions, exons, and introns. Each of these regions emits nucleotides with frequencies characteristic of that region, with these frequencies being obtained by training the HMM on data sets of many known genes.
Gene Finding

- **Genome Browser**: genome.ucsc.edu
FIGURE C  Predicting genes. Three different prediction methods (Ensembl, Fgenesh, and Genscan) were used on a region of chromosome 17 that includes the well-annotated GOSR2 gene. The black images below indicate the location of matching cDNA/EST sequences.
The Lines of Evidence Used

- Identity to a previously annotated reference sequence
- A match to one or more EST sequences from the same organism
- Similarity of the nucleotide or conceptually translated protein sequences to such sequences from other organisms in GenBank or other databases.
- Protein structure prediction that matches a domain in the PFAM database
- Association with predicted promoter sequences, including a TATA box consensus sequence, proximity to clusters of recognized transcription factor binding sites, and often (in mammals) proximity to a CpG island.
Promoter Sites for Transcription

○ Consensus sequence
The Deficiencies of Computational Gene Discovery

- *Imprecise or incomplete characterization of gene structure*

- *Characterization of false positive genes*

- *Failure to identify true genes*
Gene knockout

Modifying a gene so that it is nonfunctional and seeing what effect this has on the phenotype
Outline

- Genome annotation
  - EST Sequencing
  - Ab initio gene discovery
  - Non-protein coding genes
  - Structural features of genome sequences
Non-Protein Coding Genes

- Different from protein coding genes
- Lack of polyadenylation, constraint on sequence divergence is at the level of secondary structure (cloverleaf in tRNAs), and little information on function and distributions.
The Coverleaf Structure of Transfer RNA
Base Pairing of an Anticodon of a tRNA Molecule to an mRNA Codon
tRNAscan-SE 1.21

User Manual for command-line UNIX version of program

If you would like to run tRNAscan-SE locally, you can get the UNIX source code (compressed tar file).

NEW: Analyzing tRNAs in a published genome? See our own tRNAscan-SE analyses of completed genomes in the Genomic tRNA Database

Search Mode: Default  Source: Mixed (general tRNA model)

Format:

- Raw Sequence
- Sequence name (optional): ________ (no spaces)
- Other (FASTA, GenBank, EMBL, GCG, IG)

Paste your query sequence(s) here:

Run tRNAscan-SE
tRNA Content in the Human Genome

<table>
<thead>
<tr>
<th></th>
<th>NUN</th>
<th>NGN</th>
<th>NAN</th>
<th>NCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNU</td>
<td>46</td>
<td>18</td>
<td>44</td>
<td>45</td>
</tr>
<tr>
<td>UNC</td>
<td>54</td>
<td>22</td>
<td>56</td>
<td>55</td>
</tr>
<tr>
<td>UNA</td>
<td>8</td>
<td>15</td>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>UNG</td>
<td>13</td>
<td>6</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>CNU</td>
<td>13</td>
<td>29</td>
<td>41</td>
<td>9</td>
</tr>
<tr>
<td>CNC</td>
<td>19</td>
<td>32</td>
<td>59</td>
<td>12</td>
</tr>
<tr>
<td>CNA</td>
<td>7</td>
<td>28</td>
<td>26</td>
<td>11</td>
</tr>
<tr>
<td>CNG</td>
<td>40</td>
<td>11</td>
<td>74</td>
<td>21</td>
</tr>
<tr>
<td>ANU</td>
<td>36</td>
<td>24</td>
<td>47</td>
<td>15</td>
</tr>
<tr>
<td>ANC</td>
<td>48</td>
<td>36</td>
<td>53</td>
<td>24</td>
</tr>
<tr>
<td>ANA</td>
<td>16</td>
<td>28</td>
<td>43</td>
<td>20</td>
</tr>
<tr>
<td>ANG</td>
<td>100</td>
<td>12</td>
<td>57</td>
<td>22</td>
</tr>
<tr>
<td>GNU</td>
<td>18</td>
<td>26</td>
<td>47</td>
<td>17</td>
</tr>
<tr>
<td>GNC</td>
<td>24</td>
<td>40</td>
<td>53</td>
<td>34</td>
</tr>
<tr>
<td>GNA</td>
<td>11</td>
<td>23</td>
<td>43</td>
<td>25</td>
</tr>
<tr>
<td>GNG</td>
<td>47</td>
<td>11</td>
<td>57</td>
<td>24</td>
</tr>
</tbody>
</table>
Codon Bias

(B)

Number of tRNAs with anticodon

Percentage of codons for an amino acid
### Classes of Noncoding RNA in the Human Genome

<table>
<thead>
<tr>
<th>Class</th>
<th>Function</th>
<th>Number</th>
<th>Localization</th>
</tr>
</thead>
<tbody>
<tr>
<td>tRNA</td>
<td>Protein synthesis</td>
<td>~500</td>
<td>Dispersed large clusters</td>
</tr>
<tr>
<td>rRNA</td>
<td>Protein synthesis</td>
<td>~200 each</td>
<td>Tandem arrays</td>
</tr>
<tr>
<td>U snRNAs</td>
<td>Splicing</td>
<td>&lt;20 each</td>
<td>Dispersed in clusters</td>
</tr>
<tr>
<td>snoRNAs</td>
<td>rRNA modification</td>
<td>~100</td>
<td>Dispersed single copy</td>
</tr>
<tr>
<td>Others</td>
<td>Various</td>
<td>~20 ??</td>
<td>Single copy</td>
</tr>
</tbody>
</table>

*Source: As reported by IHGSC (2001).*
16S RNA
Ribosomes

- Contain four types of ncRNA
  - In the large subunit
    - 28S rRNA
    - 5.8S rRNA
    - 5S rRNA
    - 18S small subunit rRNA
Outline

- **Genome annotation**
  - *EST Sequencing*
  - *Ab initio gene discovery*
  - *Non-protein coding genes*
  - *Structural features of genome sequences*
Structural Features of Genome Sequences

- Repetitive sequences
- GC content
- Simple sequence repeats
- Segmental duplication
- Structure of centromeres and telomeres
Repetitive Elements

- *Transposon-derived repetitive elements*
- *Inactive mRNA-derived copies of cellular genes (pseudogenes)*
- *Simple sequence repeats*
- *Segmental duplications of up to 300 kb*
- *Blocks of noninterspersed repeats*
Database of Genome Sizes

DOGS - Database Of Genome Sizes

- Introduction - Background and source description
- Full database (flat file) (format description)
- Abbreviated database (by name) (by common name) (by size) - A table of most recent genome size estimates for all the organisms present in the database. The references are not shown.
- GenBank and genome sizes (postscript format) - A table of Top Twenty organisms in the latest GenBank release and their genome sizes. A graph showing the growth of GenBank is included.
- Nucleotide data - Molecular weight data of nucleotides and conversions from DNA basepairs to mass of genome
- Download - the files above from our FTP site

Genome size databases and genome research links

- Plant Genome Size Database - DNA C-Value Database at Royal Botanic Gardens, Kew.
- Animal Genome Size Database by T. Ryan Gregory.
- Genome Allomes for sequenced genomes by David Ussery.
- Extensive list of links to genome resources by David Ussery.
- DBA mammalian genome size database by University of Pavia, Italy.
- MAGPIE - Terry Gaasterland's extensive list of genome sequencing projects. Last updated 1998.
- SIMPLE - analysis of long genomic sequences. Relationship between genome size and sequence repetitiveness.

We want YOUR data!

Please send new data, comments and suggestions to Nikolaj Biom, E-mail: nikob@cbs.dtu.dk, or Kristoffer Rapacki, E-mail: rapacki@cbs.dtu.dk.

Acknowledgements

Our thanks are due to Dr Tom Schneider for friendly encouragement and helpful suggestions.
Use of SINES and LINES to Derive Phylogenetic Relationships

- SINES: Short Interspersed Nuclear Elements---
  70~500 bp, 10^6 copies

- LINES: Long Interspersed Nuclear Elements-7000 bp,
  10^5 copies

- Repetitive non-coding sequences that form large fractions of eukaryotic genomes
  Human chromosomal DNA: at least 30%
  Some higher plant: over 50%
Classes of Repetitive Sequences and Their Frequency In Mammalian Genomes
Transposon

Target site

Host DNA

ATGCA

TACGT

Insertion of transposon into host DNA causing a sequence of host DNA at the site of insertion to be duplicated

ATGCA 123456789

TACGT 123456789

123456789

987654321

987654321

Direct repeats of flanking DNA at the target site

Transposon, showing inverted terminal repeats of 9 base pairs, where the numbers 1 to 9 indicate a sequence of base pairs
Distribution of GC Content Along Human Chromosome 1

- Genomes show wide variation in their overall GC content
Simple Sequence Repeats

- **Microsatellites**: simple sequence repeats with a repeat length of up to 13 bases
- **Minisatellites**: longer repeats
- **SSR are important in human genetic studies**
Microsatellite-Associated Human Diseases

**TABLE 2.2 Microsatellite-Associated Human Diseases**

<table>
<thead>
<tr>
<th>Disease</th>
<th>Gene</th>
<th>Type</th>
<th>Normal</th>
<th>Disease</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fragile X syndrome</td>
<td>FMR1</td>
<td>CGG</td>
<td>6–53</td>
<td>60–230</td>
<td>5’ UTR</td>
</tr>
<tr>
<td></td>
<td>FMR2</td>
<td>GCC</td>
<td>6–35</td>
<td>60–200</td>
<td>5’ UTR</td>
</tr>
<tr>
<td>Myotonic dystrophy</td>
<td>DMPK</td>
<td>CTG</td>
<td>5–37</td>
<td>50–’000s</td>
<td>3’ UTR</td>
</tr>
<tr>
<td>Friedreich ataxia</td>
<td>X25</td>
<td>GAA</td>
<td>7–34</td>
<td>34–100</td>
<td>Intron 1</td>
</tr>
<tr>
<td>Kennedy disease</td>
<td>AR</td>
<td>CAG</td>
<td>9–36</td>
<td>38–62</td>
<td>Coding</td>
</tr>
<tr>
<td>Huntington disease</td>
<td>HD</td>
<td>CAG</td>
<td>6–35</td>
<td>36–120</td>
<td>Coding</td>
</tr>
<tr>
<td>Haw River syndrome</td>
<td>DRPLA</td>
<td>CAG</td>
<td>6–35</td>
<td>49–88</td>
<td>Coding</td>
</tr>
<tr>
<td>Spino cerebellar ataxia</td>
<td>SCA1</td>
<td>CAG</td>
<td>6–44</td>
<td>39–82</td>
<td>Coding</td>
</tr>
<tr>
<td></td>
<td>SCA2</td>
<td>CAG</td>
<td>15–31</td>
<td>36–83</td>
<td>Coding</td>
</tr>
<tr>
<td></td>
<td>SCA3</td>
<td>CAG</td>
<td>12–40</td>
<td>55–84</td>
<td>Coding</td>
</tr>
<tr>
<td></td>
<td>SCA6</td>
<td>CAG</td>
<td>4–18</td>
<td>21–33</td>
<td>Coding</td>
</tr>
<tr>
<td></td>
<td>SCA7</td>
<td>CAG</td>
<td>4–35</td>
<td>37–300</td>
<td>Coding</td>
</tr>
<tr>
<td></td>
<td>SCA8</td>
<td>CTG</td>
<td>16–37</td>
<td>110–250</td>
<td>Coding</td>
</tr>
<tr>
<td></td>
<td>SCA12</td>
<td>CAG</td>
<td>7–28</td>
<td>66–78</td>
<td>5’ UTR</td>
</tr>
</tbody>
</table>

*Source: From Cummings and Zoghbi (2000).*
Segmental Duplication

- *Up to 300 kb*
- *Common in vertebrate and plant genomes*
Structure of Centromeres and Telomeres

Two identical chromosomes

Centromere

Microtubules forming part of the spindle

Extending out of centromere are kinetochores

RNA template base pairs with end of telomeric DNA (the template is a short stretch of a longer RNA molecule but only the template part is shown for clarity)

5' RNA template
3' Telomerase

+ Deoxynucleoside triphosphates

Telomerase synthesizes telomeric DNA unit using the enzyme's inbuilt RNA template

5' RNA template
3' Telomerase

New repeat unit

The telomerse now translocates to the right and repeats the process adding unit after unit of telomeric DNA. Hundreds of such units can be added.

5' RNA template
3' (n)

This section is filled in by DNA synthesis
Outline

- Functional annotation and gene family clusters
- Clustering of genes by sequencing similarity
- Clusters of orthologous genes
- Phylogenetic classification of genes
- Gene ontology
The first goal of any genome sequencing project is to classify as many genes as possible into functional families.

Functional Annotation is an essential step toward understanding how genes and gene products interact.

To dissect the true function of genes, functional annotation must be used in conjunction with classical genetics, biochemical and cell biological methods.

First - Pass Classification:

- achieved by protein similarity searches, using software such as BLAST- p to screen for amino acid sequence matches in protein databases that are unlikely to occur at random.

- Between 1/3 and 1/2 of all of the predicted proteins are orphans. (True for multicellular eukaryotic genomes, yeast and prokaryotes).

*Orphan gene: A predicted gene that does not show any similarity to any other gene in the databases, and hence which cannot be assigned to a gene family.
Outline

○ Functional annotation and gene family clusters
  ○ Clustering of genes by sequencing similarity
  ○ Clusters of orthologous genes
  ○ Phylogenetic classification of genes
  ○ Gene ontology
Another Common Result Obtained From BLAST-p

- Each query sequence matches multiple proteins from one or more species
  - This can happen because:
    - One domain in the query is present in a family of proteins.
    - Multiple domains match different proteins.
Protein Databases Assembled by Structural Biologists

- **PFAM**
  - allows researchers to immediately access data containing biochemical properties of a set of predicted proteins.

- **InterPro**
  - goes a step further than PFAM as it classifies individual protein domains.
  - gene annotations now typically link directly to InterPro classifications.
## InterPro Classification

### TABLE 2.3  Number of InterPro Protein Domains in Some Eukaryotic Genomes

<table>
<thead>
<tr>
<th>Domain type</th>
<th>InterPro ID</th>
<th>Number of domains</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Human</td>
</tr>
<tr>
<td>Immunoglobulin</td>
<td>IPR003006</td>
<td>765</td>
</tr>
<tr>
<td>C$_2$H$_2$ zinc finger</td>
<td>IPR000822</td>
<td>706</td>
</tr>
<tr>
<td>Protein kinase</td>
<td>IPR000719</td>
<td>575</td>
</tr>
<tr>
<td>Rhodopsin-like GPCR</td>
<td>IPR000276</td>
<td>569</td>
</tr>
<tr>
<td>P-loop motif</td>
<td>IPR001687</td>
<td>433</td>
</tr>
<tr>
<td>Reverse transcriptase</td>
<td>IPR000477</td>
<td>350</td>
</tr>
<tr>
<td>Rrm domain</td>
<td>IPR000504</td>
<td>300</td>
</tr>
<tr>
<td>G-protein WD-40 repeat</td>
<td>IPR001680</td>
<td>277</td>
</tr>
<tr>
<td>Ankyrin repeat</td>
<td>IPR002110</td>
<td>276</td>
</tr>
<tr>
<td>Homeodomain</td>
<td>IPR001356</td>
<td>267</td>
</tr>
</tbody>
</table>

Source: From IHGSC 2001, Table 25.
The major classes of protein molecular function

- Enzyme
- Signal Transduction (including receptors and kinases)
- Nucleic acid binding (including transcription factors and nucleic acid enzymes)
- Channel (voltage and chemically gated)
- Structural (including cytoskeletal, extracellular matrix, and motor proteins)
Outline

- Functional annotation and gene family clusters
  - Clustering of genes by sequencing similarity
  - Clusters of orthologous genes
- Phylogenetic classification of genes
- Gene ontology
Sequence alignment serves the purpose of identifying which genes to include in a phylogenetic analysis.

Prior to genome sequencing:

- Many genes were cloned by hybridization to cloned genes from another species, or by degenerate PCR.
- Not guaranteed to identify all the potential relatives of a particular gene.

Postgenomic:

- Assume that your gene is the same as the one that shows the closest sequence match in another genome.
- Problem: When the query gene matches multiple members of a family of genes that have discrete functions (the closest match may be misleading in terms of function.)
To Help Reduce The Problem

- Start classification of molecular function with the identification of as large a set of possible family members as possible.

- **PSI - BLAST** (Position-Specific Iterated BLAST): The idea is to align the sequences obtained in an initial protein database search and use this to construct a profile, which is then used to initiate a fresh search. This process is continued until no further matches are identified.
COGs

- Cluster of orthologous genes
COGs

- **Consist of**
  - **orthologs:** A gene in another lineage that is derived from the same ancestral gene that was present prior to the lineage split.
  - **paralogs:** A duplicate copy of a gene that arose subsequent to the split between the two lineages that are being compared.
Orthologs and Paralogs

(A)

MSAAQTDCGPKRV

HuA

MSGAQTNCGPRRV

HuA'

MSGVQTDCAPRKV

MmA

(B)

(C)
Outline

- Functional annotation and gene family clusters
  - Clustering of genes by sequencing similarity
  - Clusters of orthologous genes
  - Phylogenetic classification of genes
- Gene ontology
Phylogenetic methods are considered to provide a much more reliable indicator of functional subclassifications than COGs.
Outline

- **Functional annotation and gene family clusters**
  - Clustering of genes by sequencing similarity
  - Clusters of orthologous genes
  - Phylogenetic classification of genes
  - Gene ontology
It is clear that annotation on the basis of molecular function alone cannot describe or predict biological function.

Knowledge of one species is not directly transferable to another.

The gene ontology (GO) consortium has established a project to annotate cell biological and molecular functions and/or the subcellular localization features of each protein for the complete set of genes identified by genome projects.

GO has already annotated several thousand genes from each of the major model organisms.

The GO site has classified protein function into one of 3 major categories:

- biological processes
- molecular function
- cellular components
Gene Ontology

(A) Biological process
- Behavior
- Cell communication
- Cell growth and maintenance
- Death
- Developmental processes
- Perception of external stimulus
- Physiological processes
- Viral life cycle

(M) Molecular function
- Antitoxin
- Anticoagulant
- Antioxidant
- Apoptosis regulator
- Cell cycle regulator
- Cytoskeletal regulator
- Defense/immunity protein
- Vitamin transporter
and many more

(C) Cellular component
- Cell fraction
- Cell wall
- Extracellular
- Intracellular
- Membrane
- Unlocalized

(Nucleic acid binding)

Cell adhesion (577, 1.9%)
Chaperone (159, 0.5%)
Cytoskeletal structural protein (876, 2.8%)
Extracellular matrix (437, 1.4%)
Immunoglobulin (264, 0.9%)
Ion channel (406, 1.3%)
Motor protein (376, 1.2%)
Structural protein of muscle (296, 1.0%)
Protooncogene (902, 2.9%)
Select calcium binding protein (34, 0.1%)
Intracellular transporter (350, 1.1%)
Transporter (533, 1.7%)

Major GO categories for molecular function

Molecular function unknown
### Gene Ontology

**Evidence codes**

- IMP, inferred from mutant phenotype
- IGI, inferred from genetic interaction
- IPI, inferred from physical interaction
- ISS, inferred from sequence/structural similarity
- IDA, inferred from direct assay
- IEP, inferred from expression pattern
- IEA, inferred from electronic annotation
- TAS, traceable author statement
- NAS, non-traceable author statement
- NR, not recorded
Web resources for various organisms have also been developed.

- Ex. *Flybase*: http://flybase.bio.indiana.edu/