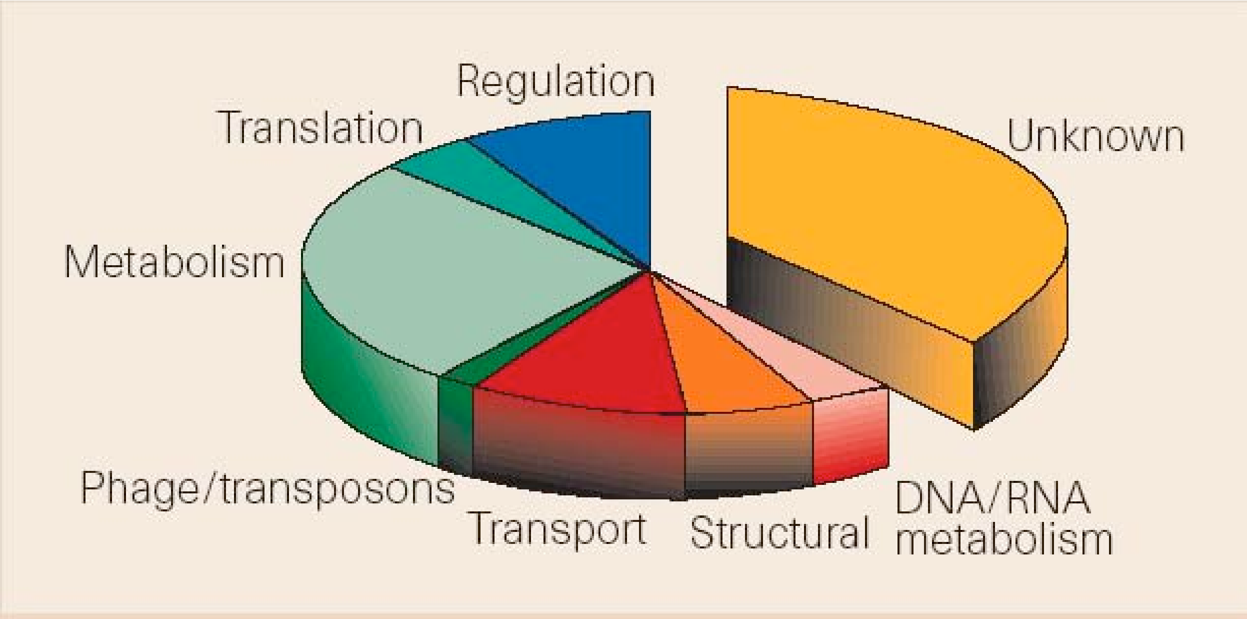
5/16/05

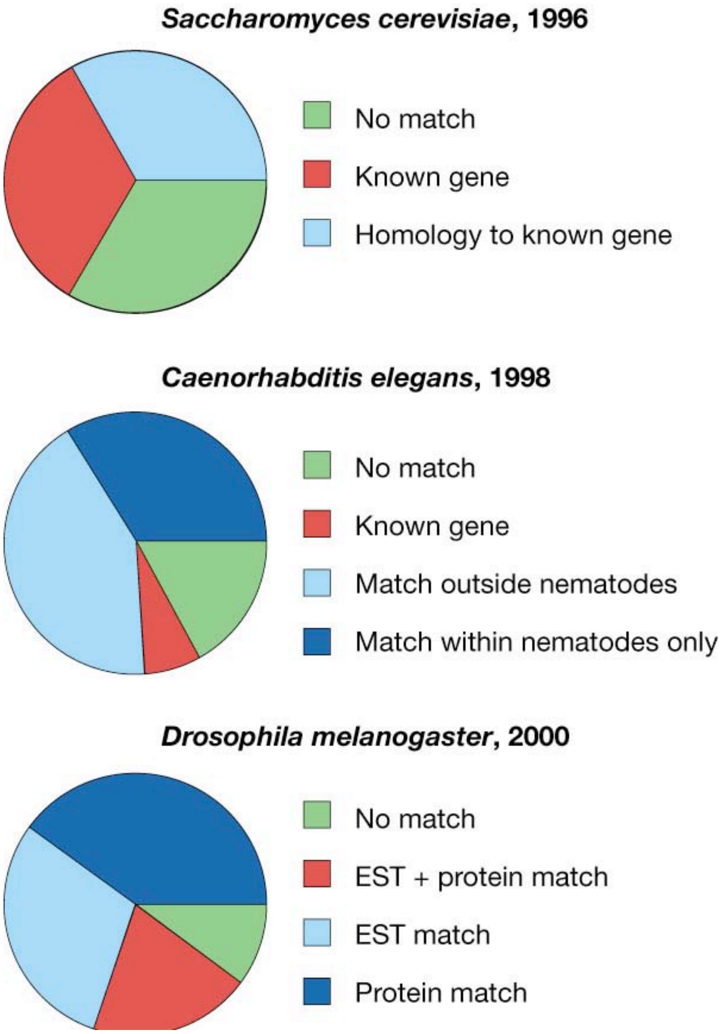
What is so sensational about having a "complete sequence" of a genome instead of only the sequences of "interesting" genes, many of which have been reported long ago by geneticists, biochemists and molecular biologists?

We don't know what we don't know:
until an entire genome is in hand, one can't be certain that we haven't overlooked certain categories of gene and protein functions using the "traditional" approaches

E. coli genome: at time of sequence completion
~30 - 40% of identified protein-coding genes had *no known function* and are *not obviously related to genes of known function*



E. coli genome pie



The distribution of genes in eukaryotic genomes. shown for three organisms are the relative number of genes that were

- previously identified
- had some homology to known genes
- had no match in any sequence database at the time of completion of the genome sequence

Defining Genes in the Genomics Era

molecular definition of a gene: **a complete chromosomal segment responsible for making a functional product**

this definition includes:

- **sequence components required for expression of a gene product (inclusion of both coding and regulatory sequences)**
- **a requirement that the product be functional**

Where to start?

How we glean the existence of a gene from this monotonous one-dimensional array of digital information?

```
TTGCAGATTAGTCCAGGCAGAAACAGTTAGATGTCCCCAGTTAACCTCCTATT
TGACACCACTGATTACCCCATTGATAGTCACACTTTGGGTTGTAAGTGACTTT
TTATTTATTTGTATTTTTGACTGCATTAAGAGGTCTCTAGTTTTTTATCTCTTGT
TTCCCAAAACCTAATAAGTAACTAATGCACAGAGCACATTGATTTGTATTTAT
TCTATTTTTAGACATAATTTATTAGCATGCATGAGCAAATTAAGAAAAACAAC
AACAAATGAATGCATATATATGTATATGTATGTGTGTATATATACACATATAT
ATATATATTTTTTTTCTTTTCTTACCAGAAGGTTTTAATCCAAATAAGGAGAA
GATATGCTTAGAACTGAGGTAGAGTTTTCATCCATTCTGTCCTGTAAGTATTT
TGCATATTCTGGAGACGCAGGAAGAGATCCATCTACATATCCCAAAGCTGAA
TTATGGTAGACAAAGCTCTTCCACTTTTAGTGCATCAATTTCTTATTTGTGTAA
TAAGAAAATTGGGAAAACGATCTTCAATATGCTTACCAAGCTGTGATTCCAA
ATATTACGTAAATACACTTGCAAAGGAGGATGTTTTTAGTAGCAATTTGTACT
GATGGTATGGGGCCAAGAGATATATCTTAGAGGGAGGGCTGAGGGTTTGAAG
TCCAACTCCTAAGCCAGTGCCAGAAGAGCCAAGGACAGGTACGGCTGTCATC
ACTTAGACCTCACCCTGTGGAGCCACACCCTAGGGTTGGCCAATCTACTCCCA
GGAGCAGGGAGGGCAGGAGCCAGGGCTGGGCATAAAAGTCAGGGCAGAGCC
ATCTATTGCTTACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAA
CAGACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCT
GTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTG
GTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTG
GAGACAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATT
GGTCTATTTTCCCACCCTTAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGT
TCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCT
AAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGG
CTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTG
TGACAAGCTGCACGTGGATCCTGAGAACTTCAGGGTGAGTCTATGGACCCT
TGATGTTTTCTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAGGAAGGGG
AGAAGTAACAGGGTACAGTTTAGAATGGGAAACAGACGAATGATTGCATCA
GTGTGGAAGTCTCAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATT
GTTTTCTTTTGTTTAATTCTTGCTTTCTTTTTTTTTCTTCTCCGCAATTTTTACTA
TTATACTTAATGCCTTAACATTGTGTATAACAAAAGGAAATATCTCTGAGATA
CATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATT
TGGAATATATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTATTTTCTT
TTATTTTTAATTGATACATAATCATTATACATATTTATGGGTTAAAGTGTAAT
GTTTTAATATGTGTACACATATTGACCAAATCAGGGTAATTTTGCATTTGTAA
TTTTAAAAAATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAA
```

One of the first steps in annotating a complete genome sequence is to try to figure out where the genes are by ORF analysis

Open Reading Frame (ORF) Analysis:

- genomic DNA is fed into a computer and translated in each of the six possible reading frames
- searches for translational frames beginning with AUG and ending with a stop codon
- ORF's (open reading frames) are identified as long runs of coding triplets without stop codons
- Any ORFs of at least 100 codons are candidates for genes

ORF analysis seems straightforward enough

Why is this approach unsatisfactory for genomes from complex eukaryotes?

ORF analysis does not work well for higher eukaryotic genomes

- Genes in higher eukaryotes may span tens or hundreds of kb with the protein-coding regions accounting for only a few percent of the total sequence
- [overhead of cystic fibrosis gene]

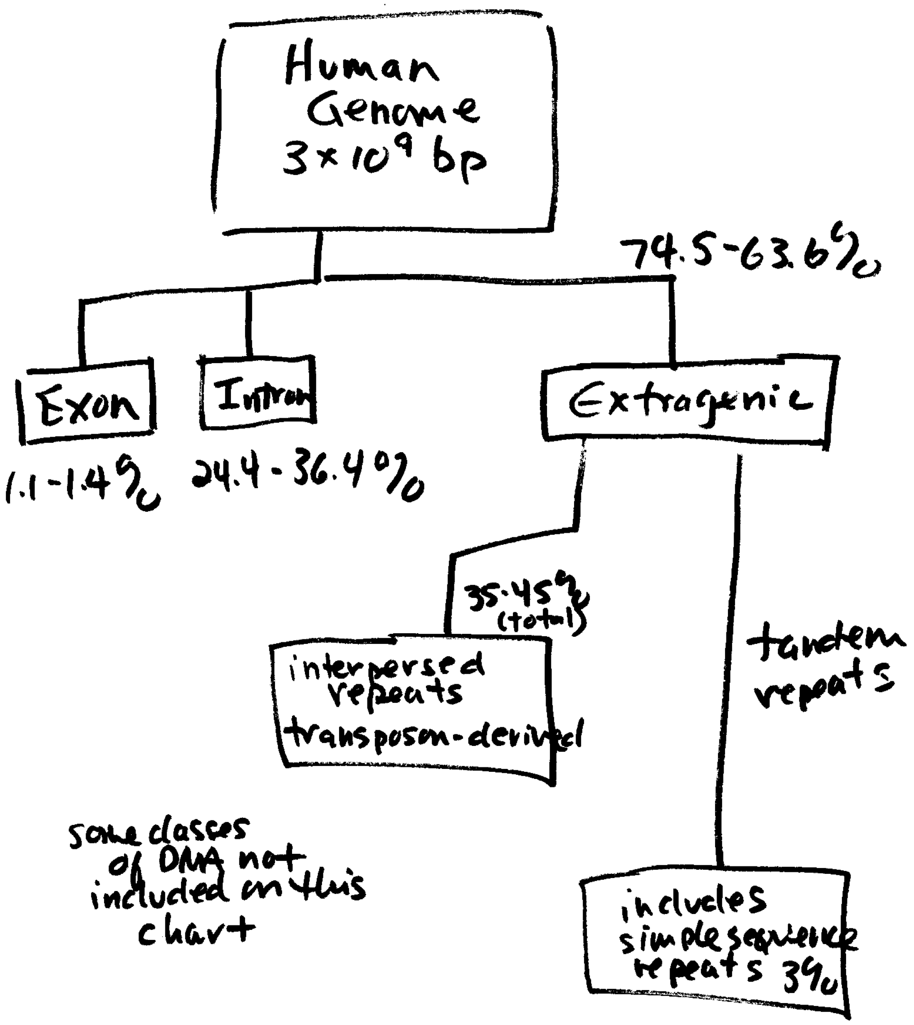
Most metazoan genes contain very short exons (average size is ~140 nucleotides)

In contrast introns are often tens of thousands of nucleotides long

Winner so far:
one intron in the human *neurexin* gene is 480,000 bp

The problem is compounded by the fact that only a

small fraction of genome codes for proteins


Human Genome 3×10^9 bp
74.5-63.6%
Exon
1.1-1.4%
Intron
24.4-36.4%
Extragenic
35.45% (total)
interpersed repeats transposon-derived
tandem repeats
includes simple sequence repeats 3%

SCIENCE 291:1305 2001

some classes of DNA not included on this chart

Another issue (relevant to all genomes - large and small)

Many genes may have ORFs shorter than 100 codons:

Yeast gene finding
# genes with ORF> 100 codons: 6,274

# “gene” with ORF> 15 codons: 100,000

Identifying genes within large regions of uncharacterized DNA is a difficult undertaking and currently the focus of many research efforts

OTHER STRATEGIES?

Other criteria for defining a gene:

Sequence "Features" ***(such as codon bias)***

Sequence Conservation: ***instead of focussing on an individual sequence, identify genes by comparing multiple sequences among organisms***

Evidence for Transcription: ***a non-sequence-based approach for identifying genes is to search for RNA or protein expression -- the hallmark of a gene!***

And a traditional, but not yet anachronistic approach ….

***Gene Inactivation:* acertaining the significance of a DNA sequence by mutating the sequence (random or targeted gene knockout) or inactivating the product of the sequence (RNAi)**

PROBLEMS with this approach??

Sequence Features

Coding vs non-coding sequence features

GENEFINDER/GRAIL/GENIE/GENSCAN

- Systematically use statistical criteria to identify likely genes within a region of genomic sequence
- Candidate genes are evaluated on the basis of “scores” that reflect their

  1. CODING POTENTIAL (coding bias detection)

  2. FUNCTIONAL SITE POTENTIAL

Computational approaches such as GRAIL combine a set of sensor algorithms to localize coding regions

CODON BIAS DETECTION

Defined coding recognition modules take into account seven sensor algorithms each designed to provide an indication of the coding potential of a region of sequence.

One example:

Frame bias matrix: nonrandom frequency with which each of the four bases occupies each of the three positions within codons:

- Due to unequal usage of amino acids
- and to preferred use of codons for particular amino acids (codon bias).
- Look at all reading frames. *If a region codes for protein, then one frame should have a signigicantly better correlation to the bias matrix than the other possible reading frames.*

Problems with codon bias:

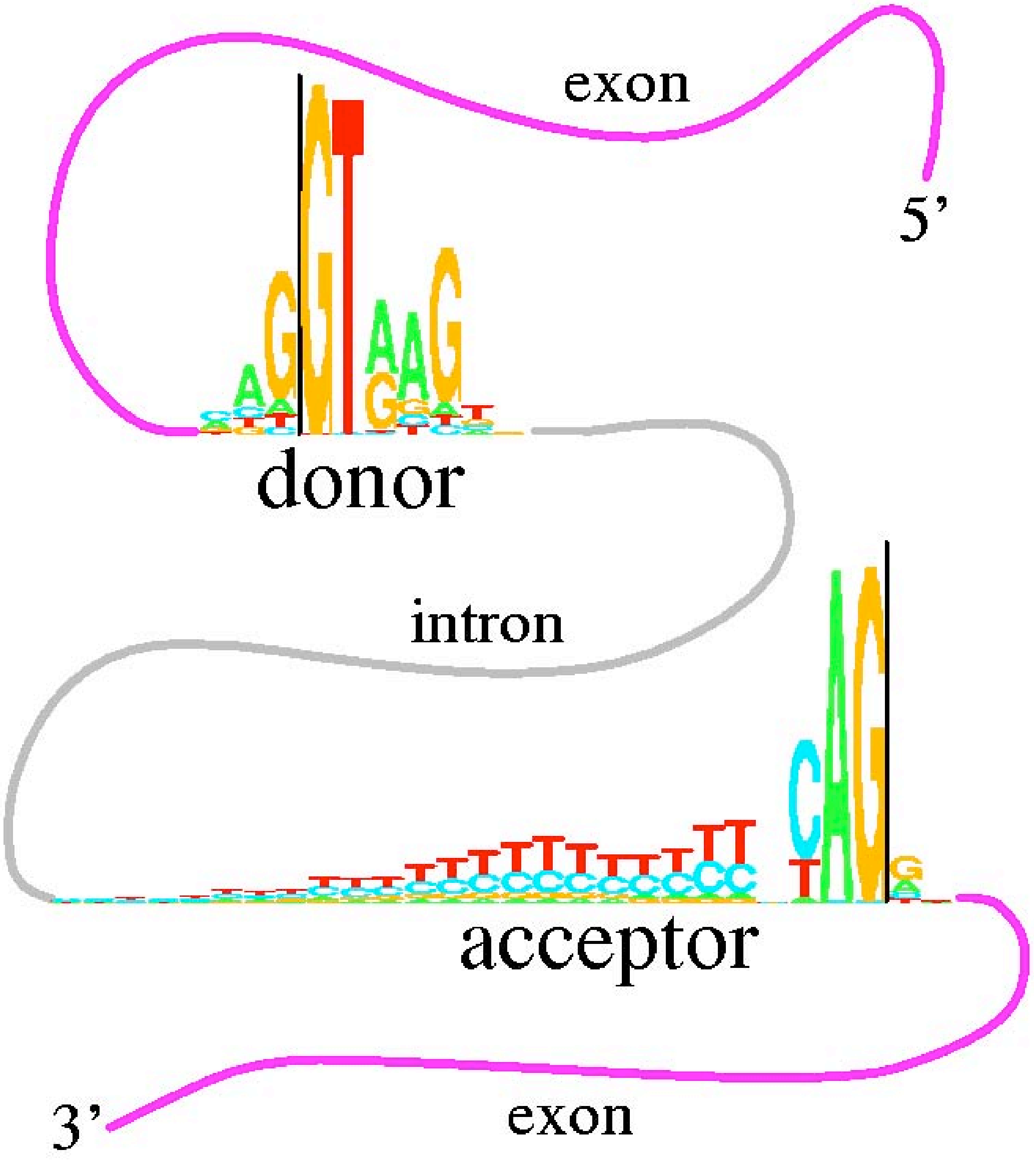
- **for many genes the basis is weak**
- **small ORFs (or exons) contain too few codons to exhibit statistically significant bias**

FUNCTIONAL SITE POTENTIAL

Focusses on recognizing those locations where the gene expression machinery interacts with the nucleic acid

What might the sensors be?

This figure shows two "sequence logos" which represent sequence conservation at the 5'(donor) and 3'(acceptor) ends of human introns. The region between the black vertical bars is removed during mRNA splicing. The logos graphically demonstrate that most of the pattern for locating the intron ends resides on the intron. This allows more codon choices in the protein-coding exons. The logos also show a common pattern "CAG|GT", which suggests that the mechanisms that recognize the two ends of the intron had a common ancestor. See R. M. Stephens and T. D. Schneider, "Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites", J. Mol. Biol., 228, 1124-1136, (1992)



Look for the concensus sequences for:

- promoters (TF binding sites): represents a significant challenge to those who write the programs -- WHY?
- intron splice site: (some absolutely conserved bases -- but over all signal is fairly degenerate)
- polyadenylation and translation termination signals (may be helpful)

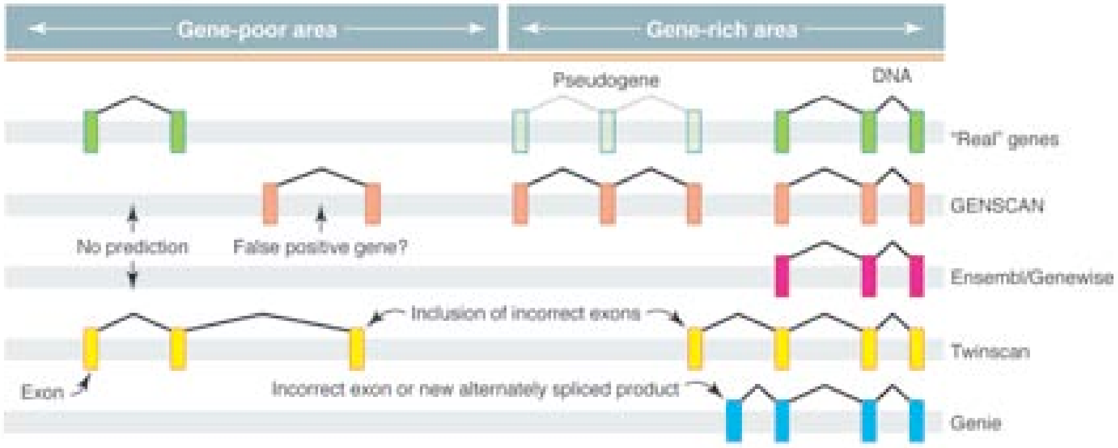
INTEGRATED GENE PARSING:

integrated gene-finding programs that:

- search first for functional sites
- then perform a coding region analysis
- integrates information: if there is a candidate splice site interrupting a coding region, is there noncoding sequence on the other side?

Computer programs that use DNA sequence features (codon bias, functional sites) alone *predict fewe*r than 50% of exons and 20% of complete genes!

Nature 301: 1040 August 22, 2003
Gene counters struggle to get the right answer:



Never perfect. No program calls all genes correctly. Some see genes (shown here as coding regions, or exons, connected by bent lines) where there are none; some miss a gene altogether; and some don't put all the gene's parts in the right places.

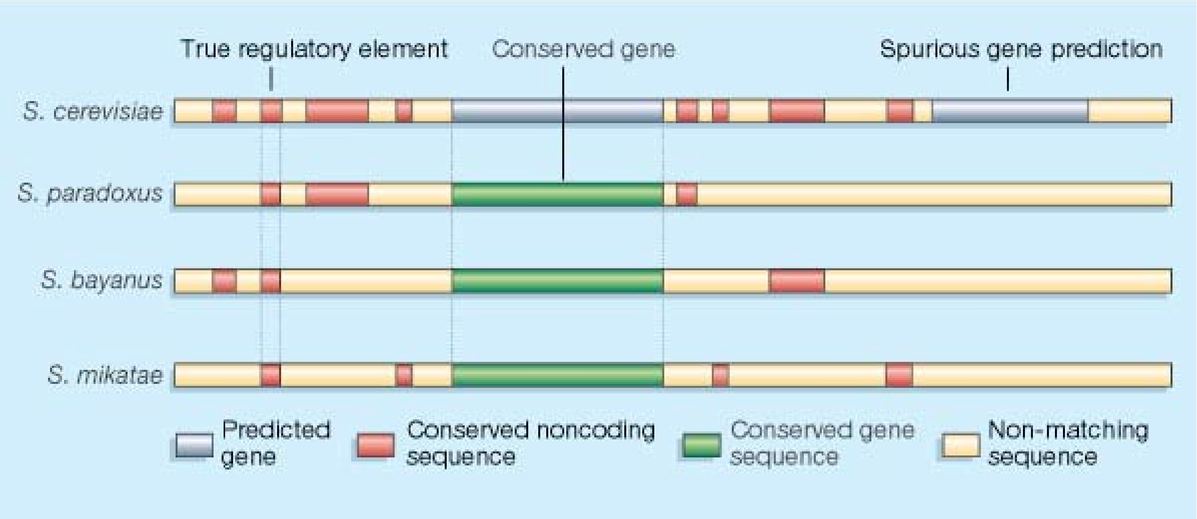
Sequence Conservation:
DNA sequence conservation among species is an effective method for gauging the importance of a specific sequence

Instead of focussing on an individual sequence, identify genes by comparing multiple sequences among organisms

(requires sequences of related organisms separated by the appropriate evolutionary distance)

Sequence Conservation

Comparative Genomics: alignment of genomic sequences from 4 Saccaromyces species



- **Comparative genomics. Comparing the DNA sequences from several species makes it possible to find regulatory regions — short sequences that turn genes on and off — and eliminate spurious gene predictions.**

- **Red boxes highlight areas of sequence similarity between at least two species.**

- **Functional sequences — genes and regulatory elements — tend to be conserved across all species.**

- **The figure shows how one true regulatory element and one correctly identified gene might emerge from a comparison of four yeast species.**

This strategy will identify conserved regulatory regions was well as coding sequences

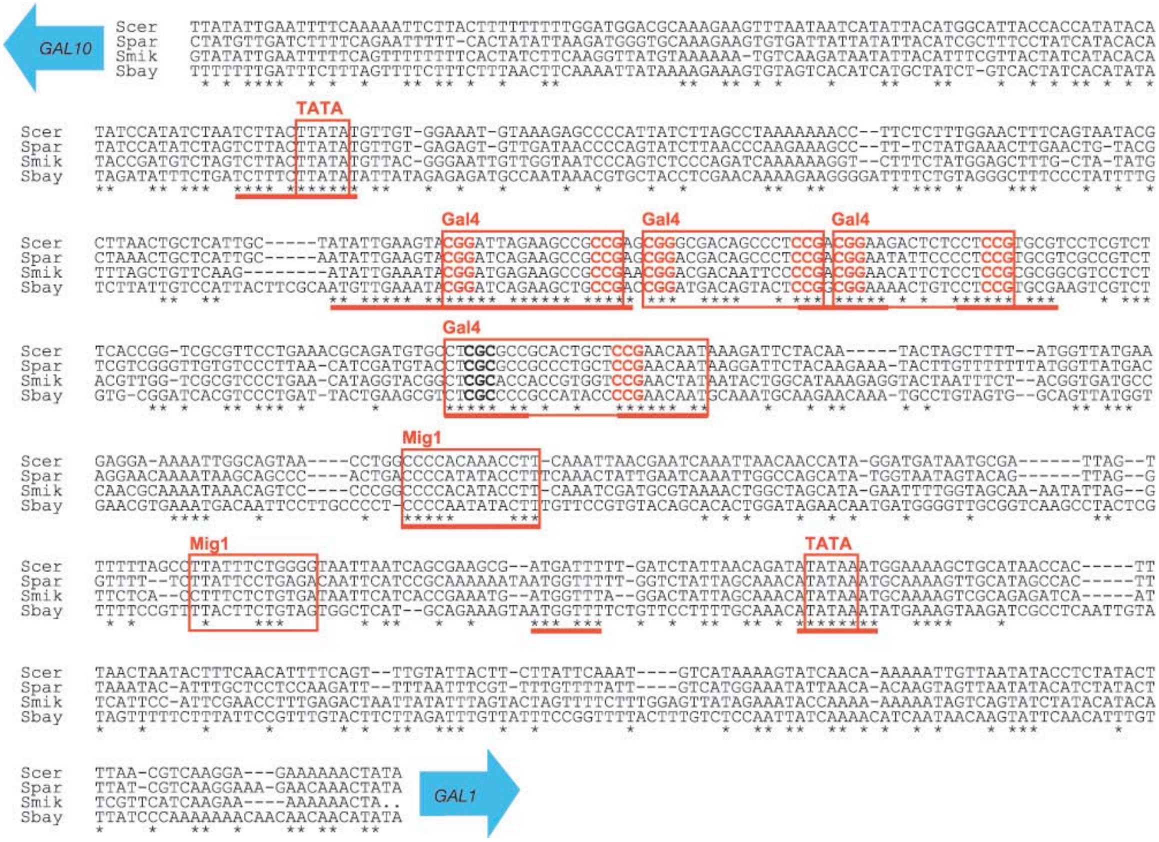
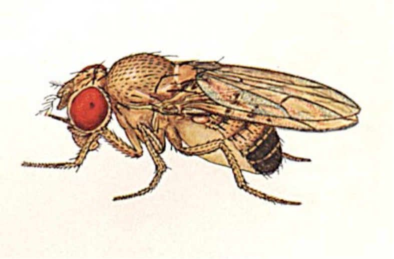


Figure 6 Conservation in the GAL1–GAL10 intergenic region.
Multiple alignment of the four species shows a strong overlap between functional nucleotides and stretches of conservation.
Asterisks denote conserved positions in the multiple alignment.
Blue arrows denote the start and transcriptional orientation of the flanking ORFs. Experimentally validated factor-binding footprints are boxed and labelled according to the bound factor. Stretches of conserved nucleotides are underlined. Nucleotides matching the published Gal4 motif are shown in red. The fourth experimentally validated site differs: it shows a longer footprint and a non-standard consensus motif (bold). This variant motif is also conserved across all four species. Scer, S. cerevisiae; Spar, S. paradoxus; Smik, S. mikatae; Sbay, S. bayanus.

PROBLEMS? with the sequence conservation approach

Organisms with sequenced genomes

Drosophila melanogaster



Caenorhabditis elegans: free-living roundworm

Saccharomyces cerevisiae: yeast

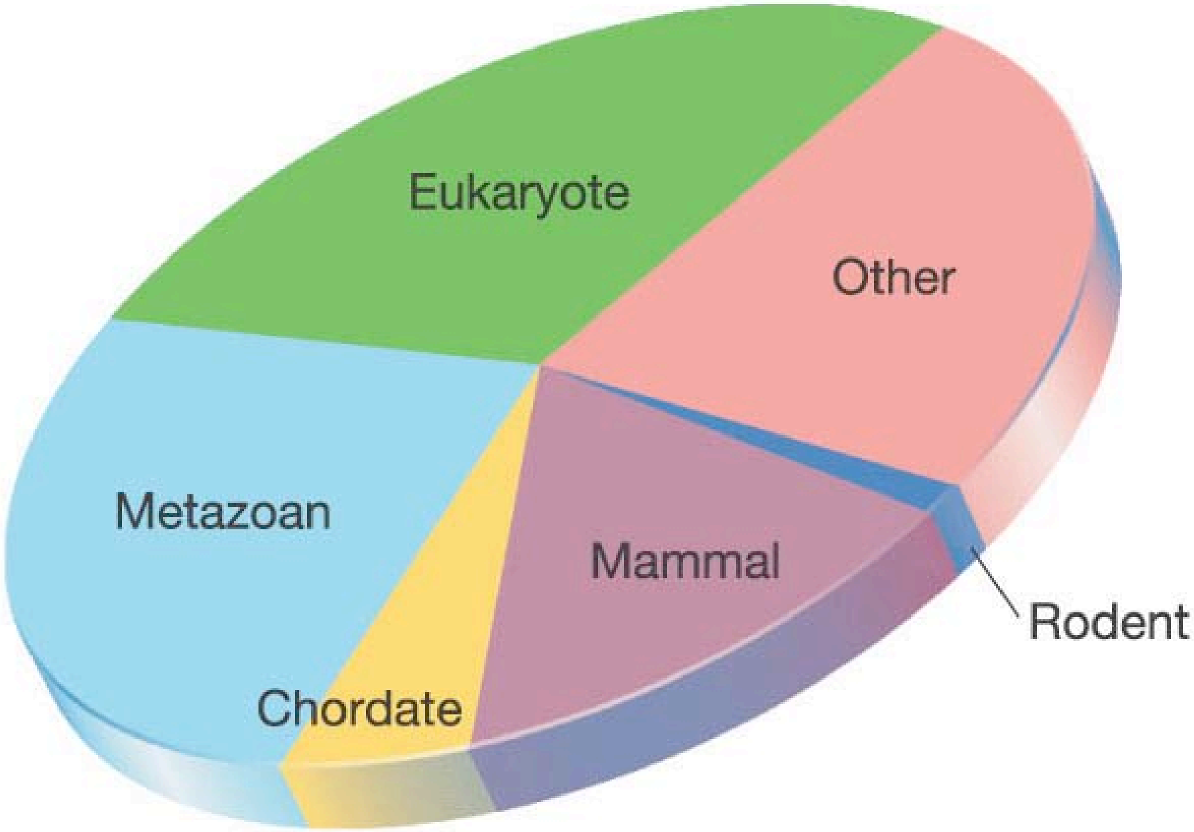
Arabidopsis thaliana: a weed

Mus musculus: a mouse

Escherichia coli

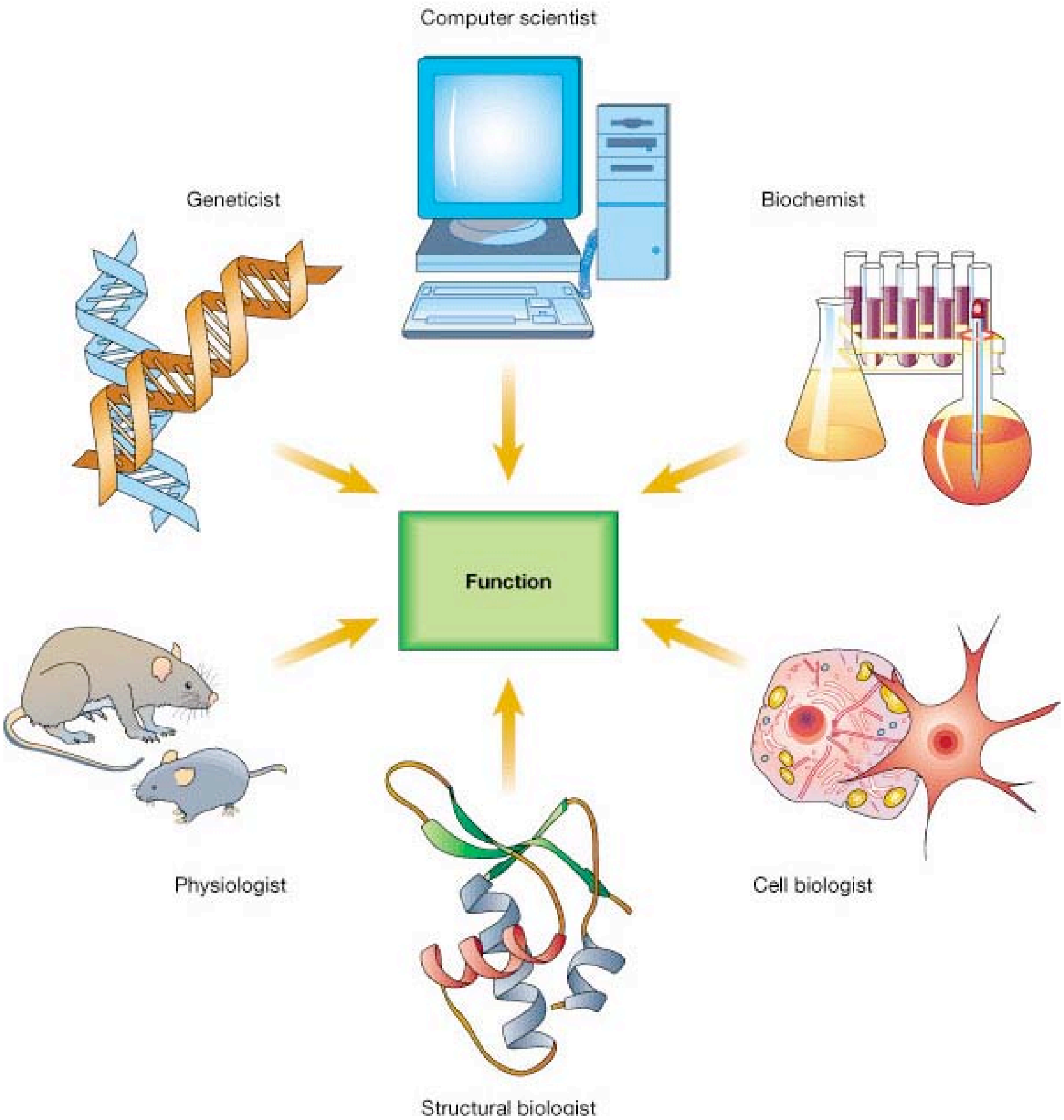
What if species being compared are very closely related?

What if species being compared are only distantly related?



Fraction of genes in the mouse genome that are shared with each taxonomic grouping:
For example, the mammalian wedge indicates fraction of genes shared with other mammals (14%) but not with other chordates
Note large wedge of genes shared by all eukaryotic organisms (29%)
What would other reflect?

WHO to compare with whom?



Understanding Gene Function. The function of a specific gene can be approached from many scientific perspectives with a variety of tools

Criteria for defining a gene:

Open Reading Frame (ORF) Analysis:

Sequence "Features" *(such as codon bias)*

Sequence Conservation: ***instead of focussing on an individual sequence, identify genes by comparing multiple sequences among organisms***

Evidence for Transcription: ***a non-sequence-based approach for identifying genes is to search for RNA or protein expression -- the hallmark of a gene!***

Gene Inactivation: ***acertaining the significance of a DNA sequence by mutating the sequence (random or targeted gene knockout) or inactivating the product of the sequence (RNAi)***