

91學年度上學期「生物資訊學」課程

---

# Genome Informatics

張傳雄

國立陽明大學 遺傳學研究所

11-18-2002





## GOLD™ Genomes OnLine Database



Contact:  
GOLD

Last Update:  
November 15, 2002

Sponsored by  
Integrated Genomics Inc.

Search GOLD: 690 genome projects

112

Published Complete Genomes  
including 2 chromosomes

343

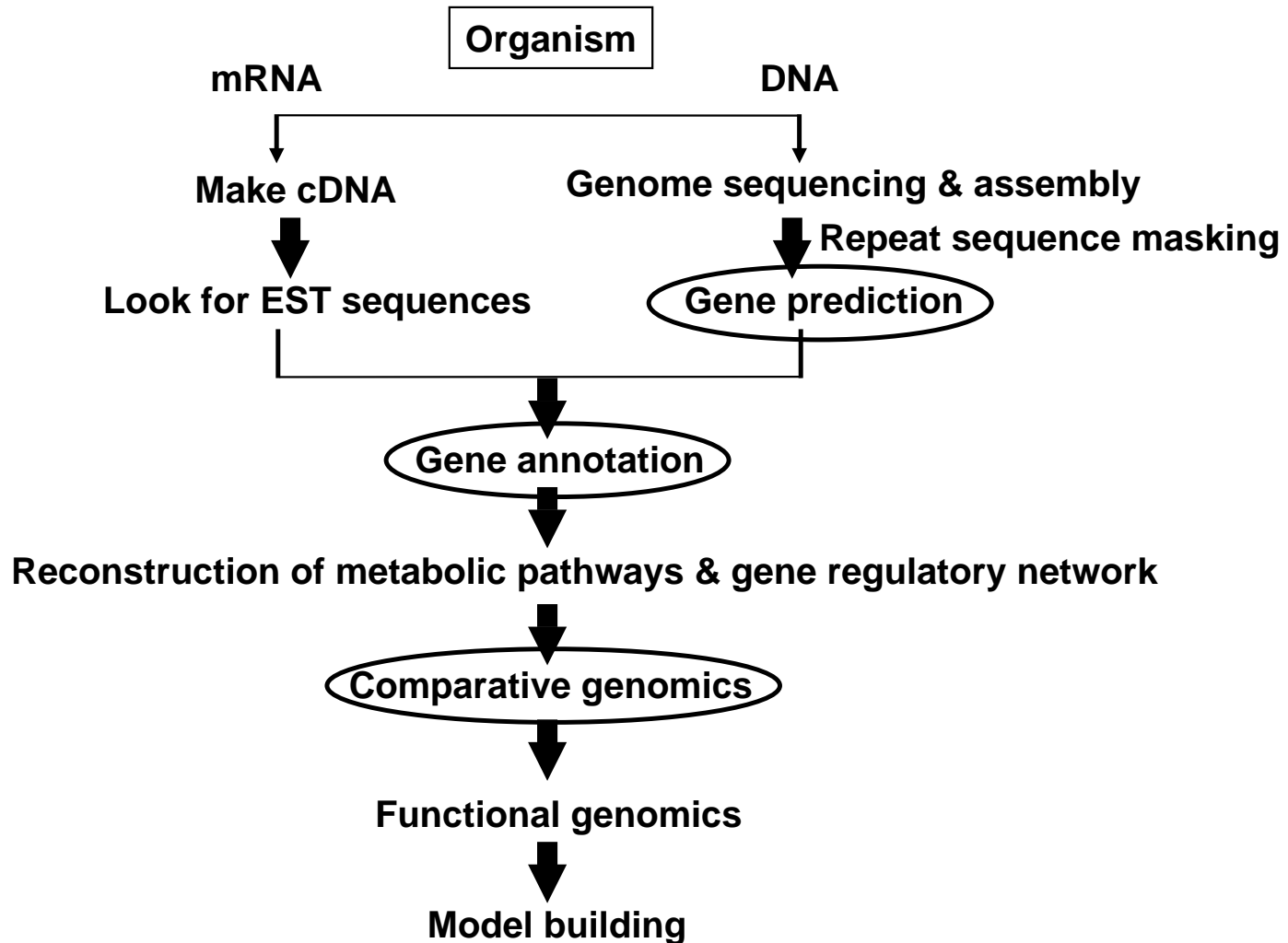
Prokaryotic Ongoing Genomes

235

Eukaryotic Ongoing Genomes  
including 8 chromosomes

<http://wit.integratedgenomics.com/GOLD/>

# Steps of Genome Analysis

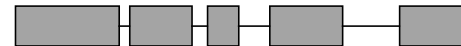


# What is gene prediction?

---

- Detecting meaningful signals in uncharacterised DNA sequences.
- Knowledge of the interesting information in DNA.
- Gene prediction is ‘recognising protein-coding regions in genomic sequence’

GATCGGTCGAGCGTAAGCTAGCTAG  
ATCGATGATCGATCGGCCATATATC  
ACTAGAGCTAGAATCGATAATCGAT  
CGATATAGCTATAGCTATAGCCTAT



# Why is gene prediction important?

---

- Increased volume of genome data generated
- Paradigm shift from gene by gene sequencing (small scale) to large-scale genome sequencing.
- No more one gene at a time. A lot of data.
- Foundation for all further investigation. Knowledge of the protein-coding regions underpins functional genomics.

Note: this presentation is for the prediction of genes that encode protein only;  
Not promoter prediction, sequences regulate activity of protein encoding genes.

# Approaches for gene prediction (Gene finding)

---

- ORF finding (simple but messy)
- *ab initio* prediction
  - Measures of codon bias
  - Simple statistical frequencies
    - Finding promoters and poly(A) sites
      - Translation initiation
      - Coding regions
      - Splice donors and acceptors
- Comparative prediction
  - Using sequence similarity data
  - Using cross-species similarities
    - Using EST data
      - Using protein sequences
      - Using profiles and HMMs

# Gene prediction - *ab initio*

---

- Advantage: can find candidate genes even if they do not resemble known genes or contain known domains
- Most reliable criteria: coding region detection coupled to splice acceptor/donor pairs, with reading frame conservation
- Additional information gained from presence of promoter elements (TATA or CAAT boxes), Kozak sequences, poly(A) signals

# Gene prediction based on similarity

---

- Sequencing projects have already produced a lot of sequence data (cDNA, EST).
- Many newly identified genes have already homologous sequences in databases.
- Use related proteins to derive exon-intron structure of genomic sequences.
- Most straightforward: align EST sequences to genome, while observing splicing rules
- For finding new genes, use BLASTX to search genome piece against protein databases
- Most sophisticated: align protein sequences or motif descriptors (e.g. HMMs) to genome sequence, while including a splicing model

# Gene prediction - complications

---

- There can be more than one gene in a given region
- There can be genes on both strands
- There are often splice variants
  - exon-intron boundaries are not easily resolved. There are conserved GT....AG bases at the ends of introns, but same bases exists throughout genomes and it is difficult to predict which of them form exon-intron junctions.
- There can be overlapping genes
- In eukaryotes most of the genomic DNA is “junk”.  
e.g., human: only 5 % of the whole genome (150 Mpb of 3 000 Mbp) is part of a coding region.

# Gene Finding via Open Reading Frame (ORF) Prediction

---

- 6 possible ORFs
  - frames 1,2,and 3 in 5' to 3' direction
  - frames 1,2, and 3 in 5' to 3' direction of complimentary strand
- Prokaryotes
  - ATG start codon
  - stop codons usually TAA, TAG, TGA
  - lack of introns in coding regions
  - find longest reading frame from ATG to stop codon
    - non-encoding regions will have stop codons, hence short reading frames
    - can genes overlap in different ORFs?? Textbook says 'sometimes'
    - good, but not perfect, prediction of protein-encoding regions
    - usually > 300 amino acids in length

# the leftmost ATG rule?

---

- The simplest way: look for long open reading frames (ORFs).
- ORF = DNA sequence that starts with a start codon and ends with a stop codon and the sequence contains only one stop codon.
- Long ORFs are potential genes:
  - Average distance between stop codons in “random” DNA is  $64/3 \approx 21$
  - Length of average protein is about 300 amino acids.
- The rule of the ‘longest ORF’ was frequently applied to annotate complete microbial genomes with gene start assigned to the 5’-most ATG codon.

# The problems of using the leftmost ATG

---

(Briefings in Bioinfo. 3:181, 2002)

- Leading to proliferation of annotation errors,
- Complicating genomic analyses that depend on intergenic distances,  
(e.g., prediction of the operon structure)
- Making it impossible to predict secreted proteins via analysis of signal peptides,
- Obstructing analysis of translational regulation.
- Dealing with frame-shifts caused by sequencing errors.

# ORF finding in prokaryotes

---

- Simplest method of finding DNA sequences that encode proteins by searching for open reading frames
- An ORF is a DNA sequence that contains a contiguous set of codons that specifies an amino acid
- Six possible reading frames
- Good for prokaryotic system (no/little post translation modification)
- Runs from Met (AUG) on mRNA → stop codon TER (UAA, UAG, UGA)
- <http://www.ncbi.nih.gov/gorf/gorf.html>  
NCBI ORF Finder

# Start and Stop codons

---

## Start codons

- Prokaryotes - 90% of the time AUG is the initiation codon, but sometimes GUG or UUG is initiation codon
- Eukaryotes - AUG is almost always the initiation codon

## Stop codons

- UAA, UAG, UGA always signal the end of translation

# ORF Finder (Open Reading Frame Finder)

[PubMed](#)
[Entrez](#)
[BLAST](#)
[OMIM](#)
[Taxonomy](#)
[Structure](#)

NCBI

Tools  
for data mining

GenBank  
sequence  
submission support  
and software

FTP site  
download data and  
software

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database.

This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

Enter GI or ACCESSION

or sequence in FASTA format

FROM:  TO:

Genetic codes

Standard

- [The Standard Code](#)
- [The Vertebrate Mitochondrial Code](#)
- [The Yeast Mitochondrial Code](#)
- [The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code](#)
- [The Invertebrate Mitochondrial Code](#)
- [The Ciliate, Dasycladacean and Hexamita Nuclear Code](#)
- [The Echinoderm Mitochondrial Code](#)
- [The Euplotid Nuclear Code](#)
- [The Bacterial and Plant Plastid Code](#)
- [The Alternative Yeast Nuclear Code](#)
- [The Ascidian Mitochondrial Code](#)
- [The Flatworm Mitochondrial Code](#)
- [Blepharisma Nuclear Code](#)
- [Chlorophycean Mitochondrial Code](#)
- [Trematode Mitochondrial Code](#)
- [Scenedesmus Obliquus Mitochondrial Code](#)
- [Thraustochytrium Mitochondrial Code](#)

# Gene finding: Prokaryotes vs. Eukaryotes

---

- Prokaryotes
  - Contiguous open reading frames (ORF)
  - Short intergenic sequences
  - Good method: detecting large ORFs
  - Complications:
    - Partial sequences
    - Sequencing errors
    - Start codon prediction
    - Overlapping genes on both strands

# Bacterial promoter

---

-35

T<sub>82</sub>T<sub>84</sub>G<sub>78</sub>A<sub>65</sub>C<sub>54</sub>A<sub>45</sub>...  
(16-18 bp)...

T<sub>80</sub>A<sub>95</sub>T<sub>45</sub>A<sub>60</sub>A<sub>50</sub>T<sub>96</sub>...(A,G)  
-10 +1

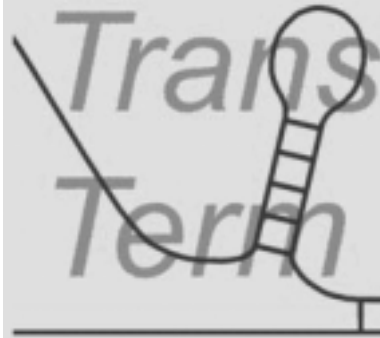
# Terminators

---

- Stem/loop
    - structural only
  - 3'-U tail
  - Rho-independent
- C-rich
  - G-poor
  - “loose” consensus
- Rho-dependent

# TransTerm from TIGR

---



TransTerm is a program that finds rho-independent transcription terminators in bacterial genomes. Each terminator found by the program is assigned a confidence value that provides an estimate of its probability of being a true terminator.

For a description on how the confidence is calculated see our paper:

Maria D. Ermolaeva, Hanif G. Khalak, Owen White, Hamilton O. Smith and Steven L. Salzberg. Prediction of Transcription Terminators in Bacterial Genomes. *J Mol Biol* 301, (1), 27-33 (2000)

The TransTerm runs under Unix and requires about ten minutes per megabase of input sequence on 550 MHz Intel processor. The computational time requirement scales linearly with genome size. The input for the TransTerm is genome sequence and genes coordinates.

You can select a Genome to View high-confidence TransTerm results:

## Obtaining TransTerm:

TransTerm is available free of charge to researchers using it for non-commercial purposes. We ask only that you fill out and return our [license agreement](#). Once you've downloaded and printed the agreement, either you or an authorized representative of your institution should sign it and mail it to the address listed on the agreement. We much prefer email! Just email the agreement from the account where you'd like us send the system, which must be a nonprofit organization's email address. (No dot.com addresses for free licenses.) Send email to [mariae@tigr.org](mailto:mariae@tigr.org), or fax to 301-838-0206, attention Maria Ermolaeva. If you fax it, please send an email message containing the email address to which we should send the system; sometimes fax copies are not clear. After you (e)mail or fax in your license agreement, we will email you instructions on how to download a Unix tar file containing the complete system, including source code.

If you represent a for-profit organization, please contact us by email at [license@tigr.org](mailto:license@tigr.org) for details on how to obtain a commercial license.

*Last modified on: July 12, 2002*

<http://www.tigr.org/software/transterm.html>

# Translation

---

Ribosome Binding Site (RBS),  
Shine-Dalgarno Site (SD sequence)

**nnGGAGGnnnnnATG...**

typical *E. coli*

**nnaaAGGnnnnnATG**

# Software Tools from TIGR

---



A system for finding genes in microbial DNA, especially the genomes of bacteria and archaea. **Glimmer** (Gene Locator and Interpolated Markov Modeler) uses interpolated Markov models (IMMs) to identify the coding regions and distinguish them from noncoding DNA.



A fast, flexible system for detecting splice sites in the

genomic DNA of various eukaryotes. The system has been trained and tested successfully on *Plasmodium falciparum* (malaria), *Arabidopsis thaliana* and human genomes. Training data sets for Human and *Arabidopsis thaliana* are included. It is fully described in Pertea M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* 2001 Mar 1;29(5):1185-90

## RepeatFinder

**RepeatFinder** is a computational system for analysis of repetitive structure of genomic sequences. The method uses suffix trees for efficient computation of exact repeats and organizes those repeats into classes. The method can be applied to individual genome sequences or sets of sequences. The output is multi-fasta file of found repeat sequences that can be used as the target of searches.



TransTerm is a program that finds rho-independent transcription terminators in bacterial genomes. Each terminator found by the program is assigned a

confidence value that provides an estimate of its probability of being a true terminator. TransTerm has been published: Prediction of Transcription Terminators in Bacterial Genomes Ermolaeva, M.D., Khalak, H.G., White, O., Smith, H.O., Salzberg, S.L. *Journal of Molecular Biology* 301, 27-33 (2000)



**RBSfinder** is a Perl script that implements an algorithm to find ribosome binding sites for genes in bacterial and archaeal genomes. It is normally run as a post-processor to the Glimmer gene finder or to other prokaryotic gene finders.



## A probabilistic method for identifying start codons in bacterial genomes

Baris E. Suzek<sup>1</sup>, Maria D. Ermolaeva<sup>2</sup>, Mark Schreiber<sup>3</sup> and Steven L. Salzberg<sup>1, 2,\*</sup>

<sup>1</sup>Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA <sup>2</sup>The Institute for Genomic Research, 9712 Medical Center Dr, Rockville, MD 20850, USA and <sup>3</sup>Department of Biochemistry, University of Otago, PO Box 56, Dunedin, New Zealand

Received on December 18, 2000; revised on April 12, 2001 and July 4, 2001; accepted on July 9, 2001

### ABSTRACT

As the pace of genome sequencing has accelerated, the need for highly accurate gene prediction systems has grown. Computational systems for identifying genes in prokaryotic genomes have sensitivities of 98–99% or higher (Delcher *et al.*, *Nucleic Acids Res.*, **27**, 4636–4641, 1999). These accuracy figures are calculated by comparing the locations of verified stop codons to the predictions. Determining the accuracy of start codon prediction is more problematic, however, due to the relatively small number of start sites that have been confirmed by independent, non-computational methods. Nonetheless, the accuracy of gene finders at predicting the exact gene boundaries at both the 5' and 3' ends of genes is of critical importance for microbial genome annotation, especially in light of the important signaling information that is sometimes found on the 5' end of a protein coding region. In this paper we propose a probabilistic method to improve the accuracy of gene identification systems at finding precise translation start sites. The new system, RBSfinder, is tested on a validated set of genes from *Escherichia coli*, for which it improves the accuracy of start site locations predicted by computational gene finding systems from the range 67–77% to 90% correct.

- RBS-Finder can be a post-processing tool for GLIMMER or GeneMark.

# Summary: Bacterial genome annotation

---

- Find ORFs (e.g., run GeneMark/Glimmer)
- Find promoter & regulatory regions
- Document operons
- Document orthologs in other species, determine function
- Verify termini by multiple alignment
- Document metabolic pathways, ensure all enzymes are present

# Gene finding: Prokaryotes vs. Eukaryotes

---

- Eukaryotes
  - Complex gene structures (exon/introns)
    - *D. melanogaster* has an average of 4 introns/gene
    - Very long genes (*D. melanogaster* X gene 160 kb)
    - Very long introns
    - Many introns
    - “Nested”, overlapping, and alternatively spliced genes
    - 5' UTRs with non-coding exons
    - Long 3' UTRs
    - Complex transcription machinery
  - ORF-finding alone is not adequate

# Approaches to finding genes

---

- First find/mask repeats and other low complexity regions
- Search by signal - find genes by identifying the sequence signals involved in gene expression
- Search by content - find genes by statistical properties that distinguish protein-coding DNA from non-coding DNA
- Combined - newest systems for gene finding combine these two strategies
- To make use of the best computational techniques, it is necessary to submit one's sequence to the analysis of several different software packages

# Gene finding: genome

---

- Genome composition
  - Long ORFs tend to be coding
  - Presence of more putative ORFs in GC rich genomes (Stop codons = UAA, UAG & UGA)
- Genome complexity
  - Simple repetitive sequences (e.g. dinucleotide) & dispersed repeats tend to be anti-coding
  - May need to mask sequence prior to gene prediction

# Complex genome DNA

---

- ~10% highly repetitive (300 Mbp)
  - NOT genes
- ~25% moderate repetitive (750 Mbp)
  - **Some** genes
- ~25% exons and introns (800 Mbp)
- 40%=?
  - Regulatory regions
  - Intergenic regions

# Repeat sequences

---

- The Human Genome contains a high proportion of repetitive DNA
- Can be grouped into tandemly repeated DNA and interspersed repetitive DNA
- An estimated 1/3 of human genome consists of interspersed repetitive DNA sequences which are primarily degenerate copies of transposable elements

# Repeat sequence types

---

- Short Interspersed Nuclear Elements (SINEs)
  - Alu repeats, GC rich, length: ~ 280 base pairs, located in untranslated intronic regions
- Tandemly repeated DNA
  - repeats often associated with disease syndromes, telomeres contain long arrays of TTAGGG repeats
- Long Interspersed Nuclear Elements (LINEs)
  - AT rich regions, length: 6-8 kb, LINEs contain internal promoters for RNA polymerase III

# Masking repetitive DNA

---

- First step is to locate and remove interspersed and simple repeats from eukaryotic sequences
- Such repeats may overlap regions transcribed by RNA polymerase, but they rarely overlap promoters or coding regions of exons
- Locations can often provide important negative information on the location of gene features
- Repeats can often confuse other analysis, especially database searches
- For local installation of programs, it may be useful to add cloning vector sequences to sequence collections that should be masked
- Repeat finding programs – CENSOR, RepeatMasker, XBLAST

---

## REPEAT/VECTOR MASKING SOFTWARE

Site Name	Description of Site
<a href="#">CENSOR</a>	The CENSOR web server allows users to have query sequences aligned against a reference collection of human, rodent, or plant repeats.
<a href="#">DNA Repeats</a>	This is a collection of DNA repeat-finding sites.
<a href="#">RepeatMasker</a>	RepeatMasker screens DNA sequences in fasta format against a library of repetitive elements and returns a masked query sequence ready for database searches as well as a table annotating the masked regions.
<a href="#">RepMask</a>	repeat masking using RepeatMasker, Xblast , XNU and SEG at TigemNet
<a href="#">VecScreen at NCBI</a>	VecScreen is a system for quickly identifying segments of a nucleic acid sequence that may be of vector origin.
<a href="#">XBLAST</a>	Reads BLAST output and masks query.

<http://home.san.rr.com/dna/darryl/nucleotideAnalysis.html>

# Knowing what to look for...

---

- Transcribed region
  - mRNA, tRNA, snoRNA, snRNA, rRNA
- Structural region
  - Exon, intron, 5' UTR, 3' UTR, ORF, cleavage product
  - Mutations: insertion, deletion, substitution, inversion, translocation
  - Functional or signal region
  - Promoter, enhancer, DNA/RNA binding site, splice site signal, poly-adenylation signal
  - Protein processing: glycosylation, methylation, phosphorylation site
- Similarity
  - Homolog, paralog, genomic overlap (syntenic region)
- Other feature types
  - Transposable element, repetitive element
  - Pseudogene
  - STS, insertion site

# DNA transcription unit features

---

- Promoter elements
  - Core promoter elements
    - TATA box
    - Initiator (Inr)
    - Downstream promoter element (DPE)
  - Transcription factor (“TF”) binding sites
    - CAAT boxes
    - GC boxes
    - SP-1 sites
    - GAGA boxes
  - Enhancer site(s)

---

## REGULATORY SITE FINDING SOFTWARE

Site Name	Description of Site
<a href="#">Eukaryotic Promoter Database</a>	The Eukaryotic Promoter Database is an annotated non-redundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally.
<a href="#">Web Promoter Scan Service</a>	Predicts Promoter regions based on scoring homologies with putative eukaryotic Pol II promoter sequences.
<a href="#">PROMOTER SCAN II</a>	PROMOTER SCAN II is a program developed to recognize and predict pol II promoters in genomic DNA sequences. Presently it is limited to mammalian promoter sequences.
<a href="#">Regulatory Sequence Analysis Tools</a>	This site provides a series of modular computer programs specifically designed for the detection of regulatory signals in non-coding sequences
<a href="#">TRANSFAC</a>	The Transcription Factor Database to search for binding sites and more
<a href="#">Welcome to TESS</a>	Transcription Element Search System

<http://home.san.rr.com/dna/darryl/nucleotideAnalysis.html>

# mRNA features

---

- Exon
  - Initial, internal, terminal
    - Codon usage, preference \* Control elements (e.g. splice enhancers)
- Intron
  - 5' splice site (“GT”), branchpoint (lariat), 3' splice site (“AG”)
  - Repeat elements
- Start codon (translation start site)
  - “Kozak” rule
- UTR (untranslated regions)
  - 5' UTR
    - Translation regulatory elements \* RNA binding sites
  - Initial, internal, terminal
    - Control elements (e.g. splice enhancers)
  - 3' UTR
    - RNA binding sites (*cis*-acting elements)
- Stop codon
- Poly-adenylation signal and site
- RNA destabilization signal

# Finding tRNA genes

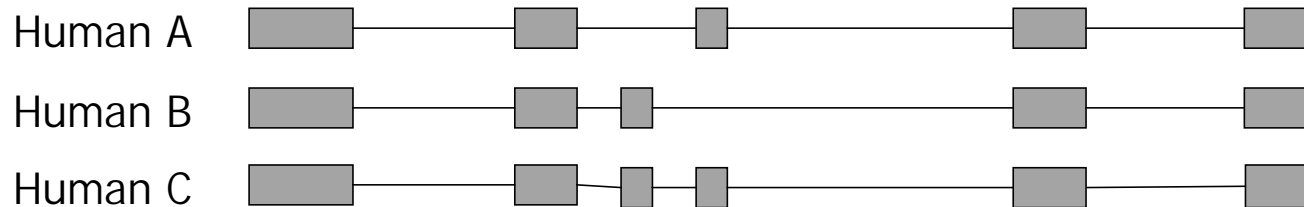
---

- Recognition of tRNA genes is easier than recognition of protein coding genes.
- Due to simpler structure of polIII promoters and the conserved secondary structure of tRNAs.
- The tRNA gene recognition problem has apparently been solved in tRNAscan-SE program, which combines the elements of several earlier programs.
- Result is a method that reportedly identifies over 99% of true tRNA genes with less than one false positive expected per genome.
- <http://www.genetics.wustl.edu/eddy/tRNAscan-SE/>

# Alternate splicing

---

- Alternate splicing (isoforms) are very difficult to predict.



# Alternative splicing – How common?

---

- Preliminary estimates – 35% of human genes display alternative splicing at 5' end

(Miranov, Genome Research 1999)

- Human genome draft – about 60% of genes display alternative splicing

(International Human Genome Sequencing Consortium, Nature 2001)

# Splice junctions

---

- Splice junctions - sites where introns are spliced out of RNA transcripts
- The splice sites on either end of the intron have different properties
- The 5' splice site is called the donor site
- The 3' splice site is called the acceptor site
- Virtually all spliceosomal introns begin with GT and end with AG – this nearly invariant rule is used by the majority of gene-finding programs to narrow the search space of exon and intron boundaries

# Useful databases for identifying alternatively spliced RNAs & regulatory elements of alternative splicing data

(*Genome Biology* 2002, 3(11):reviews0008.1–0008.16)

Name	Description	URL
Alternative Splicing Database	Alternative exon database compiled from the literature	<a href="http://csgisigna.csh.org/new_alk_exon_db/">http://csgisigna.csh.org/new_alk_exon_db/</a>
Alternative splicing databases	Several alternative splicing databases, including a database of splice variants of disease genes, a complete splice-site database, and a database of alternative splice forms for seven organisms	<a href="http://www.bioinf.mdc-berlin.de/splice/">http://www.bioinf.mdc-berlin.de/splice/</a>
AltExtron	Transcript-confirmed human introns and exons; includes alternatively spliced data subsets	<a href="http://www.bit.uq.edu.au/altExtron/">http://www.bit.uq.edu.au/altExtron/</a> or <a href="http://www.ebi.ac.uk/~thanasra/altExtron/">http://www.ebi.ac.uk/~thanasra/altExtron/</a>
ASAP	Human alternative splicing database; part of the Alternative Splicing Annotation Project	<a href="http://www.bioinformatics.ada.edu/HASDB">http://www.bioinformatics.ada.edu/HASDB</a>
ASDB	Alternatively spliced gene database; includes protein and nucleotide sequences for human, mouse, rat, <i>Drosophila</i> , <i>Caenorhabditis elegans</i> , chicken, cow and rabbit and viruses	<a href="http://cbg.ersc.gov/asdb">http://cbg.ersc.gov/asdb</a>
AsMamDB	Alternatively spliced mammalian genes; includes human, mouse, and rat	<a href="http://166.111.30.65/ASPMAMDB.html">http://166.111.30.65/ASPMAMDB.html</a>
Gene Resource Locator	Gene map database; includes information on alternatively spliced transcripts	<a href="http://grl.gk.u-tokyo.ac.jp">http://grl.gk.u-tokyo.ac.jp</a>
Intronator	Introns in <i>C. elegans</i> ; includes a catalog of alternatively spliced transcripts	<a href="http://www.cse.ucsc.edu/~keno/intronator/">http://www.cse.ucsc.edu/~keno/intronator/</a>
ISIS	All introns identified in GenBank	<a href="http://isis.bit.uq.edu.au/">http://isis.bit.uq.edu.au/</a>
PALS db	Putative alternative splicing predicted by EST alignments for mouse and human	<a href="http://palsdb.ym.edu.tw/">http://palsdb.ym.edu.tw/</a>
SELEX-DB	In vitro selected oligomers; includes SELEX sequences for splicing factors and is supplemented by SYSTEM (experimental design) and CROSS_TEST (cross-validation test) databases	<a href="http://www.mgs.bionet.nrc.ca/mgs/systems/selex/">http://www.mgs.bionet.nrc.ca/mgs/systems/selex/</a>
SpliceDB	Mammalian splice sites	<a href="http://genomic.sanger.ac.uk/spldb/SpliceDB.html">http://genomic.sanger.ac.uk/spldb/SpliceDB.html</a> or <a href="http://www.softberry.com/spldb/SpliceDB.html">http://www.softberry.com/spldb/SpliceDB.html</a>
SpliceNest	A database that maps GeneNest onto human genomic sequence and is integrated with GeneNest (EST clusters for human, mouse, zebrafish and <i>Anolis</i> ) and SYSTEMS (protein sequence clusters) databases	<a href="http://splicenest.molgen.mpg.de/">http://splicenest.molgen.mpg.de/</a>
STACK	Putative human transcripts reconstructed from ESTs; includes the context of different tissues or pathological states	<a href="http://www.scrib.ac.za/Dbases.html">http://www.scrib.ac.za/Dbases.html</a>
Yeast Intron Database	Introns in <i>Saccharomyces cerevisiae</i>	<a href="http://www.cse.ucsc.edu/research/compbio/yeast_introns.html">http://www.cse.ucsc.edu/research/compbio/yeast_introns.html</a>

# Splice site prediction programs

---

Program	Organism	Method
GeneSplicer (152)	<i>Arabidopsis</i> , human	HMM + MDD
NETPLANTGENE (42) ( <a href="http://www.cbs.dtu.dk/services/NetPGene/">http://www.cbs.dtu.dk/services/NetPGene/</a> )	<i>Arabidopsis</i>	NN
NETGENE2 (43) ( <a href="http://www.cbs.dtu.dk/services/NetGene2/">http://www.cbs.dtu.dk/services/NetGene2/</a> )	Human, <i>C.elegans</i> , <i>Arabidopsis</i>	NN + HMM
SPLICEVIEW (39) ( <a href="http://l25.itba.mi.cnr.it/~webgene/wwwspliceview.html">http://l25.itba.mi.cnr.it/~webgene/wwwspliceview.html</a> )	Eukaryotes	Score with consensus
NNSPLICE0.9 (44) ( <a href="http://www.fruitfly.org/seq_tools/splice.html">http://www.fruitfly.org/seq_tools/splice.html</a> )	<i>Drosophila</i> , human or other	NN
SPLICEPREDICTOR (40,153) ( <a href="http://bioinformatics.iastate.edu/cgi-bin/sp.cgi">http://bioinformatics.iastate.edu/cgi-bin/sp.cgi</a> )	<i>Arabidopsis</i> , maize	Logitlinear models: (i) score with consensus; (ii) local composition
BCM-SPL ( <a href="http://www.softberry.com/berry.phtml">http://www.softberry.com/berry.phtml</a> ; <a href="http://genomic.sanger.ac.uk/gf/gf.html">http://genomic.sanger.ac.uk/gf/gf.html</a> )	Human, <i>Drosophila</i> , <i>C.elegans</i> , yeast, plant	Linear discriminant analysis

HMM, hidden MM; MDD, maximal dependence decomposition; NN, neural networks.

*Nucleic Acids Research*, 2002, 30(19): 4103-4117

# Homology-based gene prediction programs

Program	Organism	Databank or required input	Alignment	Gene reconstruction
AAT (66) ( <a href="http://genome.cs.mtu.edu/aat.html">http://genome.cs.mtu.edu/aat.html</a> )	Primates, rodents, other	cDNA, protein	DDS (improved BLASTX), DPS (improved BLASTN)	NAP, GAP2
ALN (62)		Protein	Tron code, PAM 250	
CEM (81)		Two genomic sequence	BLASTX output, WMM for sites	DP
EbEST (73) ( <a href="http://ares.ifrc.mcu.edu/EBEST/ebest.html">http://ares.ifrc.mcu.edu/EBEST/ebest.html</a> )	Human, other	dbEST	BLASTN, EST clustering, Smith-Waterman-based gapped alignment	3'-UTR detection, assembly of EST-tagged exons
Est2genome (74)		EST or cDNA, preferably BLASTN output	Modified Smith-Waterman Needleman-Wunsch algorithm	No
GeneSeqer (67,68) ( <a href="http://bioinformatics.iastate.edu/cgi-bin/gseq.cgi">http://bioinformatics.iastate.edu/cgi-bin/gseq.cgi</a> )	<i>Arabidopsis</i> , maize, generic plant	dbEST or EST database or proteins	Spliced alignment, splice recognition with SplicePredictor if missing EST match	Yes
GeneWise (60) ( <a href="http://www.sanger.ac.uk/Software/Wise2/genewiseform.shtml">http://www.sanger.ac.uk/Software/Wise2/genewiseform.shtml</a> )	Human	One protein or a HMM profile	Global alignment translated ORF/protein	DP (dynamic)
GENQUEST ( <a href="http://compbio.cornell.gov/Graal-bin/EmptyGenquestForm">http://compbio.cornell.gov/Graal-bin/EmptyGenquestForm</a> )		dbEST, SwissProt, Prosite, BLOCKS, GSDb	Smith-Waterman, Blast, Fasta	
ICE (64) ( <a href="http://theory.lcs.mit.edu/ice">http://theory.lcs.mit.edu/ice</a> )		dbEST, OWL	Look-up	DP
INFO (63)		Ne	25mer look-up table, protein/protein alignments scored with PAM 40, PAM 120, PAM 250, BLO62	No
ORFgene2 (61) ( <a href="http://l25.iba.mt.cnr.it/~webgene/wwworfgene2.html">http://l25.iba.mt.cnr.it/~webgene/wwworfgene2.html</a> )	Human, mouse, <i>Drosophila</i> , <i>Aspergillus</i> , <i>Arabidopsis</i> , <i>Caenorhabditis</i>	SwissProt	BlastP, WAM for splice sites, identity score on frequencies of dipeptides	Compatibility graph, DP
PredictGenes ( <a href="http://cbg.inf.ethz.ch/Server/subsection3_1_8.html">http://cbg.inf.ethz.ch/Server/subsection3_1_8.html</a> )	Invertebrates, vertebrates, prokaryotes, plants	SwissProt	PAM 250	DP
PROCRUSTES (59) ( <a href="http://www.bio.usc.edu/software/procrustes/wwwserv.html">http://www.bio.usc.edu/software/procrustes/wwwserv.html</a> )	Vertebrates	One homologous protein	Protein/protein alignments scored with PAM 120	DP
Pro-Gen (83) ( <a href="http://www.anchorgen.com/pro_gen/pro_gen.html">http://www.anchorgen.com/pro_gen/pro_gen.html</a> )		Two genomic sequences	Alignment of translated sequences scored with PAM 120	DP
ROSETTA (80) ( <a href="http://crossspecies.lcs.mit.edu/">http://crossspecies.lcs.mit.edu/</a> )	Human, mouse	Two genomic sequences	GLASS (global alignment system), PAM 20, GenScan method for splice sites	DP
SGP-1 (82) ( <a href="http://soft.ice.mpg.de/sgp-1">http://soft.ice.mpg.de/sgp-1</a> )	Vertebrates, angiosperm	Two genomic sequences or a pairwise local alignment output	Local alignment	DP
SIM4 (69)	All eukaryotes	cDNA/genomic	HSP from Blast	No
SLAM (85) ( <a href="http://huboon.math.berkeley.edu/~synteric/slam.html">http://huboon.math.berkeley.edu/~synteric/slam.html</a> )	Human, mouse	Two genomic sequences	Generalized pair HMM	DP
Spidey (70) ( <a href="http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/index.html">http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/index.html</a> )	Vertebrates, <i>Drosophila</i> , <i>C.elegans</i> , plant	One genomic sequence/set of mRNAs	Two Blasts: high stringency and low stringency	
SYNCOD (72) ( <a href="http://l25.iba.mt.cnr.it/~webgene/wwwsyncod.html">http://l25.iba.mt.cnr.it/~webgene/wwwsyncod.html</a> )	Human, mouse, <i>Drosophila</i> , <i>Arabidopsis</i> , <i>Aspergillus</i> , <i>Caenorhabditis</i>	BLASTN output	Silent/replacement ratio, Monte Carlo simulations	No
TAP (75) ( <a href="http://sapiens.wustl.edu/~zkan/TAP/">http://sapiens.wustl.edu/~zkan/TAP/</a> )	Human, mouse, <i>Drosophila</i>	dbEST	WU-BLASTN, SIM4	Yes
Utopia (84)	All eukaryotes	Two genomic sequences	Local alignment	Yes

DP, dynamic programming; WAM, weight array matrix; WMM, weight matrix method.

Program	Organism	Gene elements	Gene model	Homology
DAGGER (91) EuGene (31) ( <a href="http://www.inra.fr/bio/T/EuGene">http://www.inra.fr/bio/T/EuGene</a> )	<i>Arabidopsis</i>	Site scores Three-periodic IMM for exons, one IMM for introns, one for intergenic regions, one for UTR. NetGene2/SplicePredictor for splice sites	Directed acyclic graphs DP	EST/cDNA, protein
GeneId3 (89) ( <a href="http://www1.imim.es/geneid.html">http://www1.imim.es/geneid.html</a> ) GENEFINDER (28): FGENE, FEX ... ( <a href="http://genomic.sanger.ac.uk/gf/gf.html">http://genomic.sanger.ac.uk/gf/gf.html</a> ; <a href="http://www.softberry.com/berry.phtml">http://www.softberry.com/berry.phtml</a> )	Vertebrates, plants Human, mouse, <i>Drosophila</i> , <i>Caenorhabditis elegans</i> , yeast, dicots, monocots, <i>Schizosaccharomyces pombe</i> , <i>Neurospora crassa</i>	Rule-based method; WAM, discriminant analysis. Linear discriminant analysis	DP DP	EST Protein
GENEFINDER (Green) GeneGenerator (92)	Maize	Log likelihood ratio score matrix on MM Logitlinear models for splice sites, start; 3rd to 5th order MM for exons and introns	DP DP	
GeneMark (29) ( <a href="http://opal.biology.gatech.edu/GeneMark/genemark24.cgi">http://opal.biology.gatech.edu/GeneMark/genemark24.cgi</a> ) GeneMark.hmm (35) ( <a href="http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi">http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi</a> )	Prokaryotes, eukaryotes Human, mouse, <i>Drosophila</i> , <i>Gallus gallus</i> , <i>Arabidopsis</i> , rice, maize, <i>Chironomus tentativus</i> , <i>C.elegans</i> , <i>Bordetia pertussis</i> , <i>Triticum aestivum</i> Eukaryotes	5th order MM (homogeneous for introns, three-periodic for exons) 5th order MM (homogeneous for introns, three-periodic for exons)	No GHMM, DP	Under development
GeneModeler (57) ( <a href="http://tp.tigr.org/pub/software/gm/">http://tp.tigr.org/pub/software/gm/</a> ) GeneParser (27) ( <a href="http://beagle.colorado.edu/~eesnyder/GeneParser.html">http://beagle.colorado.edu/~eesnyder/GeneParser.html</a> ) Genie (44,96) ( <a href="http://www.fruitfly.org/seq_tools/genie.html">http://www.fruitfly.org/seq_tools/genie.html</a> ) GenLang (154) ( <a href="http://www.chil.upenn.edu/genlang/genlang_home.html">http://www.chil.upenn.edu/genlang/genlang_home.html</a> ) GenomeScan GENSCAN (30) ( <a href="http://genes.mit.edu/GENSCAN.html">http://genes.mit.edu/GENSCAN.html</a> )	Vertebrates <i>Drosophila</i> , human, other Vertebrates, <i>Drosophila</i> , dicots Vertebrates Vertebrates, <i>Arabidopsis</i> , maize	Nucleotide and dinucleotide composition, consensus for splice sites NN NN Grammar rules, WAM, hexuple frequencies ... GenScan method, BLASTP or BLASTX WAM for acceptor; MDD for donor; 5th order MM (homogeneous for introns, three-periodic for exons) Linear combination, diodon statistic	Rule-based method DP GHMM, DP Chart parsing, DP GHMM, DP GHMM, DP	EST Protein Protein Protein, GenomeScan (99)
GENVIEW2 (25) ( <a href="http://25.itha.mi.cnr.it/~webgene/wwwgene.html">http://25.itha.mi.cnr.it/~webgene/wwwgene.html</a> ) GlimmerM (33,34) ( <a href="mailto:salzberg@cs.jhu.edu">salzberg@cs.jhu.edu</a> )	Human, mouse, diptera Small eukaryotes, <i>Arabidopsis</i> , rice	Three-periodic IMM for exons (order 0-8), IMM for introns, 2nd order MM for splice sites NN	DP DP	EST, cDNA
GRAIL/GAP3 (90,155) ( <a href="http://compbio.ornl.gov/Grail-bin/EmptyGrailForm">http://compbio.ornl.gov/Grail-bin/EmptyGrailForm</a> ) GRPL (97)	Human, mouse, <i>Arabidopsis</i> , <i>Drosophila</i> Human, <i>Drosophila</i> , <i>Arabidopsis</i>	Reference point logistic for splice sites, 5th order MM (homogeneous for introns, three-periodic for exons) Three-periodic 4th order MM for exons, 3rd order MM for introns	GHMM, DP GHMM	Protein
HMMgene (98) ( <a href="http://www.cbs.dtu.dk/services/HMMgene/">http://www.cbs.dtu.dk/services/HMMgene/</a> ) MORGAN (48) ( <a href="http://www.cs.jhu.edu/labs/compbio/morgan.html">http://www.cs.jhu.edu/labs/compbio/morgan.html</a> ) MZEF (26) ( <a href="http://argon.cshl.org/genefinder/">http://argon.cshl.org/genefinder/</a> )	Vertebrates, <i>C.elegans</i> Vertebrates Human, mouse, <i>Arabidopsis</i> , fusion yeast	Decision tree system Quadratic discriminant analysis	DP No	
SORFIND (24) Twinscan (100)	Mouse, human	Matrix method for start and splice sites, hexamer usage (Fourier measure) GenScan method; 5th order MM for UTR and intergenic, WAM for acceptor sites	No GHMM	Genomic sequence
VELL (47) ( <a href="http://www.cs.jhu.edu/labs/compbio/vell.html">http://www.cs.jhu.edu/labs/compbio/vell.html</a> ) Xpound (156) ( <a href="http://bioweb.pasteur.fr/seqanal/interfaces/xpound-simple.html">http://bioweb.pasteur.fr/seqanal/interfaces/xpound-simple.html</a> )	Vertebrates Human	HMM Three-periodic 1st order MM for exons, 1st order MM for introns and intergenic	DP HMM	

CHMM, class HMM; GHMM, generalized HMM; IMM, interpolated MM; MM, Markov model.

# Gene finding: *ab initio*

---

- What features of a ORF can we use?
  - Size - large open reading frames
  - DNA composition - codon usage / 3rd position codon bias
  - Other features:
    - Kozak sequence CCGCCAUGG
    - Ribosome binding sites
    - Termination signal (stops)
    - Splice junction boundaries

# Gene finding: comparative

---

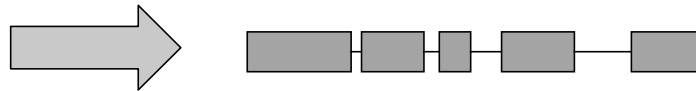
- Use knowledge of known coding sequences to identify region of genomic DNA by similarity
  - transcribed DNA sequence
  - peptide sequence
  - related genomic sequence

# *ab initio* prediction

---

- What is *ab initio* gene prediction?
  - Prediction from first principles using the raw DNA sequence only.

GATCGGTCGAGCGTAAGCTAGCTAG  
ATCGATGATCGATCGGCCATATATC  
ACTAGAGCTAGAATCGATAATCGAT  
CGATATAGCTATAGCTATAGCCTAT



Requires ‘training sets’ of known gene structures to generate statistical tests for the likelihood of a prediction being real.

# Gene finding software

---

- Signal recognition
  - Promoter prediction
  - Splice site prediction
  - Start codon prediction
  - Poly-adenylation site prediction
- Coding potential
- Coding exons
- Gene structure prediction
  - Spliced alignment
  - Neural networks
  - HMMs

## GENE-FINDING SOFTWARE

Site Name	Description of Site
<a href="#"><u>DNA Sequence Translation</u></a>	This page provides an interface to a program for identification of open reading frames in DNA sequences.
<a href="#"><u>GeneMark</u></a>	The GeneMark program relies upon an Inhomogeneous Markov Model approach combined with training datasets to predict genes.
<a href="#"><u>GenLang</u></a>	GenLang is a syntactic pattern recognition system, which uses the tools and techniques of computational linguistics to find genes and other higher-order features in biological sequence data.
<a href="#"><u>GeneParser</u></a>	GeneParser is a program for the identification of protein coding regions in genomic DNA sequences.
<a href="#"><u>GENSCAN</u></a>	This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.
<a href="#"><u>ORNL_Grail (v1.3)</u></a>	Gene prediction software
<a href="#"><u>HMMgene server</u></a>	Prediction of vertebrate and C. elegans genes. HMMgene is a program for prediction of genes in anonymous DNA.
<a href="#"><u>ORF Finder</u></a>	The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database.
<a href="#"><u>PROCRUSTES</u></a>	Similarity-Based Gene Recognition via Spliced Alignment. PROCRUSTES finds the chain of exons with the best fit to the target proteins.
<a href="#"><u>Wise2</u></a>	Wise2 can compare a single protein or a profile HMM to a genomic DNA sequence, and predict a gene structure.
<a href="#"><u>SplicePredictor</u></a>	A method to identify potential splice sites in (plant) pre-mRNA by sequence inspection
<a href="#"><u>WebGene Home Page</u></a>	Tools for prediction and analysis of protein-coding gene structure

<http://home.san.rr.com/dna/darryl/nucleotideAnalysis.html>

# A list of gene finding programs

Name	Methods	Organism	Access
ER	Discriminant Analysis	Human, Arabidopsis	<a href="#">WWW</a>
GENSCAN (seems the most accurate)	Semi Markov Model	vertebrate, caenorhabditis, arabidopsis, maize	<a href="#">WWW (Stanford)</a> , <a href="#">WWW (MIT)</a> , <a href="#">Email</a>
GRAIL	Neural Network	human, mouse, arabidopsis, drosophila, E.coli	<a href="#">WWW (ORNL or JAPAN)</a> , <a href="#">Email</a> , X-client
GenLang	Definite Clause Grammer	Vertebrate, Drosophila, Dicot	<a href="#">WWW</a> , <a href="#">Email</a>
GenView	Linear combination	Human, Mouse, Diptera	<a href="#">WWW</a>
GeneFinder(FGENEH,etc.)	LDA	Human, E.coli, Drosophila, Plant, Nematode, Yeast	<a href="#">WWW</a> , <a href="#">Email</a>
GeneID	perceptron,rules	Vertebrate	<a href="#">WWW</a> , <a href="#">Email</a>
GeneMark	5th-Markov	Almost all model organism	<a href="#">WWW</a> , <a href="#">Email</a>
GeneParser	neural networks	Human	<a href="#">ftp from the (WWW page)</a>
Genie	GHMM	Human (vertebrate)	<a href="#">WWW</a>
Glimmer	interpolated Markov models (IMMs)	microbial	<a href="#">WWW</a>
MORGAN	Decision Tree	vertebrate	<a href="#">WWW</a>
MZEF	Quadratic Discriminant Analysis	Human, mouse, Arabidopsis, Pombe	<a href="#">WWW</a> , binary
NetPlantGene	Combined Neural Networks	A. thaliana	<a href="#">WWW</a>
OCI	decision tree	Human	<a href="#">WWW</a>
PROCRUSTES	spliced alignment	vertebrate	<a href="#">WWW</a>
Soffind	rule base	Human	<a href="#">ftp</a>
VEIL	HMM	vertebrate	<a href="#">WWW</a>
Hogehoge	Wonderful method	extraterrestrial	not yet (maybe going to be announced at April 1)

**Vladimir Makarov**

is a staff scientist at the bioinformatics company BionomiX. Prior to that Dr Makarov developed bioinformatics knowledge management systems and algorithms at Paracel, a business of Celera Genomics, and at Ceres, Inc, as well as conducted research in structural and computational biology at the University of Houston and Baylor College of Medicine.

**Keywords:** gene structure, prediction, coding regions, gene finding, computational biology

# Computer programs for eukaryotic gene prediction

## Abstract

Seven popular programs for gene prediction in eukaryotic organisms are described and evaluated on the basis of availability for in-house and on-line use and prediction accuracy. This report outlines generally applicable approaches to computational gene prediction and known limitations in this field.

Computational determination of coding regions in eukaryotic genomes is an important problem<sup>1,2</sup> that has been brought into the spotlight by advances in genomic sequencing.<sup>3-5</sup> Since the publication of the human genome, public interest in gene finding has somewhat declined, but this change in fashion does not make the problem of computational gene prediction any less valid or useful. While it is clear that at present no computer algorithm would be able to annotate a newly sequenced genome in a fully automated fashion, gene-finding programs have substantial utility, in particular when they are combined with large volumes of experimental data.<sup>3</sup> The goal of this review is to describe and compare the computer programs for gene

prediction of so-called 'suboptimal' exons, paving a way for computational annotation of alternatively spliced transcripts.

In this paper, we will compare these programs on availability and accuracy as the primary criteria. Availability is evaluated in terms of:

- possibility to submit a prediction request through the web interface;
- ability to install the program locally if desired;
- ability to train and test the program independently;
- availability of source code; and

# Computer programs for eukaryotic gene prediction

Briefings in Bioinformatics, 3(2): 195-199, 2002

**Table 1:** Availability of gene prediction software

Program	URL	Web interface	Download	Source
Geneid	<a href="http://www.limim.es/software/geneid/index.html">http://www.limim.es/software/geneid/index.html</a>	yes	yes	yes <sup>a</sup>
GlimmerM	<a href="http://www.tigr.org/softlab/glimmerm/">http://www.tigr.org/softlab/glimmerm/</a>	yes	yes <sup>b</sup>	yes <sup>b</sup>
GenScan	<a href="http://genes.mit.edu/GENSCAN.html">http://genes.mit.edu/GENSCAN.html</a>	yes	yes <sup>b</sup>	no
GenomeScan	<a href="http://genes.mit.edu/genomescan/">http://genes.mit.edu/genomescan/</a>	yes	no	no
MHMGene	<a href="http://www.cbs.dtu.dk/services/HMMgene/">http://www.cbs.dtu.dk/services/HMMgene/</a>	yes	no	no
FGENES	<a href="http://genomic.sanger.ac.uk/gf/gf.shtml">http://genomic.sanger.ac.uk/gf/gf.shtml</a>	yes	no	no
Genie	<a href="http://www.cse.ucsc.edu/~dkulp/cg-bin/genie">http://www.cse.ucsc.edu/~dkulp/cg-bin/genie</a>	yes <sup>c</sup>	no	no

<sup>a</sup>Free to all under GNU licence.

<sup>b</sup>Free only to academic users. Commercial users must purchase a licence.

<sup>c</sup>Web interface to Genie uses an older version of this program, not that one referred to in publications of fruit fly and human genomes or the GASP.

**Table 2:** Applicability of gene prediction software by organism

Program	Parameters available for	May train and test
Geneid	Humans and fruit fly ( <i>Drosophila melanogaster</i> )	Yes
GlimmerM	<i>Plasmodium falciparum</i> (the malaria parasite), <i>Arabidopsis thaliana</i> , <i>Oryza sativa</i> (rice) and <i>Aspergillus thaliana</i> , maize	Yes
GenScan	General vertebrate parameter set, <sup>a</sup> <i>Arabidopsis thaliana</i> , maize	No
GenomeScan	General vertebrate parameter set, <sup>a</sup> <i>Arabidopsis thaliana</i> , maize	No
MHMGene	Humans and <i>Caenorhabditis elegans</i>	No
FGENES	Humans, fruit fly ( <i>D. melanogaster</i> ), nematode, yeast and plant <sup>b</sup>	No
Genie	Humans, fruit fly ( <i>D. melanogaster</i> )	No

<sup>a</sup>The general vertebrate parameter set is derived from the training and testing set of 570 sequences described in Buset and Guigo.<sup>13</sup> It is not limited to human sequences.

<sup>b</sup>There is no explanation on the web site for what exact species of nematodes, yeast and plants the parameter sets are offered.

**Table 3:** Comparative accuracy of gene prediction on nucleotide level

Program	Sn <sup>a</sup>	Sp <sup>a</sup>	Sn <sup>b</sup>	Sp <sup>b</sup>	Sn <sup>c</sup>	Sp <sup>c</sup>	Sn <sup>d</sup>	Sp <sup>d</sup>	Sn <sup>e</sup>	Sp <sup>e</sup>
MHMGene	97	91	93	93						
GenScan		95	90	93	93	93				
Geneid	86	83					85	92	94	92
Genie	96	92	91	90	78	84				
FGENES	89	77	86	88	77	85	92	93		
GenomeScan					Test results not available					
GlimmerM					Test results not available					

<sup>a</sup>Tested on the Adh region in *Drosophila*; <sup>16</sup> data from Pavlovic et al.<sup>14</sup> and Reese et al.<sup>16</sup>

<sup>b</sup>Tested on a set of 195 high-quality mammalian sequences (human, mouse and rat), which has been experimentally validated and includes both multiple and single gene sequences. This set is described in detail in the original paper; <sup>15</sup> data from Rogic et al.

<sup>c</sup>Tested on a set of 570 single-gene vertebrate sequences; data from Buset and Guigo.<sup>13</sup>

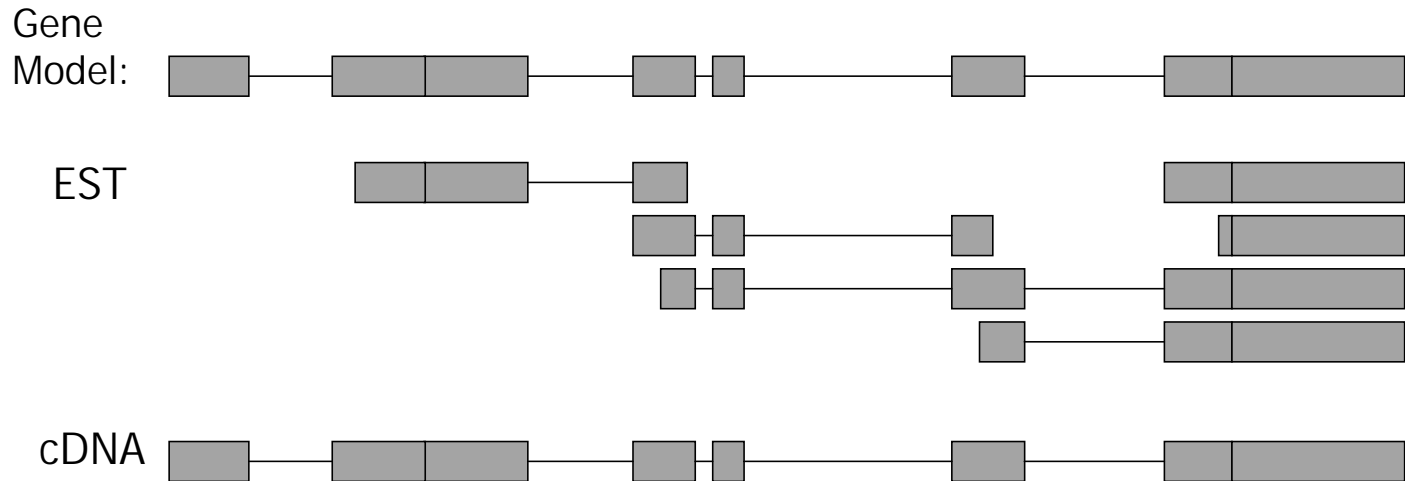
<sup>d</sup>Geneid tested by the author in-house (previously unpublished) on a set of 570 sequences.<sup>13</sup> Other performance figures collected by the author from the web sites for respective programs.

<sup>e</sup>Same as (d), except tested on the Adh region in *Drosophila*.<sup>16</sup>

# Transcript-based prediction: How it works

---

- Align transcript data to genomic sequence using a pair-wise sequence comparison



# EST databases

---

- EMBL/GenBank have separate sections for EST sequences
- ESTs are the most abundant entries in the databases (>60%)
- ESTs are separated by division in the databases
- EST sequences are submitted in bulk, but do have to meet minimal quality criteria

# Cross species sequence comparison-based gene recognition

---

## Comparative Gene Finding

- 1. Align the two sequences
- 2. Predict genes at both sequences simultaneously.
- Computational problems:
  - Aligning long (>200 Kb) sequences
  - A model for cross-species gene prediction

# Signals for exon detection

---

- Cross-species conservation
  - Protein Sequence Conservation
  - Length conservation of coding exons
- Other signals
  - Splice sites
  - Codon usage statistics
  - Length statistics
  - Gene structure/coding frame restrictions

# Tips for gene prediction

---

- Gene prediction
  - Anything found by most programs is high priority
  - Anything found by one or a few is lower priority
  - Some have been rated as more accurate than others (i.e., should look into current literatures)
  - One program is not reliable → combine different evidence.

# Genome annotation

---

- No universal criteria...
- Document repetitive elements, low-complexity regions, tRNA genes
- Document known genes (i.e. full-length mRNA known)
- Document EST hits
- Document regions with similarity to known proteins
- Run gene prediction algorithm(s)

# Bacterial genome annotation

---

- Find ORFs, run GeneMark / Glimmer
- Find promoter & regulatory regions
- Document operons
- Document orthologs in other species, determine function
- Verify termini by multiple alignment
- Document metabolic pathways, ensure all enzymes are present

# Genome annotation

---

- Annotation is the process of interpreting raw sequence data into useful biological information
- Annotations describe the genome and transform raw genome sequences into biological information by integrating computational analyses, other biological data and biological expertise.
- Old Days: One Gene done by one Lab = LOTS of INFO
- Now: Many genes = Superficial and incomplete of many genes.
- Features could be repeats, genes, promoters, protein domains.....
- Features can be linked to other databases eg Pfam/Pubmed

# Annotation is the description of:

---

- Function(s) of the protein
- Post-translational modification(s)
- Domains and sites
- Secondary structure
- Quaternary structure
- Similarities to other proteins
- Disease(s) associated with deficiency(ies) in the protein
- Sequence conflicts, variants, etc.

# Annotation sources

---

- Publications that report experimental data
- Review articles on specific protein families or groups of proteins
- Protein sequence analysis
- External experts on the organism
- Comparison with other, related sequenced organisms

# Similarity with ‘known’ proteins

---

- Single most important procedure in assigning peptide function
- Orthologous proteins that share function are similar
- Blastp against a non-redundant protein database
  - SWISSPROT – excellent annotation, limited coverage
  - TREMBL – poor annotation, full coverage
- Multiple sequence alignment can be a trigger for re-prediction
- Remember that the protein sequences and annotation could be incorrect

# Signal peptides

---

- Information regarding whether a protein is to be moved across membranes in the cell.
- Classic examples are nuclear encoded peptides which reside in the mitochondria or peptides which are to be presented on the cell surface
- One example is the SignalP program which is available as a web/mail server as well as an older implementation in the emboss/bioperl packages

# Transmembrane domains

---

- Regions of the peptide which span a membrane
- Examples include the respiratory complex in mitochondria, transporters, ion-channels etc
- A number of different methodologies to predict these
  - ‘sliding window’ algorithms based on the amino-acid composition (hydrophobic residues) e.g. TmPred
  - Hidden Markov models e.g. TMHMM

# Low-complexity regions

---

- Peptide regions which are repetitive or more formally have a low information content
- Think of these regions as peptide repeats
- Low-complexity regions can be indicative of some function, structural or molecular mimicry/host evasion or incorrect gene prediction
- Program of choice for finding low-complexity regions is seg (this is widely used in blast servers as a preliminary masking process).

# Secondary protein databases

---

- A number of secondary protein databases exist which are designed to collate similar (orthologous) proteins together
  - Pfam/SMART/Prosite/InterPro
  - COGs (Clusters of Orthologous Genes)
- Searching against these databases gives an excellent guide to assigning function

# Finding protein motif

---

- All ORFs/Genes should be tested with motif searching software, e.g.,
  - Pfam
  - Prosite
  - SMART
- Eventually run through InterPro (gives graphical representation of all motif hits)

# Clusters of Orthologous Genes (COGS)

---

- A database of orthologous genes from a range of completed genomes (currently this is weighted toward microbial genomes)
- Available from the NCBI website
- Excellent coverage of the standard metabolic enzyme families make COGS ideal for cross-referencing between genomes

# Grouping genes

---

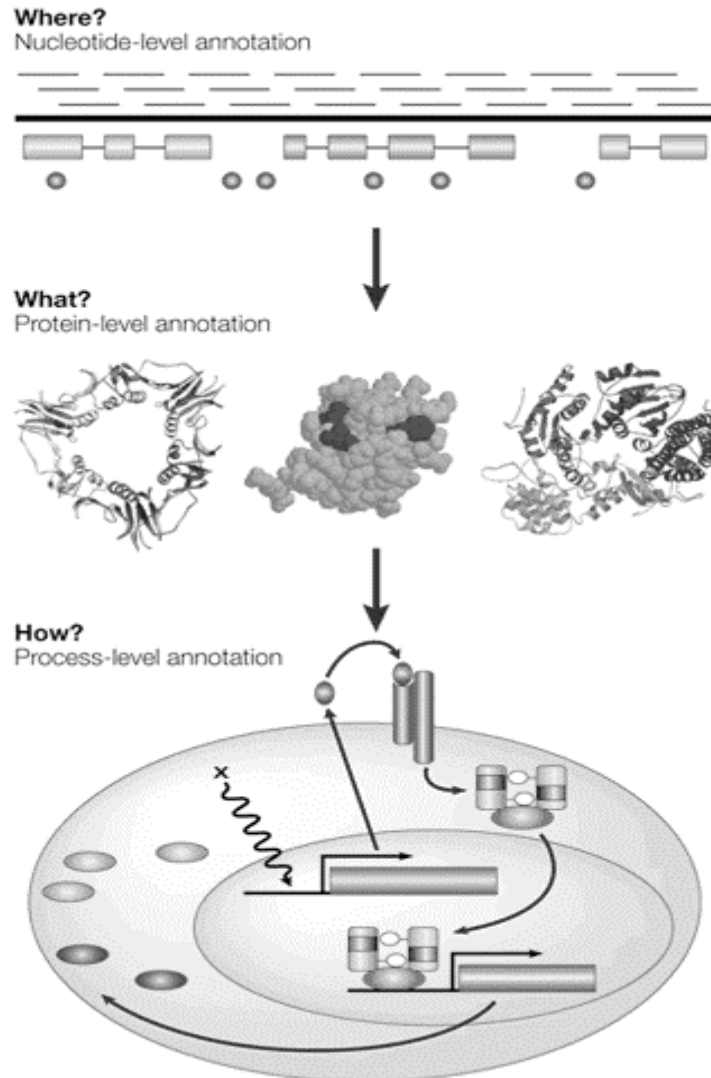
- Gene Ontology
  - based on
    - cellular component
    - molecular function
    - biological process
  - problem where prediction overlaps sets, being sorted now.

# Genome Ontology (GO)

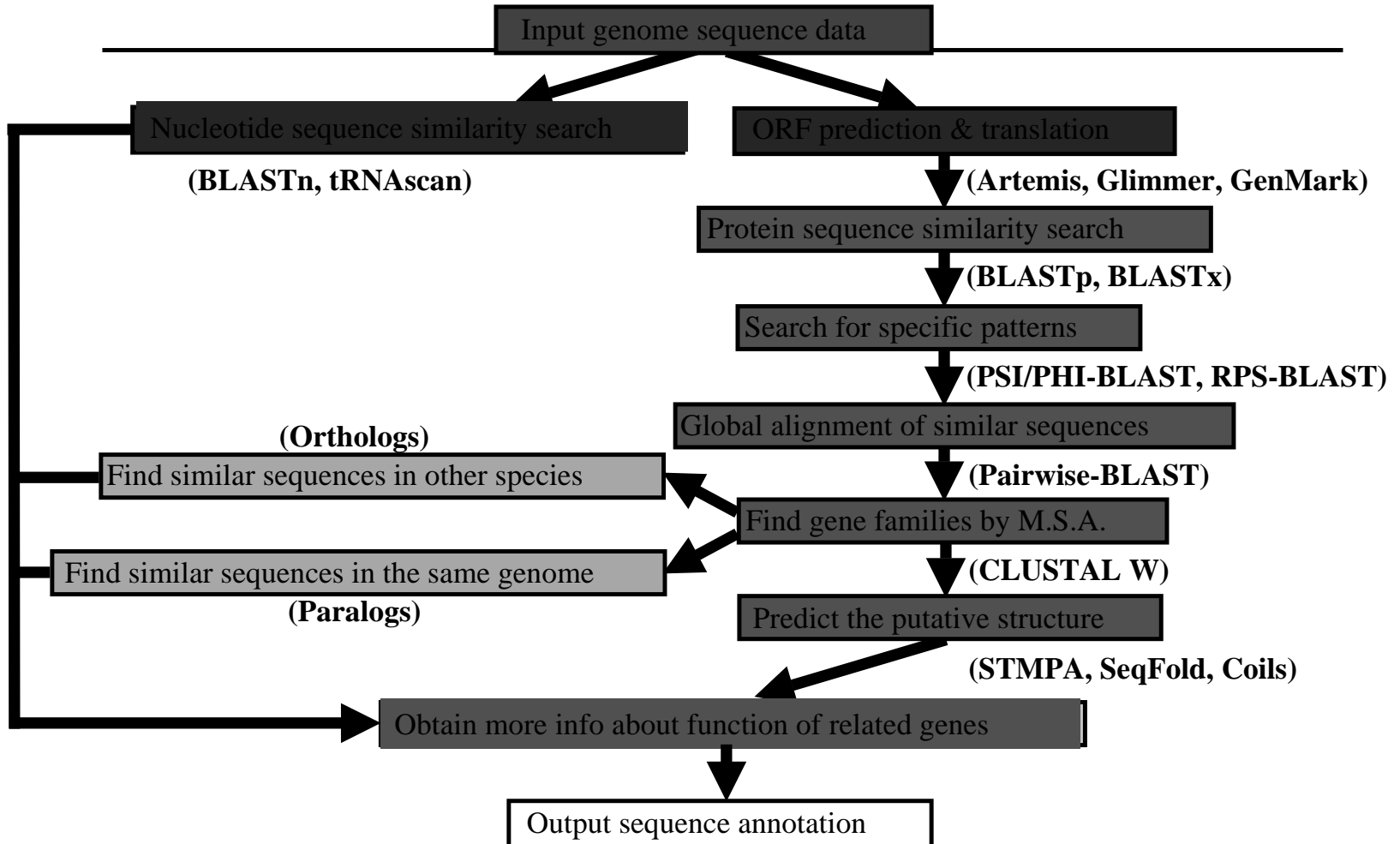
---

- An ontology is a restricted vocabulary used to describe/classify
- The GO consortium grew from efforts in the fly community to assign function/localisation/process information to the biology of the fruit fly.
- The GO consortium currently has representatives of all the major model organisms and provides a methodology of comparing genomes based on functional/biological terms
- There are several levels of assigning GO terms the lowest of which is by similarity. To this end the consortium has prepared a list of InterPro to GO mappings from which GO terms can be added to the protein based on it's InterPro matches

# The 3 Layers of Genome Annotation: Where, What & How?



# Gene Annotation Pipeline



# Predicting function from sequence similarity

---

- Orthologs- arose from speciation, same gene in different organisms  
-can have <30% homology
- Paralogs- from duplication within a genome, second copy may have new or changed function  
(difficult to distinguish between ortho- and paralogs unless whole genome is available)
- Equivalog- proteins with equivalent functions
- Analog- proteins catalyzing same reaction but not structurally related
- Some enzymes may have sequence similarity simply because common catalytic site, substrate, pathway.

# Types of homology

---

Superfamily

PROTEIN/DOMAIN

Duplication within species

Paralogs  
may have different  
functions

A

B

Speciation

Orthologs  
may have different  
functions, if same -  
Equivalent

B1

B2

# Protein families, motifs & domains.

---

- Proteins with common functions have some common features.
- Domains and motifs from conserved residues.
- Families can be grouped, profiles and HMMs derived.
- There is more to life than Blast



---

# Experimental annotation of the human genome using microarray technology

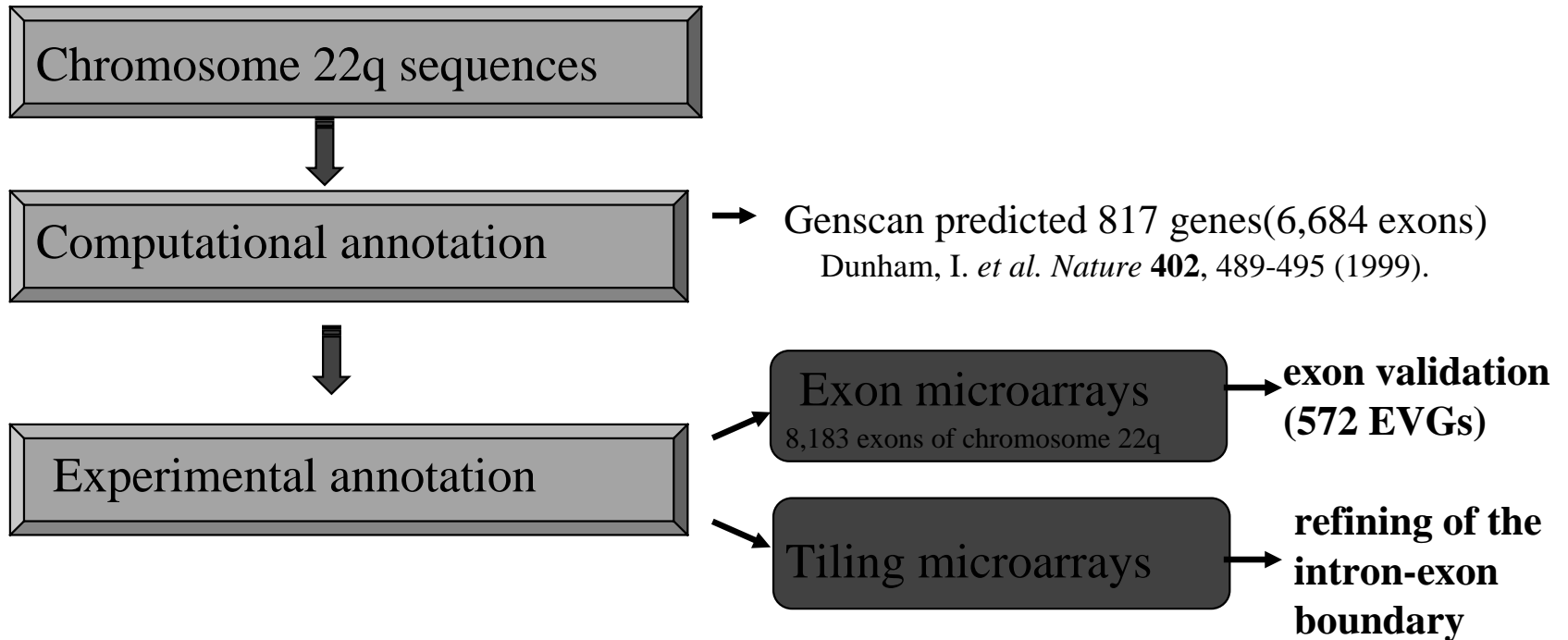
D. D. SHOEMAKER\*, E. E. SCHADT\*, C. D. ARMOUR, Y. D. HE, P. GARRETT-ENGELE, P. D. MCDONAGH, P. M. LOERCH, A. LEONARDSON, P. Y. LUM, G. CAVET, L. F. WU, S. J. ALTSCHULER, S. EDWARDS, J. KING, J. S. TSANG, G. SCHIMMACK, J. M. SCHELTER, J. KOCH, M. ZIMAN, M. J. MARTON, B. LI, P. CUNDIFF, T. WARD, J. CASTLE, M. KROLEWSKI, M. R. MEYER, M. MAO, J. BURCHARD, M. J. KIDD, H. DAI, J. W. PHILLIPS, P. S. LINSLEY, R. STOUGHTON, S. SCHERER & M. S. BOGUSKI

Rosetta Inpharmatics, Inc., 12040 115th Avenue N.E., Kirkland, Washington 98034, USA

*Nature* **409**, 922 - 927 (2001)

# Overview

---



# Metabolic databases

---

<u>Database</u>	<u>URL</u>
PFBP	<a href="http://www.ebi.ac.uk/research/pfmp">http://www.ebi.ac.uk/research/pfmp</a>
KEGG	<a href="http://www.genome.ad.jp/kegg/">http://www.genome.ad.jp/kegg/</a>
BRENDA	<a href="http://srs6.ebi.ac.uk/">http://srs6.ebi.ac.uk/</a> <a href="http://www.uni-koeln.de/math-nat-fak/biochemie/ds/dsbren_e.htm">http://www.uni-koeln.de/math-nat-fak/biochemie/ds/dsbren_e.htm</a> <a href="http://www.brenda.uni-koeln.de">http://www.brenda.uni-koeln.de</a>
PathDB	<a href="http://www.ncgr.org/research/pathdb/">http://www.ncgr.org/research/pathdb/</a>
SRS	<a href="http://srs6.ebi.ac.uk/">http://srs6.ebi.ac.uk/</a>
EcoCyc	<a href="http://ecocyc.panbio.com/ecocyc/">http://ecocyc.panbio.com/ecocyc/</a>
MPW-EMP	<a href="http://srs6.ebi.ac.uk/">http://srs6.ebi.ac.uk/</a> OR <a href="http://wit.mcs.anl.gov/WIT2/">http://wit.mcs.anl.gov/WIT2/</a>
CSNbd	<a href="http://geo.nih.gov/jp/csndb/">http://geo.nih.gov/jp/csndb/</a>
Transfac	<a href="http://transfac.gbf.de/">http://transfac.gbf.de/</a>
RegulonDB	<a href="http://www.cifn.unam.mx/Computational_Biology/regulondb/">http://www.cifn.unam.mx/Computational_Biology/regulondb/</a>
DPinteract	<a href="http://arep.med.harvard.edu/dpinteract/">http://arep.med.harvard.edu/dpinteract/</a>
ooTFD	<a href="http://www.isbi.net/">http://www.isbi.net/</a>

## Genome Annotation Quality

- What is the quality of genome annotation?
- Quality of sequence well known
- Quality of gene prediction at least roughly understood
- Functional accuracy of 99.5% claimed...  
... but not tested experimentally
- *We rely upon functional assignments for biological interpretation*

## The Annotation of *M. genitalium*

1. TIGR sequences genome and makes initial annotation
2. GeneQuiz consortium automatically annotates
3. Eugene Koonin et al (NCBI) manually make annotations
4. GeneQuiz consortium automatically re-annotates
5. Updates
  - Several groups make automated structural annotations
  - TIGR makes updates to annotation, including new genefinding

Different groups use similar methods and operated sequentially, reviewing each others' results

## Compatible Annotations

**mg463**

- TIGR:** ● high level kasgamicin resistance (*ksgA*)
- NCBI:** ● rRNA (adenosine-N6, N6-)-dimethyltransferase (*ksgA*)
- GeneQuiz:** ● Dimethyladenosine transfe [sic]

**mg010**

- TIGR:** ● DNA primase (*dnaE*)
- NCBI:** ● DNA primase (truncated version) (*DnaGp*)
- GeneQuiz:** ● DNA primase (EC 2.7.7.-)

**mg225**

- TIGR:** ● hypothetical protein
- NCBI:** ● amino acid permease
- GeneQuiz:** ● histidine permease

## Incompatible Annotations

**mg302**

- TIGR:** ● no database match  
**NCBI:** ● (glycerol-3-phosphate?) permease  
**GeneQuiz:** ● mitochondrial 60S ribosomal protein L2

**mg448**

- TIGR:** ● pilin repressor (pilB)  
**NCBI:** ● putative chaperone-like protein  
**GeneQuiz:** ● pilB protein

**mg085**

- TIGR:** ● hydroxymethylglutaryl-CoA reductase (NADPH)  
**NCBI:** ● ATP(GTP?)-utilizing enzyme  
**GeneQuiz:** ● NADH-ubiquinone oxidoredu [sic]

# Genome Annotation Quality

- **Average error rate at least 8%**
  - Actual error rate likely to be 2-3 times higher
- **Where do errors come from?**
  - Poor sequence comparison: not homology at all
  - Incorrect inferences of function from homology
  - Propagation of erroneous data
- **Solutions?**
  - Careful sequence comparison
  - Avoidance of over-annotation
  - Complete description of method in database
  - New methods for functional characterization

■ 2 OK  
■ 2 Wrong  
■ 3 OK  
■ 3 Wrong  
■ 0 or 1

