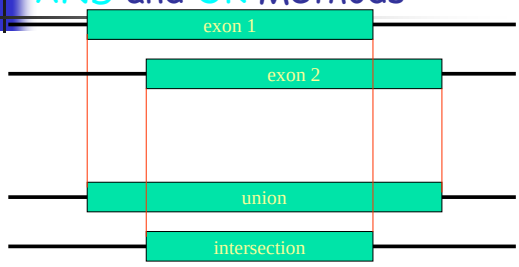
Integrated approaches for gene finding

- Programs that integrate results of similarity searches with *ab initio* techniques (GenomeScan, FGENESH+, Procrustes)
- Programs that use synteny between organisms (ROSETTA, SLAM)
- Integration of programs predicting different elements of a gene (EuGène)
- Combining predictions from several gene finding programs (combination of experts)

AND and OR Methods



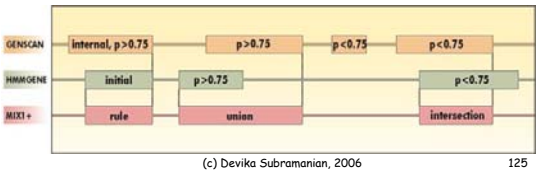
Combining Genscan and HMMgene

- High prediction accuracy as well as reliability of their exon probability make them good candidates.

- Genscan predicted 77% of exons correctly, HMMgene 75%, both 87%

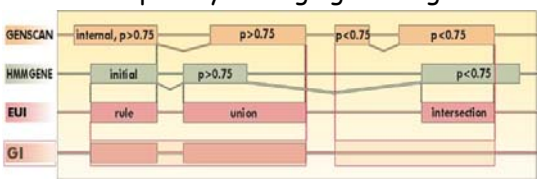
EUI Method (exon union – intersection)

1. Union of exons with $p \geq 0.75$
2. Intersection of exons with $p < 0.75$
3. Rule for initial exon



Gene intersection (GI) method

1. Intersection of genes
2. Apply EUI method to exons completely belonging to GI genes



EUI with reading frame consistency

1. Assign probabilities to GI genes. Determine position of acceptor and donor site in a reading frame.
2. GI gene with higher probability imposes the reading frame. Choose only EUI exons contained in GI genes that are in a chosen reading frame.

Results – Burset/Guigó dataset

METHODS	#no prediction	Nucleotide accuracy			Exon accuracy				
		Sn	Sp	AC	ESn	ESp	(ESn+ESp)/2	ME	WE
Genscan	8	0.94	0.93	0.92	0.78	0.81	0.80	0.09 (203)	0.05 (188)
HMMgene	38	0.93	0.94	0.92	0.81	0.83	0.82	0.14 (308)	0.04 (139)
EUI	20	0.94	**0.96**	**0.93**	**0.83**	**0.88**	**0.85**	0.12 (250)	**0.03** (96)
GI	43	0.91	**0.97**	**0.93**	**0.82**	**0.90**	**0.86**	0.18 (386)	**0.02** (67)
EUI_frame	27	0.93	**0.96**	**0.93**	**0.83**	**0.88**	**0.85**	0.13 (286)	**0.03** (87)

Summary: Eukaryotic gene finding

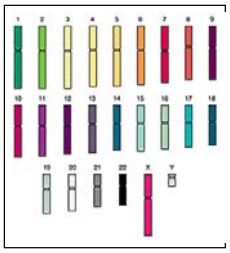
- Overall accuracy usually below 50%
  - Human gene finding is hardest
  - Very long introns, and lots of them
- Leading methods: HMMs and variants
- New ideas needed
- New opportunity: use sequence of related species

Comparison of 1196 orthologous genes

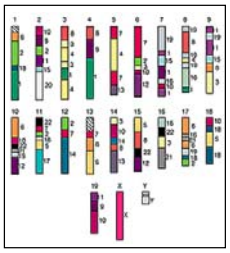
- Sequence identity between genes in human/mouse
  - exons: 84.6%
  - protein: 85.4%
  - introns: 35%
  - 5′ UTRs: 67%
  - 3′ UTRs: 69%
- 27 proteins were 100% identical.

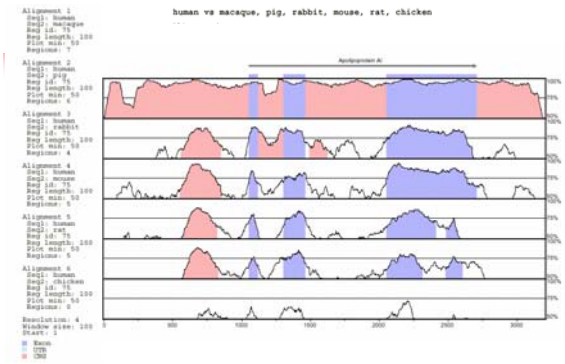
Human-mouse homology

Human

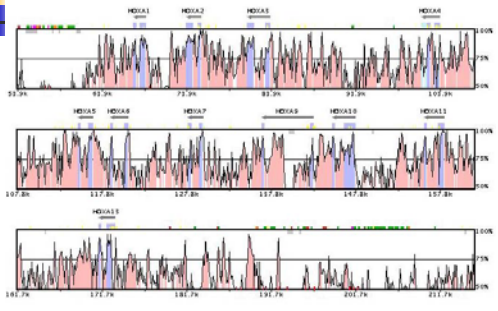


Mouse





HoxA human-mouse

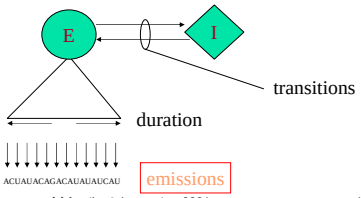


Alignment

```
 50     .    :    .    :    .    :    .    :    .    :
247 GGTGAGGTCGAGGACCCTGCA  CGGAGCTGTATGGAGGGCA   AGAGC
    |:   ||  ||||:  |||| --:||  ||| |::|   |||---||||
368 GAGTCGGGGGAGGGGGCTGCTGTTGGCTCTGGACAGCTTGCATTGAGAGG

100     .    :    .    :    .    :    .    :    .    :
292 TTC          CTACAGAAAAGTCCCAGCAAGGAGCCACACTTCACTG
    |||----------|| |   |::| |: ||||::|:||:-||  ||:| |
418 TTCTGGCTACGCTCTCCCTTAGGGACTGAGCAGAGGGCT CAGGTCGCGG

150     .    :    .    :    .    :    .    :    .    :
332               ATGTCGAGGGGAAGACATCATTCGGGATGTCAGTG
    --------------|||||||||||||||||||||:|||||||||||||
467 TGGGAGATGAGGCCAATGTCGAGGGGAAGACATCATTTGGGATGTCAGTG

200     .    :    .    :    .    :    .    :    .    :
367 TTCAACCTCAGCAATGCCATCATGGGCAGCGGCATCCTGGGACTCGCCTA
    |||||:|||||||:|||||||||||||:|| ||:|||||:||||||||
517 TTCAATCTCAGCAACGCCATCATGGGCAGTGGAATTCTGGGGCTCGCCTA
```


Twinscan

- Twinscan is an augmented version of the Genscan HMM.



Twinscan Algorithm

1. Align the two sequences (e.g. from human and mouse)
2. Mark each human base as gap (-), mismatch (:), match (|)

New "alphabet": 4 x 3 = 12 letters

Σ = { A-, A:, A|, C-, C:, C|, G-, G:, G|, U-, U:, U| }

Twinscan Algorithm

3. Run Viterbi using emissions $e_j(k)$ where $k \in$ { A-, A:, A|, …, T| }

Emission distributions $e_j(k)$ estimated from real genes from human/mouse

$e_I(x|) < e_E(x|)$: matches favored in exons
$e_I(x-) > e_E(x-)$: gaps (and mismatches) favored in introns

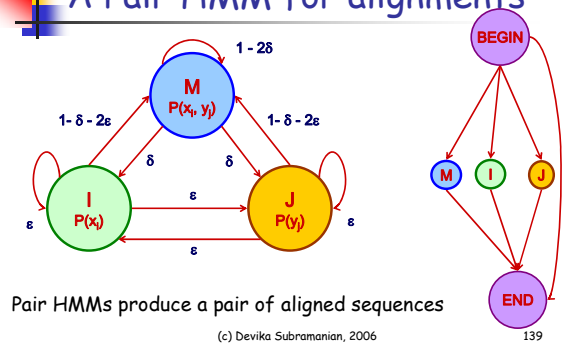
Example

Human: ACGGCGACUGUGCACGU
Mouse: ACUGUGAC GUGCACUU
Alignment: ||:|:|||-||||||:|

Input to Twinscan HMM:
A| C| G: G| C: G| A| C| U- G| U| G| C| A| C| G: U|

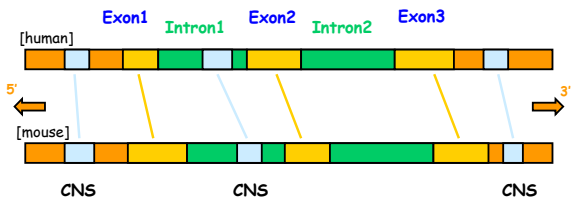
Recall, $e_E(A|) > e_I(A|)$
$e_E(A-) < e_I(A-)$

A Pair HMM for alignments



Pair HMMs produce a pair of aligned sequences

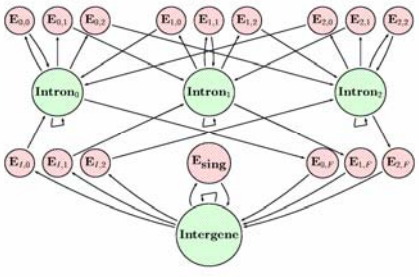
Cross-species gene finding



Exon = coding
Intron = non-coding

CNS = conserved non-coding

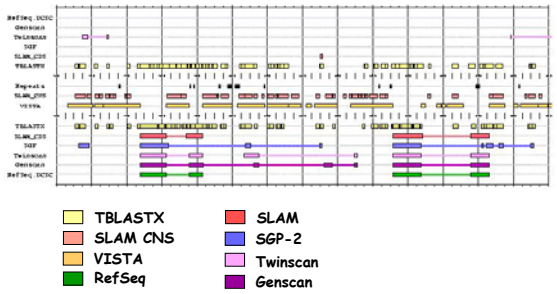
Generalized Pair HMMs



Ingredients in exon scores

- Splice site detection (VLMM)
- Length distribution (generalized)
- Coding potential (codon freq. tables)
- Isochore group

Example: HoxA2 and HoxA3



What have we learned from comparative gene finding?

- conservation is a stronger splice site indicator than consensus
- intron lengths have diverged
- gene structure conservation is more powerful than sequence conservation for prediction
- consensus for GC splice sites

Measuring Performance

	Nucleotide level			Exon level				
Test set	SN	SP	AC	SN	SP	(SN+SP)/2	ME	WE
The ROSETTA set								
ROSETTA	0.935	0.978	0.949	0.833	0.829	0.831	0.048	0.047
SGP-1	0.940	0.960	0.940	0.700	0.760	0.730	0.120	0.040
SLAM	0.951	0.981	0.960	0.783	0.755	0.769	0.038	0.057
TWINSCAN_P	0.960	0.941	0.940	0.855	0.824	0.840	0.045	0.081
TWINSCAN	0.984	0.889	0.923	0.839	0.767	0.803	0.034	0.118
GENSCAN	0.975	0.908	0.929	0.817	0.770	0.793	0.057	0.107
HoxA								
SLAM	0.852	0.896	0.864	0.727	0.533	0.630	0.000	0.333
TWINSCAN_P	0.976	0.829	0.896	0.773	0.531	0.652	0.000	0.312
TWINSCAN	0.989	0.511	0.704	0.591	0.173	0.382	0.000	0.707
SGP-2	0.640	0.637	0.619	0.409	0.173	0.291	0.091	0.596
GENSCAN	0.932	0.687	0.796	0.545	0.235	0.390	0.000	0.569
Elastin								
SLAM	0.876	0.981	0.926	0.802	0.859	0.831	0.121	0.059
TWINSCAN_P	0.942	0.950	0.945	0.879	0.889	0.884	0.066	0.056
TWINSCAN	0.933	0.877	0.903	0.835	0.826	0.831	0.110	0.120
SGP-2	0.755	0.998	0.873	0.593	0.900	0.291	0.352	0.017
GENSCAN	0.947	0.766	0.852	0.835	0.731	0.783	0.121	0.231

Priority organisms

- **Human-mouse gene finding not very high-impact**
  - lots of ancillary data gives better evidence
  - most genes now known
  - nonetheless, this problem is getting all the attention
- **Countless other species really need gene finders**:
  - Brugia malayi (causes lymphatic filariasis)
  - Toxoplasma gondii
  - Schistosoma mansoni (Schistosomiasis)
  - Entamoeba histolytica (50 million cases/year)
  - Tetrahymena thermophila (model organism)
  - Plants: potato, maize, sorghum
  - Mammals: chimp, dog, cow, pig

From the TIGR web site.

Genome scale gene finding

Strategy	Based on	Examples
Ab initio prediction	Models of gene structure/comp	Genscan, GRAIL GenLang, hmmgene
Microarray	Hybridization	Exon-scanning array
Gene inference	Homology	GenomeScan
Genomic:genomic alignment	Homology	ExoFish GLASS/Rosetta
DNA:protein alignment	Homology	GeneWise
cDNA sequencing	Sequencing	RIKEN

C. Burge Nature Genet. 27, 5-7, 2001