



Gene annotation

Alia Benkahla



Overview of this presentation

- Overview of the last “decade”
- Overview of the available tools
- The purpose of this presentation is that at the end you will know exactly what should be done for annotating your favorite sequence. We will try therefore to answer the question:
How would you annotate a genome?



What do we need to know out of genome sequence?

- Genome sequences (DNA only) are not as useful as genomes that are fully annotated
- The function and the structure of proteins/RNA is known for few sequences
- We need to know where the protein coding sequences are and what they do: this is a very big challenge in bioinformatics
- Proteins are not everything, there are also a series other un-deciphered information in DNA and RNA
- Deciphering this information will contribute to a better understanding of molecular biology



What we are looking to annotate?

- CDS
- mRNA
- Alternative RNA
- Promoter and Poly-A Signal
- Pseudogenes
- ncRNA



**We will concentrate here on
protein coding genes**



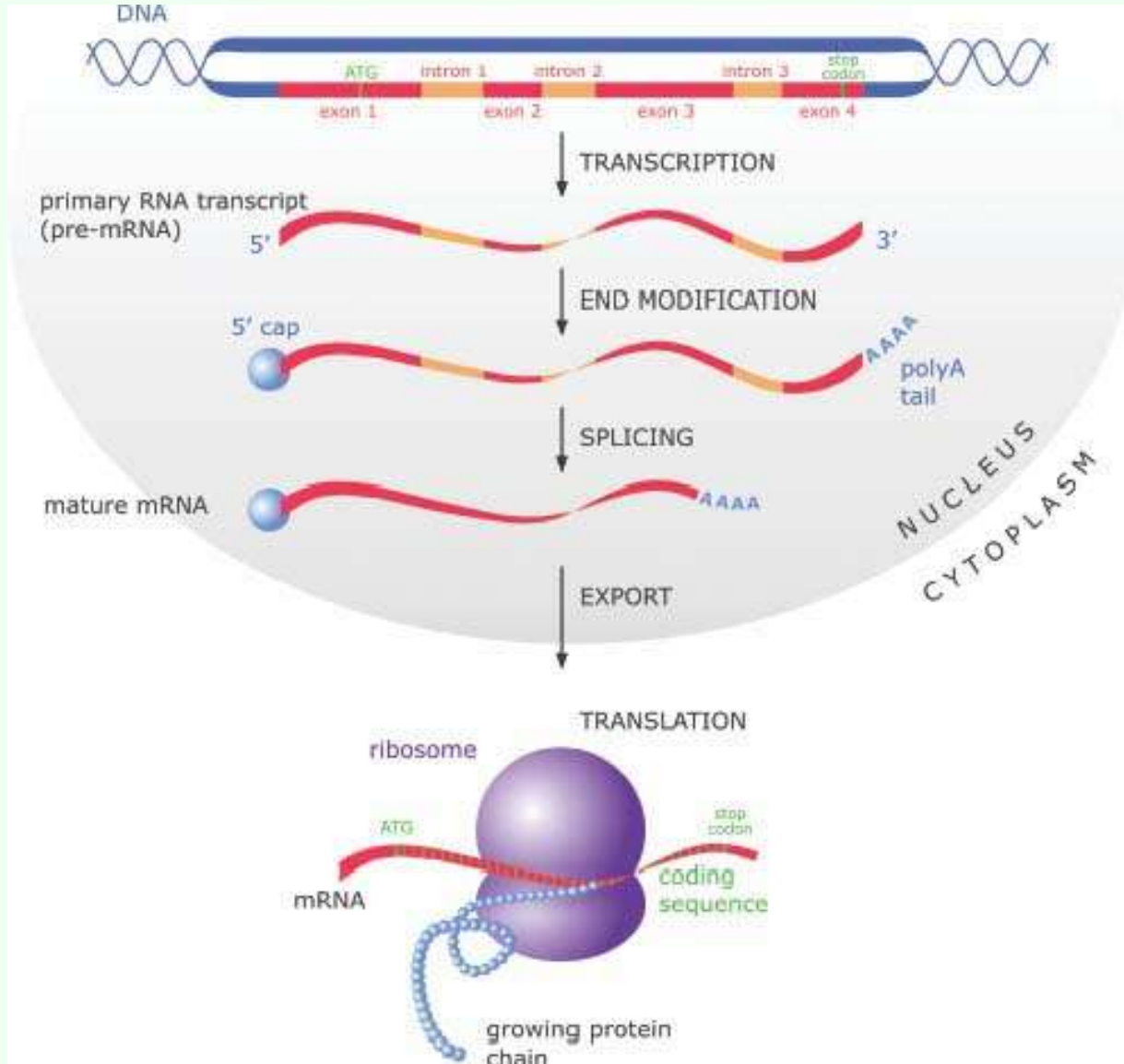
What do we have?

- Genomes sequence(s)
- Genes structures: start codon, stop codon, operons, exon, intron, 5' UTR, 3' UTR, splice-sites, length, inter-genic region, low/high gene density,
- Genes composition



Challenges

- Finding genes involves computational methods as well as experimental validation
- Computational methods are often inadequate, and often generate erroneous ‘gene’ (false positive) sequences which:
 - Are missing exons
 - Have incorrect exons
 - Over predict genes
 - Miss genes (small ones)
 - Where the 5’ and 3’ UTR are missing
- Improve the methods sensitivity (“Fraction of actual coding regions that are correctly predicted as coding”) and specificity (“Fraction of the prediction that is actually correct”)





Major steps in gene annotation

- Genome sequence



- Gene prediction



- Functional gene characterization



Overview of the last decade

- *Genome Research* began publishing complete genomes 12 years ago
- The 1996–2007 *era*, was characterized by tremendous optimism and productivity
- Genomics community exceeded its own sequencing ambitions
- Sequencing has been transformed from a major international event to a common undertaking that barely makes the covers of scientific journals
- This fact has driven a series of advances in computational genome analysis, including methods for predicting the exon-intron structures of genes

- Genomes are there to provide the “list” of their genes
- Most fundamental attribute of anyone’s “list” is the completeness of the set of translated Open Reading Frames (ORFs) and the exon–intron structures of the protein coding region of any mature mRNA

- Today we do not have a complete ORF for each locus in the genome of any higher eukaryote
- From analyzing mammalian genomes, we learned about the incredible abundance of alternatively spliced RNA and retroposed pseudogenes and about the importance of micro-RNAs, siRNAs, and other non-coding RNAs in gene regulation



Foundation of the present *era*

- Three fundamental methods for identifying ORFs in genomic sequence:
 1. *de novo* gene prediction methods;
 2. aligning cDNA and protein sequences;
 3. combining prediction methods;
- Gene discovery in prokaryotic genomes is a quite different problem from that encountered in higher eukaryotic sequences, owing to the higher gene density typical of prokaryotes and the absence of introns in their protein coding genes
 - => Tool developed for the finding of prokaryote genes are different from the ones developed for the finding of eukaryote genes



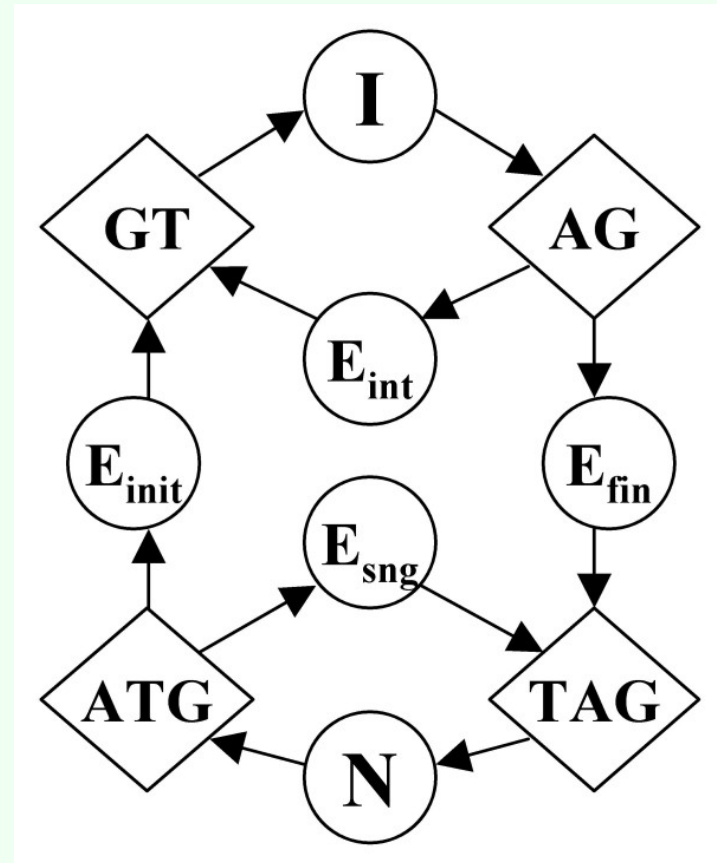
(1) de novo gene prediction methods

- Different categories:
 - Use a single genome sequence
 - Use two genome sequences
 - Use more than two genomes
- Generalized Hidden Markov model (GHMM) algorithms are very popular gene predictors
- In ordinary HMMs, the scores of features (genes) is the sums of the scores of individual bases within the feature
- In GHMMs the scores of features (exons and introns) depend globally on the entire sequence of the exon or intron
- An accurate GHMM for gene finding will assign high probabilities to correct annotations and low probabilities to incorrect annotations
- Burge and Karlin (1997) developed GENSCAN, a GHMM-based gene prediction program that could predict multiple and partial genes on both strands
- GENSCAN remained one of the most accurate and most widely used programs for many years



Example of GHMM topology

- Diamonds represent signal states (for fixed-length features) and circles represent content states (for variable-length features). Allowable transitions are shown with arrows.
 - ATG = start codon,
 - TAG = stop codon,
 - GT = donor splice site,
 - AG = acceptor splice site,
 - N = intergenic region,
 - I = intron,
 - E_{init} = initial exon,
 - E_{int} = internal exon,
 - E_{fin} = final exon,
 - E_{sng} = single exon gene.
- The denoted machine operates by transitioning stochastically from state to state, emitting a gene feature of a particular type upon entering a given state.





(2) Aligning cDNA and protein sequences

- Aligning cDNAs
 - Currently there are 42,050,137 ESTs
 - Local alignment tools like BLAST give an approximation about their location and the location of splice sites
 - To have the exact location and the exact location of splice sites, tools that allow intron gaps with no gap extension penalty were developed: EST_GENOME and SIM4
 - The efficiency of programs depends on the quality of the sequences involved
- Aligning proteins
 - Protein-oriented alignment programs were developed for discovering new members of a known protein family or discovering homologues in a new species
 - GENEWISE is the central part of the ENSEMBL gene annotation pipeline



(3) Combining prediction methods

- In the mid-1990s, the results of cDNA alignments, protein alignments, and *de novo* predictions were integrated by human experts and were often followed by RT-PCR and sequencing experiments to test the predicted exon–intron structure => Computational gene prediction using multiple sources of evidence are successful
- Automated “pipelines” for integrating evidence, such as the ENSEMBL, the NCBI, and the OTTO pipeline were published shortly before the publishing of the draft human genome sequence



Trajectory of improving accuracy

- Aligning cDNA sequences
 - quantity and quality of expressed sequences have improved
 - 415,000 ESTs in 1997 => 7,957,501 ESTs last week
 - MGC project have produced full length cDNAs from large collections of cDNA clones
 - EST_GENOME, SIM4, BLAT
- Single-genome *de novo* gene prediction
 - Burset and Guigó (1996) published the first comprehensive comparison of vertebrate prediction programs: FGENEH (61%), GeneID (51%), the rest (below 50%)
 - In 1995 Burge and Karlin reported that 78% of GENSCAN predictions are correct
 - As test sets became more realistic, estimates of programs accuracy at predicting complete human ORFs dropped
 - Most gene prediction programs, including GENSCAN tend to perform best on high GC content sequences
 - Sensitivity estimates based on the subset of genes whose structures are known should be an unbiased estimate of sensitivity on all genes, because the sets of known genes and unknown genes do not differ in ways that affect accuracy
 - It is possible that unknown genes are different from known genes but there is no reason to believe that they are
 - Determining gene boundaries is one of the most challenging aspects of ORF prediction
 - The value of these programs is not in predicting known genes but in predicting novel genes



Trajectory of improving accuracy

- Dual- and multi-genome *de novo* predictors
 - Exploit local rates and patterns of mutation inferred from alignments between two or multiple genomes
 - No other system outperformed GENSCAN on contiguous human genomic sequences until the advent of dual-genome *de novo* systems
 - Use alignments between two genomes as a rough indicator of which nucleotides are under negative selection and hence are likely to have a function that contributes to fitness: TWINSCAN and SGP2 => predict a correct ORF at about 25% of human loci with known ORFs & predict less than 75% of known human coding exons correctly
 - N-SCAN is a version of TWINSCAN with a phylogenetic conservation model that is capable of considering alignments among multiple genomes => a correct ORF at about 35% of human loci with known ORFs & predict 85% of known human coding exons correctly
 - Even though substantial progress have been made during the last 10 years, gene predictors are generally more accurate on more compact genomes; we have reached the point where *de novo* gene predictions are being used as hypotheses to drive experimental annotation via systematic RT-PCR and sequencing



Trajectory of improving accuracy

- Combining prediction methods
 - Several automated “pipelines” for integrating evidences were developed
 - The genome annotations that are most visible to the public and most widely used are created by combining predictions from different methods
 - GENOMESCAN is an enhancement of GENSCAN that modifies the scores of potential exons depending on whether they have high-scoring alignments to proteins in the databases
 - ENSEMBL gene predictions are created by GENWISE; GENSCAN is used to identify novel, and BLAST is used to align predicted proteins to proteins from other species => the purpose is to obtain a conserved set of predicted exons and genes containing few false positives
 - Pipeline predictions were compared to manual annotation by the HAVANA group:
 - Pairagon+N-SCAN_EST pipeline was substantially more specific than ENSEMBL, and equally sensitive
 - ENSEMBL predicted more transcripts per locus than Pairagon+N-SCAN_EST
 - ENSEMBL predicts at least one correct transcript at a higher percentage of the loci where it made predictions
 - Pairagon+N-SCAN_EST, which uses only human cDNA sequence aligned to its native locus, is at least as accurate as ENSEMBL, which uses cross-locus and cross-species protein alignment
 - A more recent approach to integrating predictions is to score each potential exon using a weighted combination of evidence from alignment-based predictions and *de novo* predictions. The weights are derived from estimates of the accuracy of each prediction source. Thus, if several predictors that have proven accurate in the past agree on an exon, it will receive a high score. In the case of disagreement among predictors the score will generally be lower, but more weight will be given to more accurate predictors.



Annotation “jamboree”

- In 2000, the *Drosophila* community held an annotation “jamboree,” in which fly biologists and bioinformaticists created an initial annotation of the *Drosophila* genome
- In 2002, the Sanger Institute held two Human Annotation Workshops (known as Hawk meetings). A number of groups involved in human annotation compared their annotations in order to “define a standard of annotation” and “draw up guidelines to help achieve the standard” (<http://www.sanger.ac.uk/HGP/havana/hawk.shtml>). These meetings led to the annotation standards that are used by the Sanger Institute’s Human and Vertebrate Analysis project (HAVANA, <http://www.sanger.ac.uk/HGP/havana/docs/guidelines.pdf>)
- The sensitivity and specificity catalogues output of annotation “jamboree” are the highest



de novo gene prediction

- *de novo* gene finders have been part of the standard toolbox for genome annotation and analysis
- The accuracy for compact genomes prediction has become good: one-half to two-thirds of all known genes are predicted exactly right
- Accuracy for mammalian genomes has lagged behind: large number of pseudogenes and small fraction of coding sequence
- Dual genome *de novo* systems: correctly predict about 75% of all known exons and 15–20% of known gene structures
- Pseudogene detection methods have eliminated many false positives from both *de novo* and pipeline-style annotation
- Multiple genome *de novo* systems and the elimination of many pseudogenes: have improved the *de novo* prediction accuracy

- All systems use a training set for parameter estimation



de novo gene finder: single genome sequence

- Are the first systems used to annotate newly sequenced genomes
- Requires the sequence of one genome
- Try to mimic as closely as possible the processes of transcription and RNA processing that define genes biologically
- Hidden Markov model (HMM) algorithm are very popular genes prediction
- An accurate HMM for gene finding will assign high probabilities to correct annotations and low probabilities to incorrect annotations
- GLIMMER2 is the most widely used programs in prokaryotes
- GENSCAN was for many years the most accurate and most widely used programs in eukaryotes



GLIMMER2

- The GLIMMER2 system consists of two programs:
 - build-icm, takes an input set of sequences (complete or partial ORFs from closely related organisms) and outputs an Interloped Context Model file (ICM)
 - glimmer2, uses this ICM to identify putative genes in a genome
- GLIMMER2, first identifies all ORF (>threshold value) and scores each one in all six reading frames. Those that score higher than a designated threshold in the correct reading frame are then selected for further processing.
- Selected ORFs are then examined for overlaps. If two ORFs in different reading frames overlap, the overlapping region is scored separately => The overlap region's six reading frame scores are then compared with those of the two overlapping ORFs to see which frame scores highest. When a longer ORF overlaps a shorter ORF and the overlap region scores highest in the reading frame of the longer ORF, then the shorter ORF is eliminated as a gene candidate.
- The output of the program is a list of putative gene coordinates in the genome.



Putative Genes :

1	326	1747	[+2 L=1422 r=-1.276]	
2	1816	3072	[+1 L=1257 r=-1.285]	
3	3188	3418	[+2 L= 231 r=-1.273]	
4	3437	4546	[+2 L=1110 r=-1.292]	
5	4567	4722	[+1 L= 156 r=-1.324]	
6	4851	6779	[+3 L=1929 r=-1.297]	
7	6892	6782	[-2 L= 111 r=-1.271]	
8	6969	9455	[+3 L=2487 r=-1.305]	
9	9455	9562	[+2 L= 108 r=-1.308]	
10	12138	11668	[-1 L= 471 r=-1.258]	
11	15792	14848	[-1 L= 945 r=-1.290]	
12	15847	17376	[+1 L=1530 r=-1.270]	
13	17532	18860	[+3 L=1329 r=-1.250]	
14	19060	19941	[+1 L= 882 r=-1.264]	
15	19966	20553	[+1 L= 588 r=-1.285]	
16	20752	22152	[+1 L=1401 r=-1.256]	
17	23147	22497	[-3 L= 651 r=-1.292]	
18	23767	23147	[-2 L= 621 r=-1.295]	
19	25149	23869	[-1 L=1281 r=-1.310]	
20	25764	25222	[-1 L= 543 r=-1.278]	
21	25829	26332	[+2 L= 504 r=-1.267]	
22	26812	28500	[+1 L=1689 r=-1.269]	
23	28518	28847	[+3 L= 330 r=-1.207]	
24	28865	29458	[+2 L= 594 r=-1.279]	
25	29479	29700	[+1 L= 222 r=-1.254]	
26	29758	30030	[+1 L= 273 r=-1.287]	
27	32606	32136	[-3 L= 471 r=-1.258]	
28	35843	36454	[+2 L= 612 r=-1.298]	
29	36476	37633	[+2 L=1158 r=-1.243]	
30	37718	39157	[+2 L=1440 r=-1.301]	
31	39157	39792	[+1 L= 636 r=-1.271]	
32	39869	40195	[+2 L= 327 r=-1.306]	
33	40211	40648	[+2 L= 438 r=-1.269]	
34	40663	41649	[+1 L= 987 r=-1.297]	
35	41655	42479	[+3 L= 825 r=-1.284]	
36	42497	42853	[+2 L= 357 r=-1.287]	[DelayedBy #35 L=66]
38	42915	43655	[+3 L= 741 r=-1.270]	
39	43645	43941	[+1 L= 297 r=-1.307]	[DelayedBy #38 L=45]
40	43919	44794	[+2 L= 876 r=-1.269]	



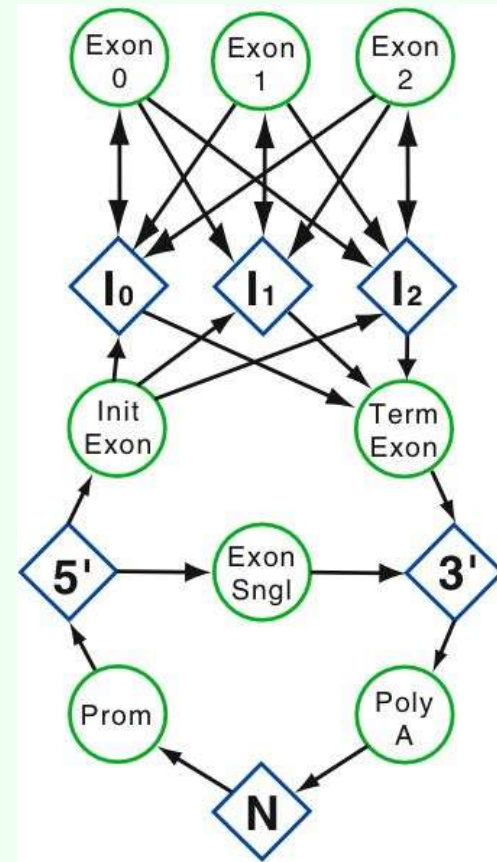
GENSCAN

- Burge and Karlin (1997) developed GENSCAN, a GHMM-based gene finder program that could predict multiple and partial genes on both strands
- GENSCAN was shown to have higher accuracy than existing methods when tested on standardized sets of human and vertebrate genes
- GENSCAN mimics the structure of eukaryote genes and predicts their complete exon/intron structures
- It is based on a general probabilistic model of genes structure
- The states of the model correspond to fundamental functional units of a eukaryotic gene, which may occur in any biologically consistent order e.g. exon, intron, intergenic region, etc.
- It includes: (1) single & multi-exon genes; (2) promoters, polyadenylation signals and intergenic sequences; and (3) genes occurring on either or both DNA strands
- Data concerning 1999 complete coding genes were used as a training set
- Overlapping transcription units and alternative splicing are not explicitly addressed



GENSCAN

- The algorithm is based on probabilistic model of gene structure similar to Hidden Markov Models (HMMs).
- A training set is used by GENSCAN in order to estimate the *HMM parameters*
 - Biological input: Codon bias in coding regions, gene structure (start and stop codons, typical exon and intron length, presence of promoters, presence of genes on both strands, etc.)
 - Covers cases where input sequence contains no gene, partial gene, complete gene, multiple genes





GENOMESCAN

- GENOMESCAN probabilistic model is based on GENSCAN's model
- Exon, intron and splice signal models with similarity to known protein sequences are integrated in the probabilistic model
- Authors have converted the information present in a set of BLASTX hits to a given human genomic sequence into the probabilistic model of GENSCAN
- GENOMESCAN focuses on vertebrate genomes



FGENESH

- Commercial software product from SoftBerry
- FGENESH is HMM-based program for predicting multiple genes in genomic DNA sequences
- In 2002 FGENESH was cited as “The most successful program” (Yu et al. (2002) Science 296:79)
- FGENESH is **50 to 100 times faster than** GENSCAN
- To improve accuracy FGENESH was trained for several taxonomic groups (human, mouse, Drosophila, Anopheles, C.elegans, SS.pombe, Plasmodium, Neurospora, Arabidopsis, Tobacco and monocot plants)
- The software FGENESH++C automate eukaryote genome annotation pipeline
 1. RefSeq mRNA mapping by EST_MAP program - mapped genes are excluded from further gene finder process
 2. *de novo* FGENESH gene finder
 3. Search of all products of predicted genes through NR database for protein homologs
- FGENESB is recommended for prokaryote genome annotation pipeline
- FGENESV is recommended for viral genome annotation pipeline



de novo gene finder: two genome sequence

- Dual-genome gene predictors rely on the fact that functional regions of a genome sequence — protein-coding genes in particular — are more conserved than non-functional ones
- Several programs have been developed that exploit sequence conservation between two genomes to predict genes
- TWINSCAN, SLAM and SGP-2 were used in the comparative analysis of the human and mouse genomes; they significantly outperformed single-genome predictors



TWINSKAN

- TWINSKAN directly extends GENSCAN allowing to exploit the homology between two related genomes
- TWINSKAN consists of a probabilistic model based on DNA conservation sequences, together with the same optimization algorithm used by GENSCAN
- It uses BLAST to search the query genome against an organism specific-genomic database
- TWINSKAN shows a notable improvement over GENSCAN in exon sensitivity and specificity and a dramatic improvement in gene sensitivity and specificity
- The algorithm was trained upon annotated genomic sequences and their homologues



de novo gene finder: from dual to multiple genomes sequence

- Dual-genome *de novo* systems use alignments between two genomes and draw inferences about the rate of evolution at each nucleotide. If the two sequences match at a particular base, that base is conserved; if they do not, it is not conserved.
- It is clearly a crude measure of evolutionary rate; multiple genome alignments provide a more precise measure of evolutionary rate
- Dual-genome predictors for mammalian genomes have had the greatest success using **relatively** distant genomes, such as mouse and human
- However, there are inherent uncertainties in reconstructing the lineages of genomic regions for two **more** distantly related organisms because so many rearrangements, segmental duplications, retro-transpositions and other events have occurred since their latest common ancestor
- One possible solution is to infer evolutionary rate from **many closely related species instead of two more distant species**; Boffelli and colleagues (Science, 2003) have observed that the collective divergence of the higher primates, as a group, is comparable to the divergence of human and mouse, yet their genomes can be aligned much more accurately than those of human and mouse



de novo gene finder: multiple genomes sequence

- Multiple genomes *de novo* gene finders are based on phylo-HMM models
- Phylo- HMMs model is a Markov process in two dimensions: (1) a substitution process over time at each site in the aligned genomes, which is guided by a phylogenetic tree; (2) and a process by which the rate of evolution changes from one site to the next
- All phylo-HMM models assume that the rate of evolution of a nucleotide depends on its function, but the most elaborate phylo-HMM models also consider the possibility that the evolutionary rate may vary from one region of a genome to another
- Furthermore, they allow the probabilities of mutation at each site to depend on the observed pattern of mutation in the previous few sites
- Example of phylo-HMM models: Evolutionary Hidden Markov Model (EHMM) and N-SCAN



N-SCAN

- N-SCAN is a version of TWINSCAN that uses multi-genome alignments to inform gene finder, rather than using alignments between the target genome and one informant genome.
- The multi-genome alignments are exploited via a tree-structured HMM that integrates a phylogenetic tree model with the hidden Markov model used for gene finding



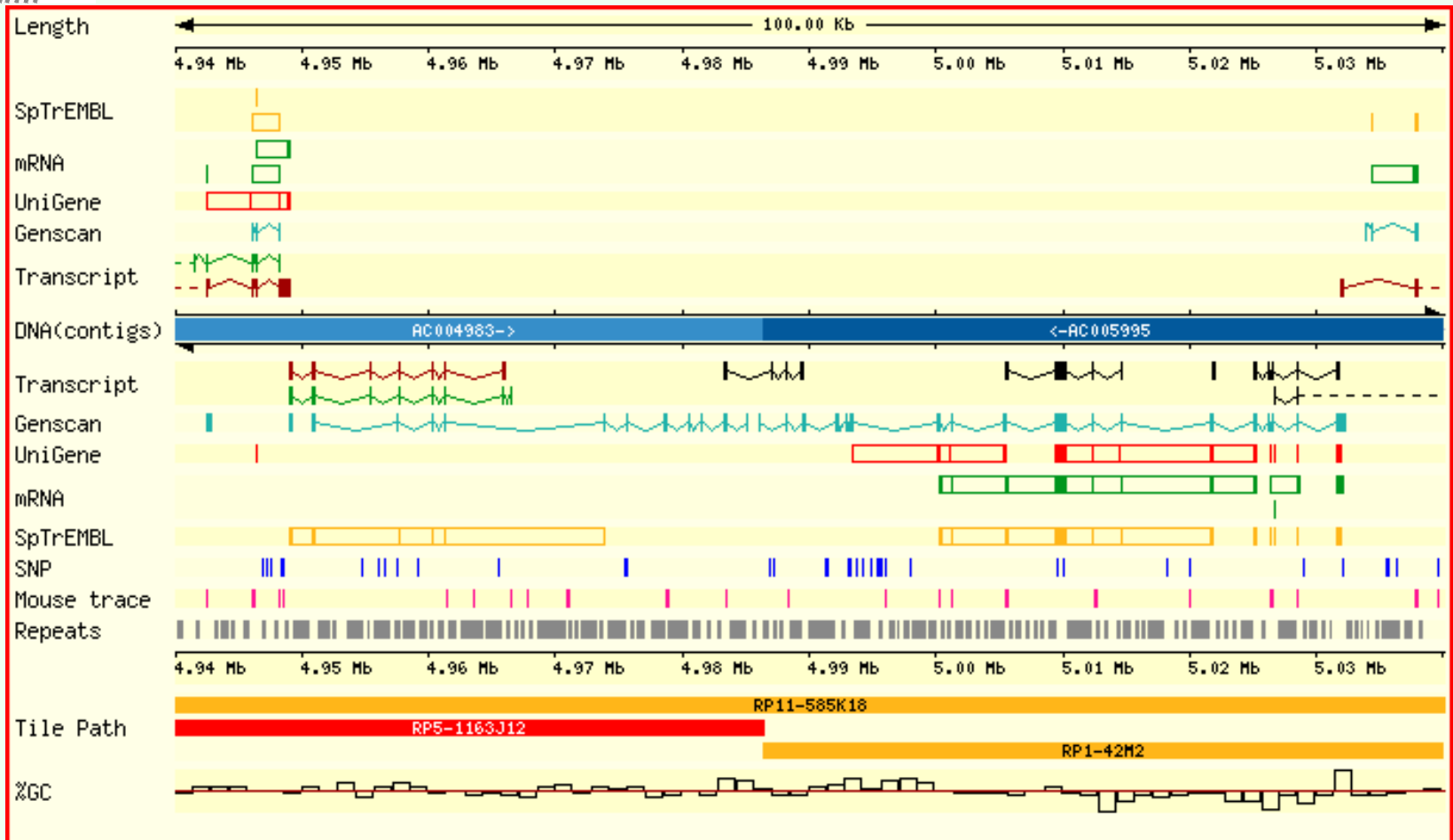
Combining the output of gene predictors

- Human annotators and automated genome annotation ‘pipelines’ (like the one of ENSEMBL) operate by combining information from cDNAs with information from one or more *de novo* gene finders
- Human annotators use their intuition and experience to synthesize the often contradictory evidence into a single gene structure, whereas pipelines generally use rules based on the intuition and experience of their designers.
- The rules are often simply priorities, for example, use an exon predicted by GENSCAN if it is supported by an EST alignment but does not overlap a GENEWISE protein alignment
- Several papers suggest that combining different *de novo* gene predictors with one another and with expression data improves accuracy; this systems is improved further by current efforts to increase the accuracy of cDNA alignment, cross-species gene-structure mapping and cross-species EST alignment.



ENSEMBL annotation pipeline

- ENSEMBL is a system providing automated genome annotation and subsequent visualization of annotated genomes
- All ENSEMBL gene finders are based on experimental evidence, which is imported via manually curated UniProt/Swiss-Prot, partially manually curated NCBI RefSeq, automatically annotated UniProt/TrEMBL records
- Un-translated regions (UTRs) are annotated to the extent supported by EMBL mRNA records
- As there is no guarantee that UTR sequences in EMBL records are complete there is similarly no guarantee that the ENSEMBL genome analysis and annotation pipeline has enough biological evidence to predict complete UTR regions





VEGA project

- The Vertebrate Genome Annotation (VEGA) database is a central repository for high quality, frequently updated, manual annotation of vertebrate finished genome sequence
- Whereas ENSEMBL concentrates on showing computationally derived gene finders on a large number of whole genomes, VEGA shows annotations arising from the labor intensive process of manual curation, on human, mouse, zebrafish and dog genomes
- Vega Human annotation was undertaken by the Havana group at the Wellcome Trust Sanger Institute and is now included in the ENSEMBL human gene build



de novo gene annotation problems

- Do not detect non-coding areas (5' and 3' UTR)
- Exon/intron boundaries are often wrong
- Do not address alternative splicing and predict 1 isoform per gene
- Predictions are for “typical” genes
 - Partial or multiple genes are often missed
 - Training sets may be biased
 - Methods are sensitive to G+C content
 - Weighting of factors may be inordinately biased



Genome annotation problems

- Assembling the genome
- Analysis & interpretation
- Lack of consistency from gene to gene
- Small genes (<300bp) are missed
- Occurrence of overlapping ORFs on opposite DNA strands often leads to ambiguities
- The best prediction programs tend to split and fuse genes, have problems to predict correctly START and STOP codons, predict single isoforms
- Lack of consistency from person to person
- Lack of controlled vocabulary
- Gene expression/molecular interactions
- Updates and maintenance



What has to be done?

- Compute the prediction
- Confirm with biological sequences (also with computational tools)
- Integrate all of this
- Annotate genome
- Validate
- Re-annotate/Update
- Check it twice
- Submit to GenBank



Some Concluding remarks

- Trust the output of gene finder programs but verify
- Beware of gene finder tools!
- Always use more than one gene finder tool and more than one genome when possible
- Active area of development of gene finder tools, so be mindful of the new literature in this



How would you annotate your favourite genome?

Bibliography about your organism

GLIMMER2 OR BLAST =====> viruses & bacteria

Three gene finders GLIMMERM, PHAT, alignments of ESTs
and full length cDNAs + manual curation =====>
COMBINER

GENSCAN/TWINSKAN/N-SCAN

Blast ORF against public database+ protein domain
programs + manual curation =====> annotate the function



General Feature Format (GFF)

- GFF lines are based on the GFF standard file format. GFF lines have nine required fields that must be tab-separated.
- Here is a brief description of the GFF fields:
 1. **seqname** - Must be a chromosome or scaffold.
 2. **source** - The program that generated this feature.
 3. **feature** - Standard feature types are "CDS", "start_codon", "stop_codon", and "exon".
 4. **start** - The starting position of the feature in the sequence.
 5. **end** - The ending position of the feature.
 6. **score** - If there is no score value, enter ".".
 7. **strand** - Valid entries include '+', '-', or '.'.
 8. **frame** - If the feature is a coding exon, frame should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be '.'.
 9. **group** - All lines with the same group (ex. same gene) are linked together into a single item.



Gene Transfer Format (GTF)

- GTF is a refinement to GFF that tightens the specification. The first eight GTF fields are the same as GFF. The group field has been expanded into a list of attributes. Each attribute consists of a type/value pair. Attributes must end in a semi-colon, and be separated from any following attribute by exactly one space
- The attribute list must begin with the two mandatory attributes:
 - gene_id value - A globally unique identifier for the genomic source of the sequence
 - transcript_id value - A globally unique identifier for the predicted transcript
- Here is an example of the ninth field in a GTF data line:

```
gene_id "Em:U62317.C22.6.mRNA"; transcript_id "Em:U62317.C22.6.mRNA"; exon_number 1
```



Bibliography

Review:

- Burge C and Karlin S. (1998) Finding the genes in genomic DNA. *Current Opinion in Structural Biology* **8**:346-354.
- Brent MR & Guigo R. (2004) Recent advances in gene structure prediction. *Current Opinion in Structural Biology* **14**:264–272.
- Brent MR. (2005) Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Res.* **15**(12):1777-86.

Paper:

- Burge C & Karlin S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* **268**(1):78-94.
- Salzberg SL, Delcher AL, Kasif S, White O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **26**(2):544-8.
- Yeh RF, Lim LP, Burge CB. (2001) Computational inference of homologous gene structures in the human genome. *Genome Res.* **11**(5):803-16.
- Korf I, Flicek P, Duan D, Brent MR. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics* **1**:S140-8.
- Solovyev V.V. (2002) Finding genes by computer: Probabilistic and discriminative approaches. *Current Topics in Computational Biology*, MIT Press, p. 365-401.

- Alexandersson M, Cawley S, Pachter L. (2003) SLAM: cross-species gene finding and alignment with a generalized pair hiddenMarkov model. *Genome Res.* **13**(3):496-502.
- Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigo R. (2003) Comparative gene prediction in human and mouse. *Genome Res.* **13**(1):108-17.
- Pedersen JS and Hein J. (2003) Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics* **19**:219–227.
- Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SMJ, Stabenau A, Storey R and M Clamp (2004) The Ensembl Analysis Pipeline. *Genome Res.* **14**(5):934-941.
- Allen JE, Perteza M, Salzberg SL. (2004) Computational gene prediction using multiple sources of evidence. *Genome Res.* **14**:142-148.
- Gross SS, Brent MR. (2006) Using multiple alignments to improve gene prediction. *J Comput Biol.* **13**(2):379-93.



Exercise 1

- We did annotate mycobacterium tuberculosis.
- How did we do?
 - We did the ICM file
 - Train the HMM with the actinobacteria ORF
 - Run glimmer2 and extract the coordinate of the ORF
====> mycobacterium ORF
 - Compared the extracted ORF to the set of available annotated protein catalogues



Exercise 2

- The strategy that was followed for annotating Leishmania genome is on the directory: `/nicolle/home/bcga/tutors/alia/leishmania`
- How did they annotate the genome?



Exercise 3

- Cut the chromosome 21 sequence that is on the directory alia/human
 - Cut windows of 1Mb starting from position 29,000,000
 - You are six groups:
 - Group 1: 29,000,000-29,200,000
 - Group 2: 29,200,000-29,400,000
 - Group 3: 29,400,000-29,600,000
 - Group 4: 29,600,000-29,800,000
 - Group 5: 29,800,000-30,000,000
 - Group 6: 30,000,000-30,200,000
 - `tutors@formation1:~/alia/human$ perl fasta-part.pl chr21.nuc 29000000 30000000 > chr21_29000000_30000000.nuc`
 - Mask the extracted sequence with RepeatMasker
 - Blast the masked sequence against RefSeq sequences. /
`export/home/gbm/dbadmin/db/refseq_all`
 - Extract the matching references
 - Align the matching sequence against the chromosome 21 sequences using
`BLAT/est2genome/sim4`