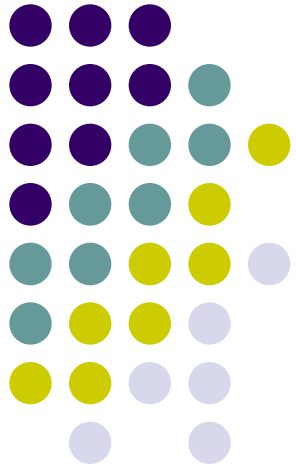
Methods and Algorithms for Gene Prediction

Chaochun Wei 韦朝春 Sc.D.
ccwei@sjtu.edu.cn
http://cbb.sjtu.edu.cn/~ccwei

Shanghai Jiao Tong University
Shanghai Center for Bioinformation Technology



Outline

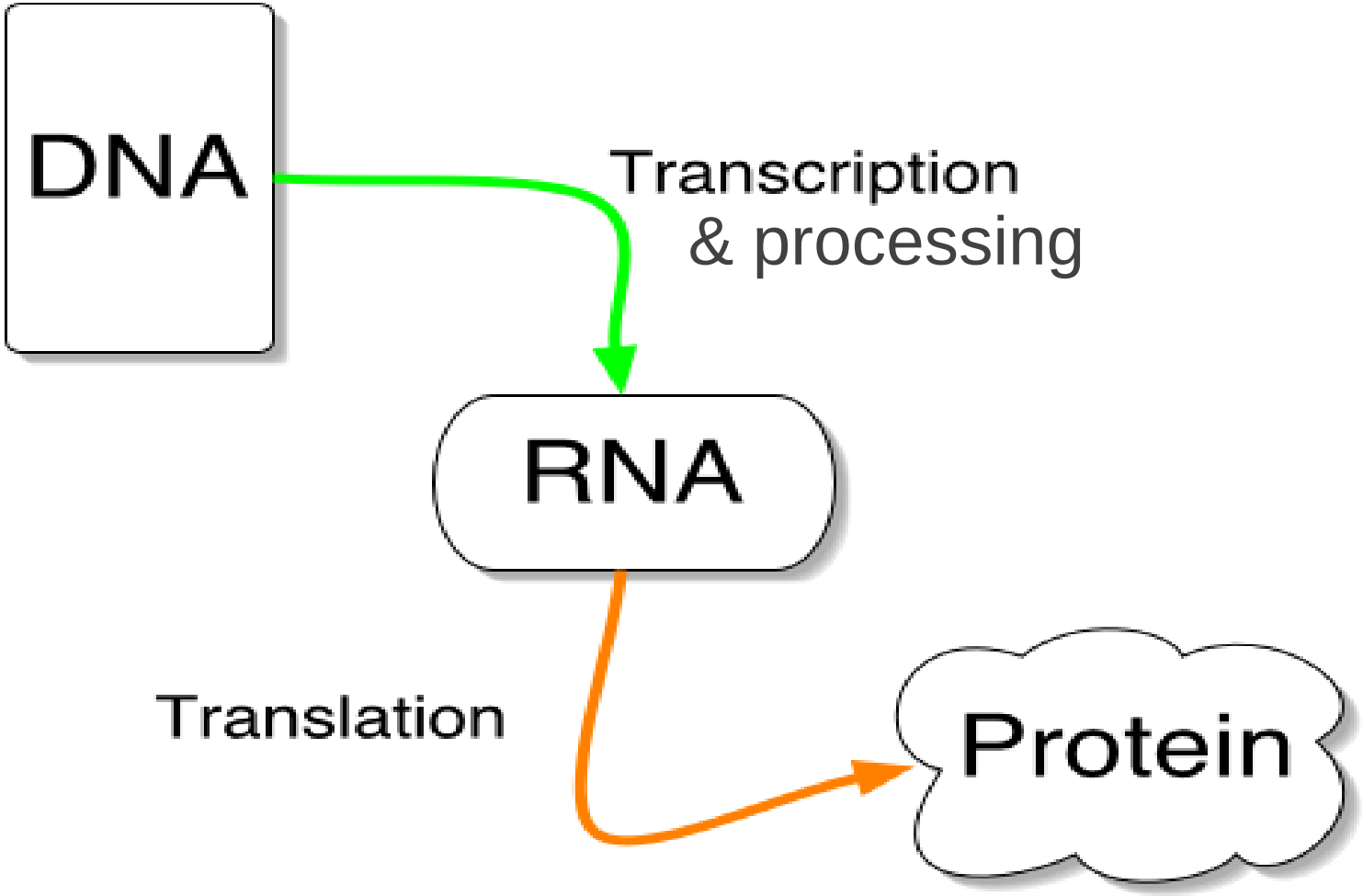
1. Introduction
2. Gene prediction methods
   - Gene prediction methods
   - HMM
   - TWINSCAN and N-SCAN
   - Using ESTs for gene prediction
   - Resources
   - Latest progress
3. Gene Prediction FAQs

1. Introduction: DNA

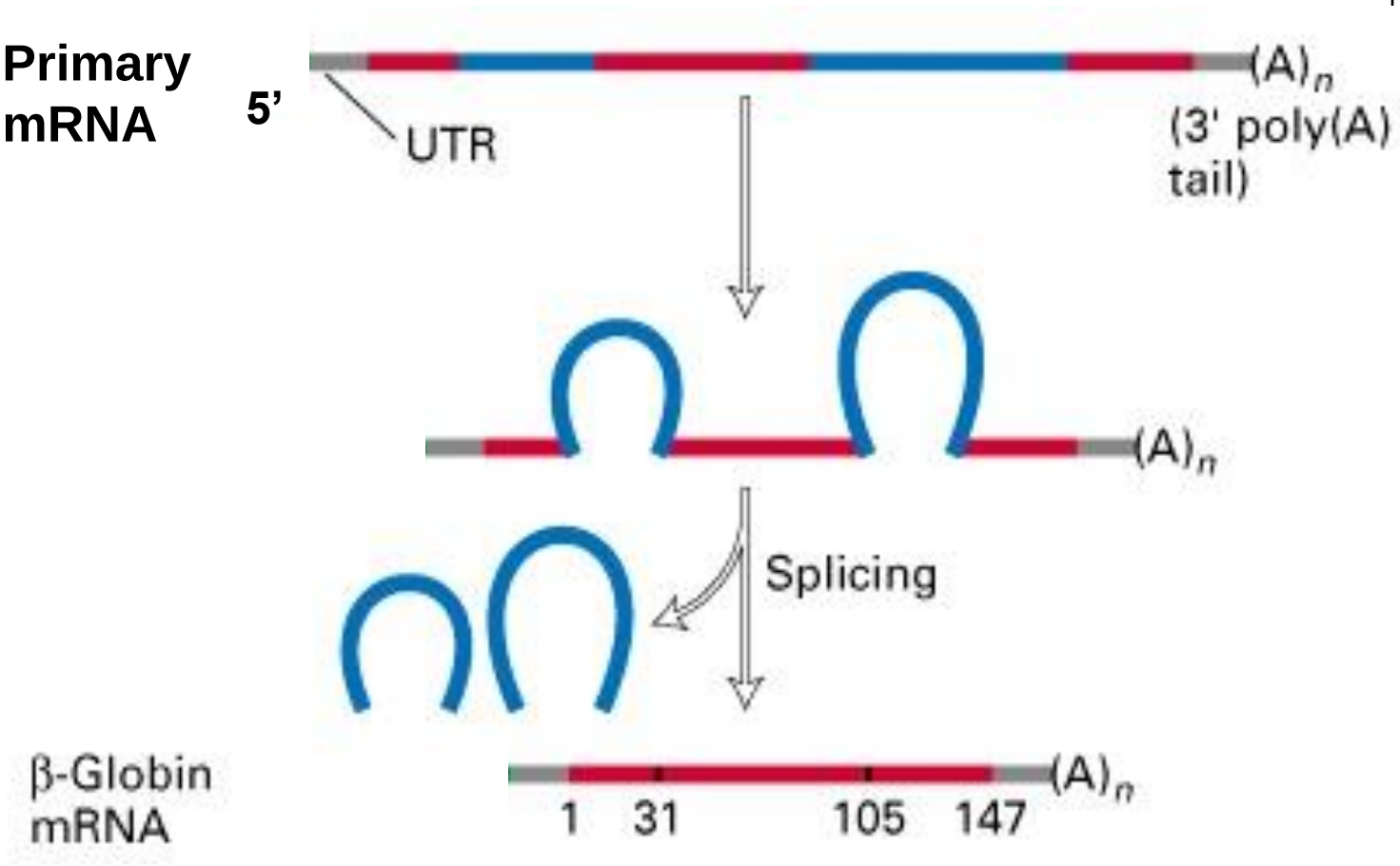
- **DNA contains genetic information.**
- **DNA can be expressed as a sequence of letters A,C,G and T.**

  Eg: ACGTTTCGAGGT

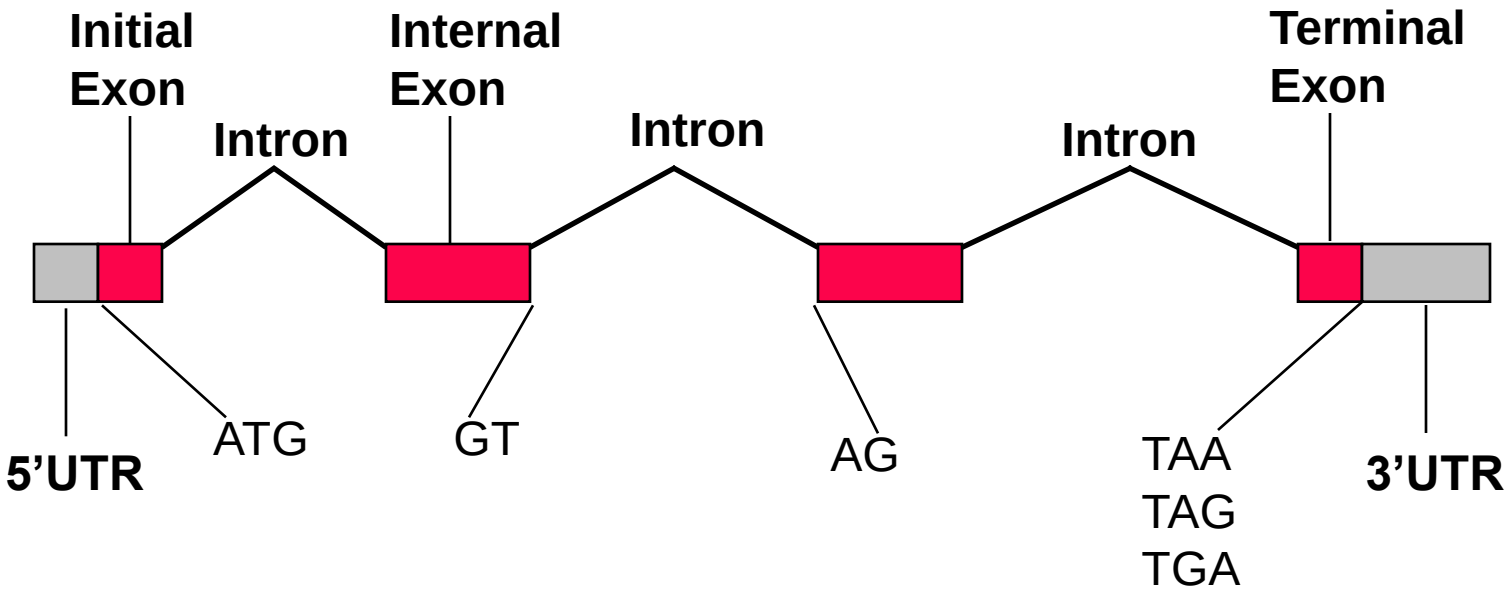
DNA→RNA→Protein



RNA Processing



Introduction: Gene Structure

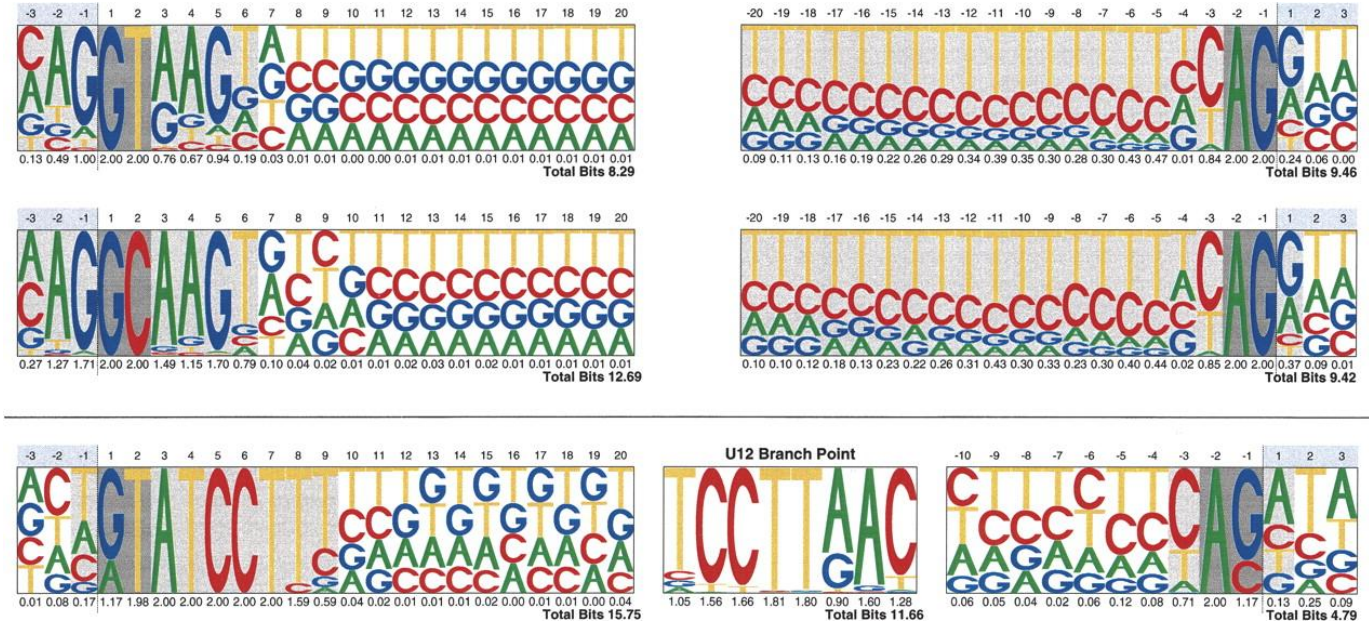


A gene is a highly structured region of DNA, it is a functional unit of inheritance.

Patterns in Splice Sites

Donor Sites

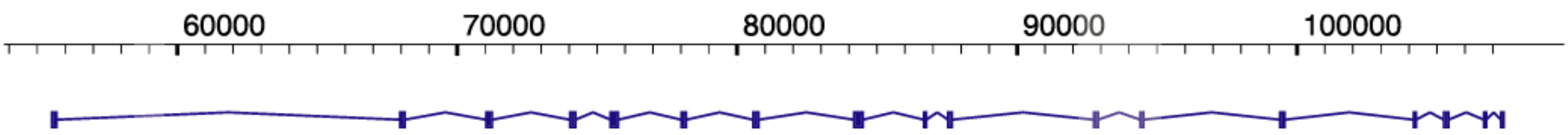
Acceptor Sites



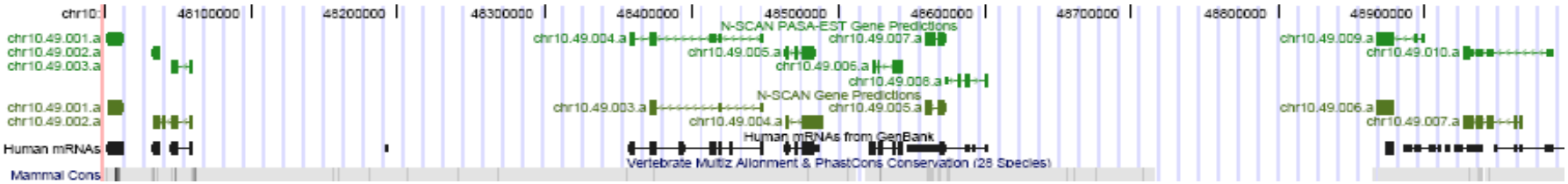
Josep F. Abril et al. Genome Res. 2005; 15: 111-119

Sequence data from RefSeq of human, mouse, rat and chicken.

A Typical Human Gene Structure



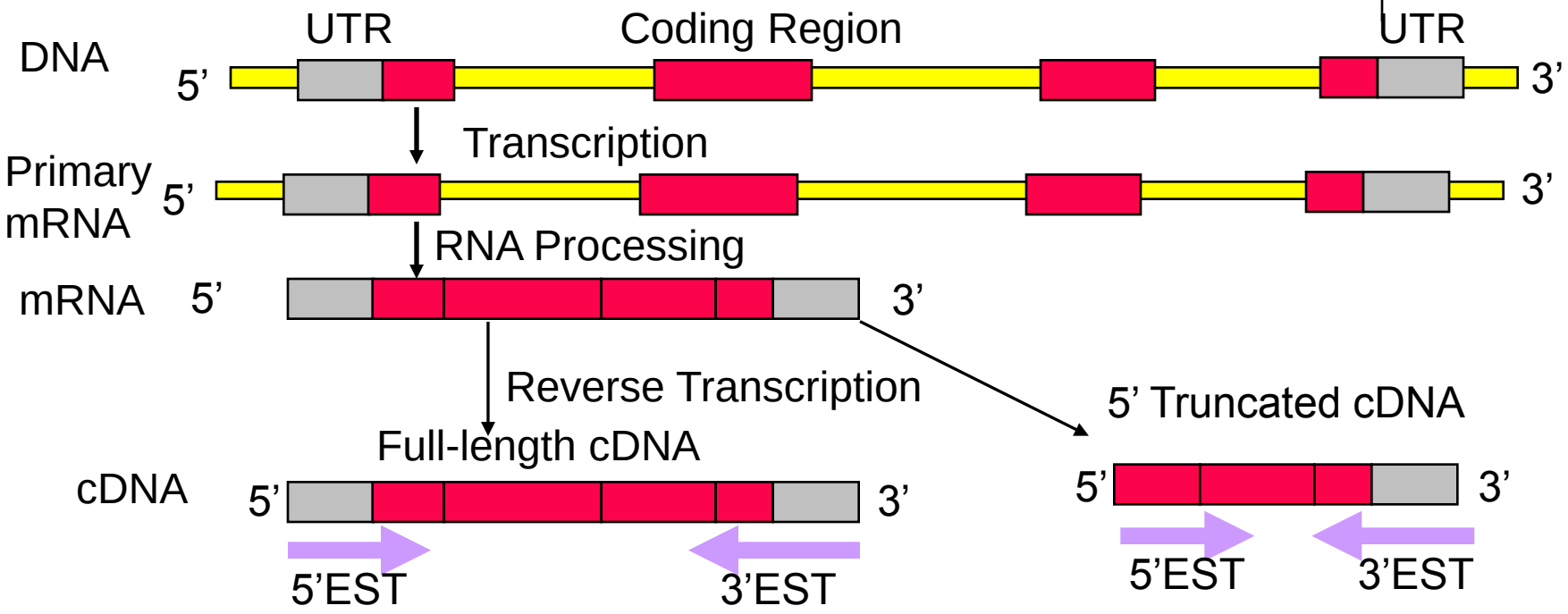
Genes in a Genome



In a Mammalian Genome

- Finding all the genes is hard
  - Mammalian genomes are large
    - 8,000 km of 10pt type
  - Only about 1% protein coding

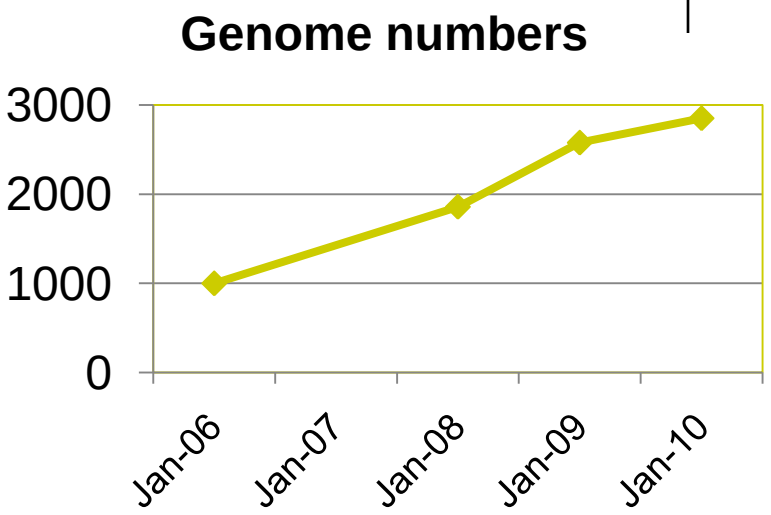
DNA, mRNA, cDNA and EST



- EST:
  - Short (~650bs)
  - High error rate (~1-5%)
  - Contains only UTRs or coding regions

The Challenge and Opportunity

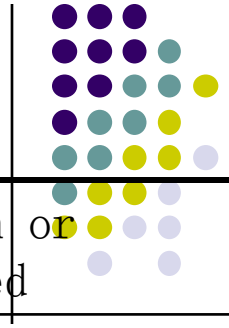
- ~3000 genomes
  - 222 animals
  - 93 plants



Better
Gene Structure
Annotation

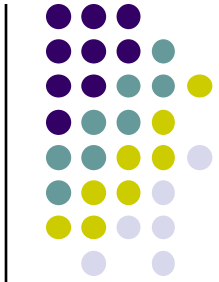
Numbers from http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html

2. Gene Prediction Method History



Generation	Date	Feature	Systems	Information or Methods used
1st	Early 1980s	Approximate ends of protein coding regions and non-coding regions	▪TestCode, Fickett 1982 ▪GRAIL, Uberbacher and Mural 1991	▪splice sites ▪promoters ▪Codon usage bias ▪Neuro-network methods
2nd	Early 1990s	A complete single gene in a short sequence	▪GeneID, Guigo et al. 1992 ▪GeneParser, Snyder and Stormo 1993 ▪FGENSH, Solovyev et al. 1994	+ ▪Translation start sites ▪Stop sites ▪Method: HMM
3rd	Mid-1990s	Multiple complete or partial genes in a long sequence	▪Genscan, Burge and Karlin, 1997	▪UTR ▪Method: Generalized HMM
4th	2000s	Complete gene structures in whole genomes	▪Twinscan, Korf, et al.2001 ▪N-SCAN, Brown, et al. 2006 ▪Twinscan_EST, N-SCAN_EST, Wei and Brent, 2006	▪Multiple-genomes ▪Transcript products ▪Method:Generalized HMM ▪Bayesian approach

Gene Prediction Methods (1)

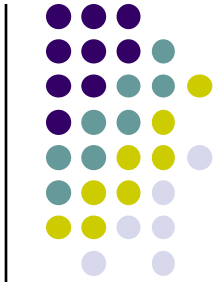


- Categorization: by input information
  1. Ab initio methods
     - Only need genomic sequences as input
       - GENSCAN (Burge 1997; Burge and Karlin 1997)
       - GeneFinder (Green, unpublished)
       - Fgenesh (Solovyev and Salamov 1997)
     - Can predict novel genes
  2. Transcript-alignment-based methods
     - Use cDNA, mRNA or Protein similarity as major clues
       - ENSEMBL (Birney, Clamp, et. al. 2004)
     - Highly accurate
     - Can only find genes with transcript evidences
       - cDNA coverage 50-60%
       - + EST coverage up to 85%

Gene Prediction Methods (2)

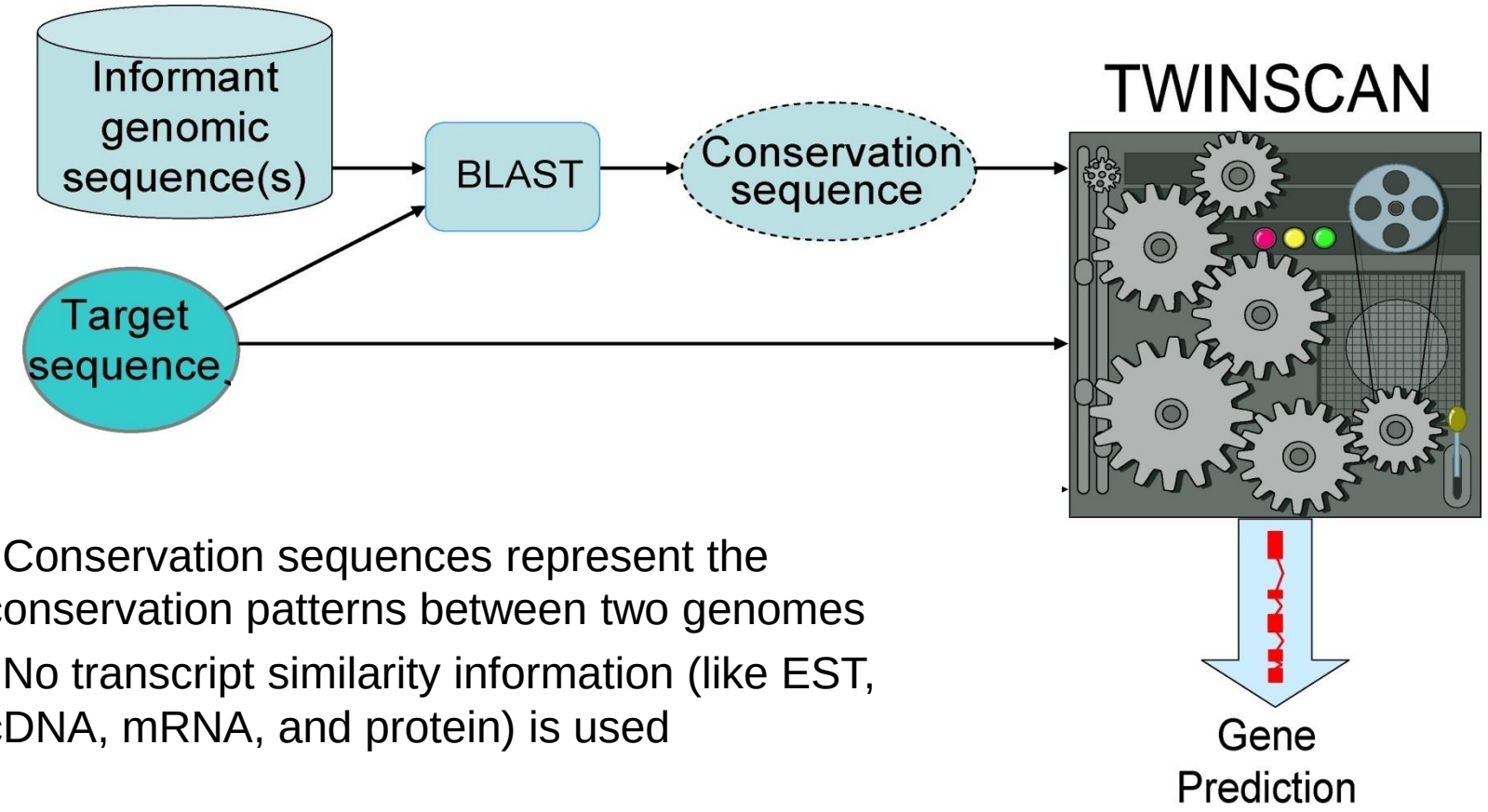
- Categorization: by input information
  3. Hybrid Methods
     - Integrate cDNA, mRNA, protein and EST alignments into ab initio methods
       - Genie (Kulp, Haussler et al. 1996)
       - Fgenesh+ (Solovyev and Salamov 1997)
       - Genomescan (Yeh, Lim et al. 2001)
       - GAZE (Howe, Chothiea et al. 2002)
       - AUGUSTUS+ (Stanke, Schoffmann et al. 2006)

Gene Prediction Methods(3)



- Comparative-Genomics-Based Methods
  - TWINSCAN and N-SCAN
    - De novo
    - Assumption:
      - Coding regions are more conserved.
    - No transcript similarity information (like EST, cDNA, mRNA, or protein) is used
  - TWINSCAN-EST and N-SCAN_EST
    - Hybrid
    - Use EST to improve prediction accuracy

TWINSCAN: A Novel Gene Prediction System Using Dual Genomes

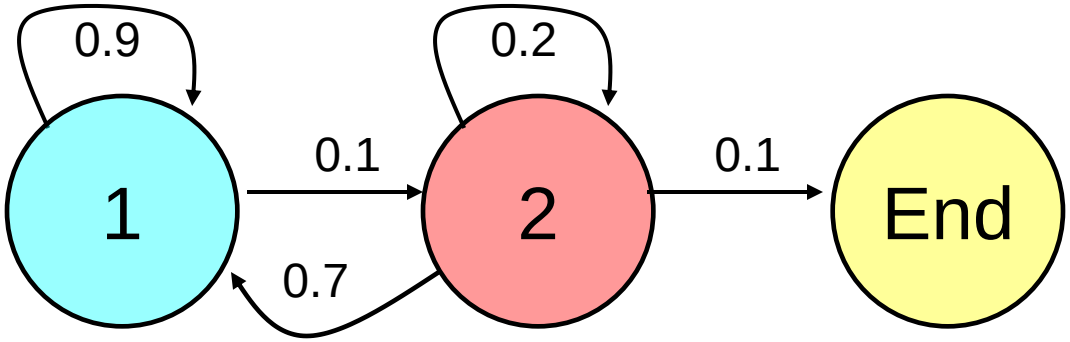


- Conservation sequences represent the conservation patterns between two genomes
- No transcript similarity information (like EST, cDNA, mRNA, and protein) is used

Hidden Markov Model:

Model behind gene predictors

HMM for two biased coins flipping



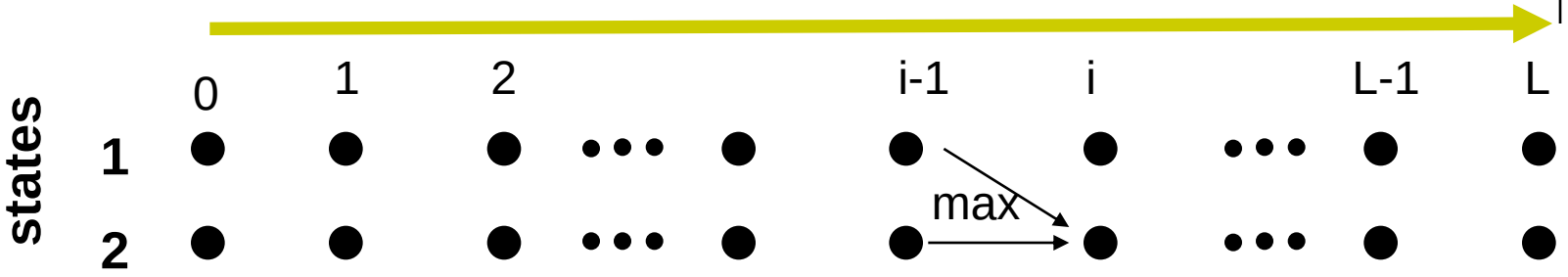
$$e_1(H) = 0.8, e_1(T) = 0.2, e_2(H) = 0.3, e_2(T) = 0.7$$

TTHHTTHTTTTTHTHHHHHTHTH **Observed sequence x**

1122111112222111112222 **Hidden state sequence π**

$$\pi^* = \arg\max_{\pi} P(x, \pi)$$

Most Probable Path and Viterbi Algorithm



Let $$f_l(i) = \max_{\{\pi_0,\ldots,\pi_{i-1}\}} (\Pr(x_0,\ldots,x_{i-1},\pi_0,\ldots,\pi_{i-1},\pi_i = l))$$

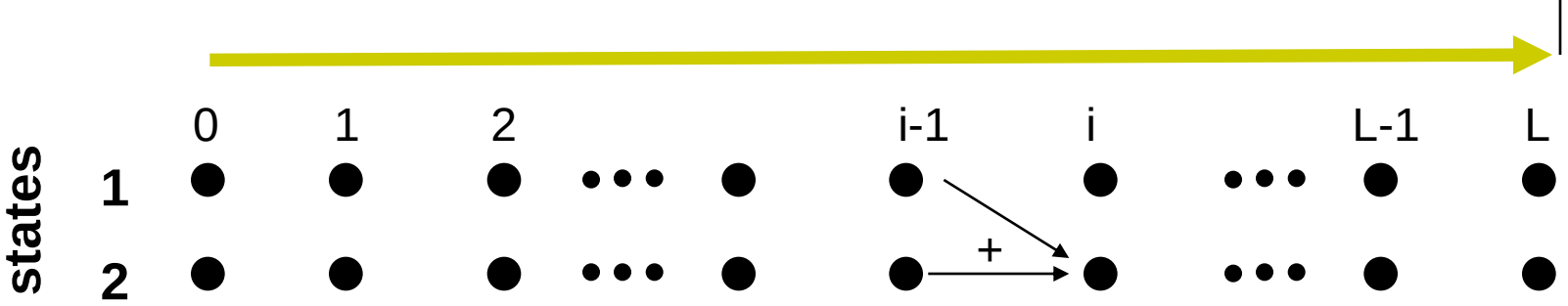
Recursion (i=1…L)

$$f_l(i) = e_l(x_i)\max_k(f_k(i-1)a_{kl});$$

$$ptr_i(l) = \arg\max_k(f_k(i-1)a_{kl}).$$

Time complexity $O(N^2L)$ space complexity $O(NL)$

Probability of All the Possible Paths and Forward Algorithm

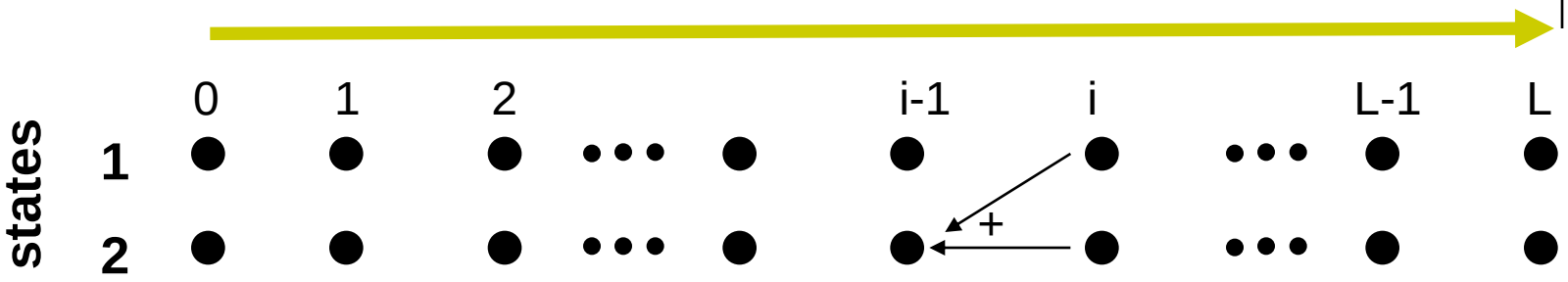


Let $$f_l(i) = \Pr(x_0, \ldots, x_{i-1}, \pi_i = l)$$

Recursion (i=1…L) $$f_l(i) = e_l(x_i) \sum_k (f_k(i-1) a_{kl})$$

Probability of all the probable paths $$P(x) = \sum_\pi P(x, \pi) = \sum_k f_k(L)$$

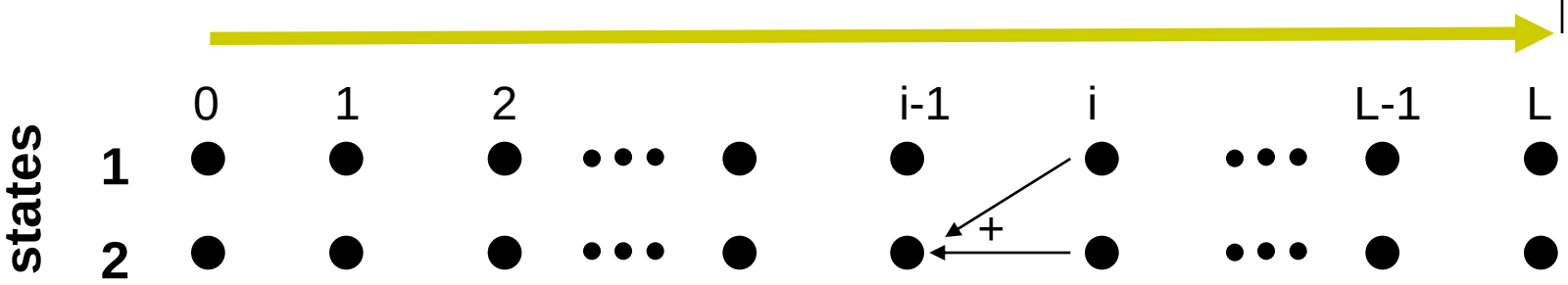
Posterior Probability and Forward and Backward Algorithm



Posterior Probability

$$P(\pi_i = k \mid x) = \frac{P(\pi_i = k, x)}{P(x)}$$

Posterior Probability and Forward and Backward Algorithm



Let $b_k(L) = \Pr(x_{i+1},\ldots,x_L \mid \pi_i = k)$

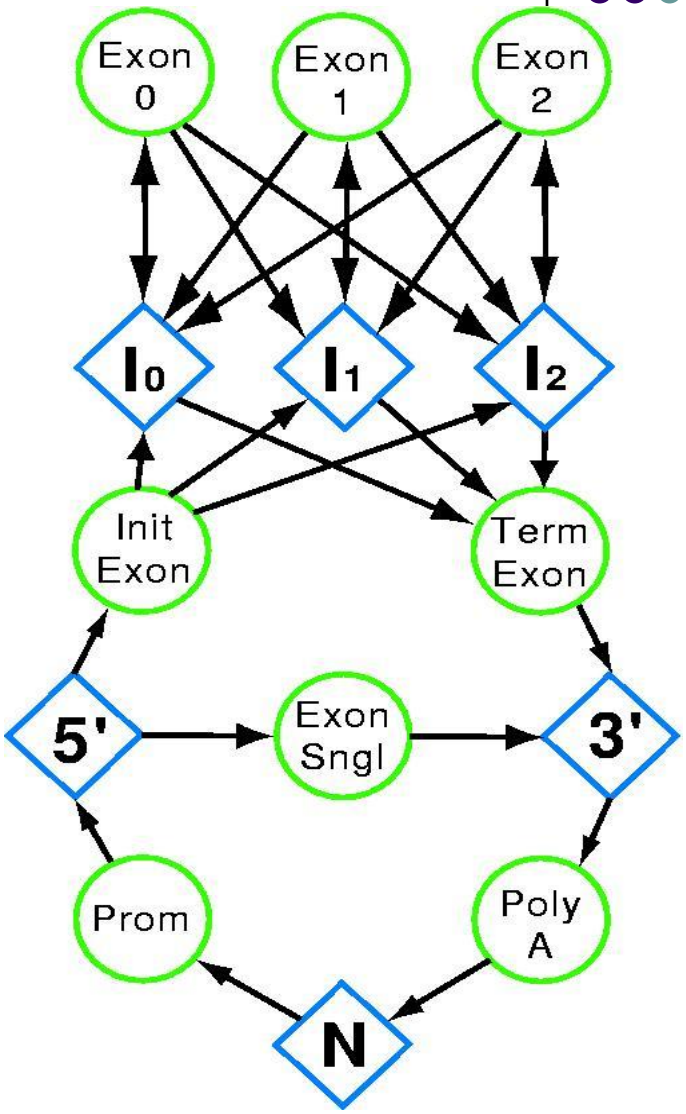
Recursion (i=L-1,…,1) $b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i+1)$

$$P(x) = \sum_\pi P(x,\pi) = \sum_l a_{0l} e_l(x_1) b_l(1)$$

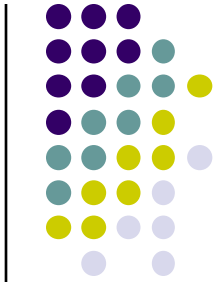
Posterior Probability $P(\pi_i = k \mid x) = \frac{P(\pi_i = k, x)}{P(x)} = \frac{f_k(i) b_k(i)}{P(x)}$

TWINSCAN Model

- **Generalized HMM**
- Each feature in a gene structure corresponds to one state.
- State-specific length models.
- State-specific sequence models
- Use Conservation information



Conservation Sequence



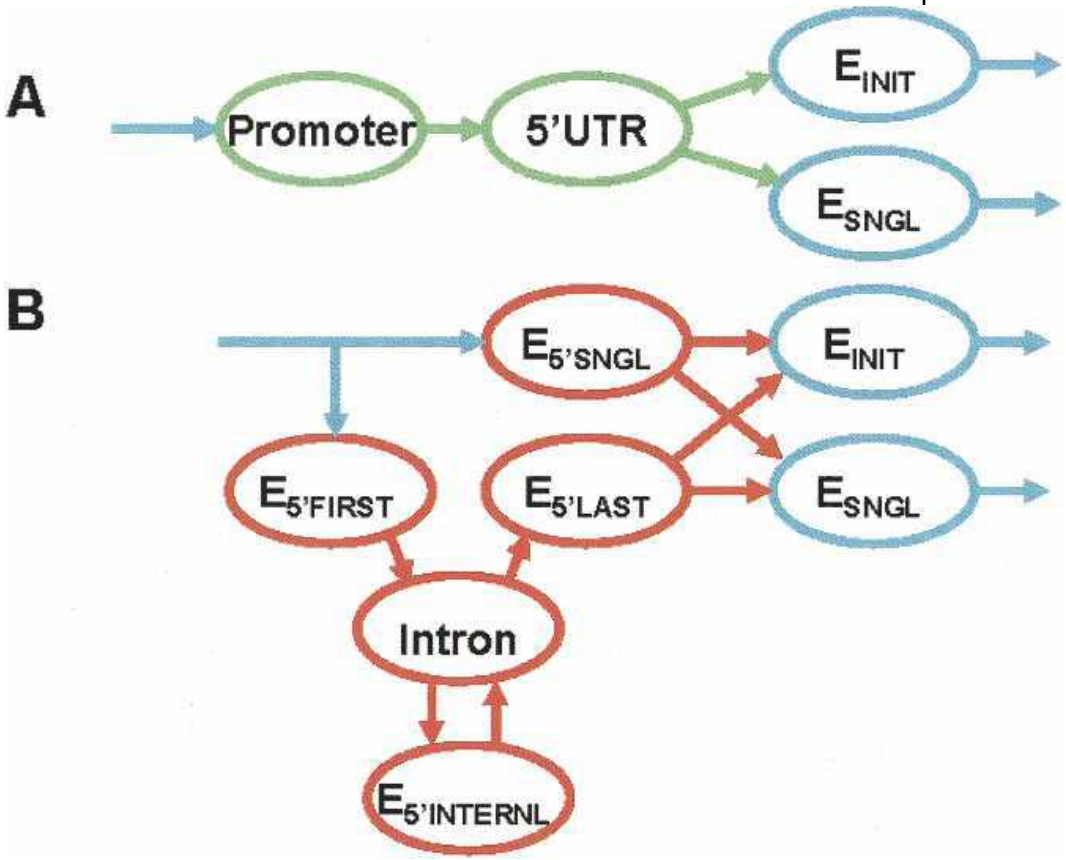
Generated by projecting local alignments to the target sequence

```
human CTAGAGATGCAAAAGAAACAGGTACCGCAGTGC---CCC
      ||||||.........|:|||||||||::||:||...||:
mouse CTAGAG---------AGACAGGTACCATAGGGCTCTCCT
```

- Pair each nucleotide of the target with
  - "|" if it is aligned and identical
  - ":" if it is aligned to mismatch
  - "." if it is unaligned

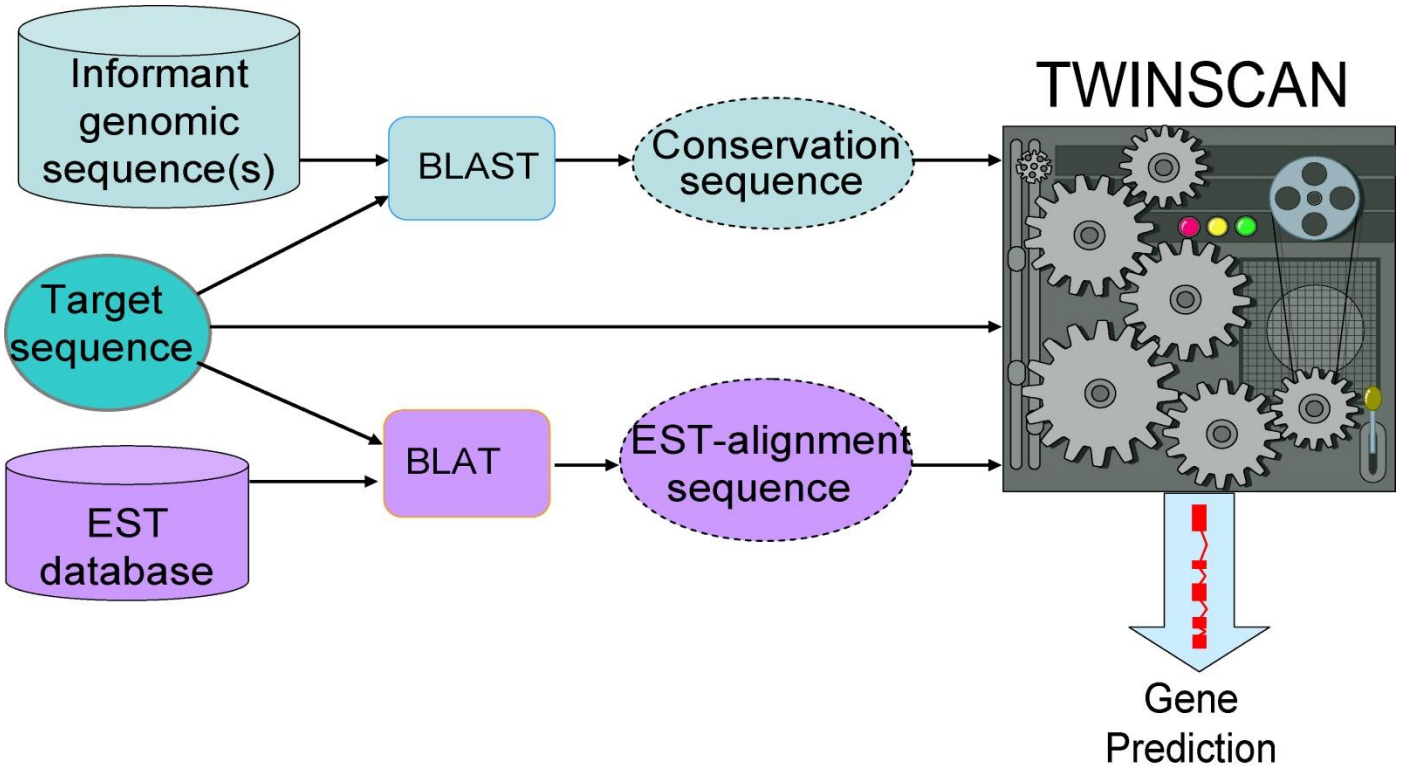
N-SCAN: A Novel Gene Prediction System Using Multiple Genomes

- Uses Bayesian model to include phylogenetic tree information
- Predicts introns in 5’UTR
- Has Conserved non-coding regions



(Brown, Gross and Brent, *Genome Res*. 2005)

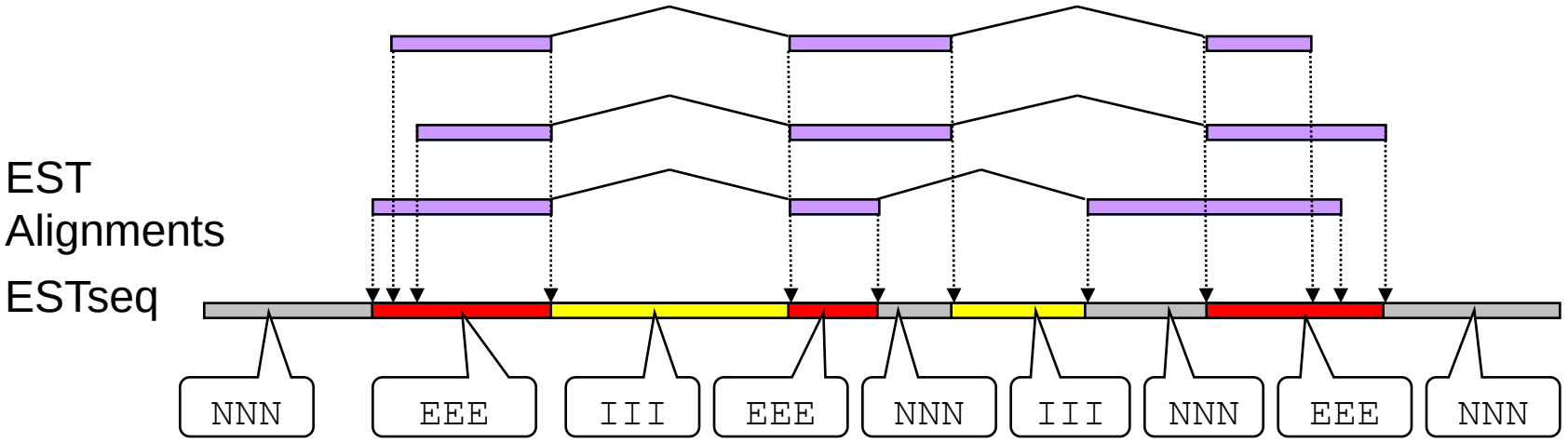
Using ESTs for Gene Prediction: TWINSCAN_EST



- Integrating EST alignment information into TWINSCAN to improve its accuracy where EST evidence exits and not to compromise its ability to predict novel genes.

Sequence Representation of EST Alignments

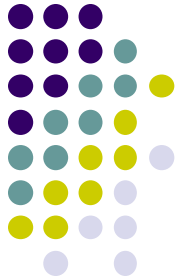
1. Use EST-to-genome alignment programs
    - BLAT (Kent 2002)
2. Project the top alignment for each EST to the target genomic sequence

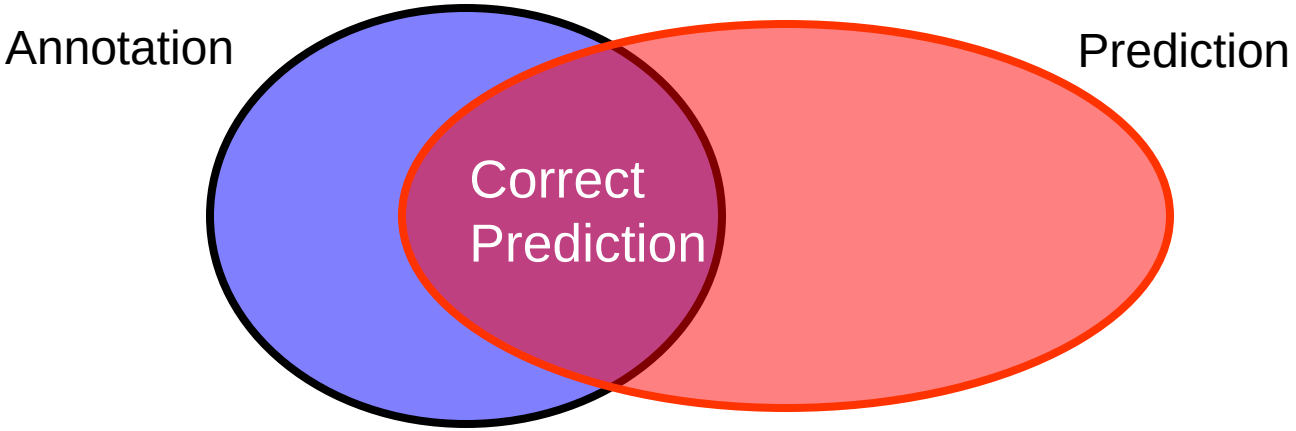


Accuracy Measurement

- Annotated data sets for training/testing
  - RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/)
  - CCDS (http://www.ncbi.nlm.nih.gov/CCDS/)
- Accuracy in different levels
  - Nucleotide level
  - Exon level
  - Gene level
  - Transcript level
- Sensitivity and specificity

Accuracy Measurement (continue)

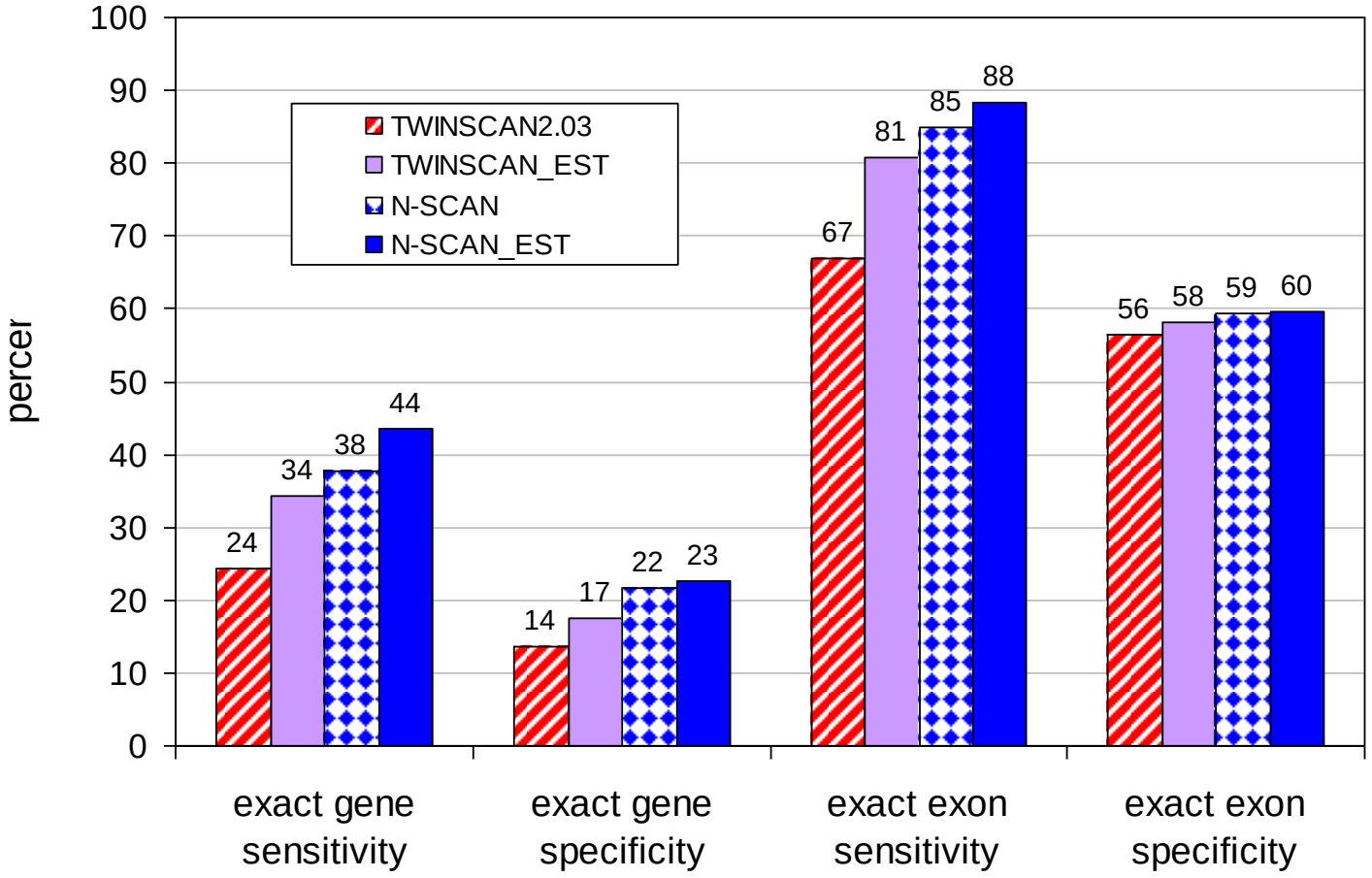




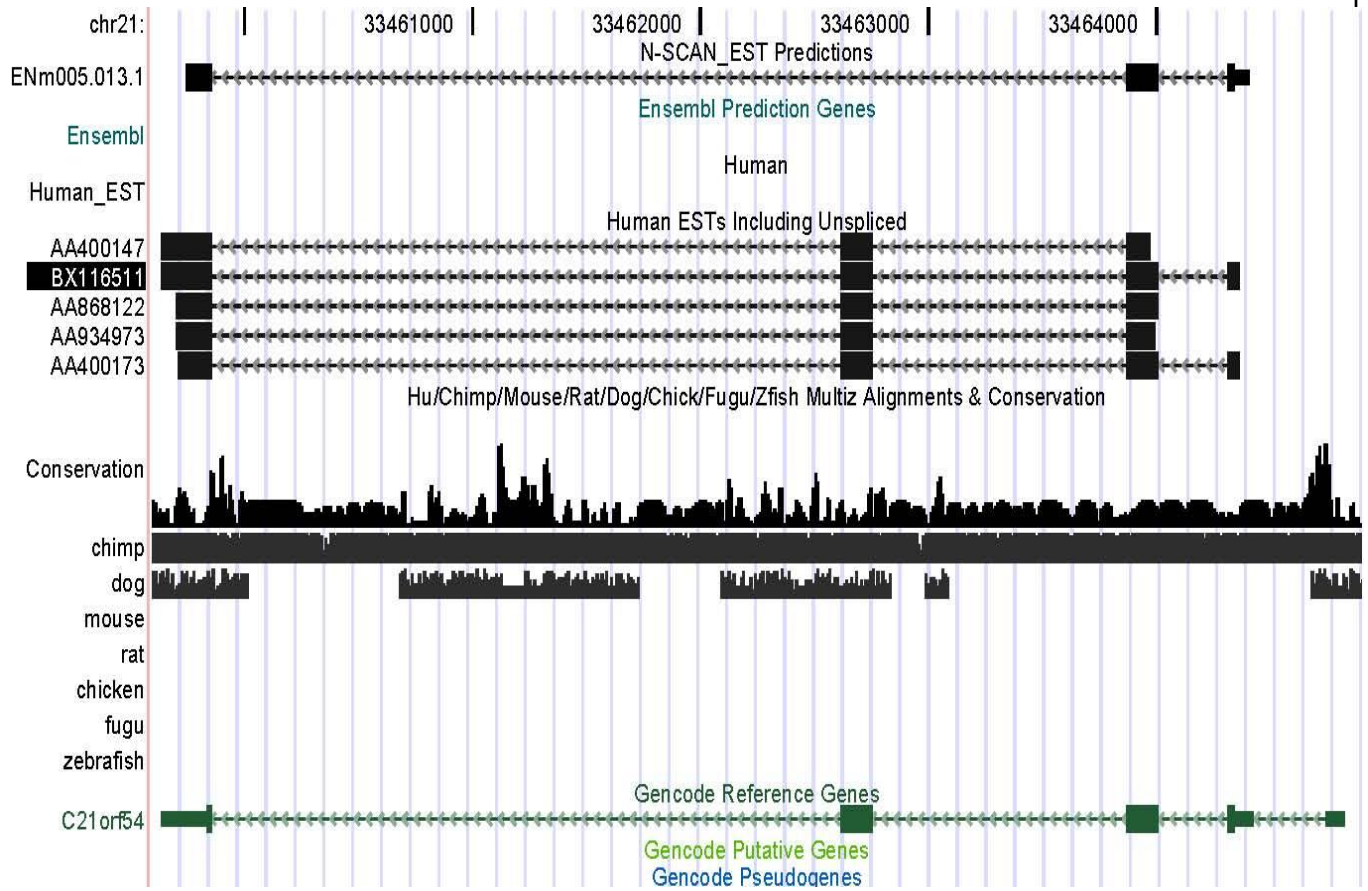
$$Sensitivity = \frac{Correct\_Prediction}{Total\_Annotation}$$

$$Specificity = \frac{Correct\_Prediction}{Total\_Prediction}$$

TWINSCAN_EST and N-SCAN_EST on the Whole Human Genome

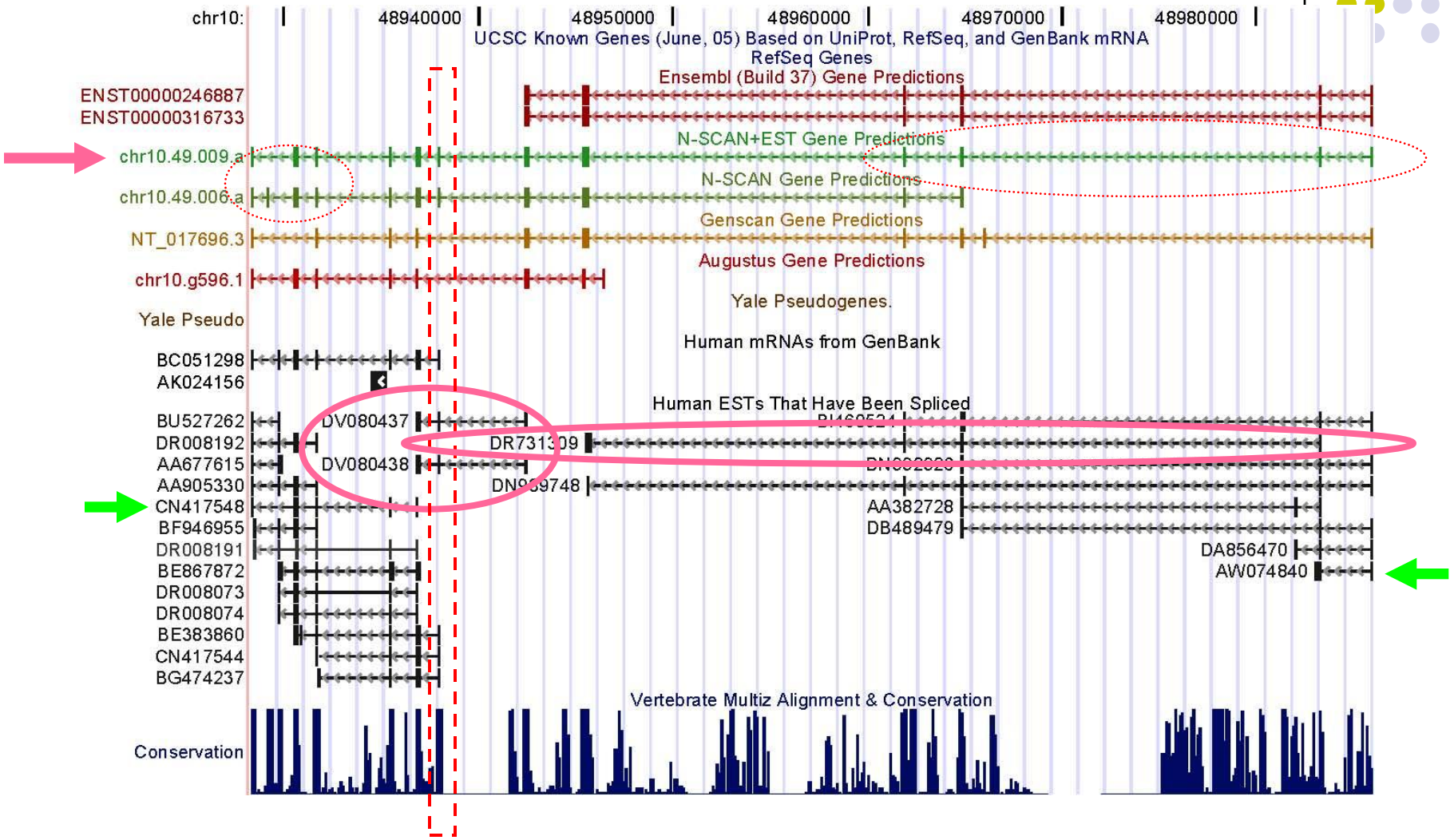


An Example of N-SCAN_EST Prediction

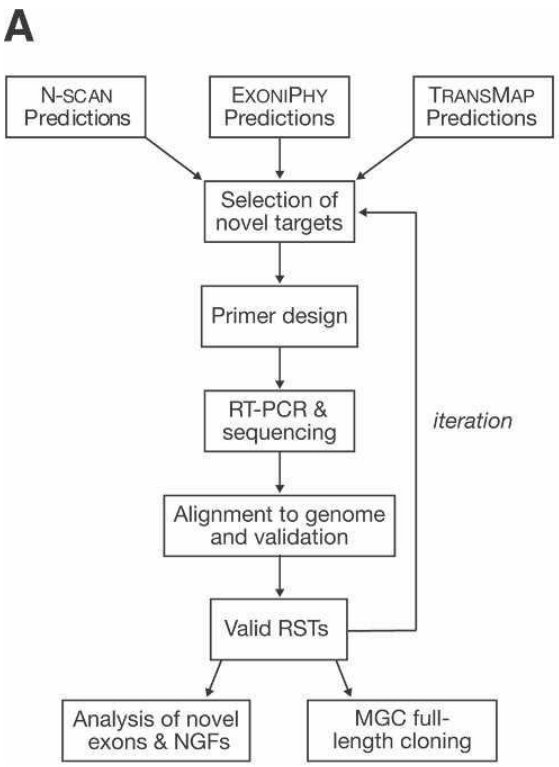


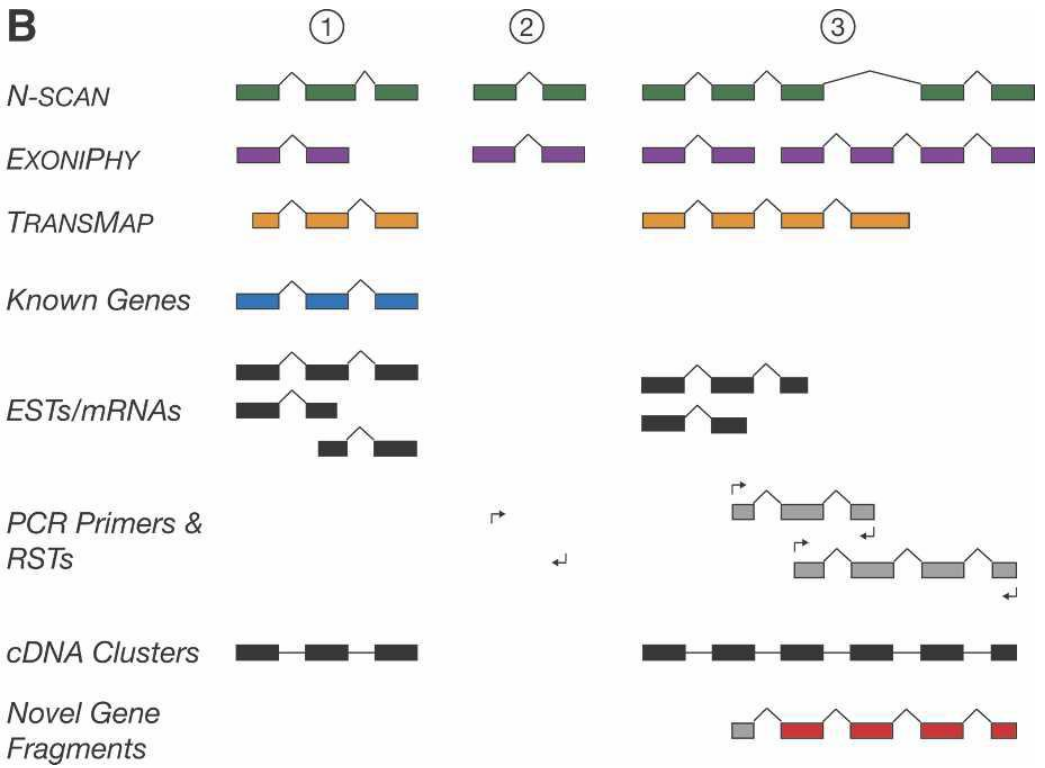
(Hg17, chr21:33,459,500-33,465,411)

An Example of N-SCAN_EST Prediction



Experimental Validation of Predictions



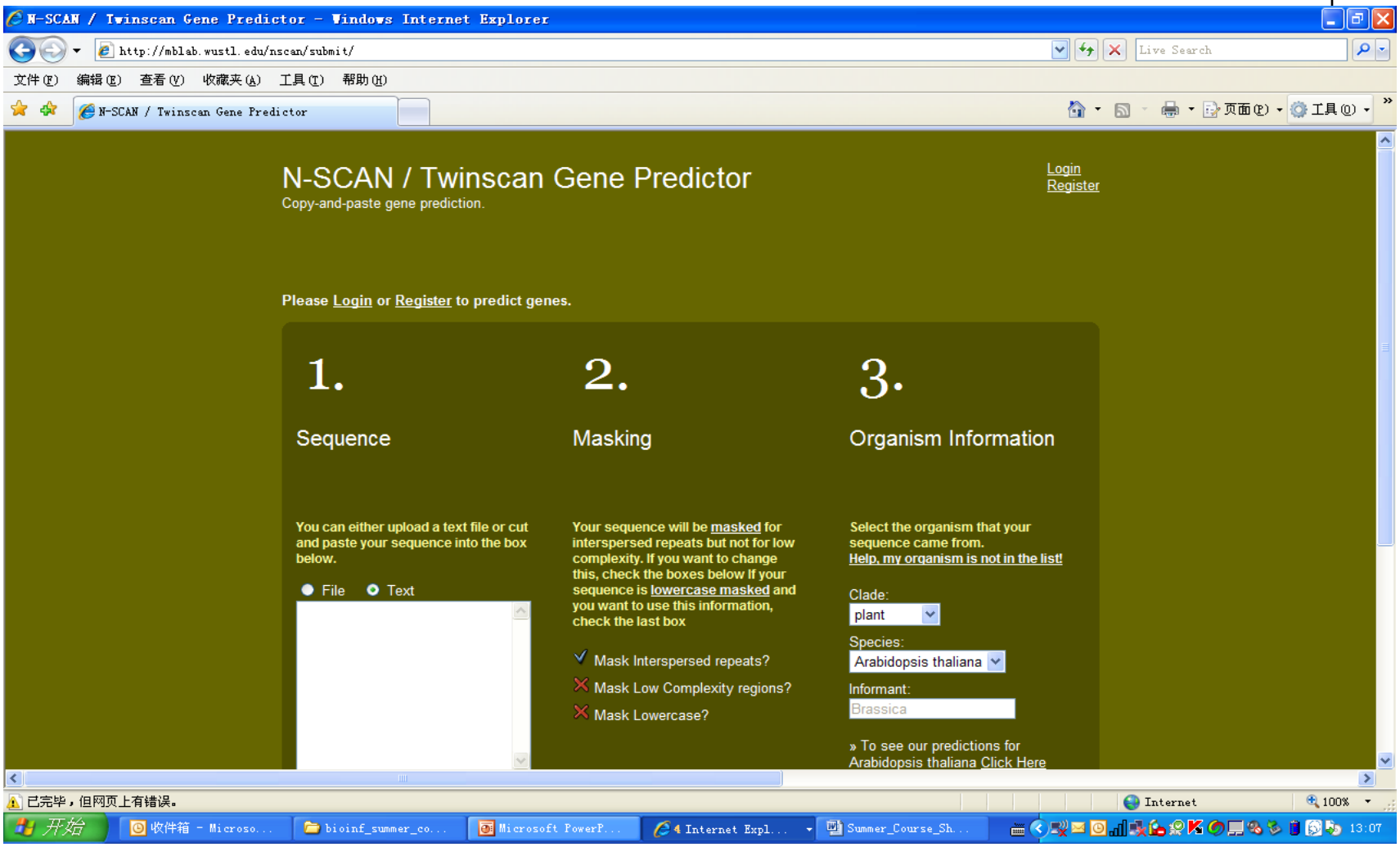


Siepel, *Genome Research*, 2007

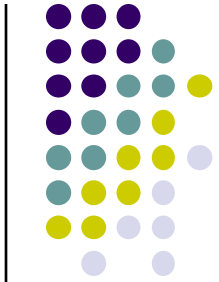
Experimental Validation of Predictions

- See
  - The MGC Project Team, "The Completion of the Mammalian Gene Collection (MGC)", Genome Research, 2009, 19:2324-2333
  - Wei, C., et al., "Closing in on the C.elegans ORFeome by Cloning TWINSCAN predictions", Genome Research, 2005, 15:577-582.
  - Tenney, A. E. et al., "Gene prediction and verification in a compact genome with numerous small introns", Genome Research, 2004

N-SCAN/TWINSCAN Webserver: http://mblab.wustl.edu/nscan/submit

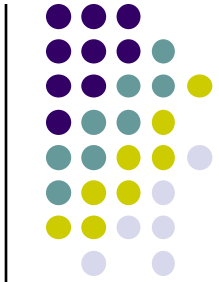


Resources for Gene Prediction



- Sequence data sets
  - Nucleotide Sequences (NCBI)
  - dbEST
  - mRNA
  - cDNA
- Annotations
  - RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/)
  - CCDS (http://www.ncbi.nlm.nih.gov/CCDS/)
- Genome Browser
  - UCSC Genome Browser(http://genome.ucsc.edu/)

Latest Progress in Gene Prediction



New Methods

- **Conrad: gene prediction using conditional random fields**. Decaprio et al., *Genome Res.* 2007 Sep;17(9):1389-98.
  - Not working for vertebrate genomes
- **SVM** for splice site

  Sonneburg et al., *BMC Bioinformatics*. 2007;8 Suppl 10:S7.
- **CONTRAST**: Gross et al., *Genome Biology* 2007, **8:**R269
  - Best **de novo** gene predictor for human (gene level accuracy **~50%)**
  - **Used SVM and conditional random field**

3. Gene Prediction FAQs (from Ian Korf)

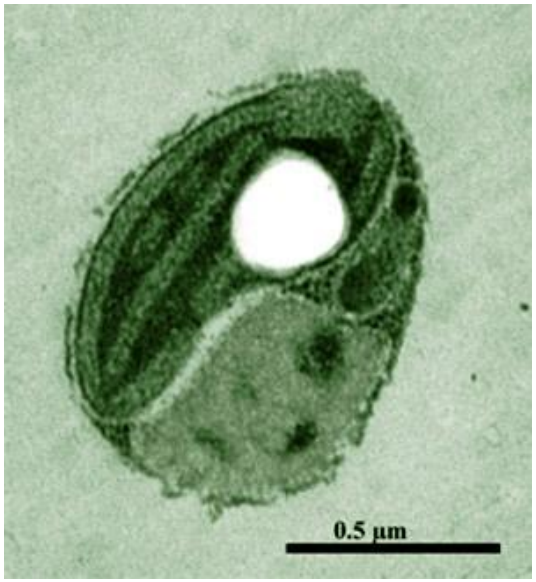
- Algorithms vs. experts
  - Q: are expert biologists better than computer programs?
  - A: Yes and no.
- Next-generation sequencing
  - Q: Will next-gen transcript sequencing replace gene prediction?
  - A: No. Rare transcripts may require directed experiments to validate.
- Prediction accuracy
  - Q: Why are gene prediction programs inaccurate?
  - A: We don’t always know why.

Gene Prediction FAQs (continue)

- Genes in my favorate genome…
  - Q: There is no gene predictor for it, what should I do?

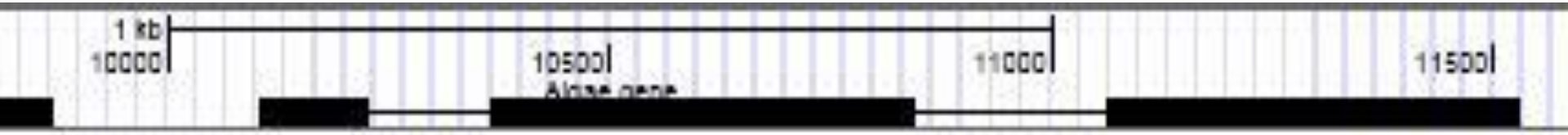
Gene prediction for algal genomes

Macroalgae

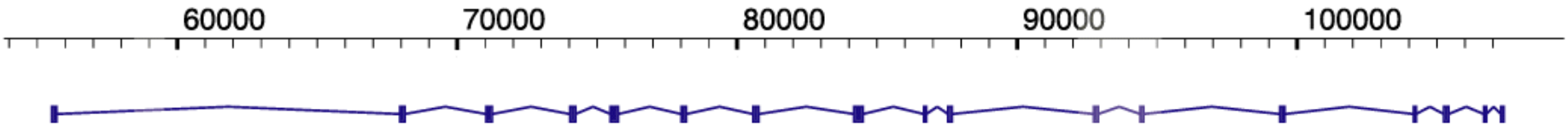


Microalgae

A Typical Alga Gene Structure



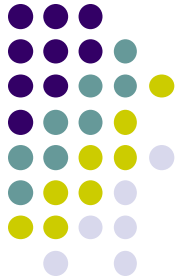
A Typical Human Gene Structure



Gene Prediction FAQs (continue)

- Genes in my favorate genome…
  - Q: There is no gene predictor for it, what should I do?
  - A: Training a gene predictor or use one that is for another organism that is close to this genome. But it may be inaccurate.
- Difficult genes
  - Q: why some genes are not predicted by any program?
  - A: They are statistical outliers.

Gene Prediction FAQs (continue)



- Just coding exons…
  - Q: why other parts are not predicted, such as non-coding exons, alternative isoforms, non-canonical splice sites, gene within genes?
  - A: There are trade offs.
- Pseudogenes
  - Q: why do some gene predictions have tiny introns?
  - A: Retro-pseudo genes often have very strong coding signals, because they are derived from highly expressed genes.

Gene Prediction FAQs (end)

- How can I tell a good gene prediction from a bad one?
- Scores have been assigned to every exon and intron of a gene. People can tell if a gene prediction is good or not by the scores of exons and introns of this gene.

  You may have to run the program on your own computer to figure them out!

Acknowledgement

- Training program organizers
- Former lab members (WashU)
  - Dr. Michael Brent
  - Dr. Paul Flicek
  - Dr. Ian Korf
  - Samuel Gross