Finding functional elements in genomes with statistical models

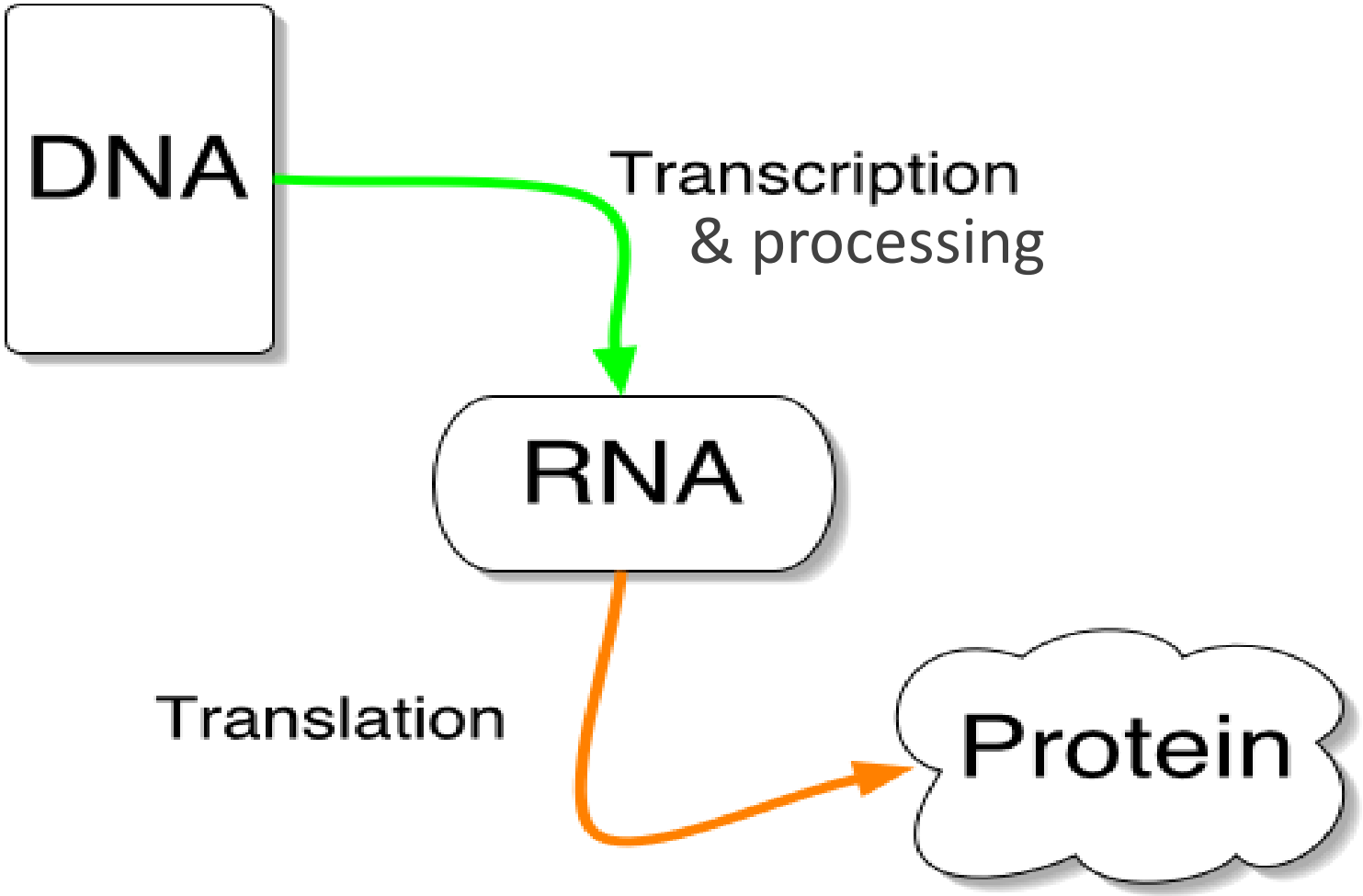
Chaochun Wei （韦朝春）
Department of Bioinformatics and Biostatistics
Shanghai Jiao Tong University

2013.6.17

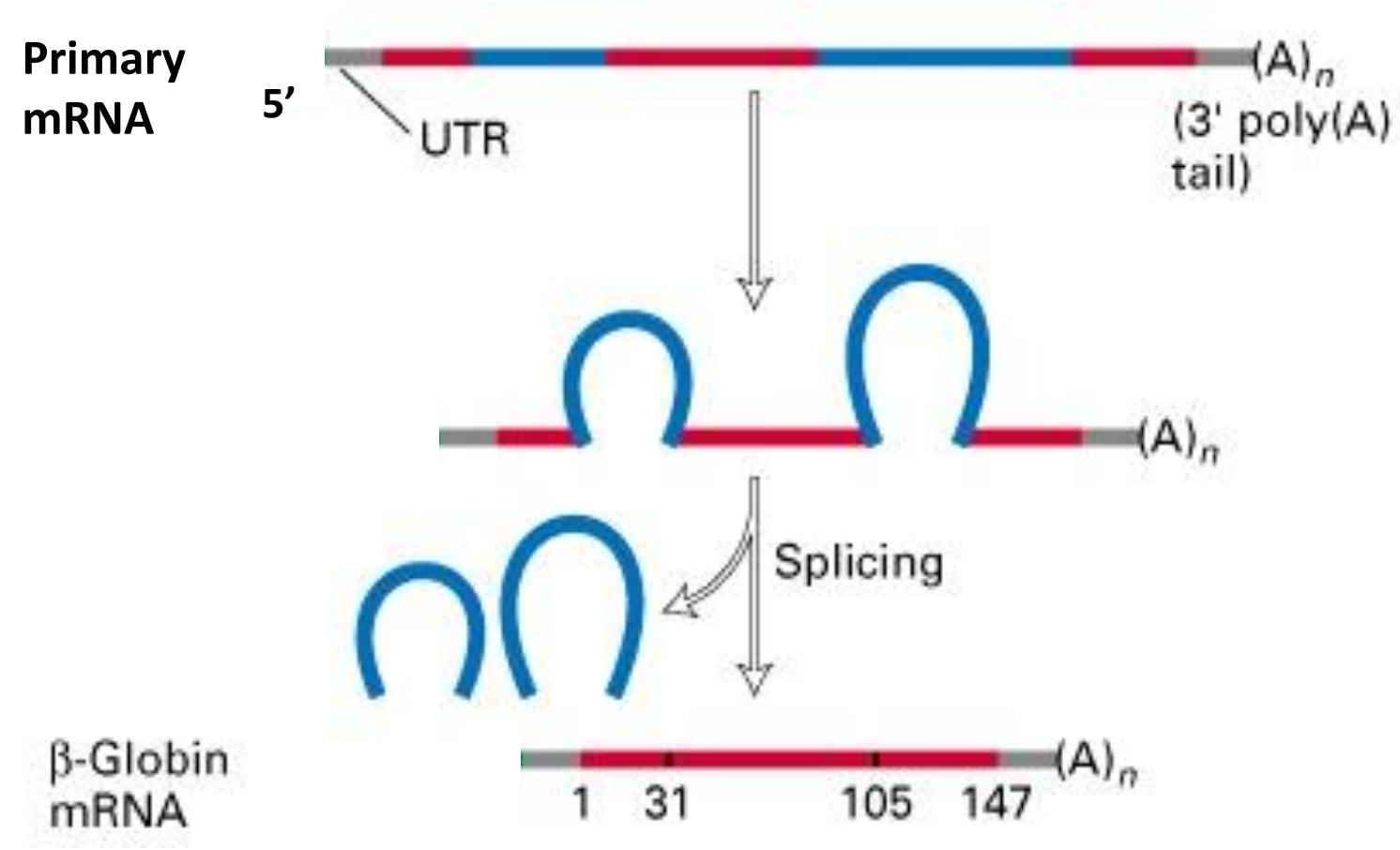
Contents

- Background
  - Functional elements in genomes
  - Statistical models
    - Hidden Markov Model (HMM)
    - Conditional Random Field (CRF)
- Finding functional elements in genomes
  - Gene structure prediction
  - Transcription factor binding site finding
  - Protein acetylation site prediction
- Summary

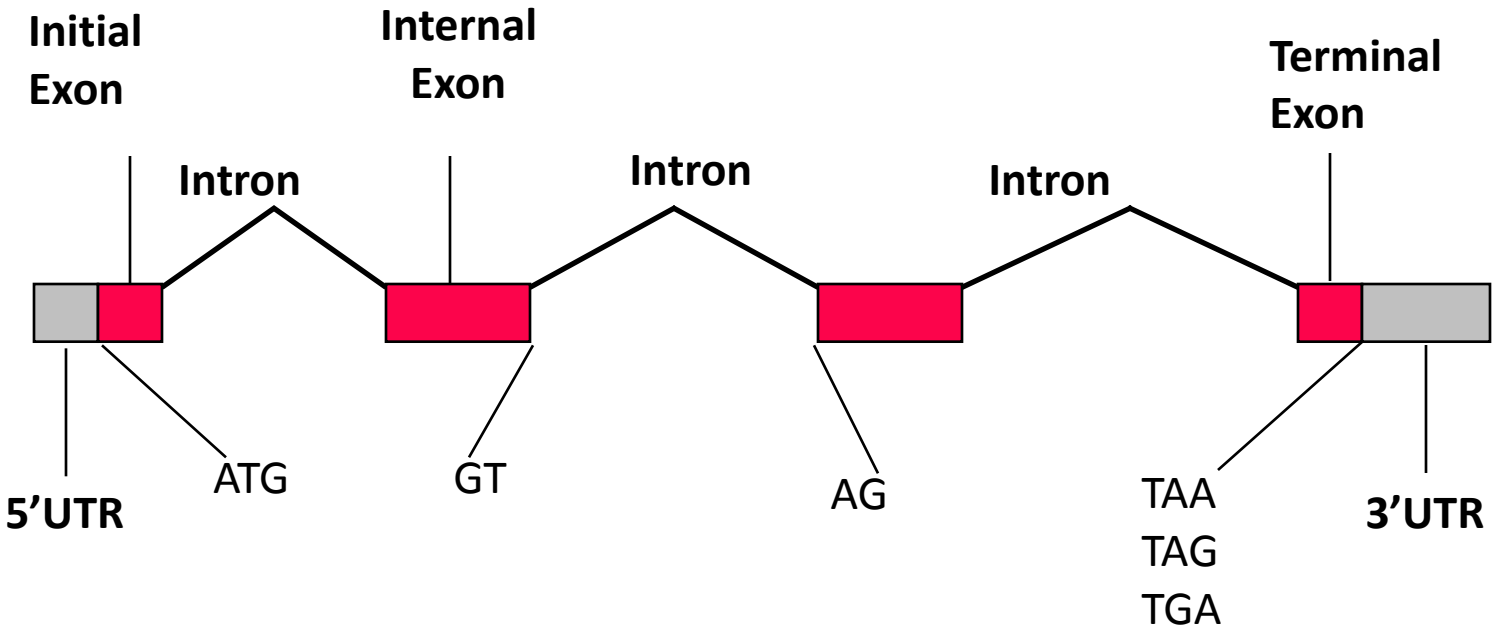
DNA→RNA→Protein



RNA Processing



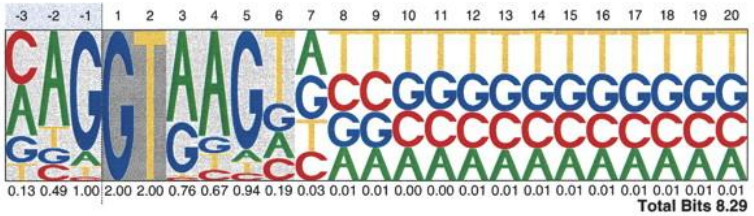
Gene Structure

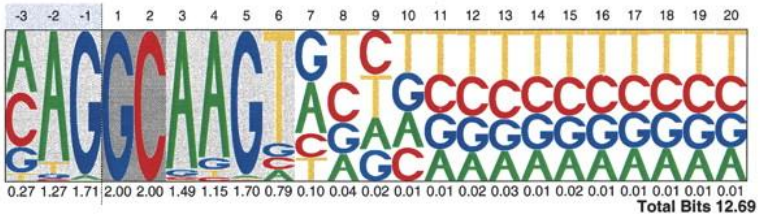


A gene is a highly structured region of DNA, it is a functional unit of inheritance.

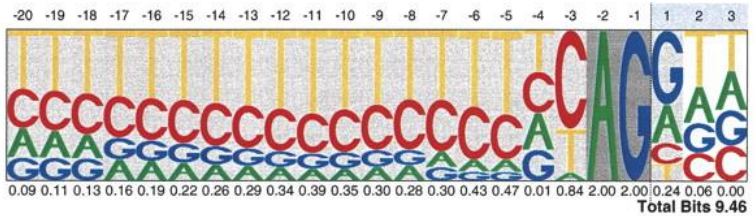
Patterns in Splice Sites

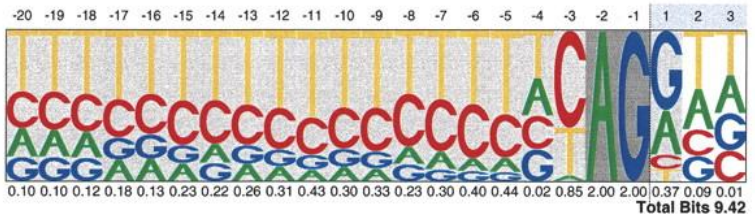
Donor Sites

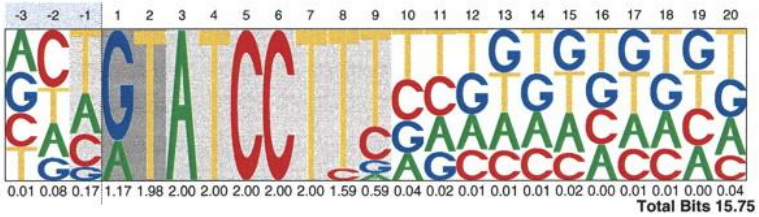


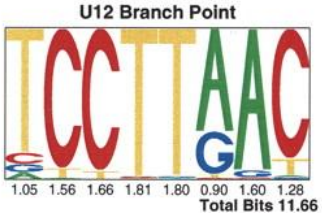


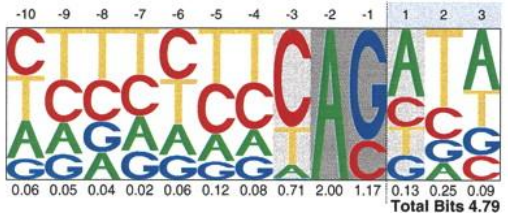
Acceptor Sites







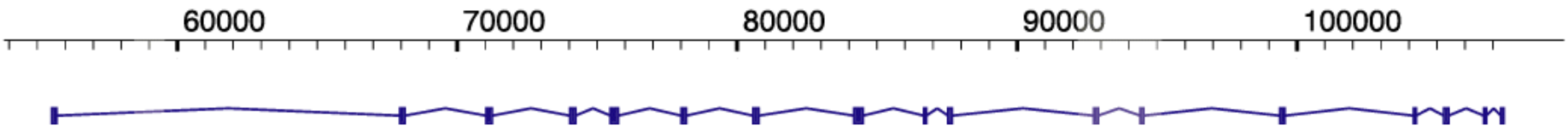




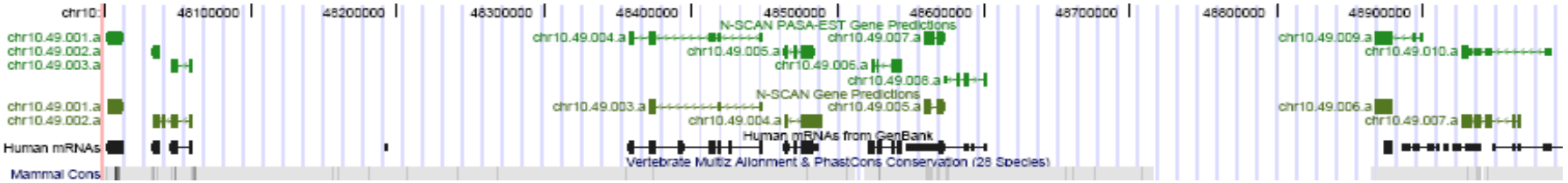
Josep F. Abril et al. Genome Res. 2005; 15: 111-119

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

A Typical Human Gene Structure

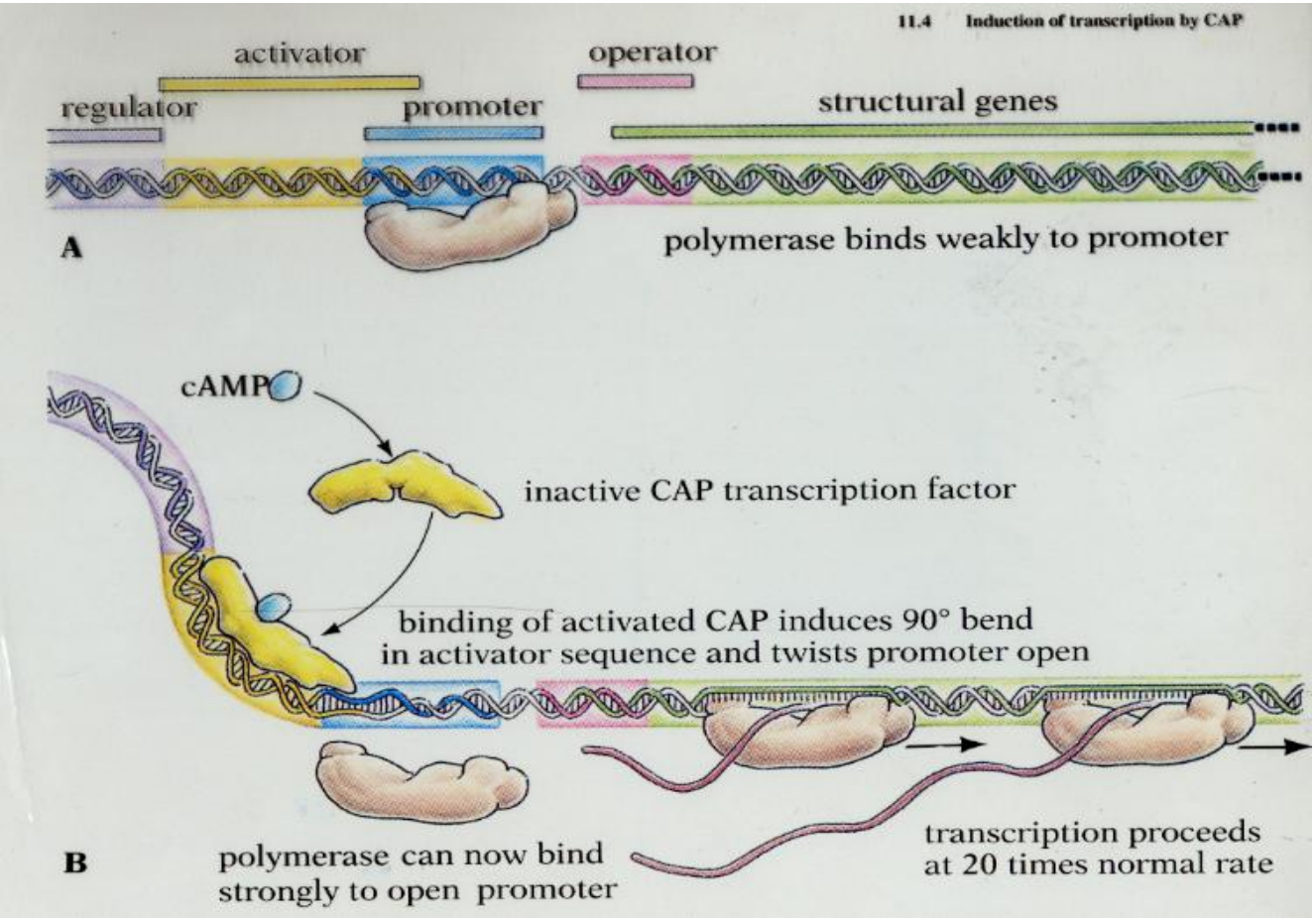


Genes in a Genome



Functional elements in genomes

- Genes
  - Protein coding genes
  - Non-coding genes
- Regulatory elements
  - Regulators
  - TFBSs
  - Promoters
  - Enhancers



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

- Finding all functional elements in genomes is still in its infant stage

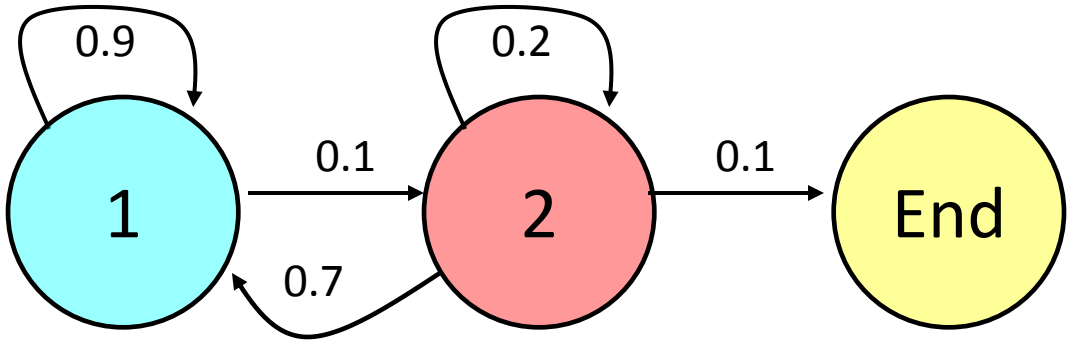
Statistical Models

- Hidden Markov Model (HMM)
- Conditional Random Field (CRF)

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**

11221111122221111112222 **Hidden state sequence** π

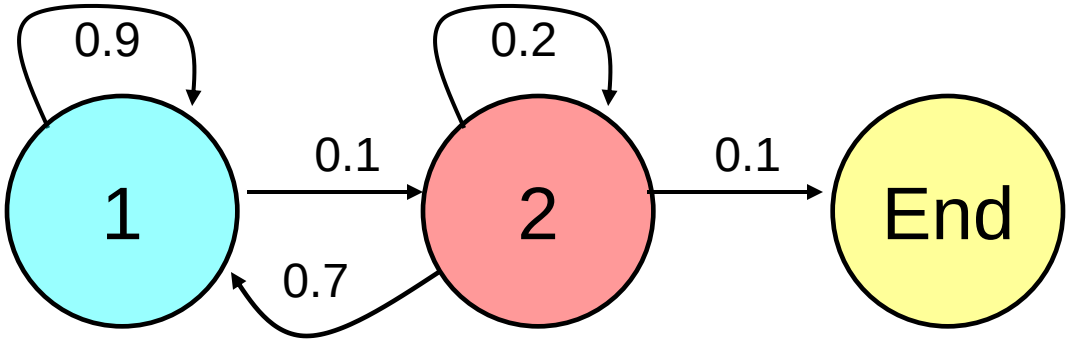
$$\pi^* = \arg\max_{\pi} P(x, \pi)$$

Hidden Markov Model

- Elements of an HMM (N, M, A, B, Init)
  1. N: number of states in the model
     - $S=\{S_1, S_2, \ldots, S_N\}$, and the state at time t is q_t.
  2. M: alphabet size (the number of observation symbols)
     - $V=\{v_1, v_2, \ldots, v_M\}$
  3. A: state transition probability distribution
     - $A=\{a_{ij}\}$ where $a_{ij}=P[q_{t+1}=S_j \mid q_t=S_i]$, $1 \le i,j \le N$
  4. E: emission probability
     - $E=\{e_j(k)\}$ (observation symbols probability distribution in state j), where $e_j(k)=P[v_k \text{ at } t \mid q_t = S_j]$, $1 \le j \le N$, $1 \le k \le M$
  5. Init: initial state probability, π_i
     - Init=$\{\pi_i\}$, whereπ_i=$P[q_1=S_i]$, $1 \le i \le N$.

HMM is a generative model

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**

$$P(x, \pi \mid \lambda) = Init_{\pi_0} * e_{\pi_0}(x(0)) * \prod_{0 \le i \le T} (a_{\pi_i \pi_{i+1}} e_{\pi_{i+1}}(x(i))$$

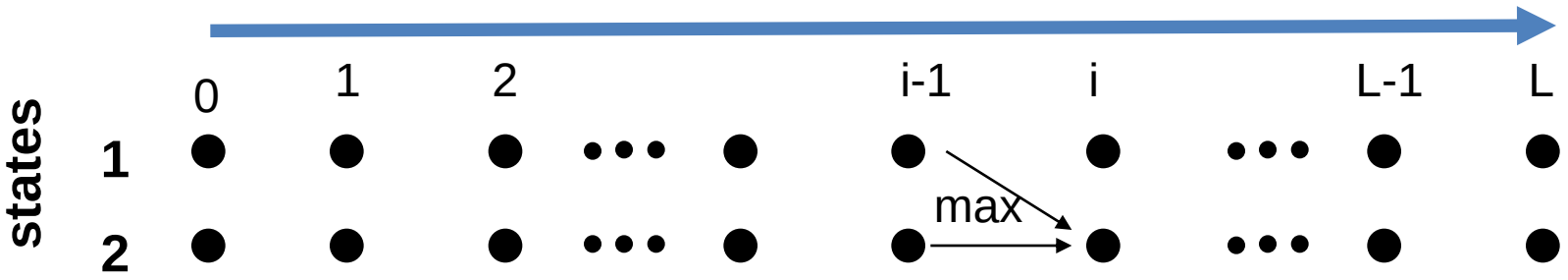
Hidden Markov Model

- HMM： λ={A, B,Init}
- Three basic problems for HMMs
  - Problem 1: From the observation $O=O_1O_2\ldots O_T$, and a model λ={A, B, Init}, how to compute P(O| λ)?
  - Problem 2: From the observation $O=O_1O_2\ldots O_T$, and a model λ={A, B, Init}, how to choose a state sequence π^*, so that

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

  - Problem 3: how to estimate model parameters λ={A, B, Init} to maximize P(O| λ).

Most Probable Path and Viterbi Algorithm



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

Initialization (j=1…N) $f_j(0) = \pi_j e_j(x_0)$

Recursion (i=1…L)

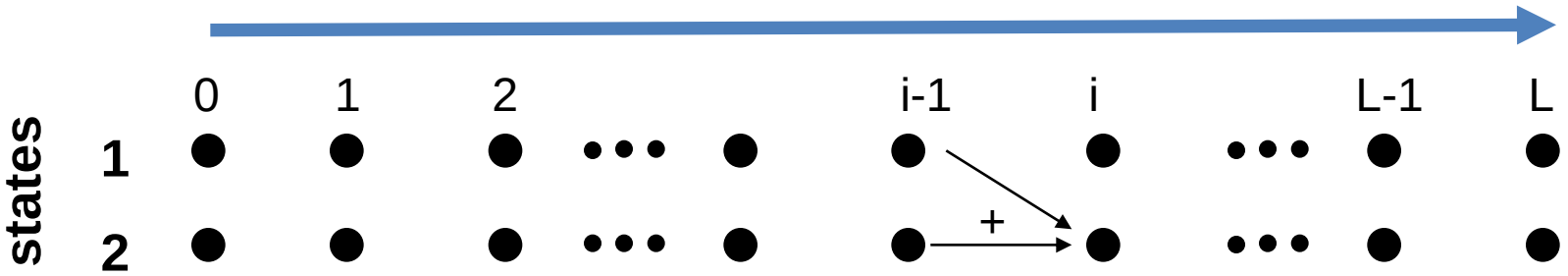
$$f_j(i) = e_j(x_i)\max_k(f_k(i-1)a_{kj});$$

$$ptr_j(i) = \arg\max_k(f_k(i-1)a_{kj}).$$

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

Solution to problem 2

Probability of All the Possible Paths and Forward Algorithm



Let $f_j(i) = \Pr(x_0, ..., x_i, \pi_i = j)$

Initialization (j=1…N) $f_j(0) = \pi_j e_j(x_0)$

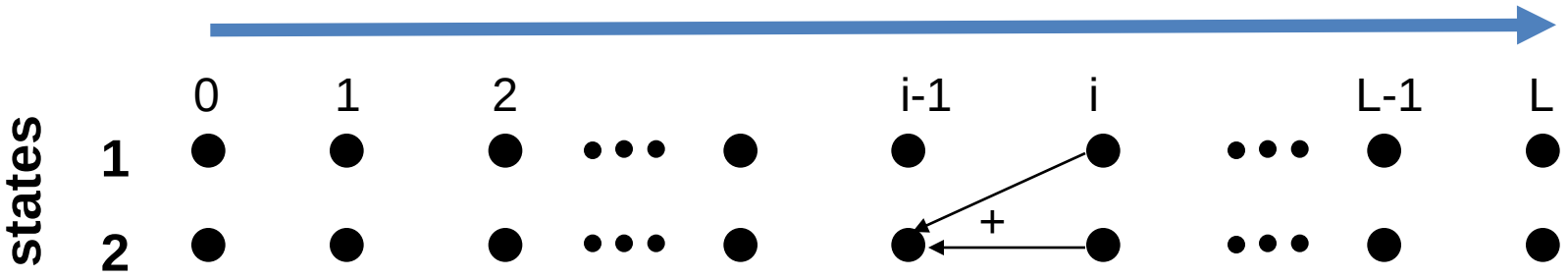
Recursion (i=1…L; $f_j(i) = e_j(x_i) \sum_k (f_k(i-1) a_{kj})$

j = 1, .., N)

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

Solution to problem 1

Backward Algorithm



Let $b_j(i) = \Pr(x_{i+1}, x_{i+2}..., x_L, \pi_i = j)$

Initialization (j=1…N) $b_j(L) = 1$

Recursion (i=L-1, L-2, …, 0, j=1, …, N) $b_j(i) = \sum_k (a_{jk} e_k(x_{i+1})) b_k(i+1)$

Probability of all the probable paths $P(x) = \sum_\pi P(x, \pi) = \sum_k b_k(0)$

Problem 3: Optimize the model parameters from the observation

- HMM： λ={A, B, Init}
- With annotations
  - Maximum likely-hood ratio
- Without annotations
  - Baum-Welch algorithm (EM algorithm)

Baum-Welch method (EM method)

- HMM： λ={A, B, Init}, Without annotations

Let $\xi_t(i, j) = P(\pi_t = i, \pi_{t+1} = j \mid x, \lambda)$

then $\xi_t(i, j) = \dfrac{f_i(t) a_{ij} e_j(x_{t+1}) b_j(t+1)}{\sum_{i}^{N} \sum_{j}^{N} (f_i(t) a_{ij} e_j(x_{t+1}) b_j(t+1))}$

Let $\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j)$

then $\sum_{t=0}^{L} \gamma_t(i)$ = expected number of transitions from S_i

$\sum_{t=0}^{L} \xi_t(i, j)$ = expected number of transition S_i to S_j

Baum-Welch method (EM method) (2)

- HMM：λ={A, B, Init}, Without annotations

Then, $\overline{Init}_i =$ expected frequency in S_i at time 0 = $\gamma_0(i)$

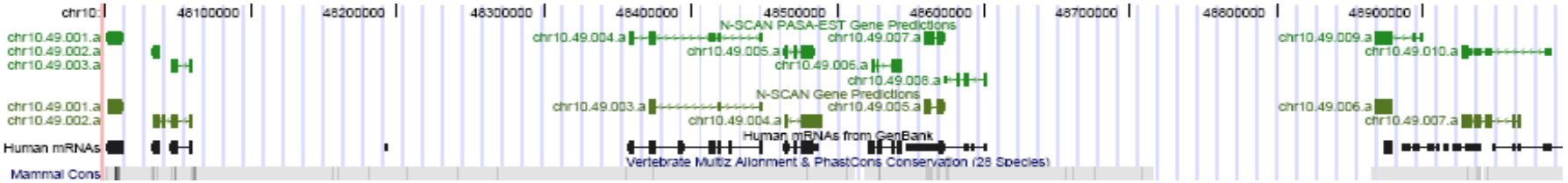
$$\bar{a}_{i,j} = \frac{\text{expected number of tranistions from } S_i \text{ to } S_j}{\text{expected number of tranistions from } S_i}$$

$$= \frac{\sum_{t=0}^{L} \xi_t(i,j)}{\sum_{t=0}^{L} \gamma_t(i)}$$

$$\bar{e}_i(k) = \frac{\text{expected number of times in state } j \text{ and oberving symbol } v_k}{\text{expected number of times in state } j}$$

$$= \frac{\sum_{t=0, s.t. x_t = v_k}^{L} \gamma_t(i)}{\sum_{t=0}^{L} \gamma_t(i)}$$

Genes in a Genome

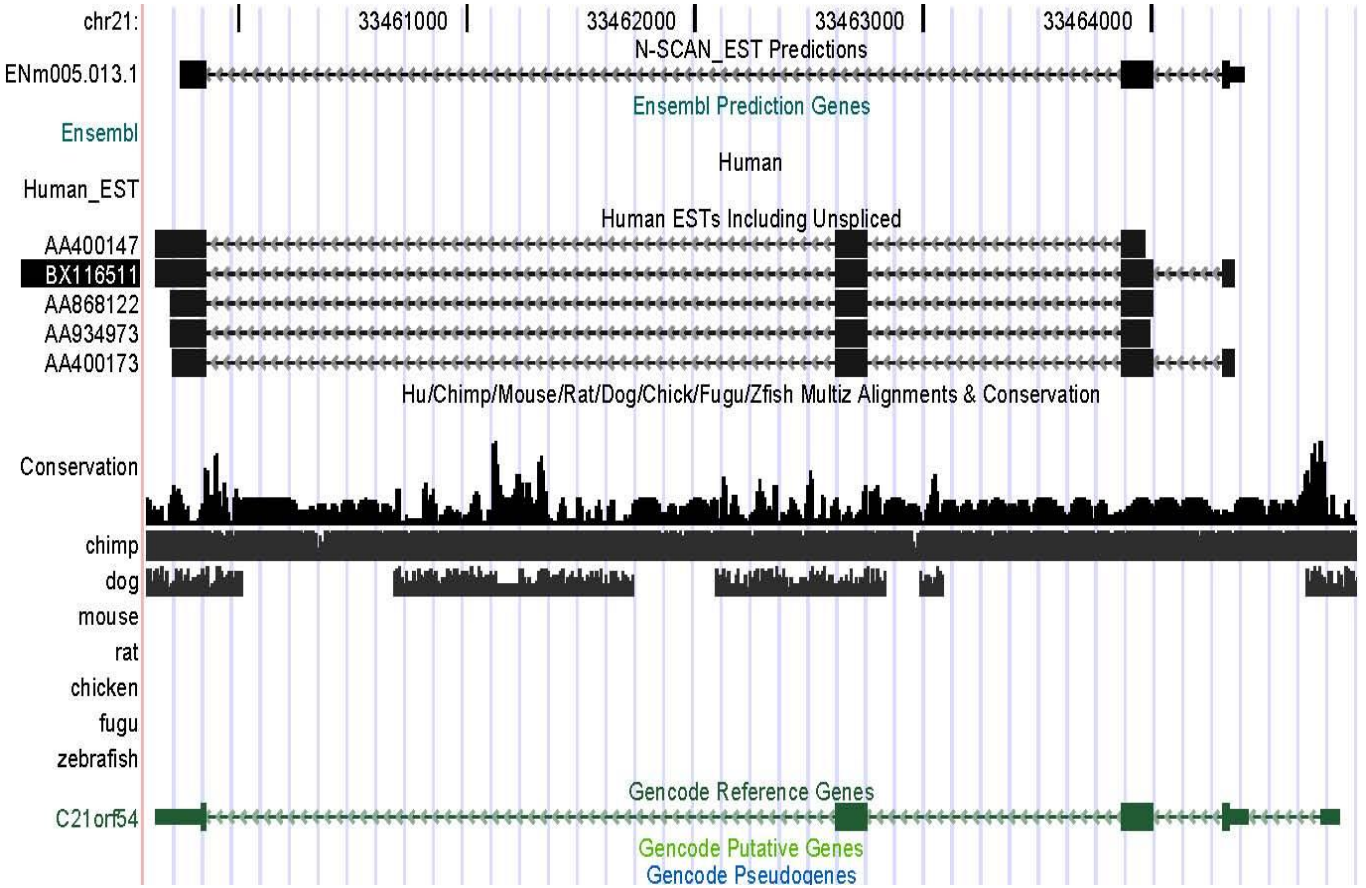


Different information for gene prediction

DNA sequence

ESTs

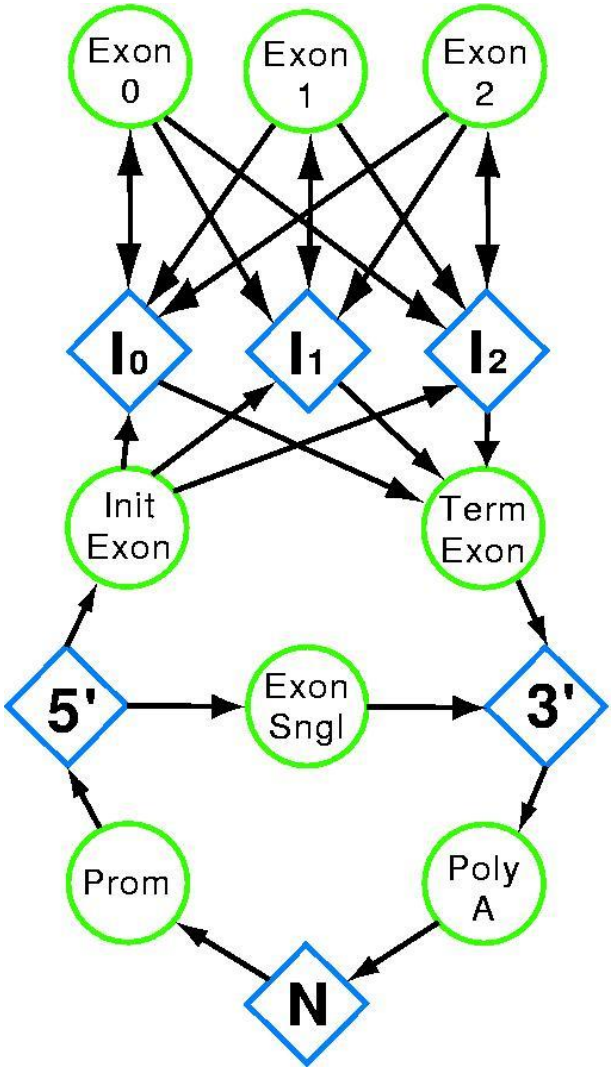
Conservation



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

TWINSCAN_EST Model

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



Conservation Sequence

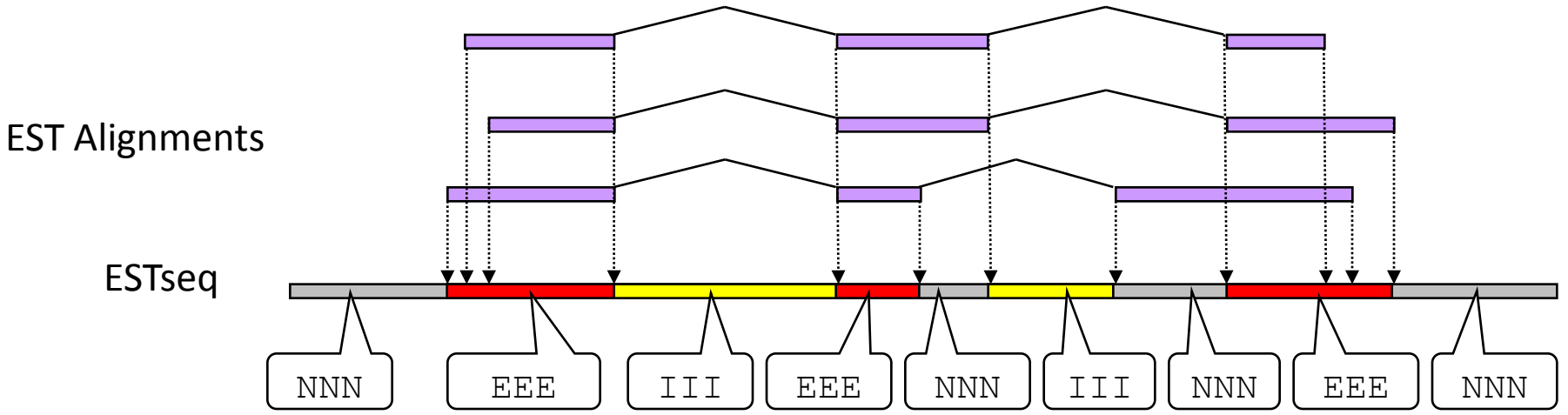
Generated by projecting local alignments to the target sequence

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

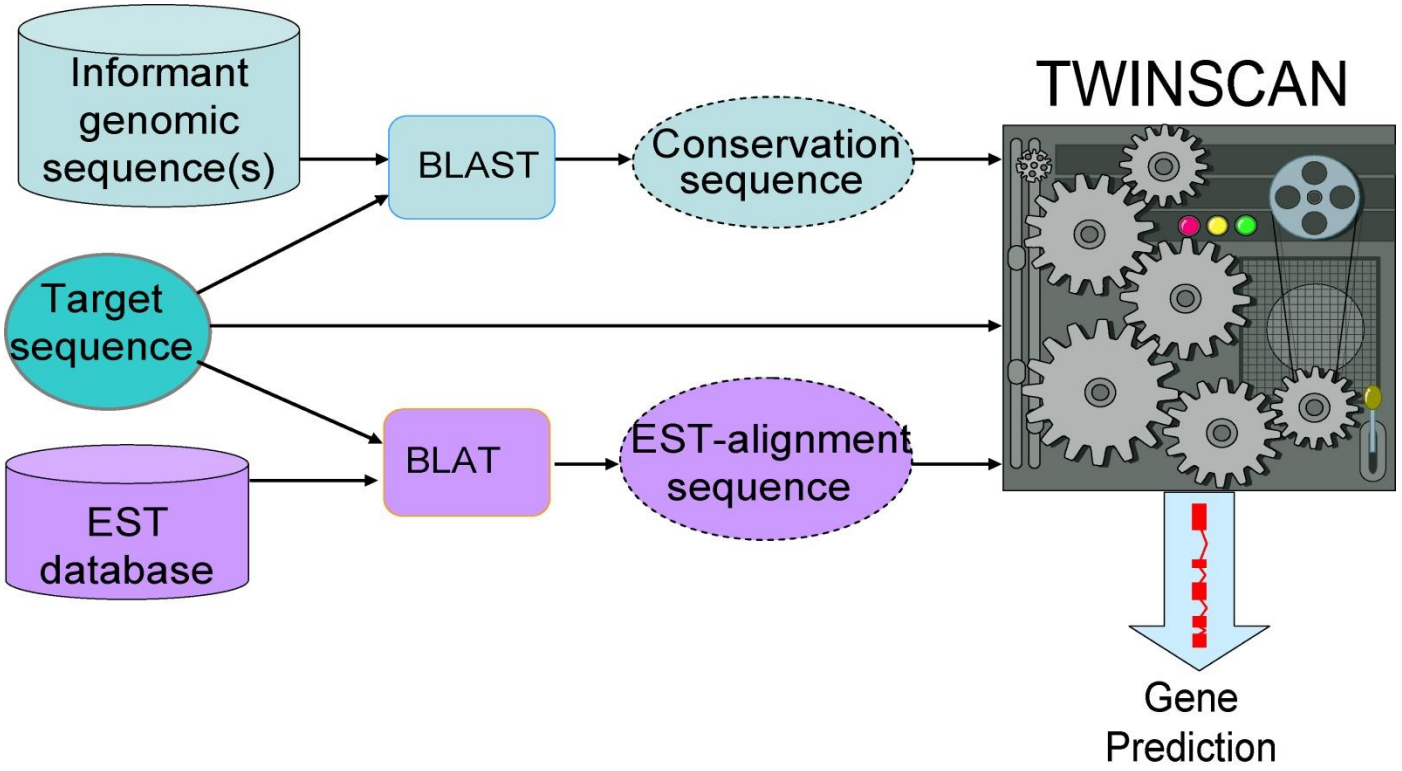
- Pair each nucleotide of the target with
  - "|" if it is aligned and identical
  - ":" if it is aligned to mismatch
  - "." if it is unaligned

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



Using ESTs for Gene Prediction: TWINSCAN_EST

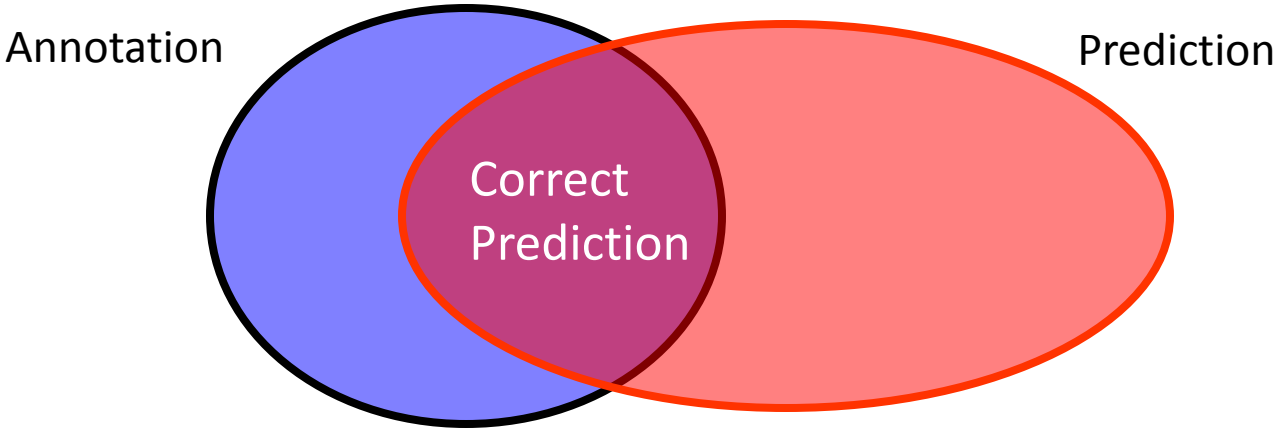


- Integrating EST alignment information into TWINSCAN to improve its accuracy where EST evidence exits and not to compromise its ability to predict novel genes.

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

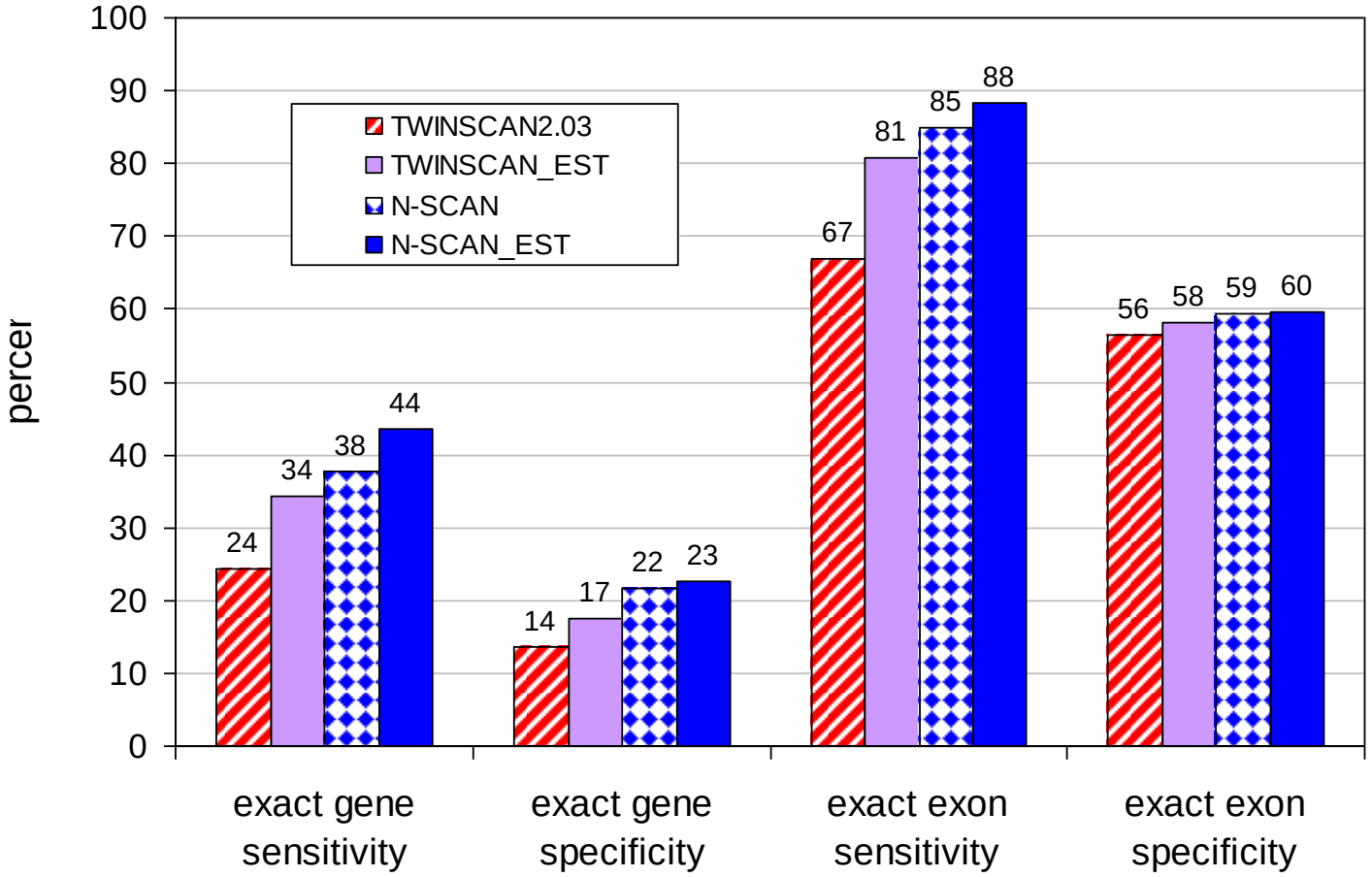
Accuracy Measurement (continue)



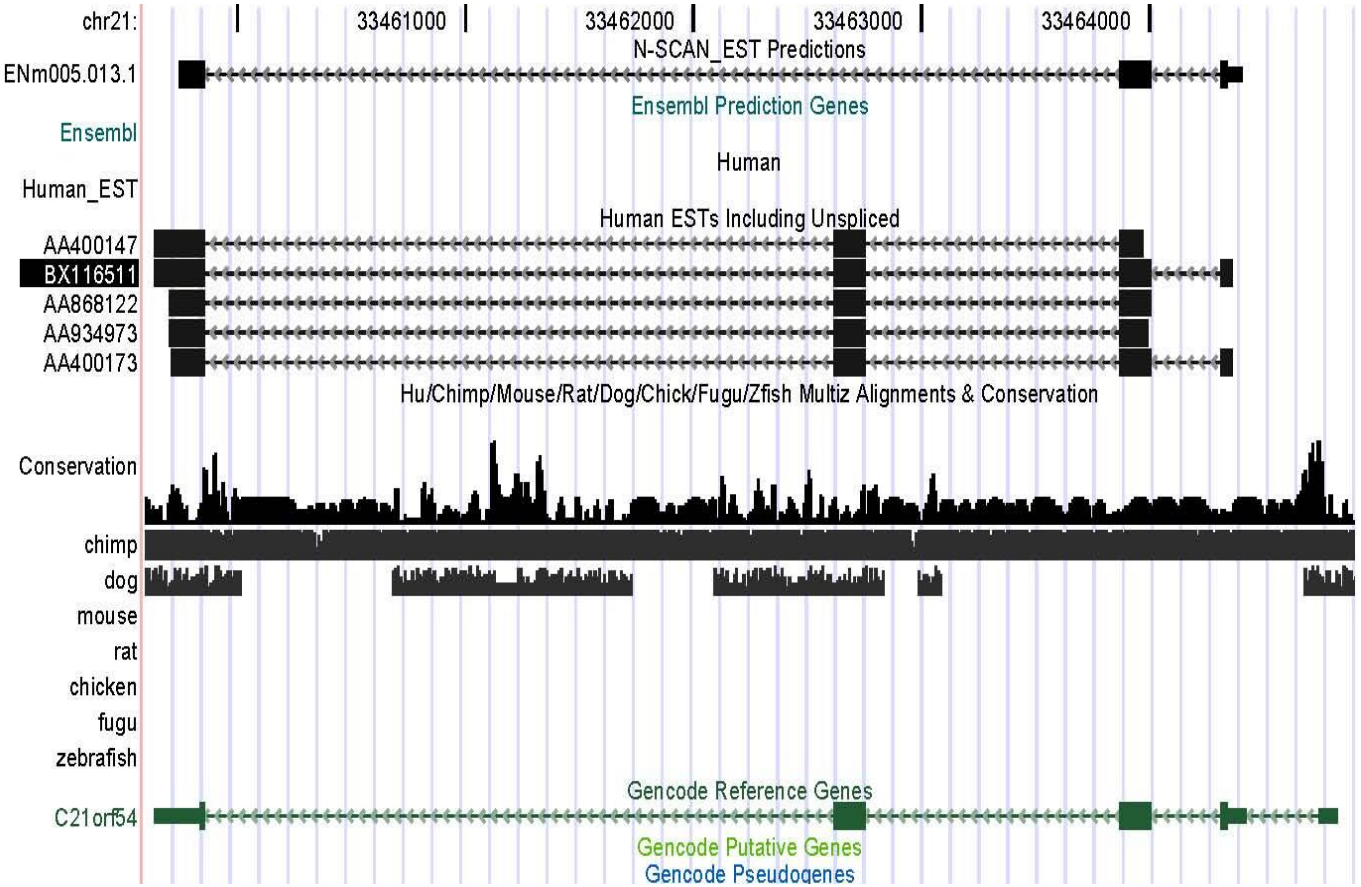
$$Sensitivity = \frac{Correct\_Prediction}{Total\_Annotation}$$

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

TWINSCAN_EST and N-SCAN_EST on the Whole Human Genome

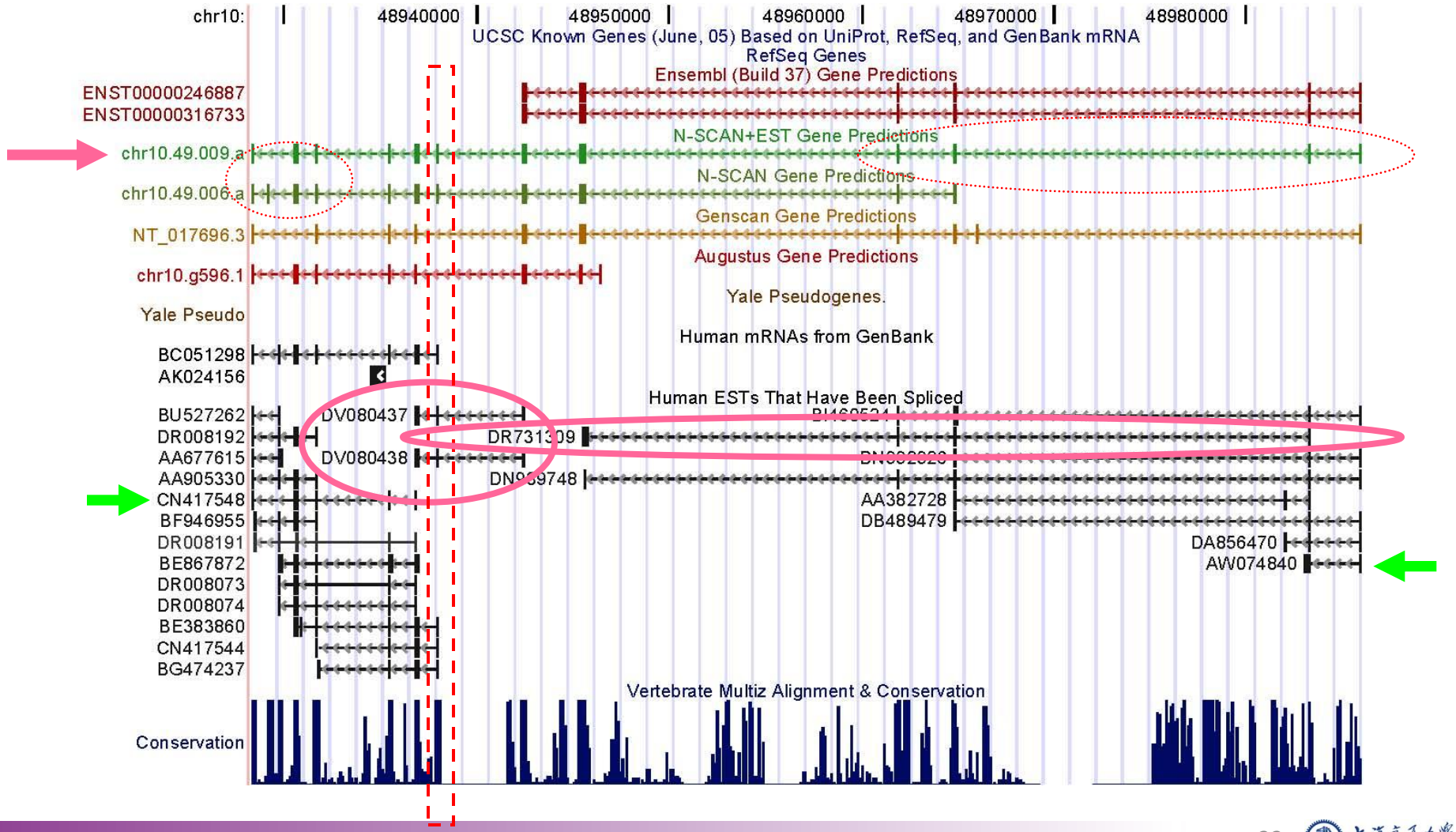


An Example of N-SCAN_EST Prediction

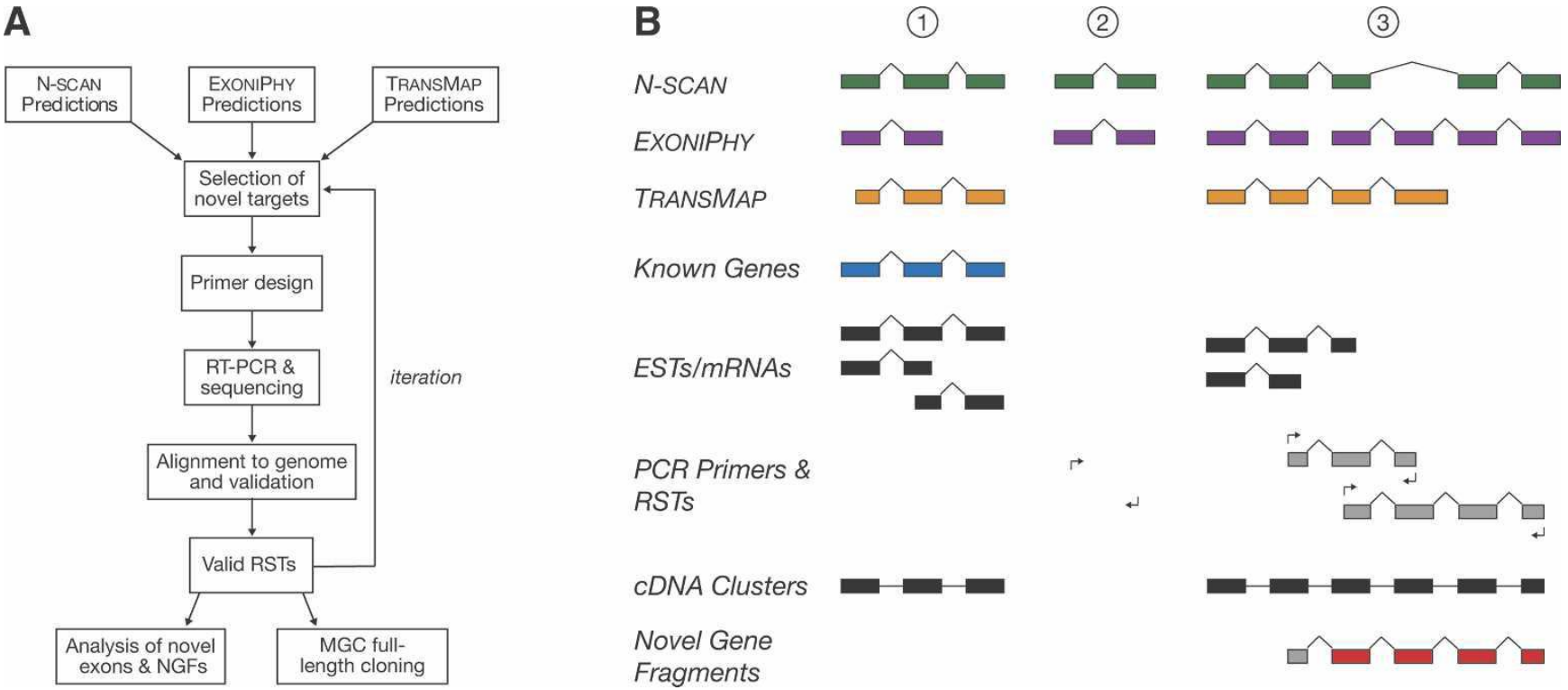


(Hg17, chr21:33,459,500-33,465,411)

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, 14, 2330-2335

Limits of HMM

- A strict statistical model
- All features need to be independent

Conditional Random Fields

- The conditional probability-like score of a label sequence (TFBS and non-TFBS) given an observation sequence *x* can be computed as follows

$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\lambda}) = \frac{\exp\left(\sum_{t=1}^{L} \sum_{k=1}^{K} \lambda_k f_k \left(y_t, y_{t-1}, t, \mathbf{x} \right) \right)}{\sum_{\mathbf{y}'} \exp\left(\sum_{t=1}^{L} \sum_{k=1}^{K} \lambda_k f_k \left(y'_t, y'_{t-1}, t, \mathbf{x} \right) \right)}$$

where *y* is *the* label sequence or annotation of all bins, *x is the* observed genomic sequence, f_k is the k-th feature functions and λ_k is the corresponding weight. The feature function can be an arbitrary function on *x* and *y'* is any label sequence.

Training and Prediction

- Training
  - To estimate the parameter vector λ, we use a Regularized Maximum Conditional Log Likelihood method.

$$\lambda_{ML} = \arg\max_{\lambda} \left(\ln(p(y \mid x; \lambda) \right)$$

That is

$$\lambda_{ML} = \arg\max_{\lambda} \left(\sum_{t=1}^{L} \lambda_k f_k - \ln\left(Z(\mathbf{x}) \right) - \frac{\|\boldsymbol{\lambda}\|^2}{2\sigma^2} \right)$$

This can be done by numerical computing.

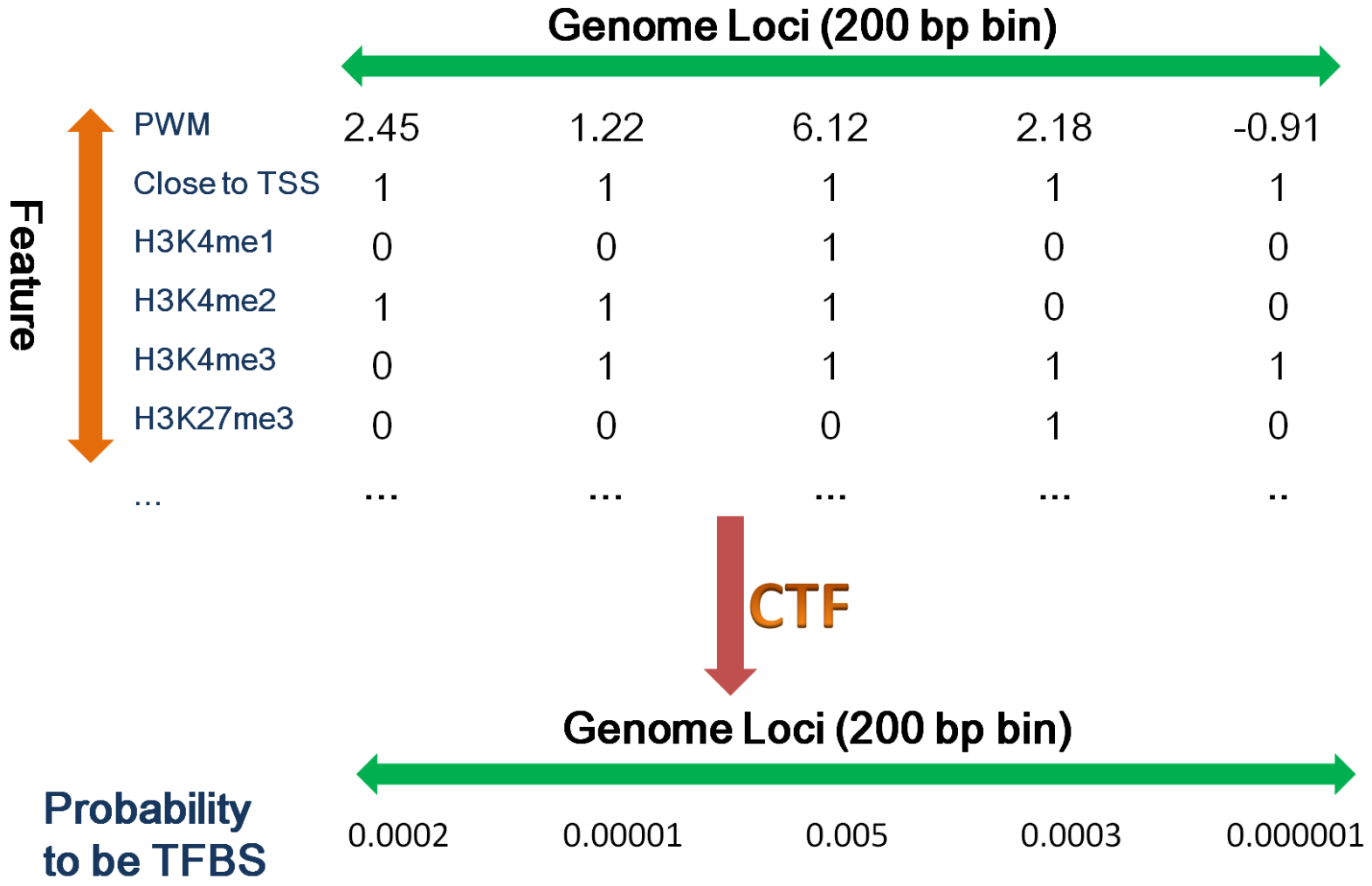
- Prediction
  - the marginal probability of *j*-th bin to be TFBS as follows

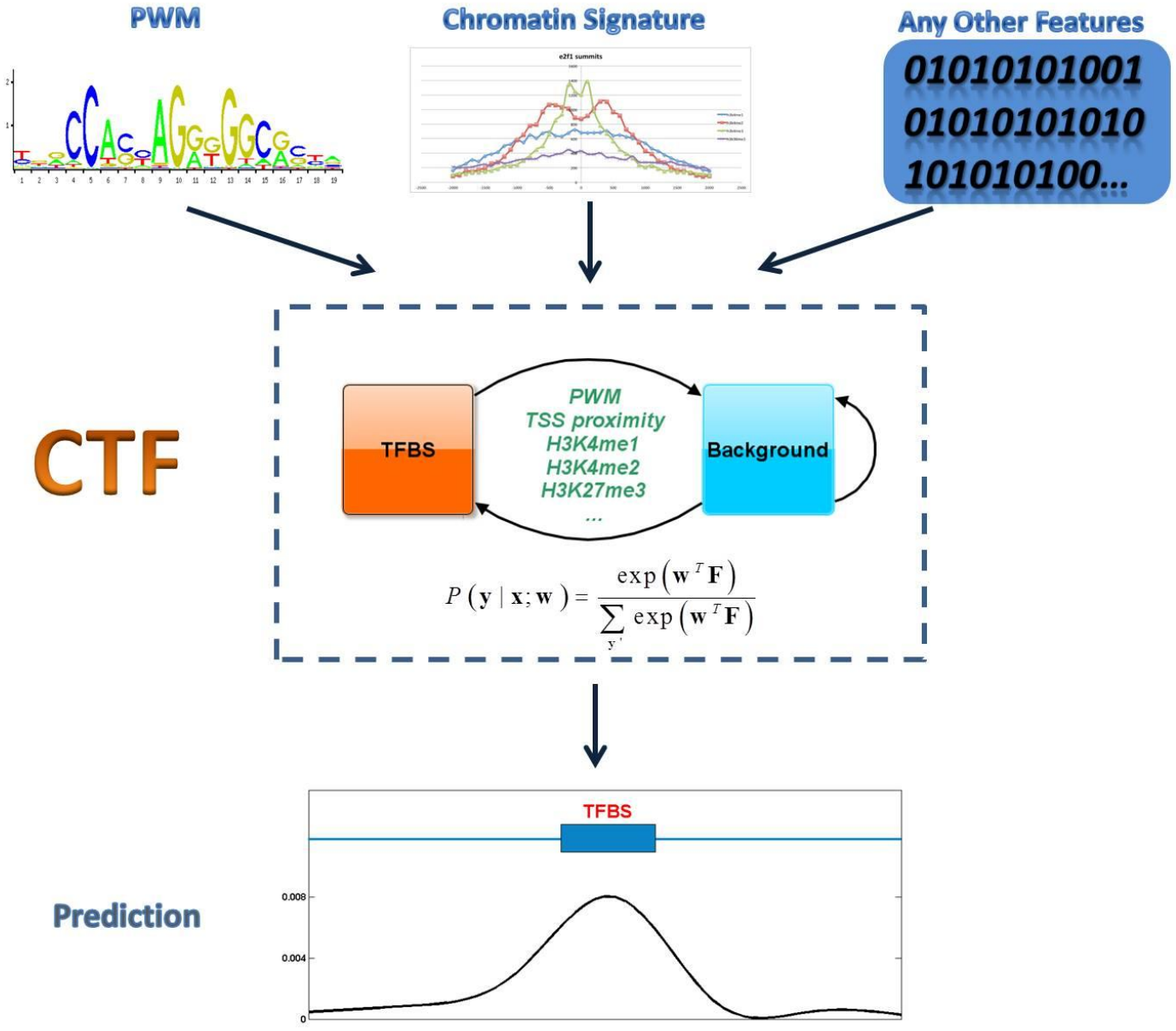
$$s_j = p\left(y_j = 1 \mid \mathbf{x}; \boldsymbol{\lambda} \right)$$

TFBS finding

- Transcript factor binding site finding is challenging , because
  - TFBSs are short, 6bps and up
    - High false positive prediction
  - No gold standard data until recent Chip-seq data.

CTF: **A novel integrated TFBS prediction system based on Conditional Random Fields**





The system diagram of CTF

Dataset

- The binding sites of 13 TFs in mouse ES cells from the ChIP-seq data from Chen et al, Cell, 2008.
- The 13 TFs:
  - c-Myc, CTCF, E2f1, ESrrb, Klf4, Nanog, n-Myc, Oct4, Smad1, Sox2, STAT3, Tcfcp2l1 and Zfx

Performance evaluation

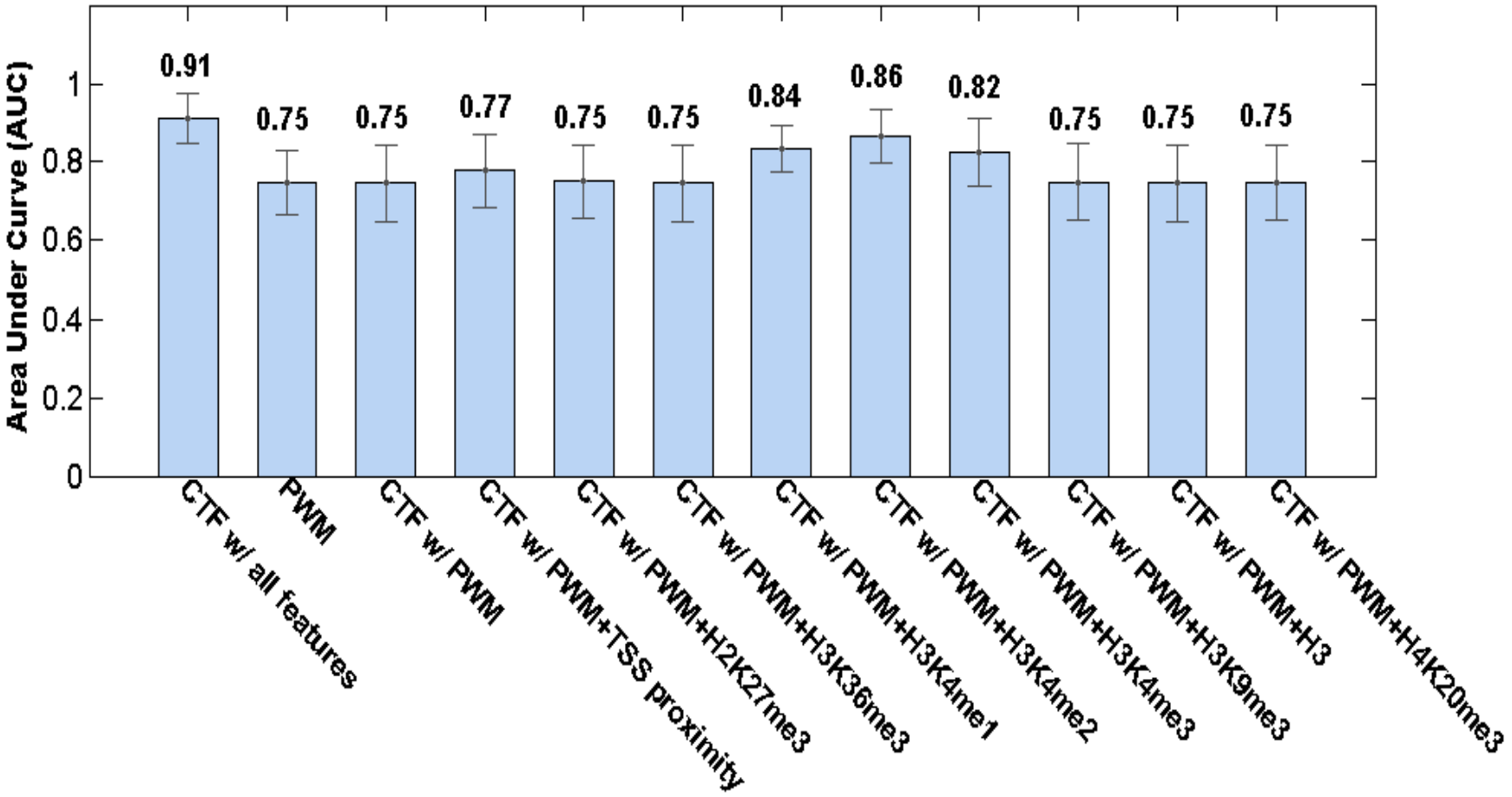
- Gold-standard TFBS dataset
  - “Peak-centric” method
  - Divide the genome into bins of 200bps
  - Those bins with Chip-seq peaks are gold standard TFBSs
- Evaluation
  - TFBSs with their centers overlapping with a bin with Chip-seq peaks are TPs
- 10-fold cross validation
  - Divide 19 chromosomes into 10 folds
  - 1 for test and 9 for training

Accuracy Comparison

	CTF	Chromia	PWM
c-Myc	**0.98**	0.94	0.84
CTCF	**0.76**	0.69	0.76
E2f1	**0.96**	0.94	0.75
Esrrb	**0.89**	0.84	0.77
Klf4	**0.96**	0.92	0.83
Nanog	**0.83**	0.82	0.62
n-Myc	**0.97**	0.94	0.86
Oct4	**0.92**	0.88	0.61
Smad1	**0.92**	0.89	0.66
Sox2	**0.90**	0.87	0.70
STAT3	**0.91**	0.86	0.72
Tcfcp2l1	**0.88**	0.83	0.79
Zfx	**0.97**	0.96	0.82
Average	**0.91**	0.88	0.75

He, Zhang, Zheng and Wei, 2012, ***BMC Genomics***

Accuracy (AUC) for PWM and CTF with different features



He, Zhang, Zheng and Wei, 2012, ***BMC Genomics***

Representative Publications

Functional element finding in genomes

- CTF: a CRF-based TFBS prediction system

  ***BMC Genomics*, 2012, 13(Suppl 8):S18**

- Interactions between TFs and their DNA targets in Mammals

  ***BMC Genomics**, 2012, 13:388

- Predicted and validated 734 novel human genes (MGC)

  ***Genome Research**, 2009, 19:2324-2333

- Using ESTs to improve gene prediction accuracy

  ***BMC Bioinformatics**, 2006, 7:327

- Gene prediction for *C.elegans, find >1,000 novel genes*

  ***Genome Research**, 2005, 15:577-582.

  Reported by ***Nature Reviews Genetics*** as "***Research highlight***"

- Gene prediction for *C. briggsae*

  ***PLoS Biology**, 2003, 1(2): E45

On going work

- “Finding and characterization of >30,000 novel human transcripts”, Zhiqiang Hu and Chaochun Wei, submitted
- “CPA: a CRF-based protein acetylation site prediction system”, Ting Hou, Guangyong Zheng, and Chaochun Wei, in preparation

Summary

- Statistical models are powerful for genomic functional element finding.
- We have
  - built statistical models
    - HMMs, CRFs
  - applied those models to
    - Protein-coding gene structure prediction
    - TFBS finding
    - Protein acetylation site prediction
    - Alternative splicing prediction
- They can also be applied in other areas in bioinformatics
  - Sequence alignment
  - Sequence classification
  - ……

Acknowledgement

- Students
  - Yupeng He
  - Ting Hou
  - Zhiqiang Hu
  - Guangyong Zheng
- Funding
  - NSFC
  - 863
  - Shanghai Pujiang Program