Guts of Dragon Promoter Finder & Mapper

Limsoon Wong

(Based on work with Vladimir Bajic & Rajesh Chowdhary)

Plan

- **Promoter & Promoter Modeling**
- **Dragon Promoter Finder**
- **Dragon Promoter Mapper**
  - Specific modeling of histone promoters
  - Whole-genome scan for genes co-regulated with histones

Promoter & Promoter Modeling

Promoter Modeling

- **What does it involve?**
  - Characterization of known promoters
  - Recognition of promoters in uncharacterized genome
- **Why is it important?**
  - Unravel gene's regulatory mechanism
  - Discover new genes
  - Define potential regulatory networks
    - **i.e. genes with similar regulatory behaviour/ promoter structure as target gene group**
- **Why is it difficult?**
  - High variability in length of promoter: Hundreds to thousands of bases
  - High variability in promoter features which themselves are difficult to predict
    - **A set of features can't be universally applied for all types of promoters**
  - TFBS occur in numerous combinations & order. Location, orientation, and mutual distance vary
  - Incomplete information about TF and TFBS

Types of Promoter Modeling Studies

- **General promoter modeling**
  - Wide applicability
  - But too general
    - **one size doesn't fit all**
- **Specific promoter modeling**
  - Better suitability in some situations

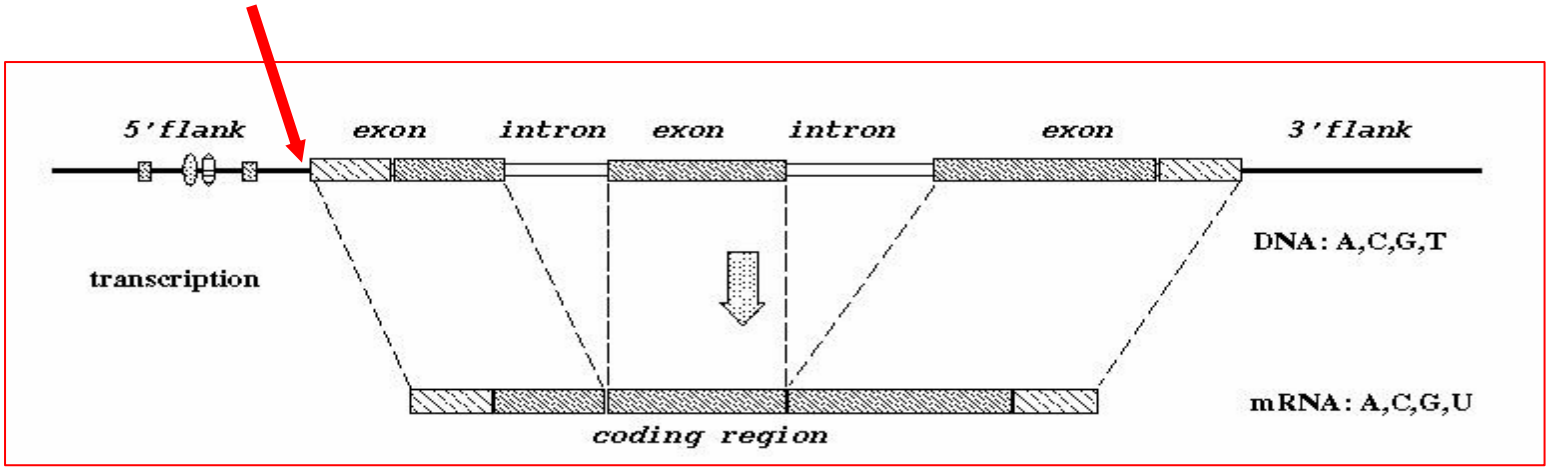
- **Advantages of specific promoter modeling**
  - Promoter structure comparison betw target & query seq give more info
    - **Increased sensitivity & specificity**
  - Identify co-regulated genes
  - determine tissue specificity of genes
  - predict function of genes

A General Promoter Finder:

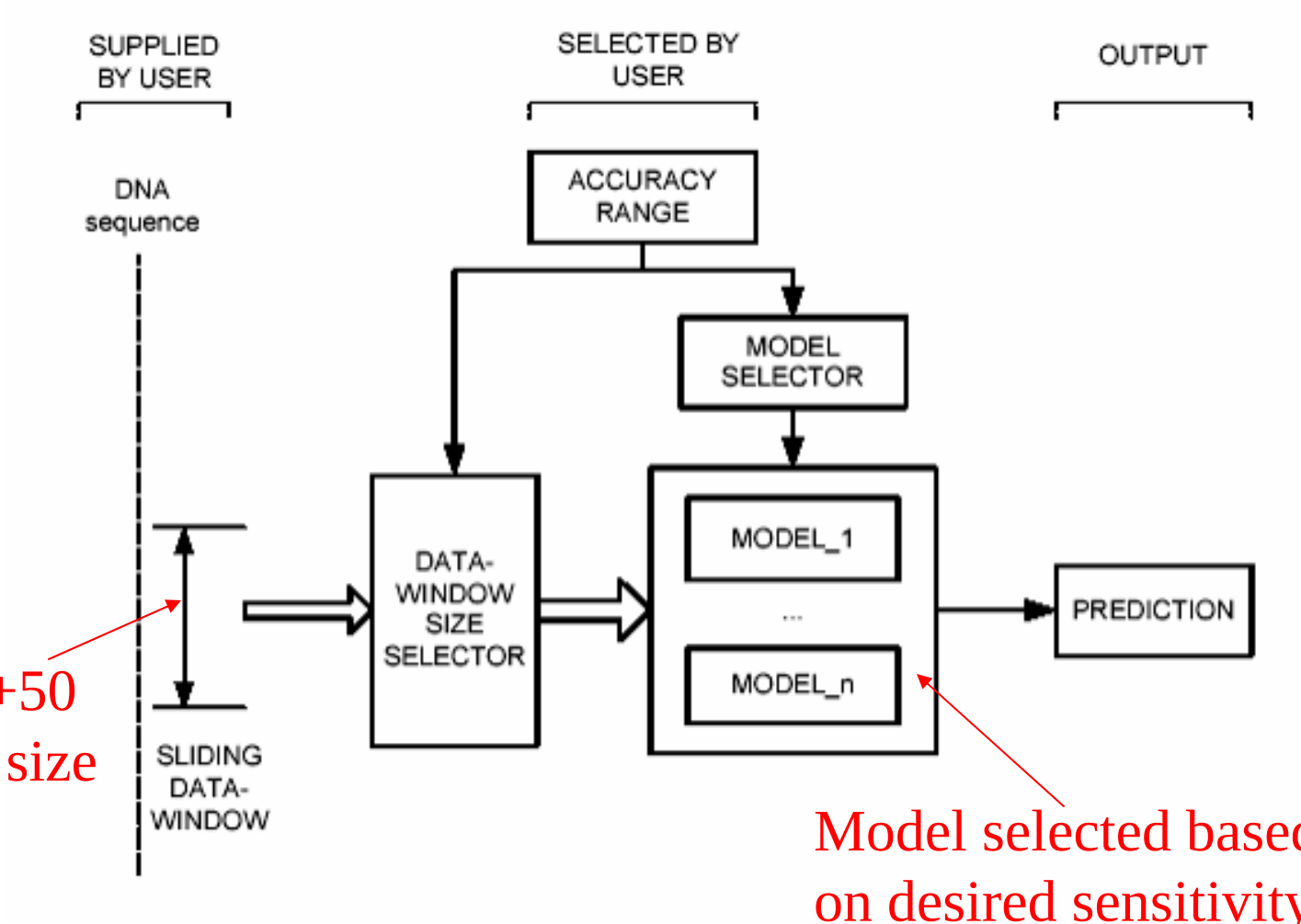
Dragon Promoter Finder

Dragon Promoter Finder

- **Multi-sensor integration via ANNs**
- **Multi-model system structure**
  - for different sensitivity levels
  - for GC-rich and GC-poor promoter regions



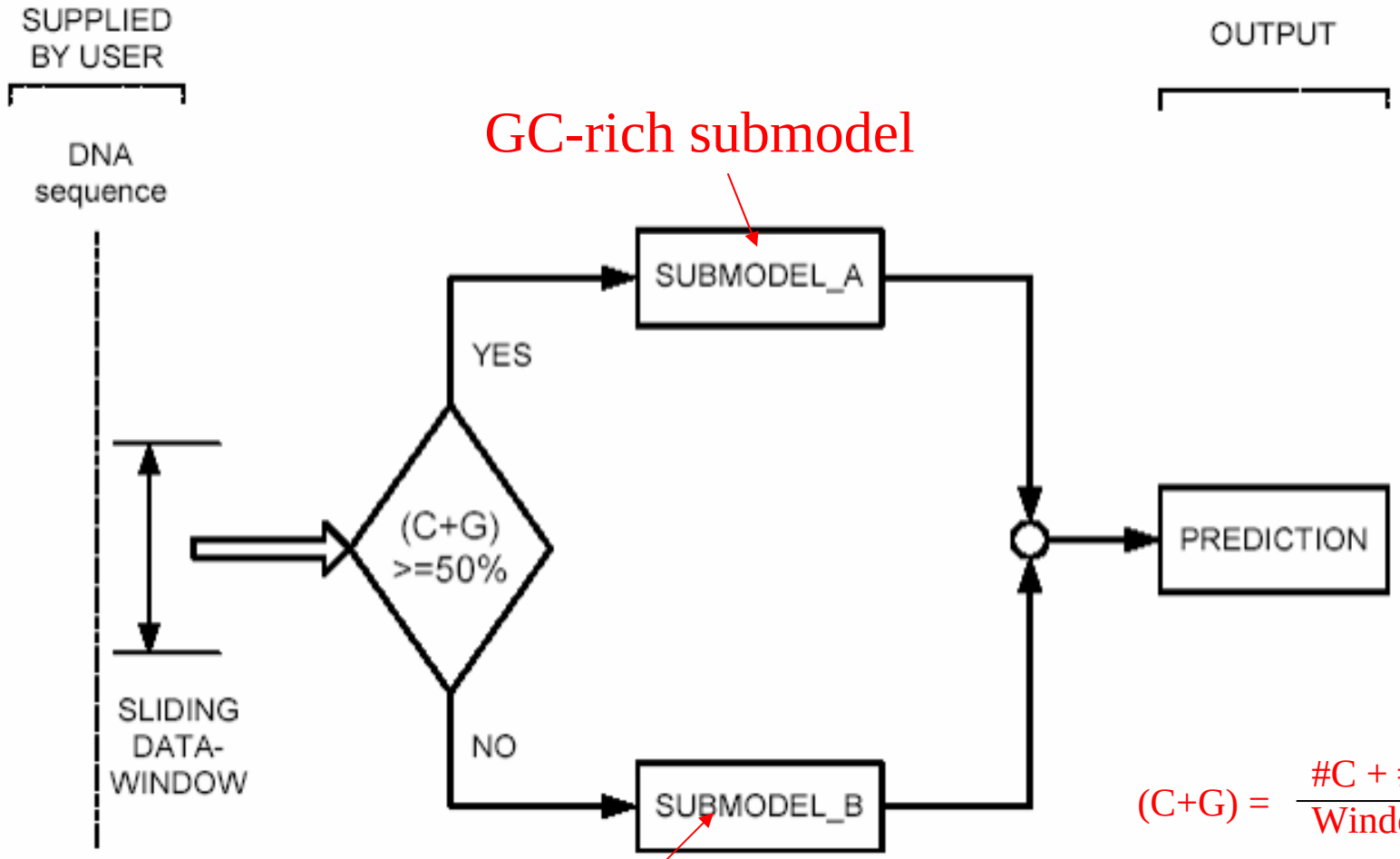
Structure of Dragon Promoter Finder



-200 to +50 window size

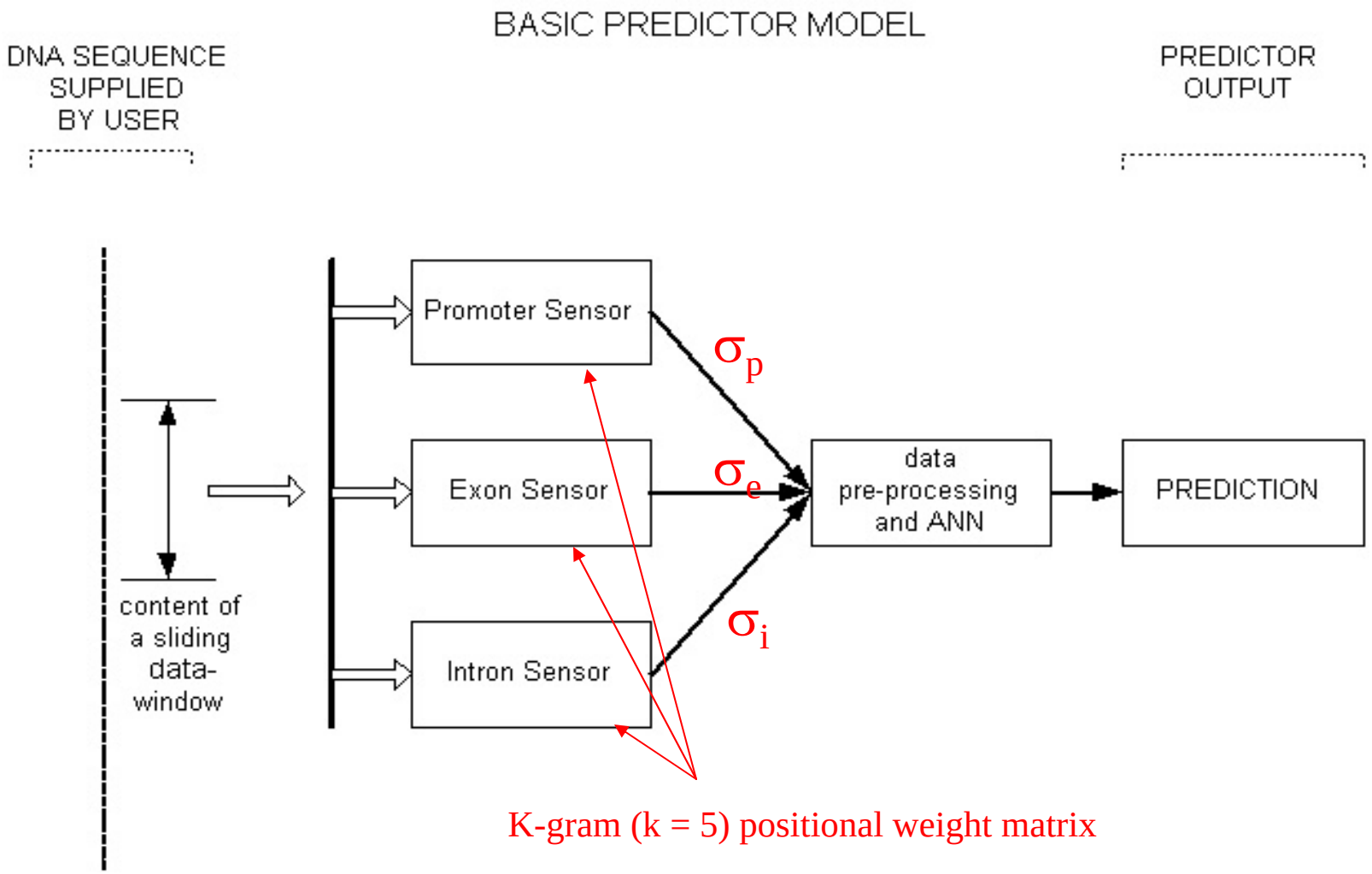
Model selected based on desired sensitivity

Each Model Has Two Submodels Based On GC Content



$$(C+G) = \frac{\#C + \#G}{\text{Window Size}}$$

Data Analysis Within Submodel



Promoter, Exon, Intron Sensors

- **These sensors are positional weight matrices of k-grams, k = 5 (aka pentamers)**
- **They are calculated as σ below using promoter, exon, intron data respectively**

Window size

$$\sigma = \frac{\left(\sum_{i=1}^{L-4} p_j^i \otimes f_{j,i} \right)}{\left(\sum_{i=1}^{L-4} \max_j f_{j,i} \right)}, \qquad p_j^i \otimes f_{j,i} = \begin{cases} f_{j,i}, \text{ if } p_i = p_j^i \\ 0, \text{ if } p_i \neq p_j^i \end{cases},$$

Pentamer at i^{th} position in input

Frequency of jth pentamer at ith position in training window

j^{th} pentamer at i^{th} position in training window

Data Preprocessing & ANN

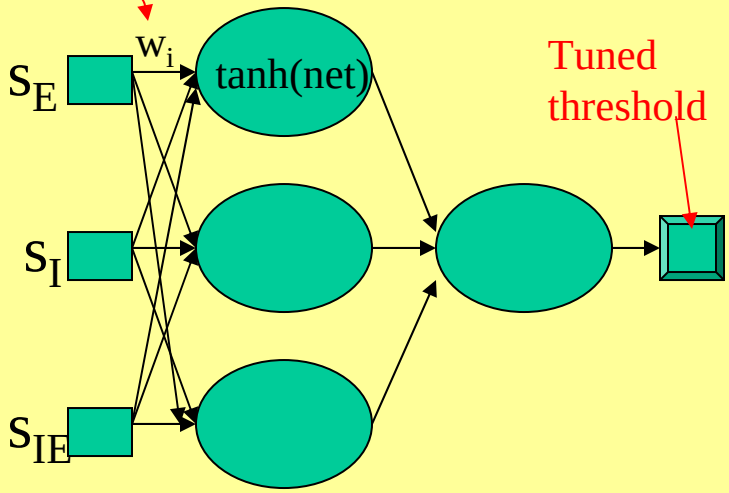
Tuning parameters

$$s_E = sat(\sigma_p - \sigma_e, a_e, b_e),$$
$$s_I = sat(\sigma_p - \sigma_i, a_i, b_i),$$
$$s_{EI} = sat(\sigma_e - \sigma_i, a_{ei}, b_{ei}),$$

where the function sat is defined by

$$sat(x,a,b) = \begin{cases} a, & if \quad x > a \\ x, & if \quad b \le x \le a. \\ b, & if \quad b > x \end{cases}$$

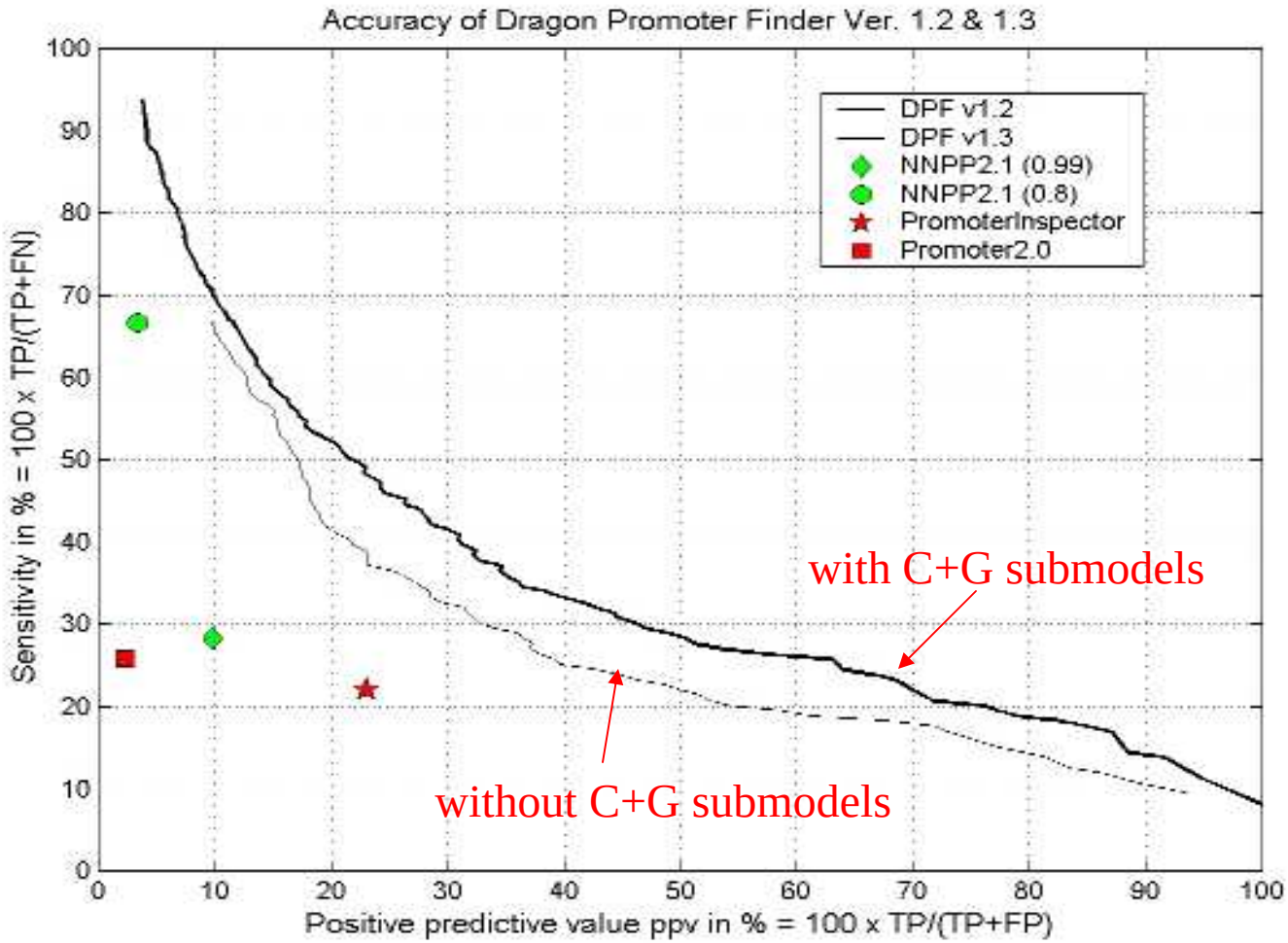
Simple feedforward ANN trained by the Bayesian regularisation method



$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$net = \Sigma\, s_i * w_i$$

Accuracy

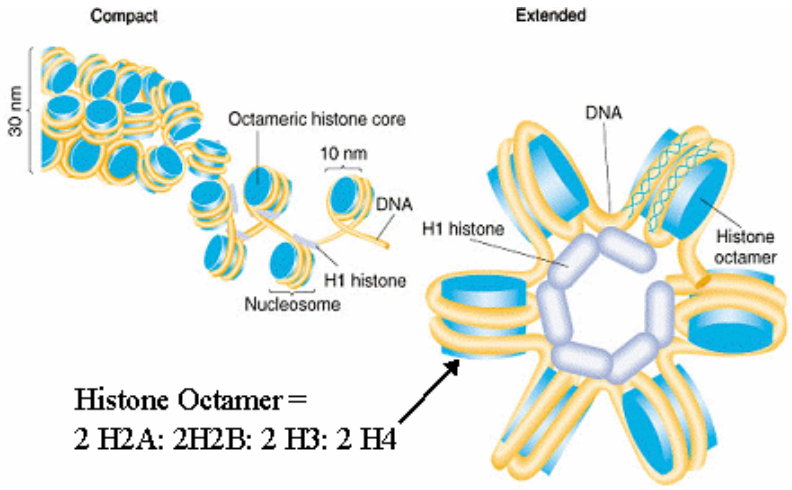


Specific (Histone) Promoter Modeling:
Dragon Promoter Mapper

Histones

- **What are histones?**
  - Basic proteins of eukaryotic cell nucleus
  - Form a major part of chromosomal proteins
  - Help in packaging DNA in the chromatin complex
- **Five types, namely H1, H2A, H2B, H3 and H4**
  - Several subtypes of the main types
- **Highly conserved**
  - H1 least conserved, H3 & H4 most conserved

- **Play essential role in chromosomal processes**
  - gene transcription, regulation
  - chromosome condensation, recombination, replication



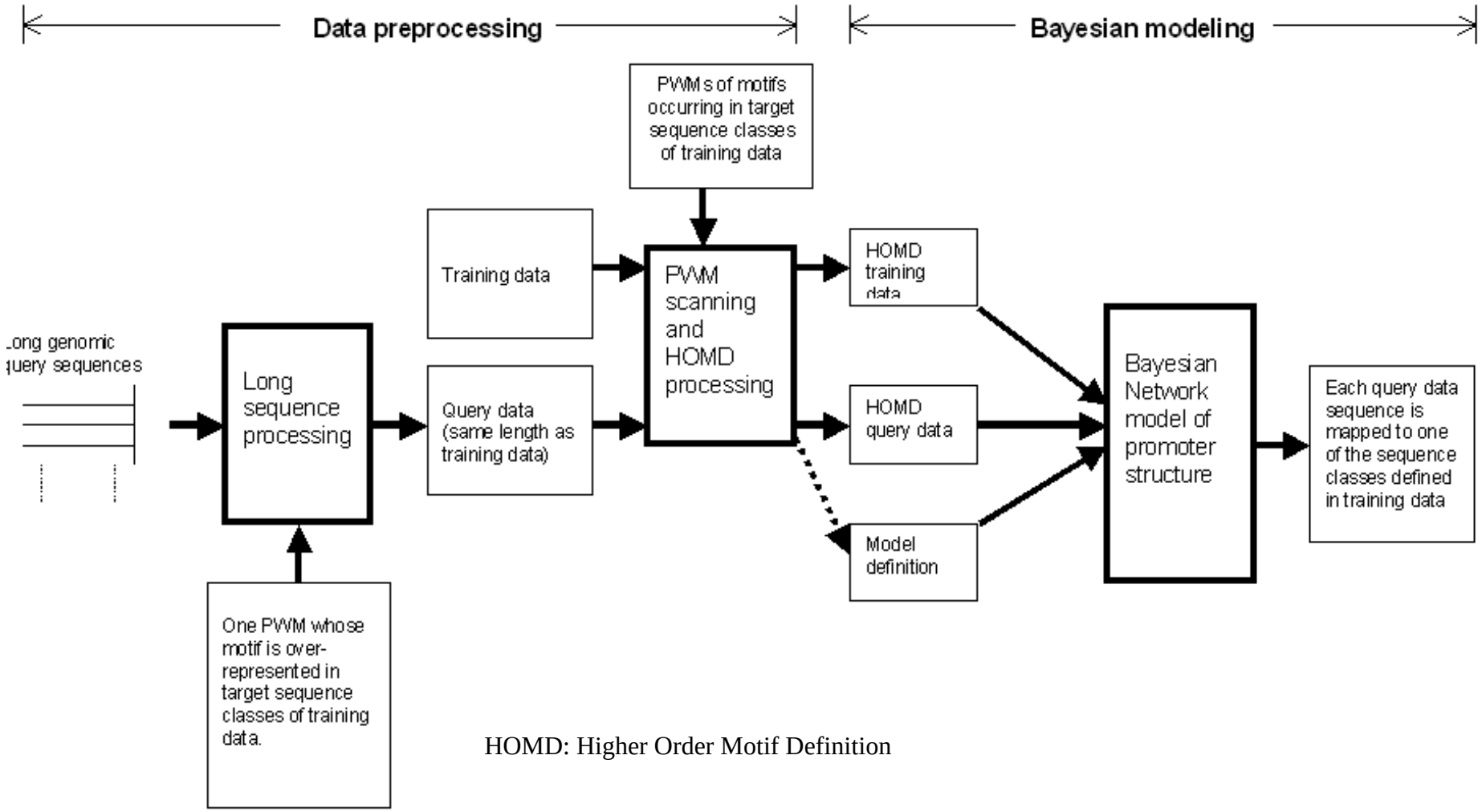
Dragon Promoter Mapper

- **Model histone promoter structure by probabilistically combining information on**
  - motif
  - its position
  - its strand
  - mutual spacer length between adjacent motifs

- **Guiding principle: A promoter is known by the binding sites it keeps**

Adapted from Chowdhary's PhD thesis presentation

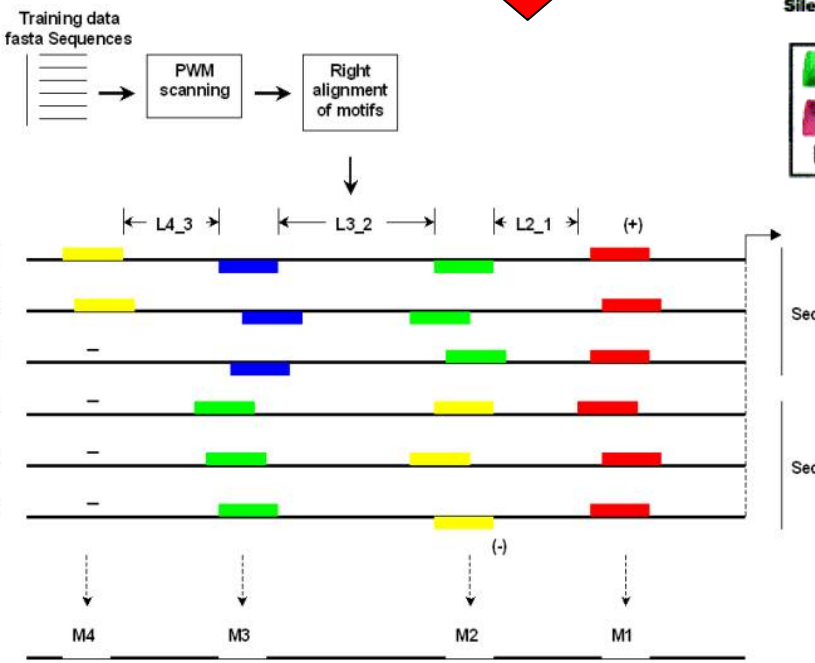
DPM Workflow

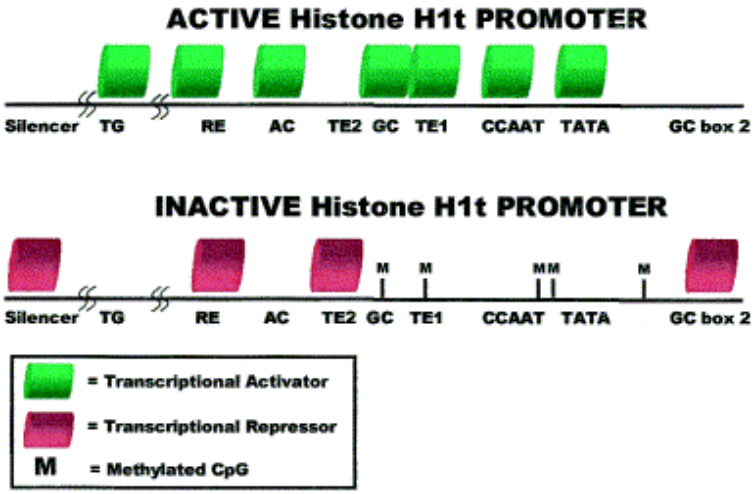


HOMD: Higher Order Motif Definition

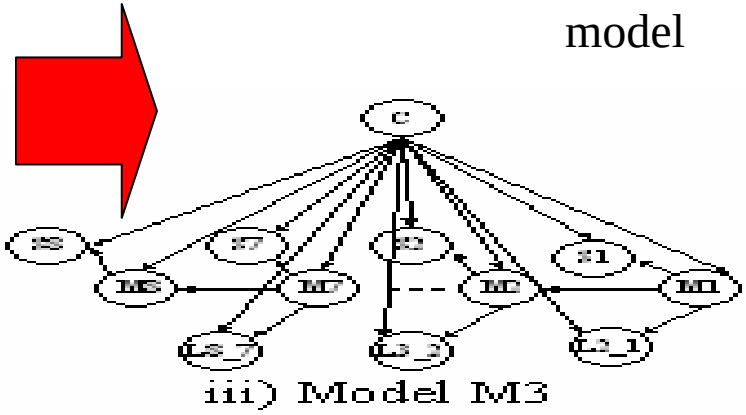
Higher-Order Motif Definition

Higher-order motif definition





User-defined model



Comparison w/ Similar Programs

	COMET	Cluster-Buster	Meta-MEME	MCAST	DPM
# of hits in 68 promoter sequences (number in brackets represent Se)	46 (0.677)	55 (0.809)	67 (0.985)	49 (0.721)	64 (0.941)
# of FP hits (number in brackets represent ppv)	2 (0.958)	2 (0.965)	38 (0.638)	35 (0.583)	5 (0.928)
Correlation coefficient	0.677	0.790	0.508	0.212	0.868

Compared programs	Motif distribution/arrangement
COMET	HIST1H1A: [+AC]13[+TATA] HIST1H1B: [+AC]56[+AC]13[+TATA] HIST1H1C: [+AC]49[+AC]77[+TATA] HIST1H1D: [+AC]52[+CAAT]16[+TATA] HIST1H1E: [+AC]78[+TATA]
Cluster-Buster	HIST1H1A: [+AC]3[-GC]-8[-TATA]-9[+TATA] HIST1H1B: [+TG]-10[-AC]105[-CAAT]45[+CAAT]170[+AC]56[+AC]6[-TATA]-7[+TATA] HIST1H1C: [+TG]-10[-AC]9[+AC]122[+AC]71[-AC]52[+AC]5[+E2F]37[+AC]54[+AC]6[-TATA]-12[-TATA]-9[+TATA] HIST1H1D: [+AC]-10[-TG]51[+CAAT]9[-TATA]-7[+TATA] HIST1H1E: [+TG]-10[-AC]189[-E2F]146[+AC]55[+AC]3[-GC]-10[-TATA]-12[-TATA]-9[+TATA]
Meta-MEME	HIST1H1A: [-AC]34[+RT1]9[-E2F]53[+AC]12[+TATA]11[+Oct1]35[+TG]17[-GC] HIST1H1B: [+AC]31[+GC]10[+AC]12[+TATA] HIST1H1C: [-TG]1[+TG]8[+AC]4[-Oct1]48[-GC]44[+AC]70[-AC]51[+AC]4[+E2F]36[+AC]53[+AC]7[-TATA]2[-GC] HIST1H1D: [+TG]8[+AC]9[+Oct1]13[+TATA]38[-Oct1]121[+AC]93[+AC]51[+CAAT]15[+TATA] HIST1H1E: [+TG]188[-E2F]145[+AC]33[+GC]7[+AC]12[+TATA]
MCAST	HIST1H1A: [-AC]34[+RT1]9[-E2F]53[+AC]12[+TATA]11[+Oct1]35[+TG]17[-GC] HIST1H1B: [+AC]31[+GC]10[+AC]12[+TATA] HIST1H1C: [+AC]7[-TATA]2[-GC] HIST1H1D: [+CAAT]15[+TATA] HIST1H1E: [+AC]12[+TATA]
DPM	HIST1H1A: [-AC]104[+GC]13[+CAAT]19[+TATA]68[+TG] HIST1H1B: [+TG]348[+AC]36[+GC]3[-CAAT]9[+CAAT]19[+TATA] HIST1H1C: [+TG]349[+AC]56[+CAAT]19[+TATA] HIST1H1D: [+TG]348[+AC]58[+CAAT]19[+TATA] HIST1H1E: [+TG]348[+AC]57[+CAAT]12[-GC]3[+TATA]
Meergans et. al., (1998)	Known binding sites in H1 histone promoters: HIST1H1A: [+CAAT]19[+TATA] HIST1H1B: [+TG]364[+AC]56[+CAAT]19[+TATA] HIST1H1C: [+TG]340[+AC]58[+CAAT]19[+TATA] HIST1H1D: [+TG]372[+AC]57[+CAAT]19[+TATA] HIST1H1E: [+TG]354[+AC]58[+CAAT]19[+TATA]
Duncliff el. al., (1995)	Mutual distance between TG-box and AC-box: HIST1H1B: [+TG]359[+AC] HIST1H1D: [+TG]355[+AC]
Duncliff et. al., (1995), Osley (1991), Gallinari et. al., (1989)	General structure of H1 histone promoter, drawn from information in the reference: [TG]350[AC]34[GC]10[CAAT]19[TATA]

Comparison w/ Similar Programs

rom Chowdhary's PhD thesis presentation

Comparison w/ General Promoter Prediction Programs

	Dragon Promoter Finder	Eponine	DPM
# of hits in 68 promoter sequences (number in brackets represent Se)	36 (0.529)	17 (0.250)	64 (0.941)
# of FP hits (number in brackets represent ppv)	2 (0.947)	0 (1.000)	5 (0.928)
Correlation coefficient	0.509	0.378	0.868

Dragon Gene Start Finder and **FirstEF** - not applicable on analyzed data.

Human Genome Scan Expt

- **Genome scanning**
  - GC-content in a genomic segment > 0.37
  - Motif for initial scan = CAAT-box
  - Length of segment upstream of CAAT-box = 425, downstream = 175
  - Min spacer betw CAAT-boxes = 0
  - Min no. of motifs in seq = 3
- **Each extracted segment classified to Histone vs Non-promoter class based on their structures**
- **Predictions classified as "Histone" were further analyzed:**
  - Annotation available?
    - **Use RefSeq gene data (HG17, May 2004)**
    - **These may be co-regulated genes**
  - Are available annotations co-expressed with histone genes?
    - **Use Gene Sorter utility (with GNF Gene Expression Atlas2 data) of UCSC Genome browser**

Human Genome Scan Results

	DPM predictions			DPM predictions mapped with annotated RefSeq genes (including histone genes)			DPM predictions mapped with histone co-expression data (including histone genes)		
Chromosome	# Predictions with CAAT-box (A)	# (A) with motifs => 3 (B)	# (B) predicted as histone class (C)	# (C) mapped with known genes	# Gene transcripts mapped with (C) (redundantly)	# Unique genes mapped with (C)	# (C) mapped with known genes	# Gene transcripts mapped with (C) (redundantly)	# Unique genes mapped with (C)
1	108973	39360	10669	1627	2220	659	473	710	215
2	109843	40786	10427	1190	1641	450	292	406	129
3	90473	34231	8642	1009	1391	372	306	373	115
4	78265	31004	8316	741	967	264	184	260	63
5	82101	31490	8179	869	1355	346	225	333	88
6	76965	29486	8007	1000	1249	384	318	389	140
7	70615	26053	6895	1011	1440	299	224	291	91
8	66855	25329	6543	744	953	238	172	222	56
9	56715	20600	5542	684	945	244	167	211	67
10	65527	24104	6534	956	1317	293	242	353	88
11	63993	23468	6037	805	1074	323	197	282	100
12	62499	23626	6466	800	981	348	275	331	122
13	39684	15488	4292	396	484	131	102	119	35
14	41734	15367	4001	530	757	201	181	229	73
15	41196	14726	4023	482	676	198	123	195	58
16	41963	14779	4064	641	834	245	152	213	66
17	39083	12970	3709	645	867	306	231	279	126
18	34174	12987	3431	359	472	115	88	107	28
19	28738	10202	3472	681	950	350	252	370	113
20	32508	11636	3087	420	646	147	92	124	49
21	15138	5803	1578	214	408	86	80	172	22
22	17620	5730	1743	315	521	128	126	188	59
M	18	5	0	0	0	0	0	0	0
X	74725	29971	7598	807	1344	285	336	446	108
Y	12531	4869	1371	52	89	20	10	11	5
Total	**1351936**	**504070**	**134626**	**16978**	**23581**	**6432**	**4848**	**6614**	**2016**

- **No. of regions w/ histone promoter-like struct = 134626**
- **Found 62 histone promoter w/ CAAT-box**
  - 53 were training seq out of 60 that had CAAT-box
- **Found that histone genes are co-regulated w/ many genes**

Conclusions

- **DPM performs well**
- **User can implement any type of correlations betw motif features based on his background knowledge**
- **Explicitly classify a segment w/ cluster of binding sites to one of target classes**
- **Handle multiple target classes of seq**
- **Create well-annotated data set of histone promoters**
- **Comprehensive model of histone promoter struct**
- **Discover regions in human genome w/ similar struct as histone promoters**
  - May be promoters co-regulated w/ histones
  - We are verifying these experimentally w/ collaborators in Germany

References

- V.B.Bajic et al., “Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates”, *J. Mol. Graph. & Mod*. 21:323--332, 2003
- V.B.Bajic et al., “Dragon Promoter Finder: Recognition of vertebrate RNA polymerase II promoters”, *Bioinformatics* 18:198--199, 2002.
- R.Chowdhary et al., “Dragon Promoter Mapper: A Bayesian framework for modeling promoter structures”, *Bioinformatics*, 2006. To appear.
- R Chowdhary et al., “Finding functional promoter motifs by computational methods: A word of caution”, *International Journal of Bioinformatics Research and Applications*, 2:282—288, 2006.
- R Chowdhary et al,. “Promoter modeling: The case study of mammalian histone promoters”, *Bioinformatics*, 21:2623--2628, 2005.