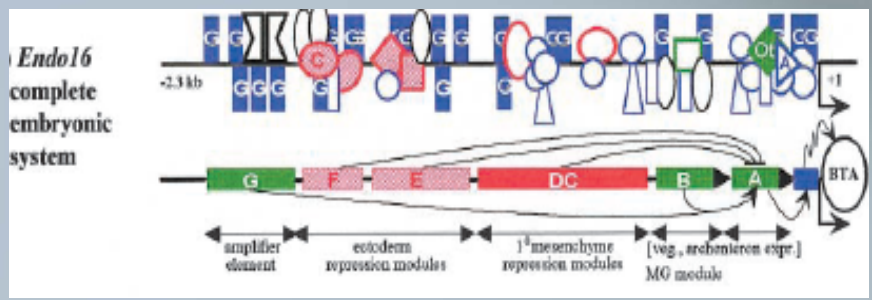
Using hexamers to predict cis-regulatory motifs in Drosophila

Authors: Bob Y. Chan and Dennis Kibler

Presenter: Ji Lee

Cis-Regulatory Modules

- CRMs are clusters of TFBS
- Two Types
  - Promoters
    - Proximal promoters
      - TATA box, CAAT box, TSS, DPE
  - Enhancers
    - Can be far away from regulated gene

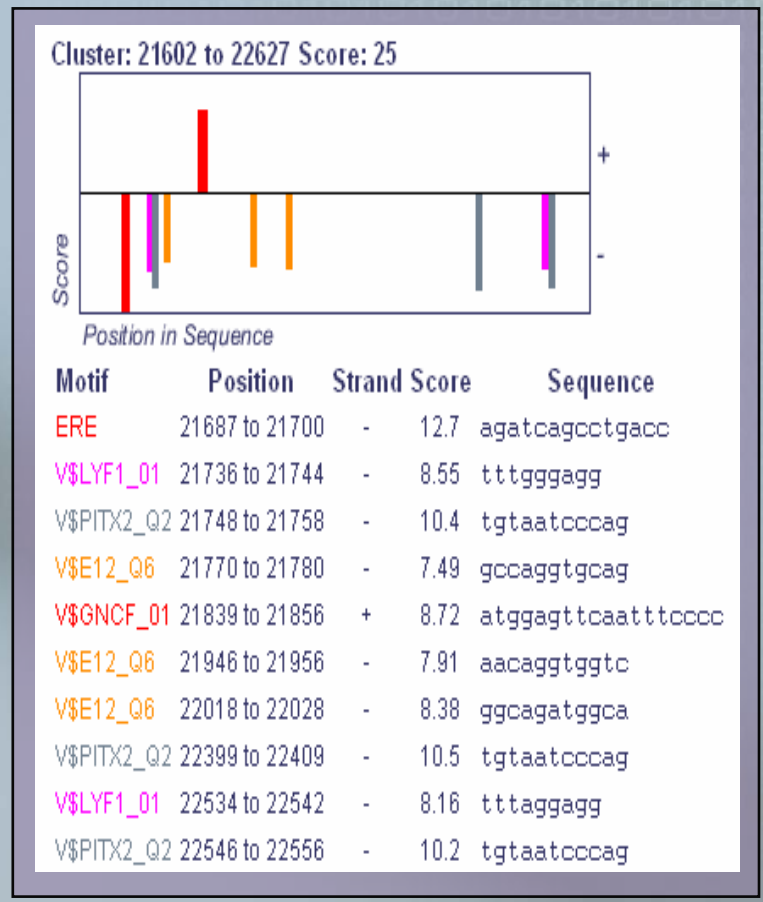


Predicting CRMs

- Classified by information used
- Searching by signal
  - Example: Identification of clustered motifs
- Phylogenetic footprinting
  - Conservation of regulatory regions between species
- Searching by content (ab initio)
  - Differentiating between CRM and non-CRM sequences based on sequence characteristics

Searching By Signals: Cluster Buster

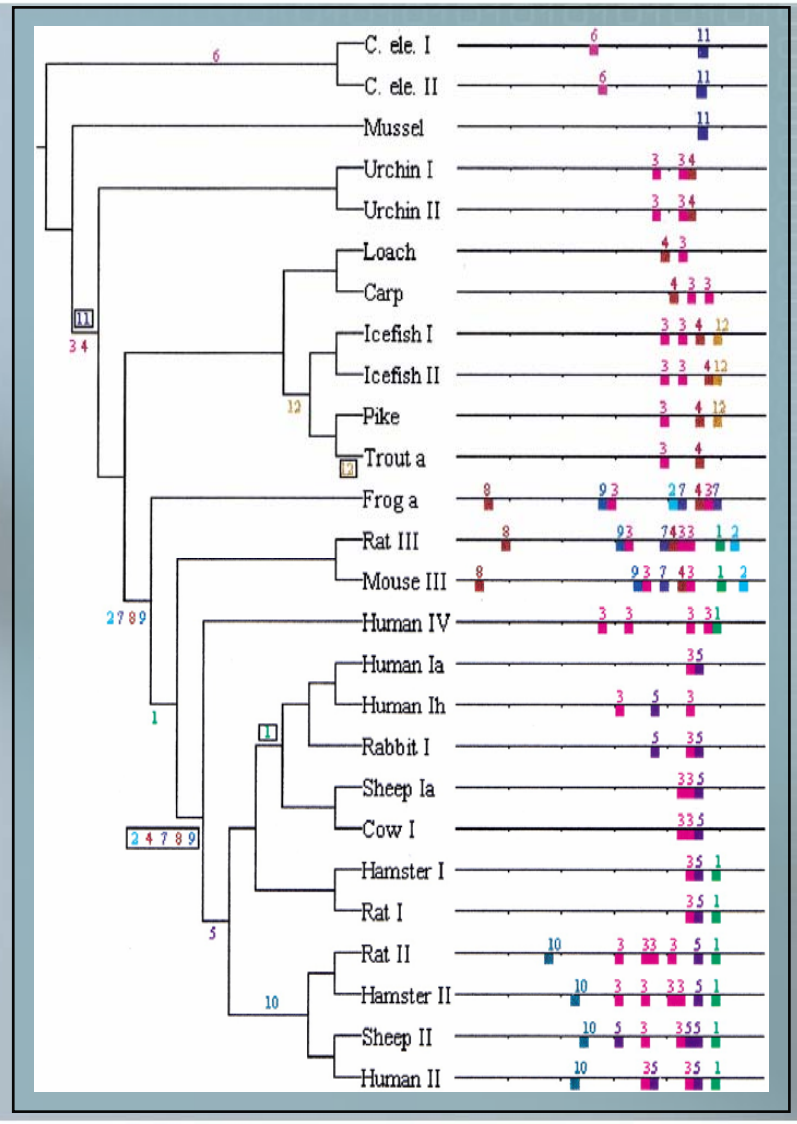
- Example of a “Search by signal” method
- Tries to identify motif-dense regions
- Log-likelihood scores
- Optimize clusters



Motif	Position	Strand	Score	Sequence
ERE	21687 to 21700	-	12.7	agatcagcctgacc
V$LYF1_01	21736 to 21744	-	8.55	tttgggagg
V$PITX2_Q2	21748 to 21758	-	10.4	tgtaatcccag
V$E12_Q6	21770 to 21780	-	7.49	gccaggtgcag
V$GNCF_01	21839 to 21856	+	8.72	atggagttcaatttcccc
V$E12_Q6	21946 to 21956	-	7.91	aacaggtggtc
V$E12_Q6	22018 to 22028	-	8.38	ggcagatggca
V$PITX2_Q2	22399 to 22409	-	10.5	tgtaatcccag
V$LYF1_01	22534 to 22542	-	8.16	tttaggagg
V$PITX2_Q2	22546 to 22556	-	10.2	tgtaatcccag

Motif Recognition Using Phylogenetic Footprinting

- ClustalW
  - Problematic when looking for shorter sequences
- Dialign
  - Improvement over ClustalW...
  - ...But still problematic
- MEME
  - Motif discovery program
- FootPrinter



Searching by Content Algorithms

- Fluffy-tail test
  - Statistical analysis of nucleotide in lists of variant length words
- LWF – Local Word Frequency
  - Analyzes word frequencies within a sliding window (local)
  - Disadvantage: Depends on word frequencies not on the words
- PromFind
  - Tries to find similar hexamer frequencies of known promoters in target sequences
  - Restrictive in nature- one promoter per input sequence but not so for enhancers

Comparison of Algorithms

Table 1: Key aspects of HexDiff and other algorithms. The table shows the knowledge used and the parameters required by the different algorithms.

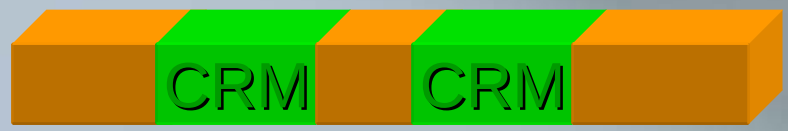
Algorithm	Knowledge Used	Parameters
HexDiff	CRM Locations	Number of hexamers in H_d Window size Window score threshold
Ahab	PWMs	Window size Free energy cutoff Order of background model
Cluster Buster	PWMs	Motif score threshold Gap parameter Cluster score threshold Residue abundance range
MSCAN	PWMs	Motif score threshold Window size Minimum hits Maximum hits
MCAST	PWMs	Motif score threshold Maximum allowed distance between adjacent hits Pseudocount weight
LWF	CRM Locations	String length Number of mismatches Detection window size Maximum number of channels Channels equalized Profile cutoff Peak width cutoff Smoothing window

HexDiff Summary

- CRM sequences vs. non-CRM sequences
- Model
  1. Training set built with sequences containing known CRMs
  2. Calculate word frequencies for all 4^6 hexamers
  3. Calculate an enrichment score for each hexamer
  4. Extract set H_d of highly represented hexamers
  5. Calculate a window score for each position *i* in a target sequence
  6. Filter window scores against a chosen threshold score
  7. Filter out “impossibly short” CRM predictions

Training HexDiff: Building

- Use sequences with known CRMs
- Split sequences into two subsets
  - Positive training set
    - Aggregate of all known CRMs extracted from sequences
  - Negative training set
    - Everything not in the positive set



Training HexDiff: Processing

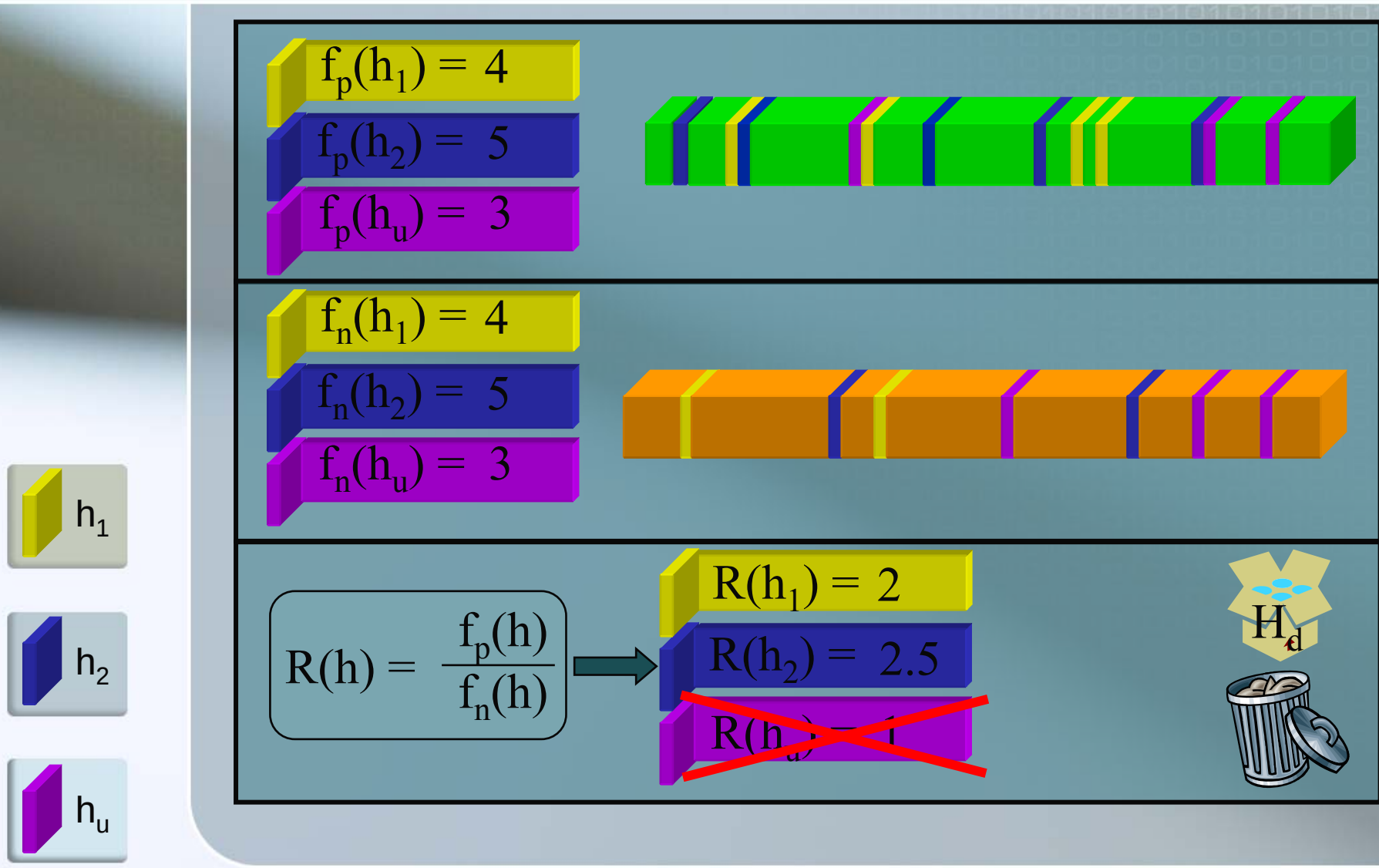
- Calculate frequency of all possible hexamers (4^6 total) on both strands
- Calculate enrichment score *R* for each hexamer

$$R(h) = \frac{f_p(h)}{f_n(h)}$$

- Select only the hexamers with the highest enrichment scores for set H_d
  - Assumption:

Increased representation $\approx$ Determinant between CRM and non-CRM sequences

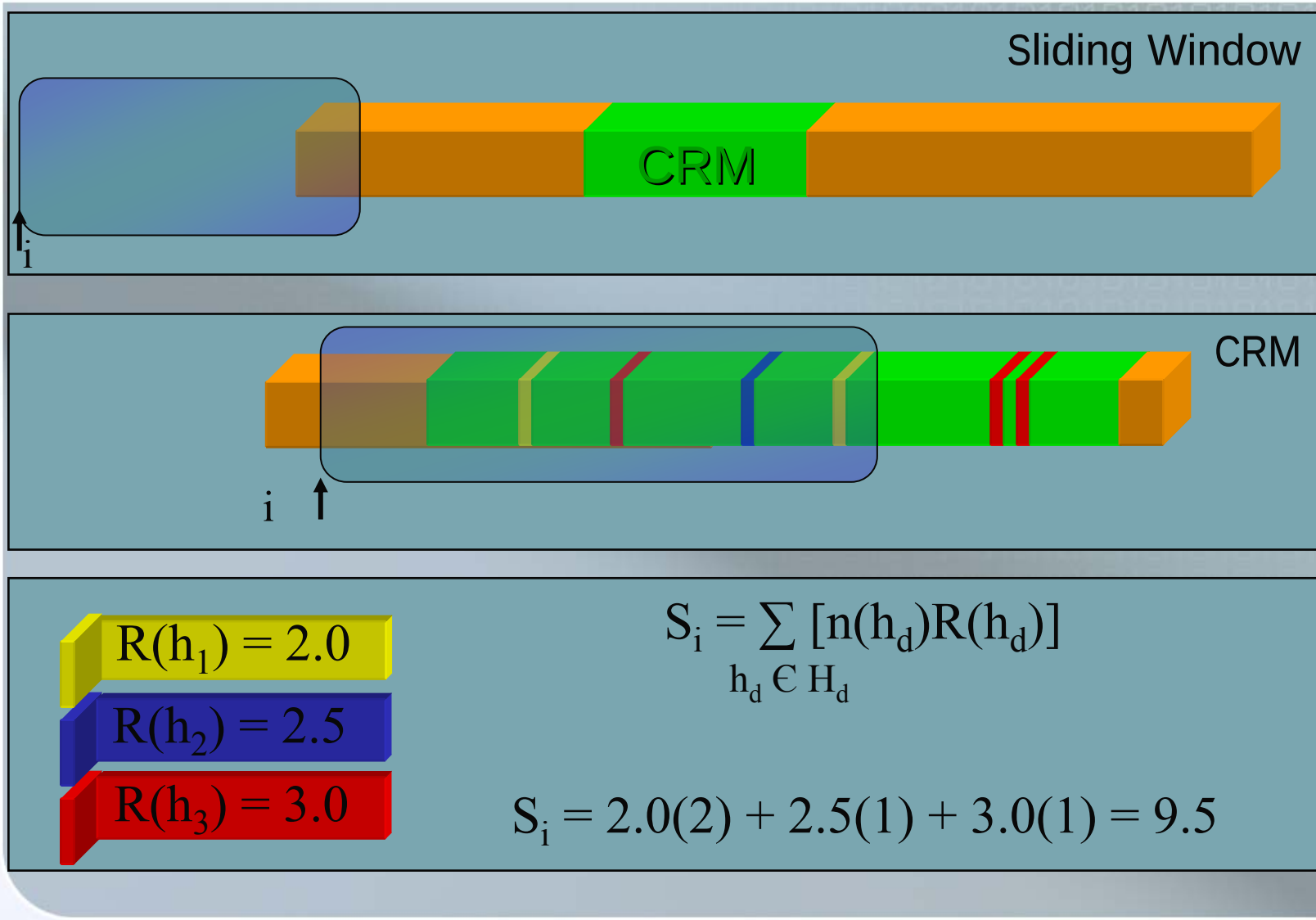
Training HexDiff: Processing



HexDiff At Work

- Sliding window of size w starting at a base i
- Count all occurrences of each h_d in H_d for the current window, $n(h_d)$
- Multiply $n(h_d)$ by $R(h_d)$
- Sum all component scores to find the score S_i for the current window
- Repeat for all i, advancing 1 base at a time

HexDiff At Work



Evaluation: LOOCV

- Leave-One-Out Cross-Validation
- Input set of 16 CRM-characterized sequences
  - 16 runs of algorithm, 1 per input sequence
  - “Test” sequence systematically changed each iteration
  - Rest of set becomes the basis for the training set

Choosing the Design and Parameters

- Designed to minimize the number of mandatory user-inputted parameters
  - Breeds conceptual simplicity
  - Avoids overfitting
- Test run uses LOOCV-optimized parameters
  - Size of H_d
  - Size of sliding window
  - Threshold score
- N-mer size and mismatches

Evaluation: Algorithm Comparison

- Assessing the accuracies of each algorithm
  - Sensitivity
    - TP/(TP + FN)
  - Specificity
    - TN/(TN + FP)
  - Positive Predictive Values (PPV)
    - TP/(TP + FP)
  - Matthews Correlation Coefficient

$$C = \frac{t_p t_n - f_p f_n}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}}$$

Evaluation: Algorithm Comparison

	TP	FP	TN	FN	Sensitivity	Specificity	PPV
HexDiff	22548	40007	602501	35751	38.68%	93.77%	36.05%
Ahab	12862	10488	632020	45437	22.06%	98.37%	55.08%
Cluster Buster	19883	33339	609169	38416	34.11%	94.81%	37.36%
MSCAN	15771	58679	583829	42528	27.05%	90.87%	21.18%
MCAST	28009	194633	447875	30290	48.04%	69.71%	12.58%
LWF	7436	61165	581343	50863	12.75%	90.48%	10.84%

Evaluation: Algorithm Performances on Test Set

- Test run on a 16 sequence set containing 52 characterized CRMs
  - Cumulative scores are the sum of all CCs

Gene	CRMs	HexDiff	Ahab	Cluster Buster	MSCAN	MCAST	LWF
btd	1	0.70	0.57	0.19	0.01	0.07	0.10
ems	3	0.00	0.00	-0.03	0.12	-0.01	-0.01
eve	6	0.55	0.63	0.65	0.50	0.41	0.06
fkh	1	-0.03	-0.02	-0.02	-0.04	-0.02	-0.01
ftz	5	0.40	0.28	0.28	0.07	0.16	0.08
gt	1	0.27	0.42	0.33	0.35	0.15	0.03
h	5	0.71	0.63	0.53	0.30	0.37	0.08
hb	2	0.35	0.63	0.39	0.34	0.24	0.04
hkb	1	0.51	0.00	-0.02	-0.02	-0.08	0.09
kni	3	0.55	0.55	0.39	0.37	0.23	-0.05
kr	3	0.43	0.00	0.77	0.20	0.11	-0.03
oc	2	0.70	-0.02	0.00	0.11	0.02	0.07
prd	7	0.01	-0.07	0.16	0.07	-0.04	0.05
run	6	0.27	0.16	0.08	0.08	0.02	0.07
slp1	3	-0.07	0.15	-0.04	0.00	0.07	0.01
tll	3	0.35	0.56	0.58	0.19	0.12	-0.04
Total	52	5.71	4.48	4.24	2.64	1.81	0.52

Evaluation: Novel CRMs

- 1 – Ahab, 2 – ClusterBuster, 3 – MSCAN, 4 – MCAST, 5 – LWF

Gene	Arm	Begin	End	Length	1	2	3	4	5	Matched
btd	X	9534921	9535192	271				*	*	
eve	2R	5492385	5493575	1190				*	*	eve_late2_mel
fkh	3R	24421705	24422385	680				*	*	
ftz	3R	2683060	2683406	346			*	*		
gt	X	2268347	2270179	1832		*	*	*		
gt	X	2290228	2290685	457	*	*	*	*	*	gt_23-bcd_mel
hb	3R	4503375	4503962	587			*	*	*	
hb	3R	4519805	4520172	367			*	*		
kni	3L	20628230	20628504	274	*	*	*	*		kni_+1_mel
prd	2L	12080435	12082316	1881	*			*	*	prd_bcd_mel
prd	2L	12089627	12089847	220				*	*	prd_1_mel
run	X	20488169	20488643	474	*	*	*	*		
run	X	20524260	20524722	462		*	*	*	*	
slp1	2L	3811050	3812092	1042				*	*	
slp1	2L	3822581	3823049	468				*	*	
slp1	2L	3824891	3825039	148	*	*	*	*	*	slp_A-bcd_mel
slp1	2L	3833433	3834671	1238		*	*	*	*	slp2_-3_mel
tll	3R	26680559	26683175	2616		*		*	*	tll_bcd_mel

Conclusion

- HexDiff utilizes local word frequencies in a biological context to predict CRMs
- Implementation of the method is in its infancy
  - More testing can only be catalyzed when implemenation is more robust
- May spawn variations of the method
- Many ways currently used to predict CRMs, but in the end there is a long way to go.

References

- Arnone, M. I. & Davidson, E. H. The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124**, 1851-1864 (1997).
- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16**, 412-424 (2000).
- Blanchette, M. & Tompa, M. Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting. *Genome Res.* **12**, 739-748 (2002).
- Frith, M. C., Li, M. C. & Weng, Z. Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucl. Acids Res.* **31**, 3666-3668 (2003).
- Nazina, A. & Papatsenko, D. Statistical extraction of Drosophila cis-regulatory modules using exhaustive assessment of local word frequency. *BMC Bioinformatics* **4**, 65 (2003).