Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting

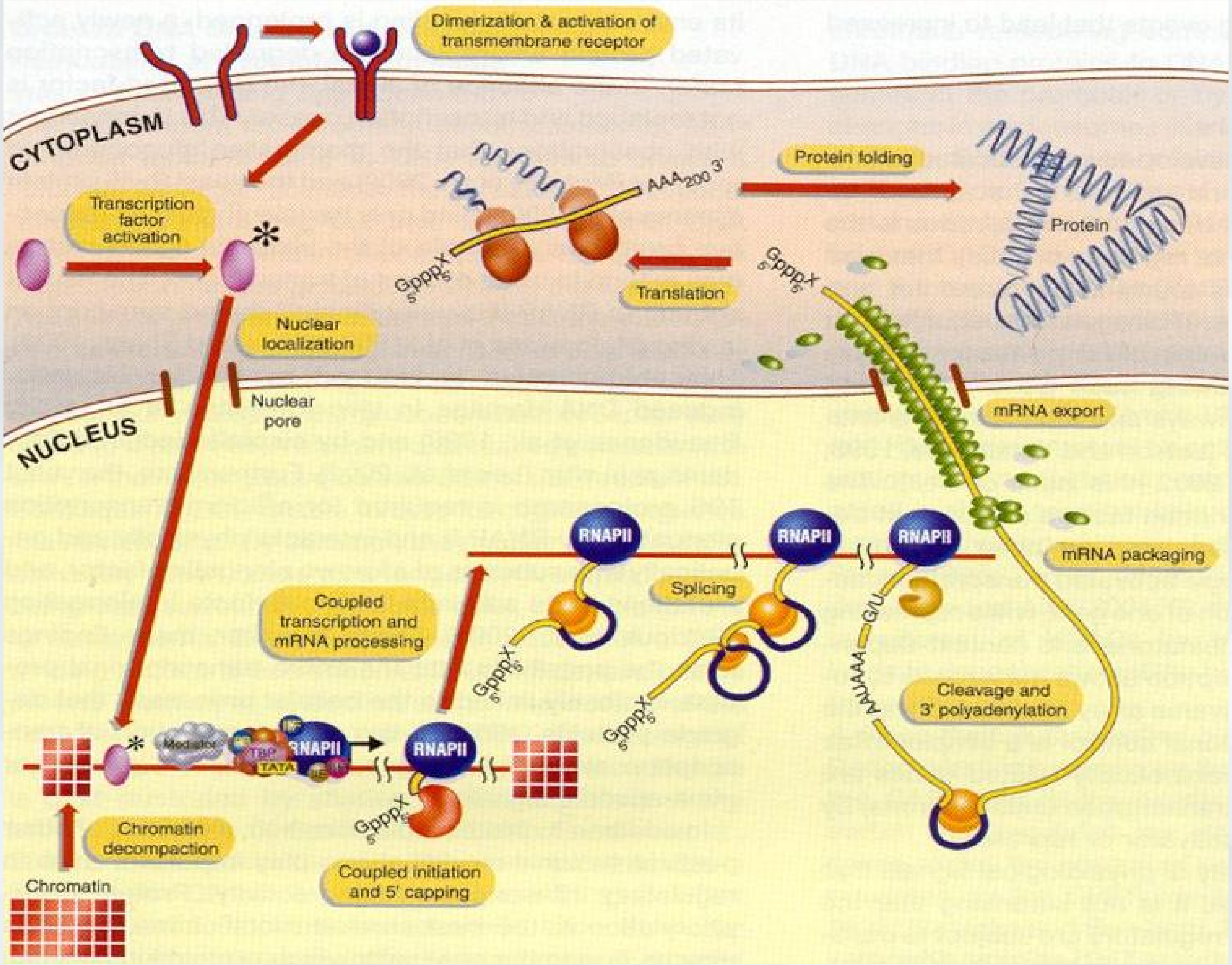
Mathieu Blanchette and Martin Tompa

Presented by Ben Bachman

What is a regulatory element?

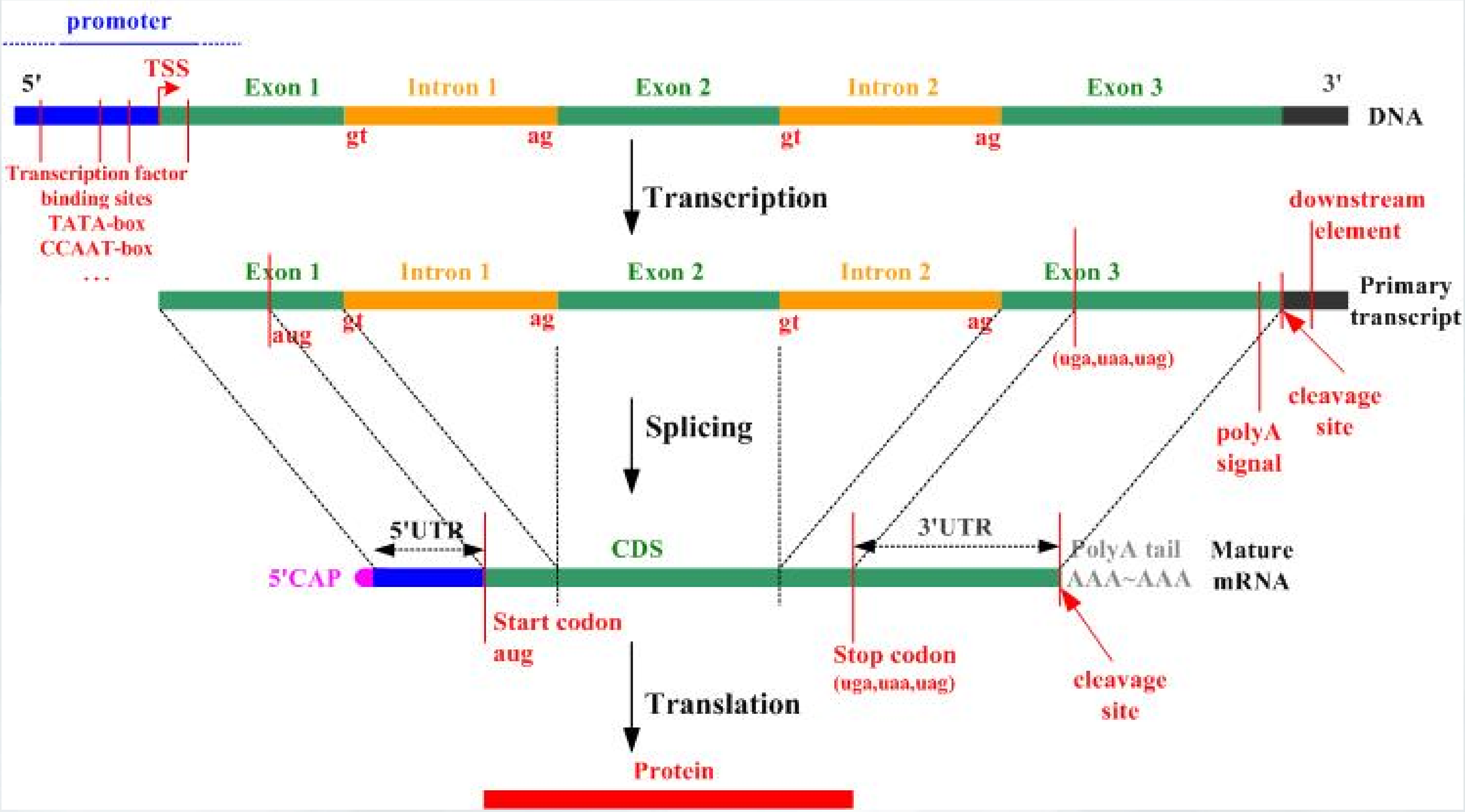
- In promoter region upstream of transcription
  - sometimes in introns/UTR
- Regulates gene expression
- Not expressed itself
- Are conserved through evolution
- Implicated in many diseases:
  - Asthma
  - Thallassemia - reduced hemoglobin
  - Rubinstein - mental and physical retardation
  - Many cancers
- Problem: different properties than exons

How does this fit into biology?



G. Orphnides and D. Reinberg (2002) A Unified Theory of Gene Expression. Cell 108: 439-451.

How does this fit into biology?



http://kachkeis.com/img/essay3_pic1.jpg

Goal: Detection of TF Binding Site

- Currently - analyze multiple promoters from coregulated genes, find conserved sequences
    - Problems?
        - Must find the coregulated genes
        - Not all genes are coregulated with another
- Instead - look at orthologous and paralogous genes in different species
    - Also uses evolutionary tree
    - Advantages:
        - Can work on single genes

Existing tools for the job?

- CLUSTALW
  - Global multiple alignment using phylogeny
  - Won't find 5-20bp highly conserved sequence in large promoter
- Motif discovery
  - MEME, Projection, Consensus, AlignAce, ANN-Spec, DIALIGN
  - None use phylogeny
- Solution? New tool "FootPrinter"

Method - Algorithm

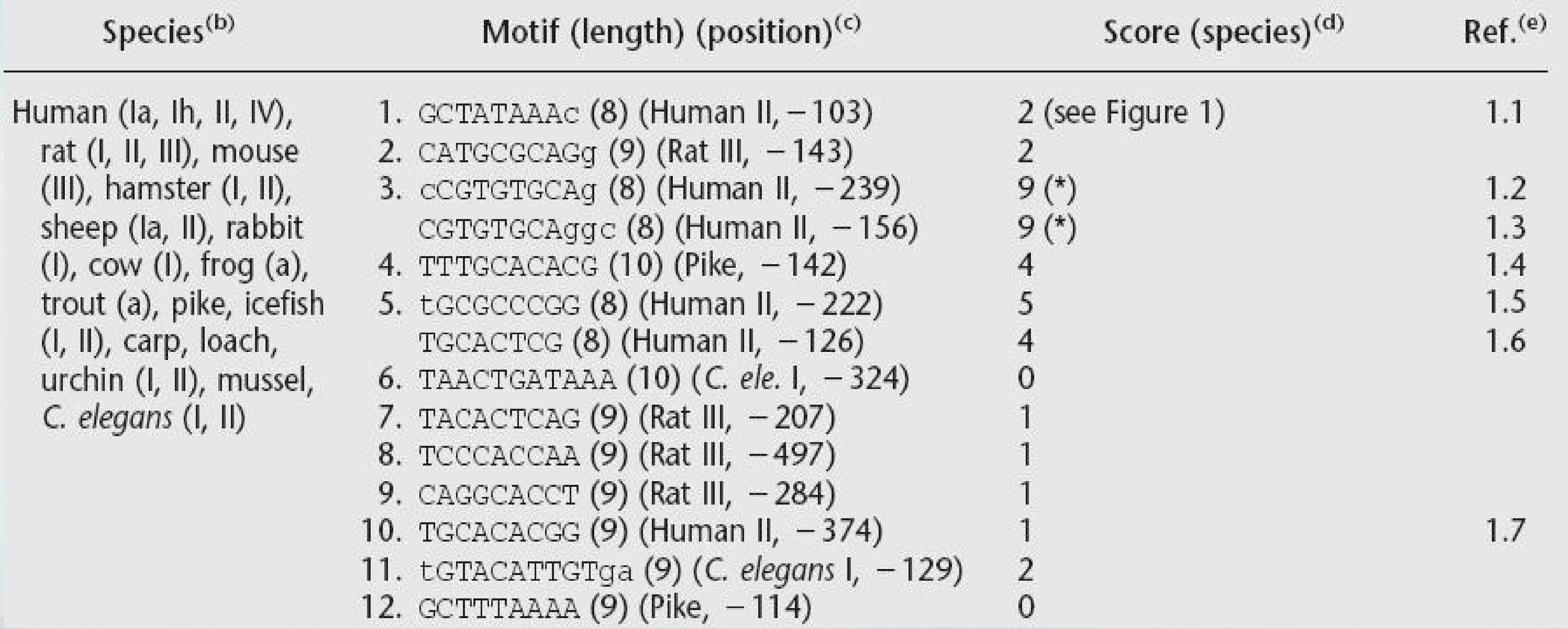
- Dynamic programming
- For two related leaves, find the most parsimonious way to have all possible k-mers (4^k) for some value of k
    - Continue up the tree
    - Return k-mers under max parsimony score for clade
    - Work back to find locations
- Only allowed point mutations
- Allows motif loss - part of parsimony score
- Requires little movement
- Validation - simulated data with no conserved sequences

Method - Data Collection

- Taken from public databases
  - Genbank
    - Some already annotated
    - Some build by authors
  - ACUTS database
- Took phylogeny from previous papers
  - Taken to be gene tree (no lateral transfer)
  - Could also base it on global alignment
  - Multifurcating trees

Results

- Metallothionein Gene family
    - many known regulatory elements
    - proteins bind heavy metals and detoxify, antioxidant
    - Ran FootPrinter on 590 bp region upstream

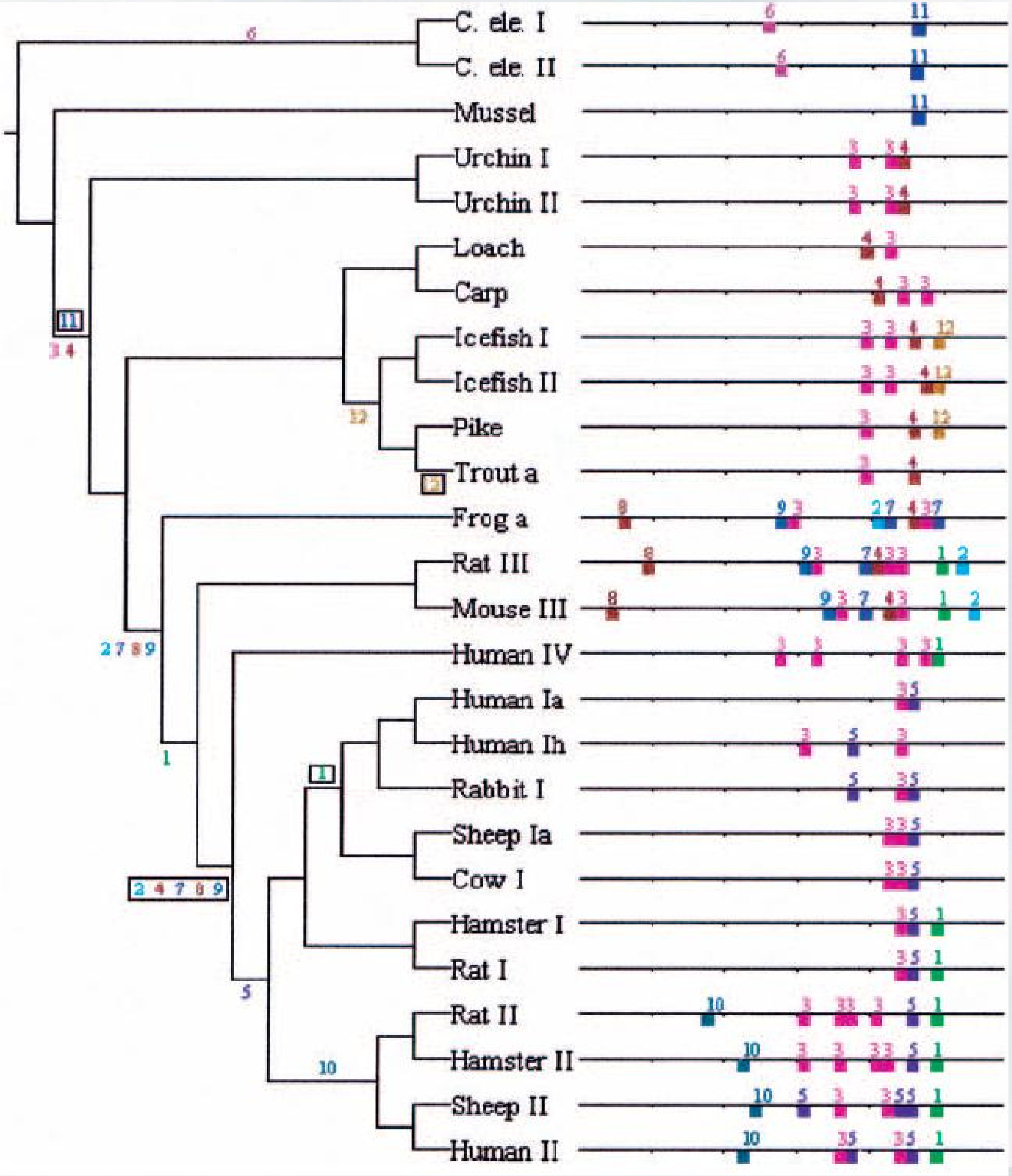


Species[b]	Motif (length) (position)[c]	Score (species)[d]	Ref.[e]
Human (Ia, Ih, II, IV), rat (I, II, III), mouse (III), hamster (I, II), sheep (Ia, II), rabbit (I), cow (I), frog (a), trout (a), pike, icefish (I, II), carp, loach, urchin (I, II), mussel, *C. elegans* (I, II)	1. GCTATAAAc (8) (Human II, −103)	2 (see Figure 1)	1.1
	2. CATGCGCAGg (9) (Rat III, −143)	2	
	3. cCGTGTGCAg (8) (Human II, −239)	9 (*)	1.2
	CGTGTGCAggc (8) (Human II, −156)	9 (*)	1.3
	4. TTTGCACACG (10) (Pike, −142)	4	1.4
	5. tGCGCCCGG (8) (Human II, −222)	5	1.5
	TGCACTCG (8) (Human II, −126)	4	1.6
	6. TAACTGATAAA (10) (*C. ele.* I, −324)	0	
	7. TACACTCAG (9) (Rat III, −207)	1	
	8. TCCCACCAA (9) (Rat III, −497)	1	
	9. CAGGCACCT (9) (Rat III, −284)	1	
	10. TGCACACGG (9) (Human II, −374)	1	1.7
	11. tGTACATTGTga (9) (*C. elegans* I, −129)	2	
	12. GCTTTAAAA (9) (Pike, −114)	0	

Results - Metallothionein

- Looked for 7, 8, 9, and 10-mers
  - vary parameters for statistical significance
- Found MREs, which bind to MTF-1 (required)
  - Metal Response Elements often have multiple copies
- Found TATA box, expected in most promoters
- Seven novel sequences
  - Non-mammallian

Results


6
11
3 4
12
12
2 7 8 9
1
1
2 4 7 8 9
5
10
C. ele. I
C. ele. II
Mussel
Urchin I
Urchin II
Loach
Carp
Icefish I
Icefish II
Pike
Trout a
Frog a
Rat III
Mouse III
Human IV
Human Ia
Human Ih
Rabbit I
Sheep Ia
Cow I
Hamster I
Rat I
Rat II
Hamster II
Sheep II
Human II
6 11
6 11
11
3 3 4
3 3 4
4 3
4 3 3
3 3 4 12
3 3 4 12
3 4 12
3 4
8 9 3 2 7 4 3 7
8 9 3 7 4 3 3 1 2
8 9 3 7 4 3 1 2
3 3 3 3 1
3 5
3 5 3
5 3 5
3 3 5
3 3 5
3 5 1
3 5 1
10 3 3 3 3 5 1
10 3 3 3 3 5 1
10 5 3 3 5 5 1
10 3 5 3 5 1

Results - Insulin

- Smaller set of data available
  - Closely related
- Lower max parsimony score
  - 8,9,10-mers with 0,1,1 parsimony
- Found four known true positives
- Missed others
  - Too many mutations
- Claim: with more data, more would be found
- Claim: search for longer motifs, more mutations
  - 12 and 15-mers didn't find anything

Results

- C-myc Intron 2
  - Intron previously shown to have part in regulation
  - ~1000 bp
  - Found 10 conserved motifs
  - All but 1 are novel
- Other regions
  - C-myc Promoter - four known, five novel
  - C-fos Promoter - five known, three novel
  - C-fos first intron - three motifs found
  - Growth hormone 1 - five known, three novel

Discussion - False Negatives

- Binding Sites found only in a small clade
    - would need many samples
- Too short
    - would result in many false positives
- Insertions/Deletions
    - FootPrinter can handle, but false positives
- High parsimony score
- Implementation can't handle:
    - Two conserved sequences with unconserved in between - TGACnnnnnGCGG
    - Different substitution costs

Discussion - Other Methods

- CLUSTALW fails at highly diverged sequences
- DIALIGN mostly gets the same results as FootPrinter
  - FootPrinter did much better in metallothionein
  - Claim: Footprinter does better with weakly conserved motifs
- MEME
  - ignores phylogeny
  - ignores motif order
  - did well in general, but not for metallothionein
- Quantitative comparison difficult
  - No definitive list to determine false positives/negatives

Comments

- No experimental validation of novel motifs
    - Can't tell if novel sequences are true positives
- Several unproven claims
    - would beat others in certain situations
- No quantification of performance
- Parameter selection unexplored

Conclusions

- May be better than other algorithms
  - Not definitively shown
  - Only one data set showed improvement
- More sequence would be useful
  - Most sequencing focuses on exome

Six years later...

- google scholar results
  - 168 Citations on this Footprinter paper, 92 on a later one
  - 228 hits for "FootPrinter"
- No where near CLUSTALW or DIALIGN hit counts, but seeing usage

- Example -
  - http://genome.cs.mcgill.ca/cgi-bin/FootPrinter3.0/FootPrinterInput2.pl