Transcriptional Regulatory Elements
Detection and Evolutionary Analysis

Wolfgang Otto

Institute of Computer Science
University of Leipzig

November 29, 2011

The Beginning

The Beginning

Question:

The Beginning

Question:

- "What and why am I?"

The Beginning

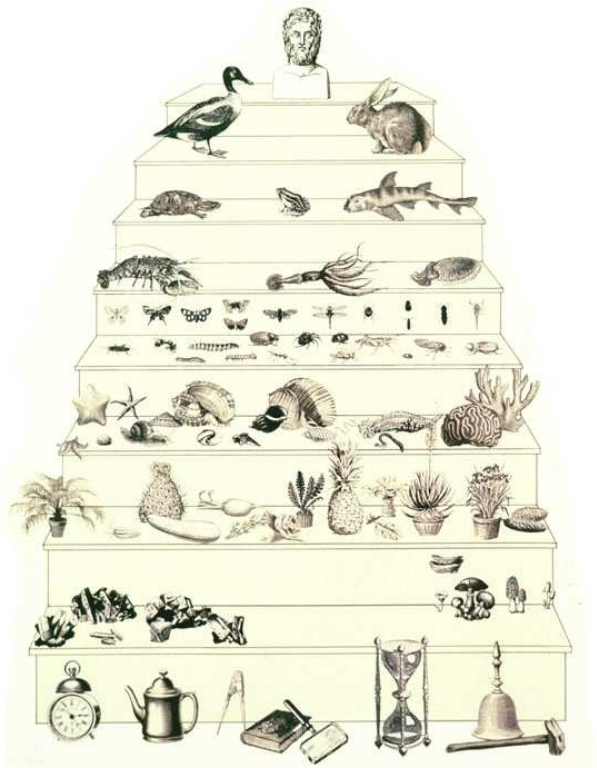
Question:

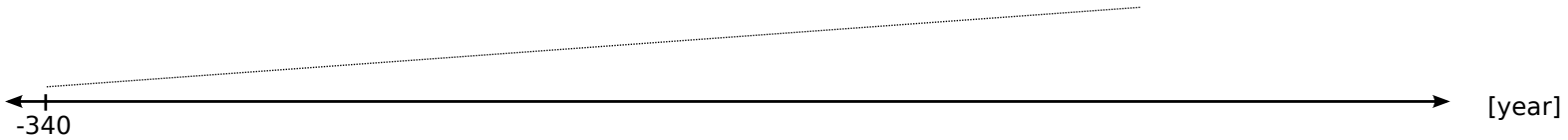
- "What and why am I?"
- or less phylosophical: "How works evolution and development?"

The Evolution of Evolution

$\sim 2,350$ years ago:

- "Scala Naturae" of Aristotle (384 BC – 322 BC)
- classification from imperfect inanimate matter to most perfect form

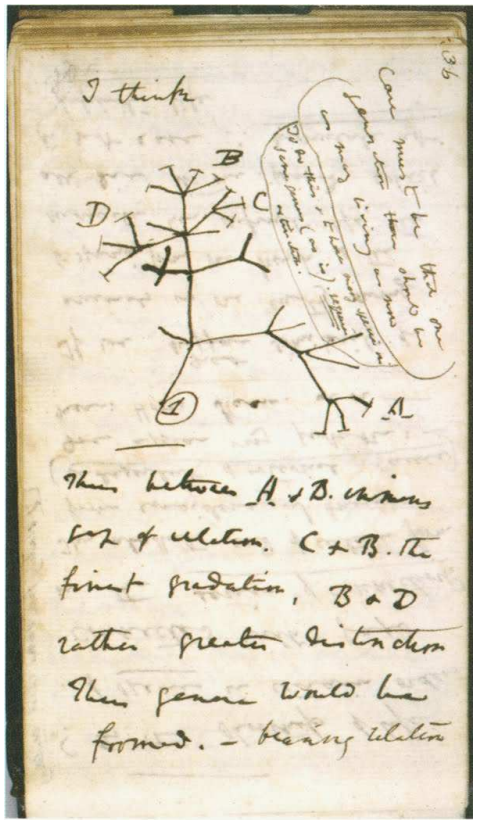


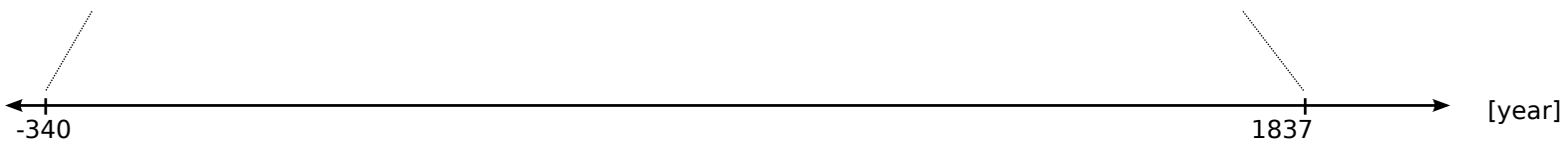


The Evolution of Evolution

174 years ago:

- "Transmutation of Species" of Charles Darwin (1809 – 1882)
- first evolutionary tree

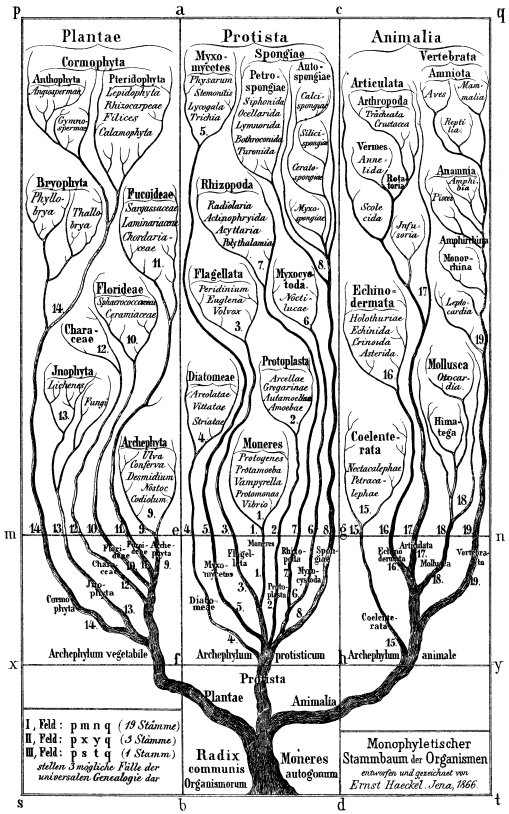


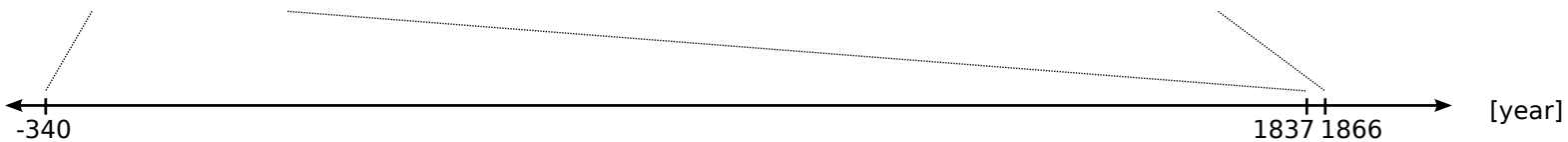


The Evolution of Evolution

145 years ago:

- "Generelle Morphologie der Organismen" of Ernst Haeckel (1834 – 1919)
- first single tree of all forms of life

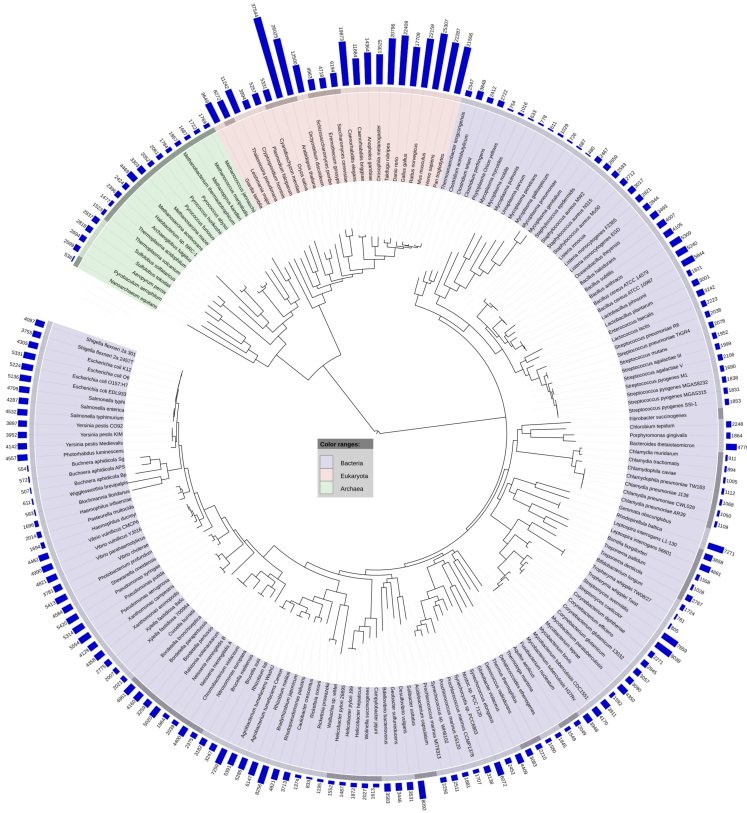


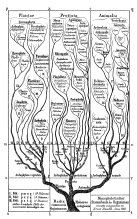


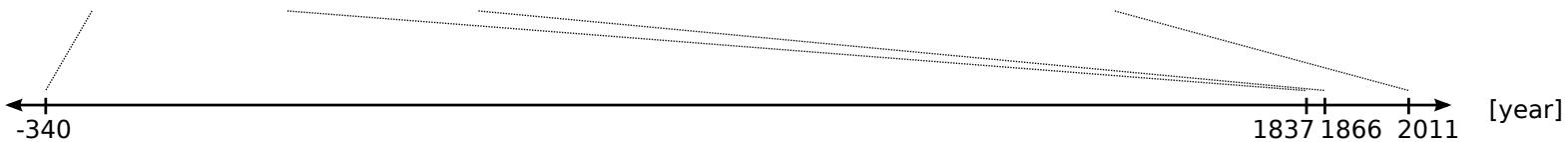
The Evolution of Evolution

today

- "Tree of Life" from European Molecular Biology Laboratory
- based on known genomes







Mechanisms of Evolution and Development

- end of 1950s: body plan of each life form is encoded in genomes

Mechanisms of Evolution and Development

- end of 1950s: body plan of each life form is encoded in genomes
- genome projects: differences between genomes are rather small

Mechanisms of Evolution and Development

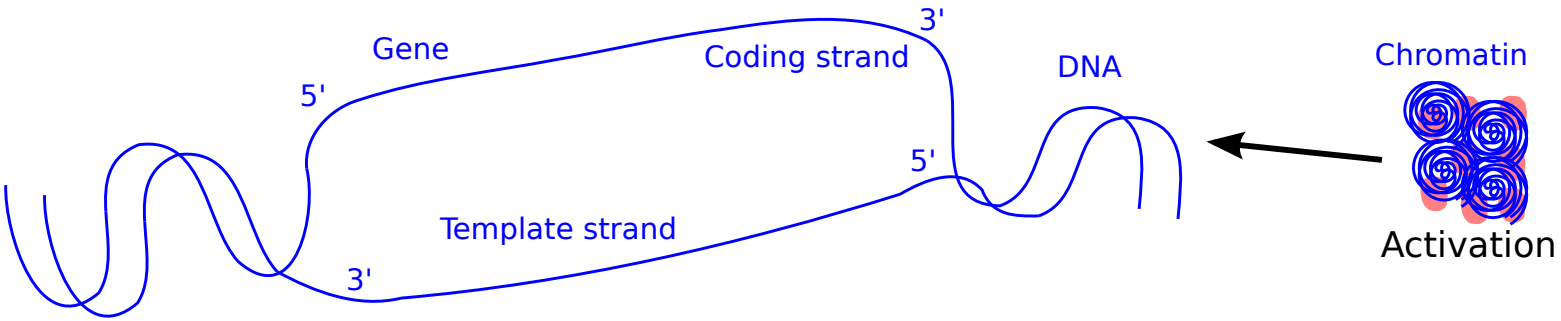
- end of 1950s: body plan of each life form is encoded in genomes
- genome projects: differences between genomes are rather small
- today: time, location and amount of gene products are responsible for causal differences

Mechanisms of Evolution and Development

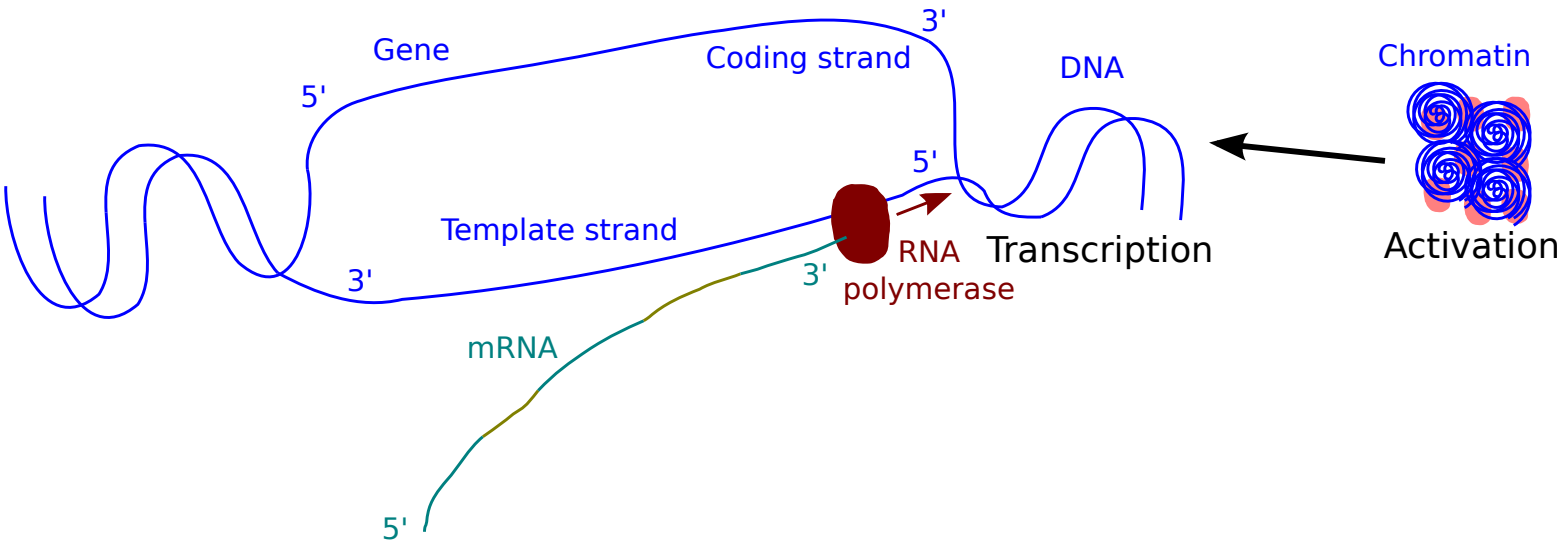
- end of 1950s: body plan of each life form is encoded in genomes
- genome projects: differences between genomes are rather small
- today: time, location and amount of gene products are responsible for causal differences
- $\Rightarrow$ regulation of gene expression is basis for differentiation, morphogenesis and versatility and adaptability of any organism

Gene Expression and Regulation

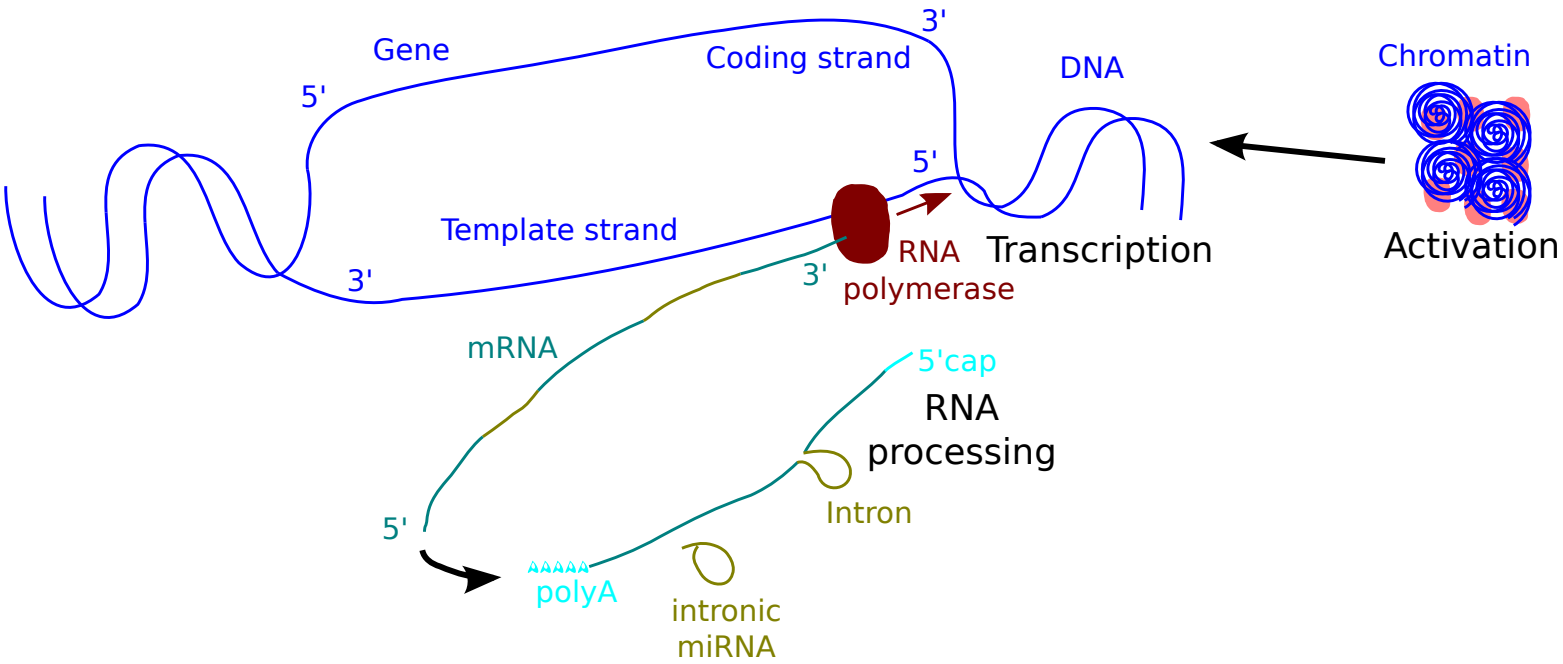
Gene Expression and Regulation



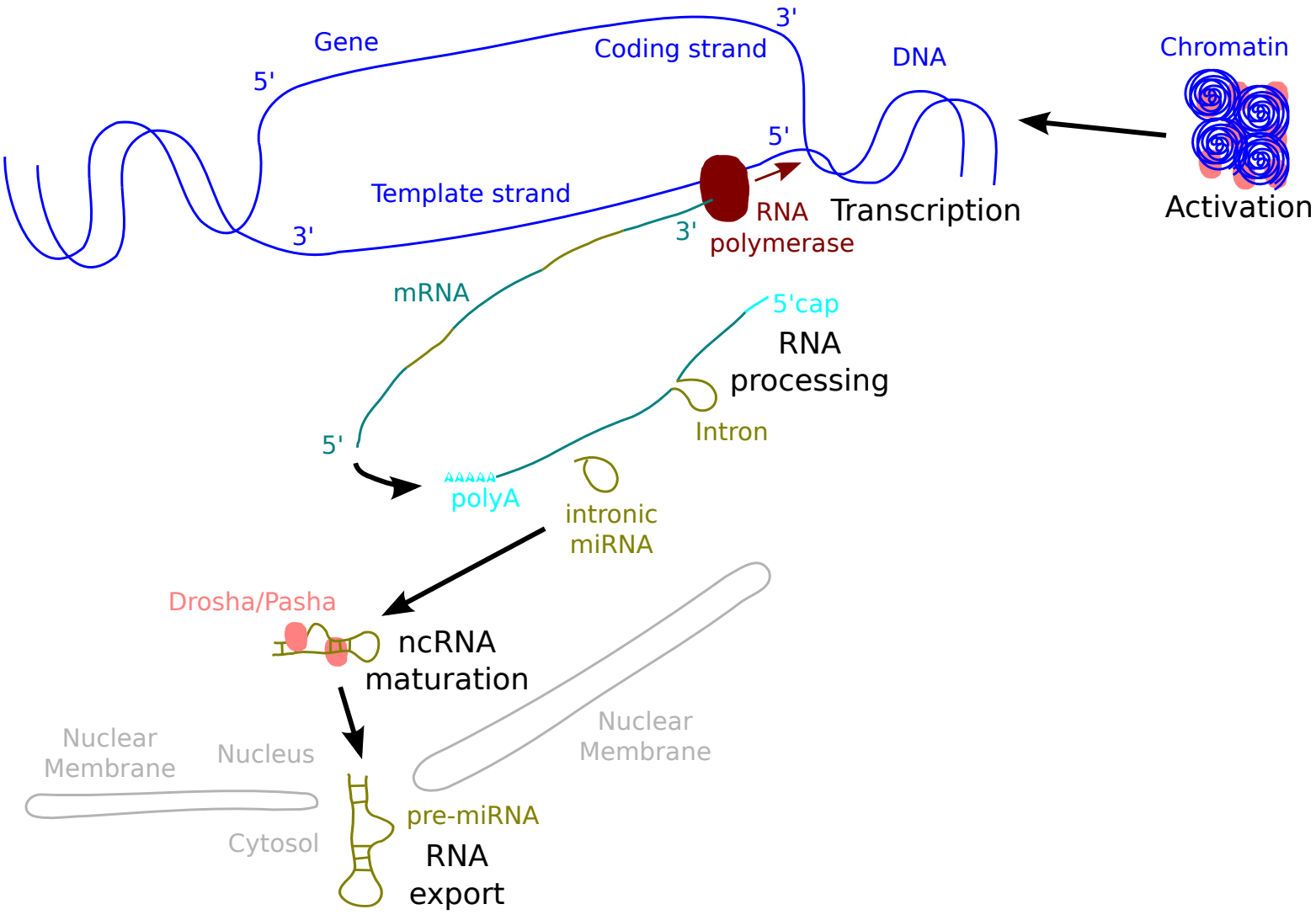
Gene Expression and Regulation



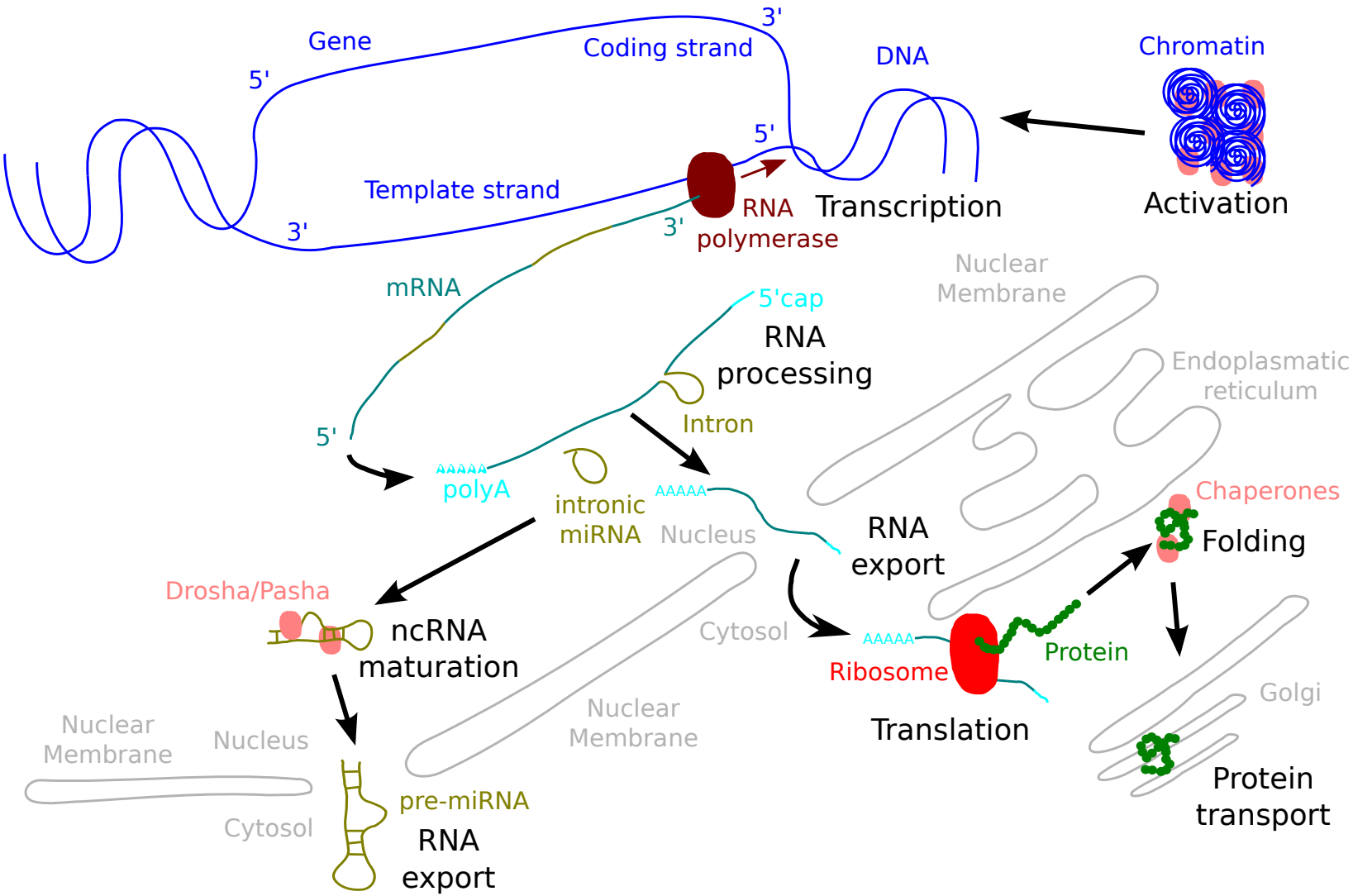
Gene Expression and Regulation



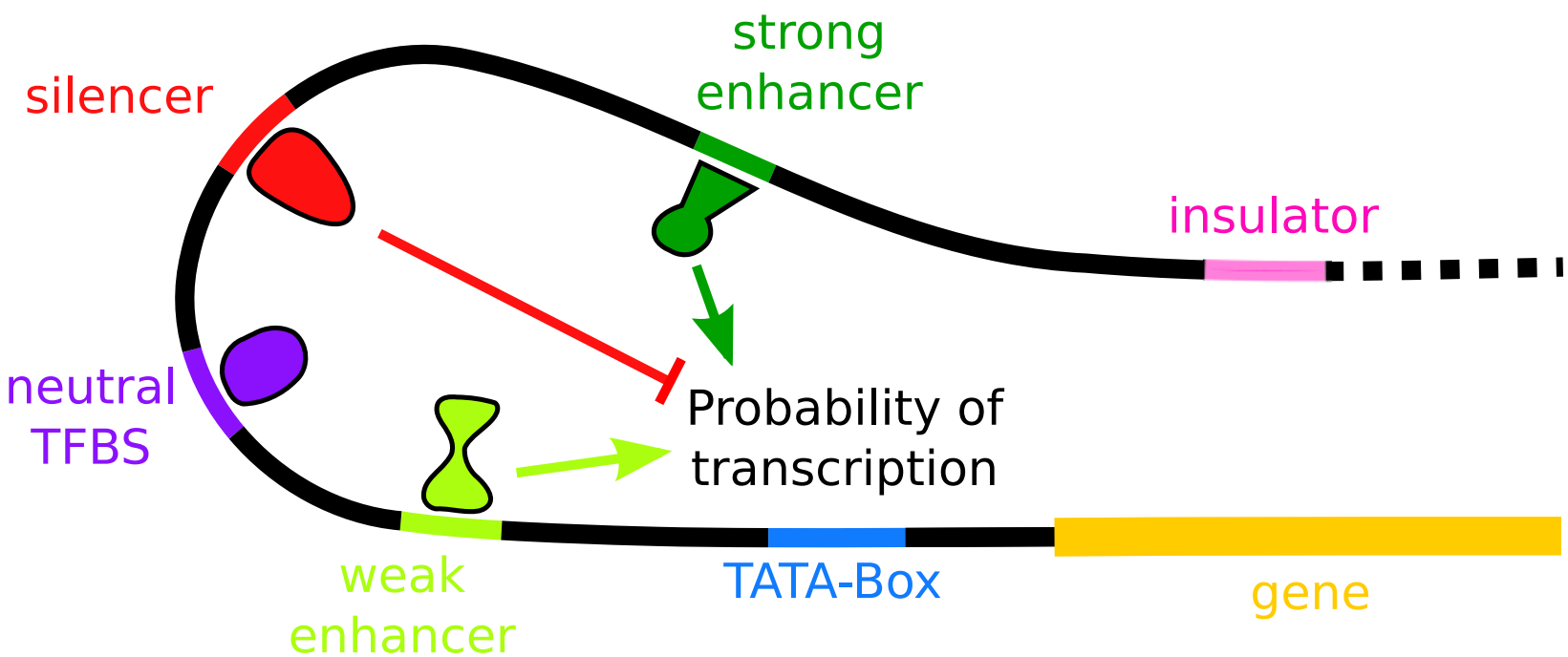
Gene Expression and Regulation



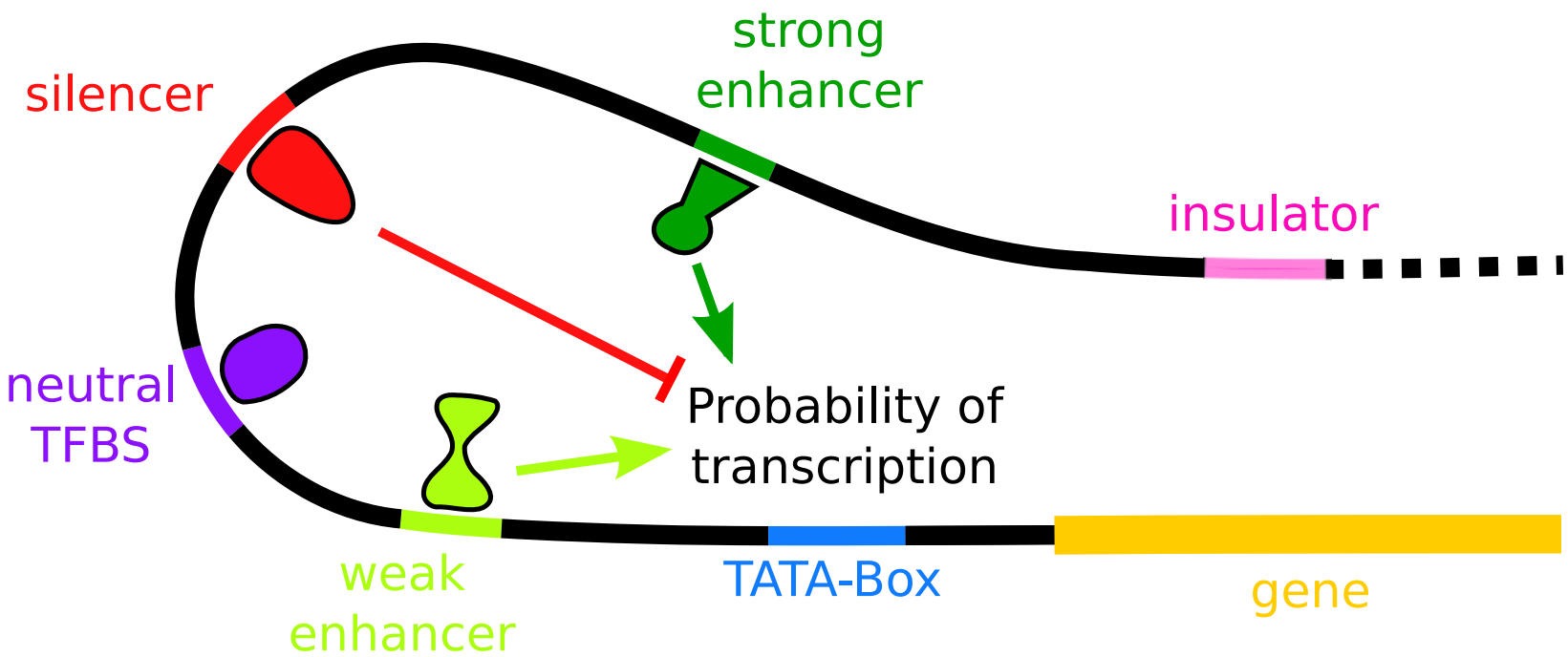
Gene Expression and Regulation



Transcriptional Regulation

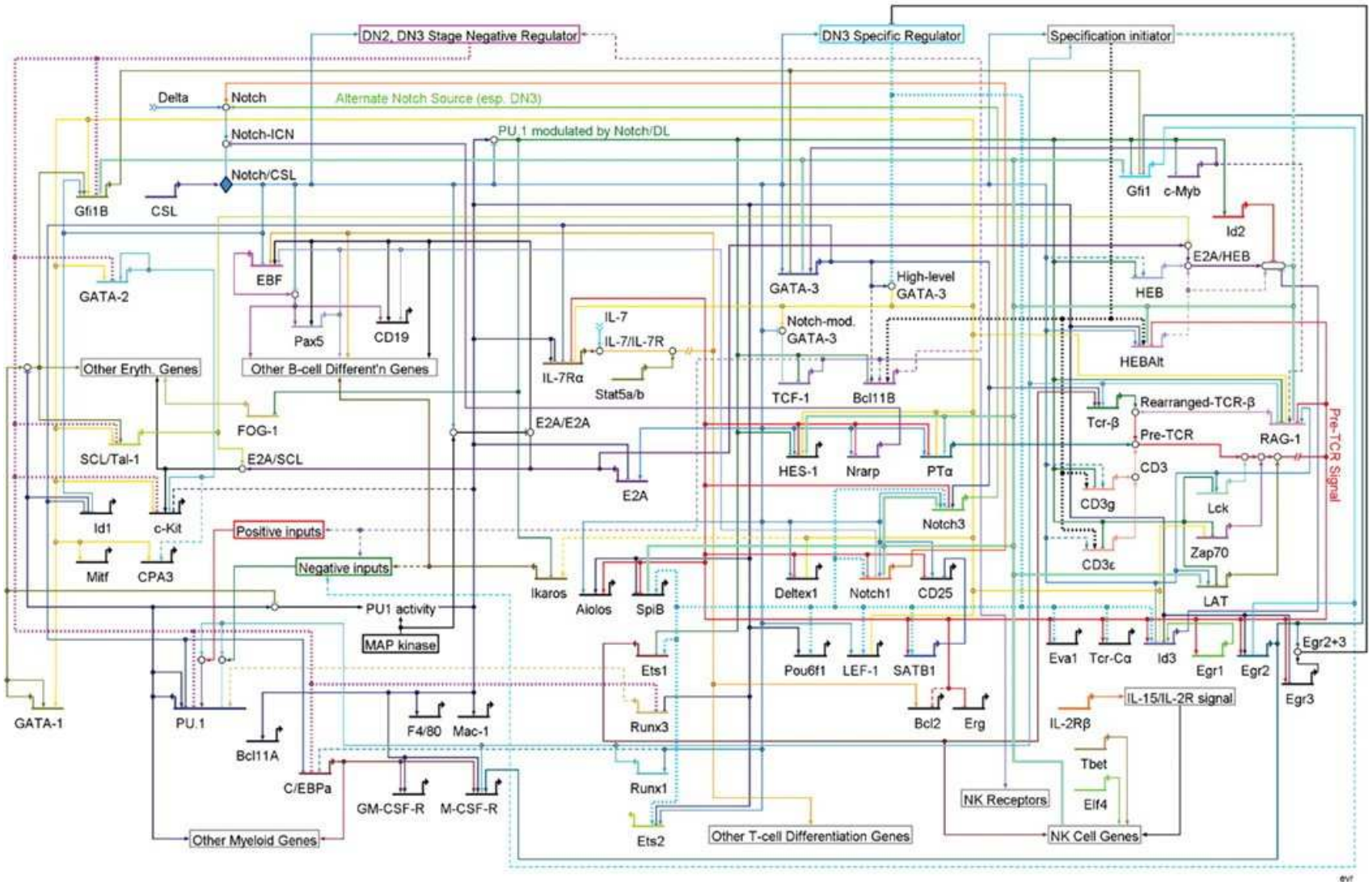


Transcriptional Regulation



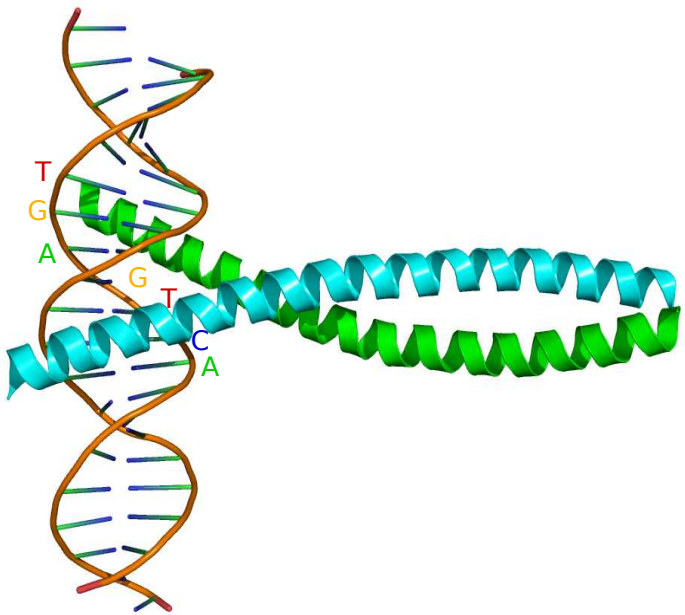
$$|\text{gene}| = f(\text{silencer}, \text{weak enhancer}, \text{strong enhancer}, \text{neutral TFBS})$$

Transcriptional Regulation



Regulatory Elements

TF / DNA Complex



Regulatory Elements

TF / DNA Complex

Binding Sites

Pos	1	2	3	4	5	6	7
Site 1	T	G	A	G	T	C	A
Site 2	T	G	A	C	T	A	A
Site 3	T	A	A	G	T	T	A
Site 4	T	G	A	C	C	C	A

Regulatory Elements

TF / DNA Complex

Binding Sites

Pos	1	2	3	4	5	6	7
Site 1	T	G	A	G	T	C	A
Site 2	T	G	A	C	T	A	A
Site 3	T	A	A	G	T	T	A
Site 4	T	G	A	C	C	C	A

Jundm2
Variable spacer length

TGACGTCA ✓

TGACTCA ✓
TGAGTCA ✓

ERRalpha
Position interdependence

CAAGGTCA ✓
AGGGGTCA ✓

CAGGGTCA ✗
CGGGGTCA ✗

Zfp187
Alternate recognition interfaces

ATGTACTAAT ✓

GCCCTTGTCC ✓

Summary: Gene Regulation

- basis for differentiation, morphogenesis and versatility and adaptability of any organism

Summary: Gene Regulation

- basis for differentiation, morphogenesis and versatility and adaptability of any organism
- mainly performed by binding of *trans*-regulatory factors (transcription factors, TF) to *cis*-regulatory elements (transcription factor binding sites, TFBS)

Summary: Gene Regulation

- basis for differentiation, morphogenesis and versatility and adaptability of any organism
- mainly performed by binding of *trans*-regulatory factors (transcription factors, TF) to *cis*-regulatory elements (transcription factor binding sites, TFBS)
- mutations at TFBS have potential for immense changes

Summary: Gene Regulation

- basis for differentiation, morphogenesis and versatility and adaptability of any organism
- mainly performed by binding of *trans*-regulatory factors (transcription factors, TF) to *cis*-regulatory elements (transcription factor binding sites, TFBS)
- mutations at TFBS have potential for immense changes
- $\Rightarrow$ research of regulatory elements is one of main fields in life sciences

Outlook

Detection of Regulatory Elements

Evolutionary Analysis of Regulatory Elements

Outlook

Detection of Regulatory Elements

- knowledge about location is important for understanding motif preference, tissue-specific regulation and evolution

Evolutionary Analysis of Regulatory Elements

Outlook

Detection of Regulatory Elements

- knowledge about location is important for understanding motif preference, tissue-specific regulation and evolution
- experimental detection of TFBS is limited to small number of cases, need for computer based methods

Evolutionary Analysis of Regulatory Elements

Outlook

Detection of Regulatory Elements

- knowledge about location is important for understanding motif preference, tissue-specific regulation and evolution
- experimental detection of TFBS is limited to small number of cases, need for computer based methods
- $\Rightarrow$ `Tracker`-Algorithm

Evolutionary Analysis of Regulatory Elements

Outlook

Detection of Regulatory Elements

- knowledge about location is important for understanding motif preference, tissue-specific regulation and evolution
- experimental detection of TFBS is limited to small number of cases, need for computer based methods
- $\Rightarrow$ `Tracker`-Algorithm

Evolutionary Analysis of Regulatory Elements

- TFBS undergo slightly divergence and turnover, transcriptional output remains conserved

Outlook

Detection of Regulatory Elements

- knowledge about location is important for understanding motif preference, tissue-specific regulation and evolution
- experimental detection of TFBS is limited to small number of cases, need for computer based methods
- $\Rightarrow$ `Tracker`-Algorithm

Evolutionary Analysis of Regulatory Elements

- TFBS undergo slightly divergence and turnover, transcriptional output remains conserved
- investigation of molecular evolution of TFBS can reveal timing/kind of evolutionary changes that affect gene regulation

Outlook

Detection of Regulatory Elements

- knowledge about location is important for understanding motif preference, tissue-specific regulation and evolution
- experimental detection of TFBS is limited to small number of cases, need for computer based methods
- $\Rightarrow$ `Tracker`-Algorithm

Evolutionary Analysis of Regulatory Elements

- TFBS undergo slightly divergence and turnover, transcriptional output remains conserved
- investigation of molecular evolution of TFBS can reveal timing/kind of evolutionary changes that affect gene regulation
- $\Rightarrow$ `creto`-Algorithm

Detection of Regulatory Elements

Detection of Regulatory Elements

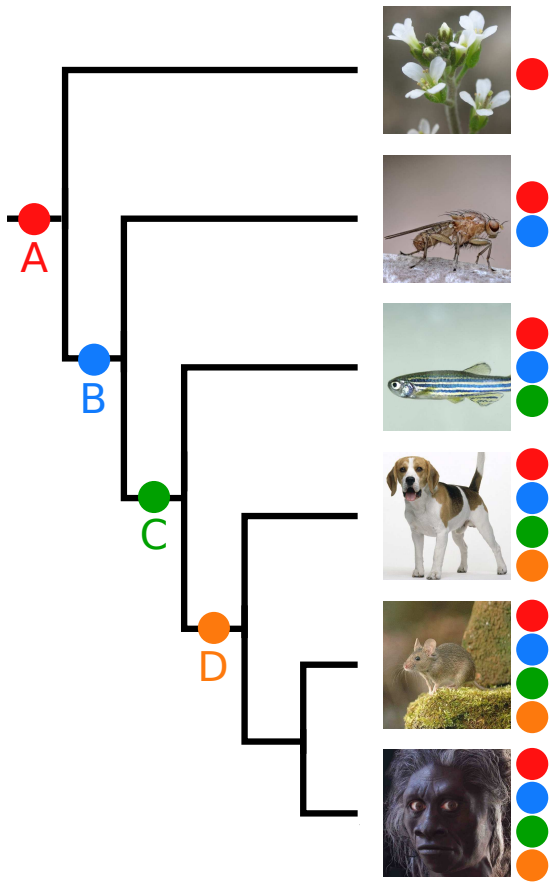
- regulatory elements are crucial for all life processes

Detection of Regulatory Elements

- regulatory elements are crucial for all life processes
- mutations are mostly lethal and are not passed to next generation (stabilizing selection)

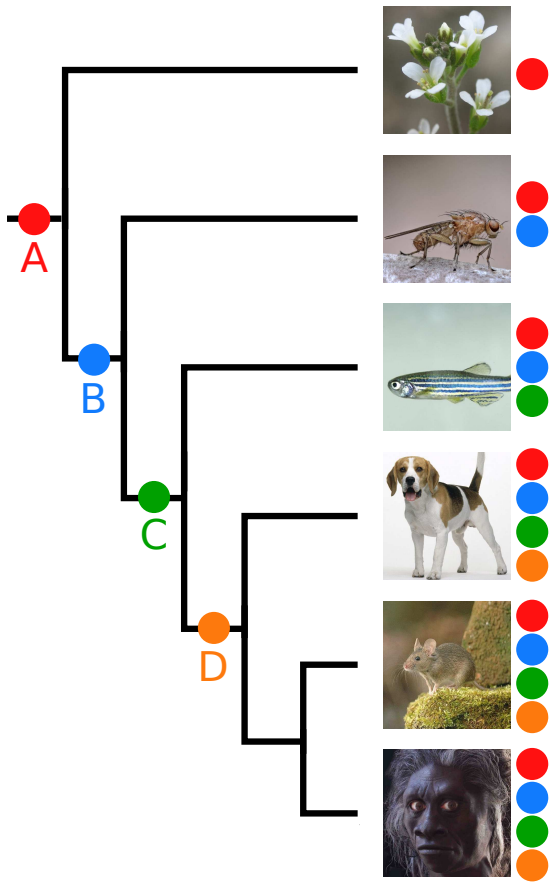
Detection of Regulatory Elements

- regulatory elements are crucial for all life processes
- mutations are mostly lethal and are not passed to next generation (stabilizing selection)
- regulatory elements evolve much slower than adjacent non-functional DNA (*phylogenetic footprints*)



Detection of Regulatory Elements

- regulatory elements are crucial for all life processes
- mutations are mostly lethal and are not passed to next generation (stabilizing selection)
- regulatory elements evolve much slower than adjacent non-functional DNA (*phylogenetic footprints*)
- detectable by comparative sequence analysis $\Rightarrow$ phylogenetic footprinting



Problematic Characteristics

- search for short, variable motifs (down to only 6nt)

Problematic Characteristics

- search for short, variable motifs (down to only 6nt)
- located in large regulatory region (1000nt and more), in front, behind or inside gene

Problematic Characteristics

- search for short, variable motifs (down to only 6nt)
- located in large regulatory region (1000nt and more), in front, behind or inside gene
- unconserved surrounding areas, variable distances possible

Problematic Characteristics

- search for short, variable motifs (down to only 6nt)
- located in large regulatory region (1000nt and more), in front, behind or inside gene
- unconserved surrounding areas, variable distances possible
- problems:

Problematic Characteristics

- search for short, variable motifs (down to only 6nt)
- located in large regulatory region (1000nt and more), in front, behind or inside gene
- unconserved surrounding areas, variable distances possible
- problems:
  - can easily be overseen

Problematic Characteristics

- search for short, variable motifs (down to only 6nt)
- located in large regulatory region (1000nt and more), in front, behind or inside gene
- unconserved surrounding areas, variable distances possible
- problems:
  - can easily be overseen
  - not statistically significant

Problematic Characteristics

- search for short, variable motifs (down to only 6nt)
- located in large regulatory region (1000nt and more), in front, behind or inside gene
- unconserved surrounding areas, variable distances possible
- problems:
  - can easily be overseen
  - not statistically significant
  - outweighted by surrounding random similarities

Problematic Characteristics

- search for short, variable motifs (down to only 6nt)
- located in large regulatory region (1000nt and more), in front, behind or inside gene
- unconserved surrounding areas, variable distances possible
- problems:
  - can easily be overseen
  - not statistically significant
  - outweighted by surrounding random similarities

Bioinformatic Challenge

- use of low stringency because of insignificant motifs

Bioinformatic Challenge

- use of low stringency because of insignificant motifs
- $\Rightarrow$ vast number of alignments between random similarities in unrelated areas

Bioinformatic Challenge

- use of low stringency because of insignificant motifs
- $\Rightarrow$ vast number of alignments between random similarities in unrelated areas

Bioinformatic Challenge

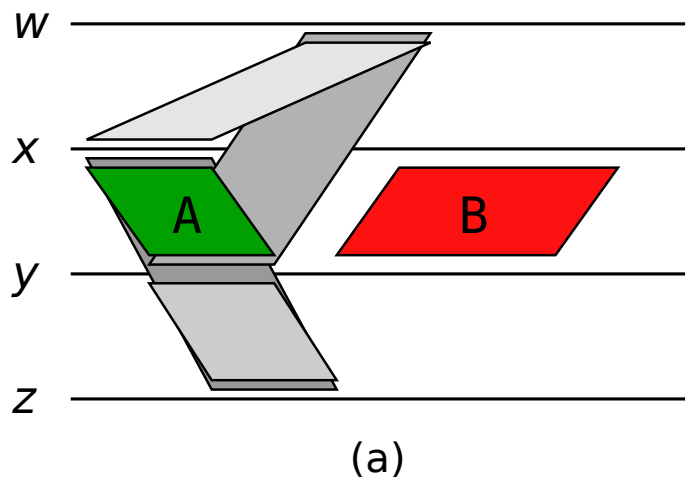
- use of low stringency because of insignificant motifs
- $\Rightarrow$ vast number of alignments between random similarities in unrelated areas
- aim: determining alignments between evolutionary related areas

Evolutionary Information

- (a) functional motifs are widely conserved $\Rightarrow$ support

Evolutionary Information

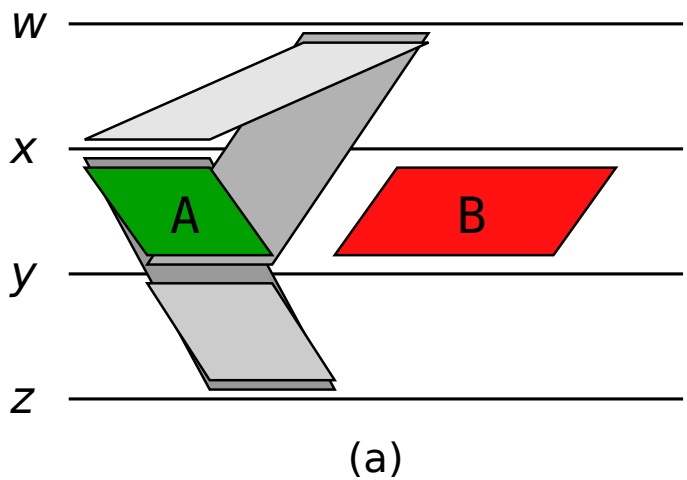
- (a) functional motifs are widely conserved $\Rightarrow$ support



(a)

Evolutionary Information

- (a) functional motifs are widely conserved $\Rightarrow$ support
- (b) order of motifs defines windows for new motifs $\Rightarrow$ consistence



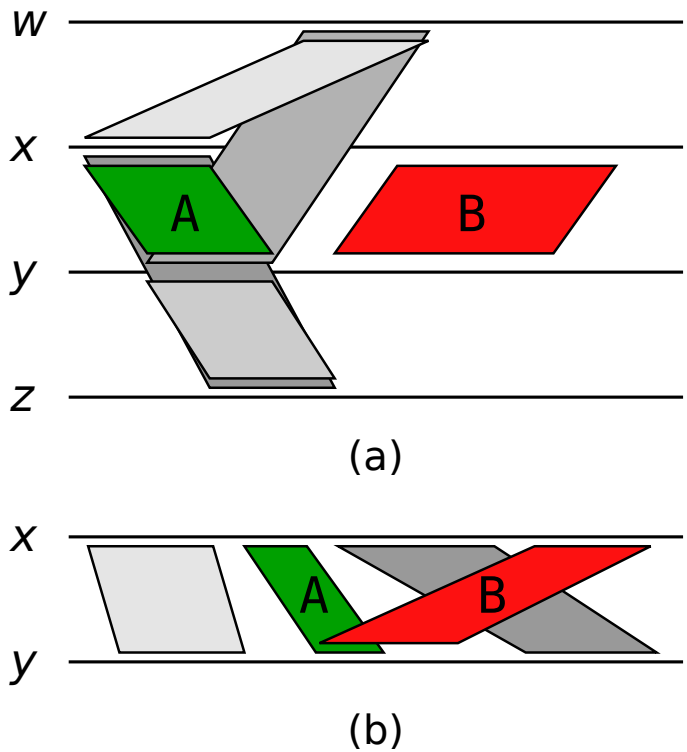
(a)

Evolutionary Information

- (a) functional motifs are widely conserved $\Rightarrow$ support
- (b) order of motifs defines windows for new motifs $\Rightarrow$ consistence

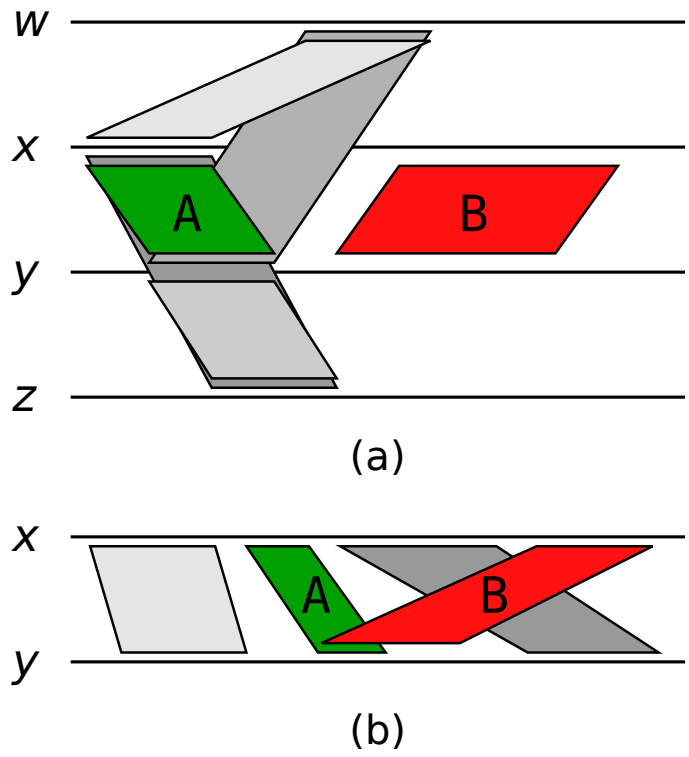
Evolutionary Information

- (a) functional motifs are widely conserved $\Rightarrow$ support
- (b) order of motifs defines windows for new motifs $\Rightarrow$ consistence
- existing multiple alignment approaches: global alignments disregard support of segments, local alignments disregard order information



Evolutionary Information

- (a) functional motifs are widely conserved $\Rightarrow$ support
- (b) order of motifs defines windows for new motifs $\Rightarrow$ consistence
- existing multiple alignment approaches: global alignments disregard support of segments, local alignments disregard order information
- idea: calculate pairwise local alignments with low stringency, determine maximal consistent subsets based on support



Alignment Definitions

	1	2	3	4
S_1	A	C	G	A
S_2	A	T	G	A
S_3	A	T	A	

Alignment Definitions

	1	2	3	4
S_1	A	C	G	A
S_2	A	T	G	A
S_3	A	T	A	

	A	C	G	A
	A	T	G	A
	A	T	–	A

Alignment Definitions

	1	2	3	4
S_1	A	C	G	A
S_2	A	T	G	A
S_3	A	T	A	

	A	C	G	A
	A	T	G	A
	A	T	–	A

[1,1] [1,2] [1,3] [1,4]
[2,1] [2,2] [2,3] [2,4]
[3,1] [3,2] [3,3]

Alignment Definitions

	1	2	3	4
S_1	A	C	G	A
S_2	A	T	G	A
S_3	A	T	A	

	A	C	G	A
	A	T	G	A
	A	T	–	A

[1,1] [1,2] [1,3] [1,4]
[2,1] [2,2] [2,3] [2,4]
[3,1] [3,2] [3,3]

$$\begin{pmatrix}[1,1]\\[2,1]\\[3,1]\end{pmatrix}\ \begin{pmatrix}[1,2]\\[2,2]\\[3,2]\end{pmatrix}\ [1,3]\ \begin{pmatrix}[1,4]\\[2,4]\\[3,3]\end{pmatrix}$$

Alignment Definitions

	1	2	3	4
S_1	A	C	G	A
S_2	A	T	G	A
S_3	A	T	A	

	A	C	G	A
	A	T	G	A
	A	T	–	A

[1,1] [1,2] [1,3] [1,4]
[2,1] [2,2] [2,3] [2,4]
[3,1] [3,2] [3,3]

$$\begin{pmatrix}[1,1]\\[2,1]\\[3,1]\end{pmatrix} \begin{pmatrix}[1,2]\\[2,2]\\[3,2]\end{pmatrix} [1,3] \begin{pmatrix}[1,4]\\[2,4]\\[3,3]\end{pmatrix}$$

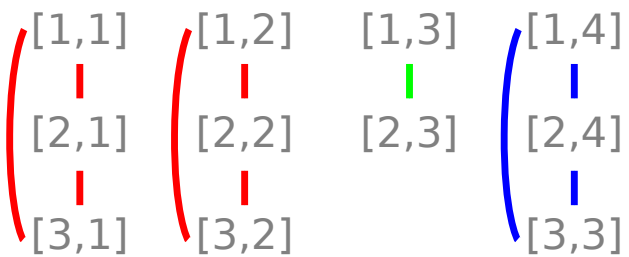
$$\begin{pmatrix}[1,1]\\[2,1]\\[3,1]\end{pmatrix} \begin{pmatrix}[1,4]\\[2,4]\\[3,3]\end{pmatrix}$$

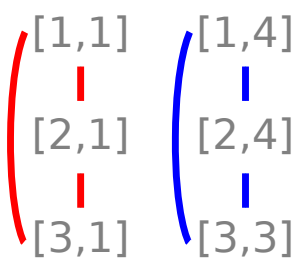
Alignment Definitions

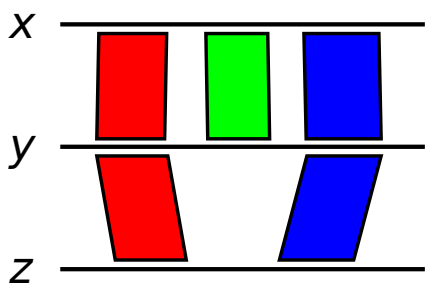
	1	2	3	4
S_1	A	C	G	A
S_2	A	T	G	A
S_3	A	T	A	

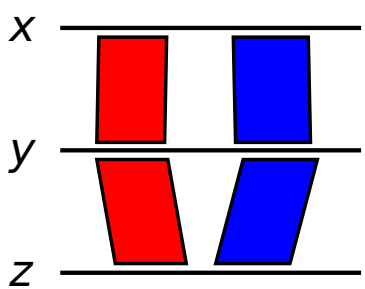
A	C	G	A	
A	T	G	A	
A	T	–	A	

[1,1] [1,2] [1,3] [1,4]
[2,1] [2,2] [2,3] [2,4]
[3,1] [3,2] [3,3]

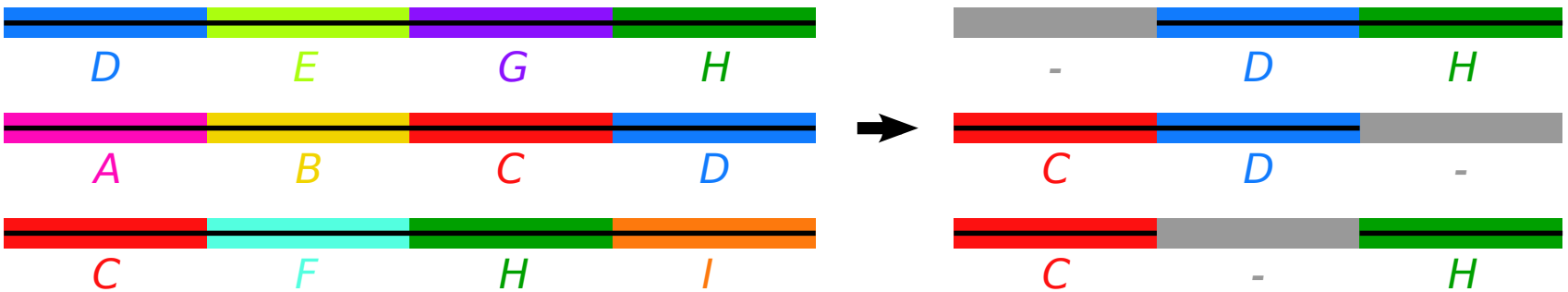








Motif Alignment

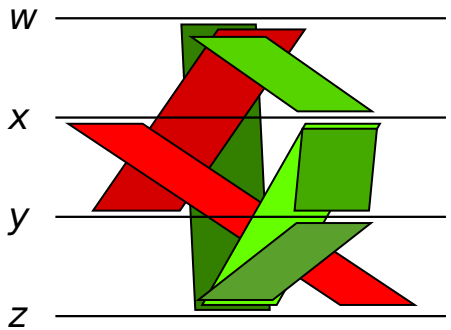


Consistent Alignments

Definition (Consistency)

An alignment collection $\mathcal{A} = \{A_1, \ldots, A_n\}$ over sequences $\mathcal{S} = \{S_1, \ldots, S_m\}$ is *consistent* $\Leftrightarrow$ it exists a multiple alignment M over $\mathcal{S}$ so that all pairs of nucleotides aligned by alignments in $\mathcal{A}$ are also aligned in M.

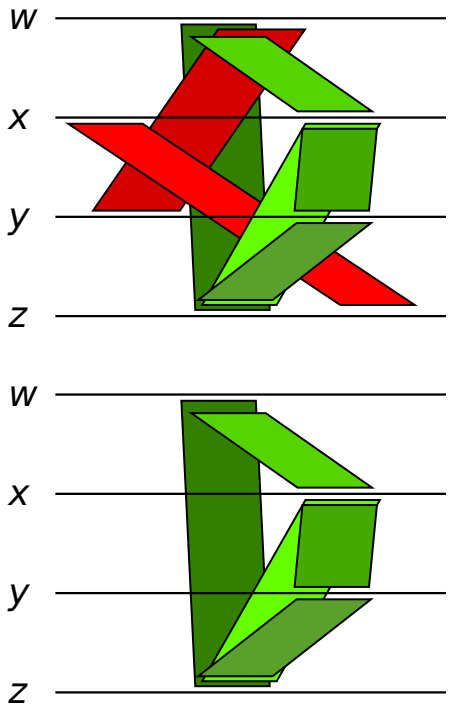
Consistent Alignments



Definition (Consistency)

An alignment collection $\mathcal{A} = \{A_1, \ldots, A_n\}$ over sequences $\mathcal{S} = \{S_1, \ldots, S_m\}$ is *consistent* $\Leftrightarrow$ it exists a multiple alignment M over $\mathcal{S}$ so that all pairs of nucleotides aligned by alignments in $\mathcal{A}$ are also aligned in M.

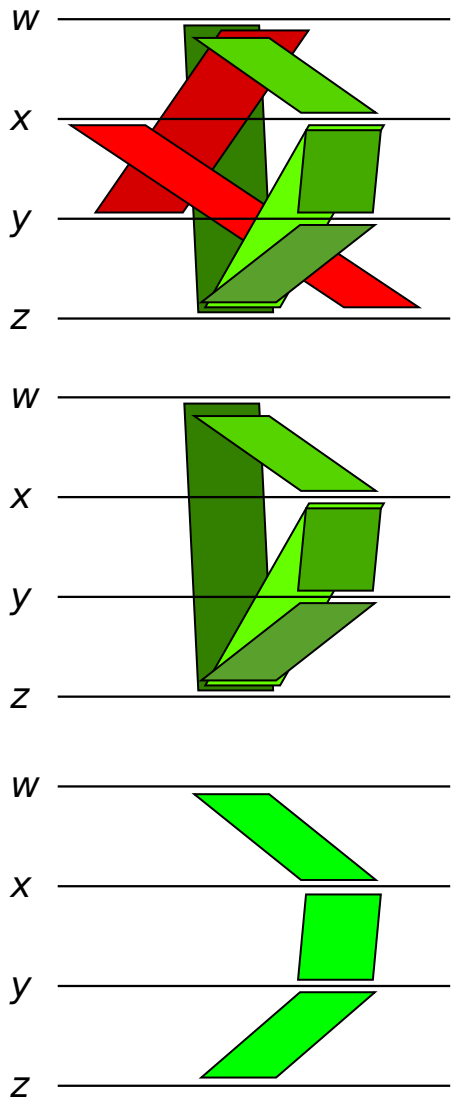
Consistent Alignments



Definition (Consistency)

An alignment collection $\mathcal{A} = \{A_1, \ldots, A_n\}$ over sequences $\mathcal{S} = \{S_1, \ldots, S_m\}$ is *consistent* $\Leftrightarrow$ it exists a multiple alignment M over $\mathcal{S}$ so that all pairs of nucleotides aligned by alignments in $\mathcal{A}$ are also aligned in M.

Consistent Alignments



Definition (Consistency)

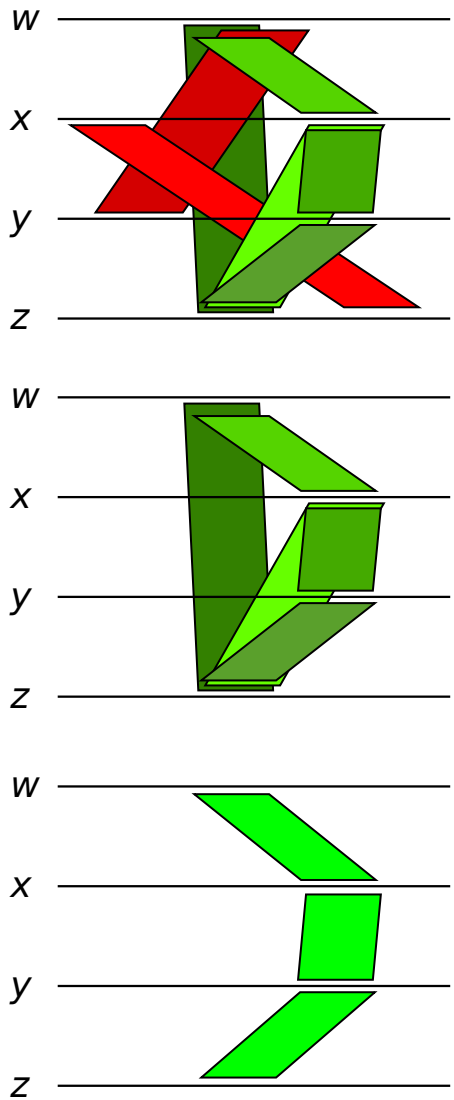
An alignment collection $\mathcal{A} = \{A_1, \ldots, A_n\}$ over sequences $\mathcal{S} = \{S_1, \ldots, S_m\}$ is *consistent* $\Leftrightarrow$ it exists a multiple alignment M over $\mathcal{S}$ so that all pairs of nucleotides aligned by alignments in $\mathcal{A}$ are also aligned in M.

Optimization Problem

Definition (Maximal Consistent Alignment Subset Problem)

Given an alignment collection $\mathcal{A} = \{A_1, \ldots, A_n\}$ over sequences $\mathcal{S} = \{S_1, \ldots, S_m\}$, find a maximal subset $\mathcal{A}'$ of $\mathcal{A}$ that is consistent.

Optimization Problem



Definition (Maximal Consistent Alignment Subset Problem)

Given an alignment collection $\mathcal{A} = \{A_1, \ldots, A_n\}$ over sequences $\mathcal{S} = \{S_1, \ldots, S_m\}$, find a maximal subset $\mathcal{A}'$ of $\mathcal{A}$ that is consistent.

Complexity of MCASP

- MCASP $\in$ NP (reducible to 'Multiple Alignment Problem' in P)

Complexity of MCASP

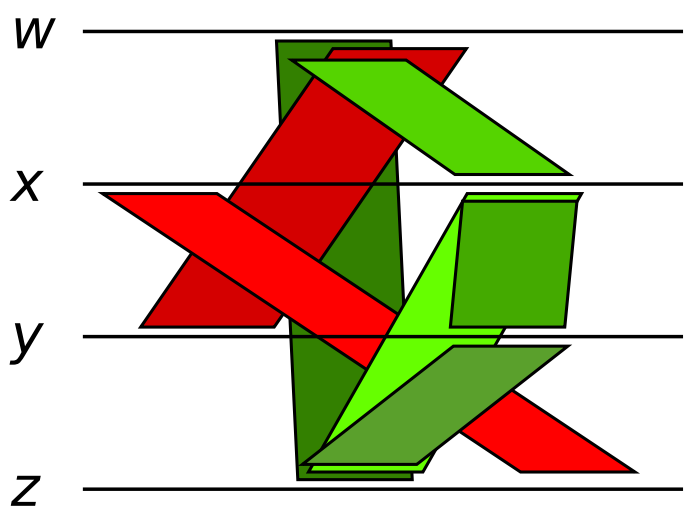
- MCASP $\in$ NP (reducible to 'Multiple Alignment Problem' in P)
- optimal solution: check each subset of $\mathcal{A}$ for consistency

Complexity of MCASP

- MCASP $\in$ NP (reducible to 'Multiple Alignment Problem' in P)
- optimal solution: check each subset of $\mathcal{A}$ for consistency
- exponential growth

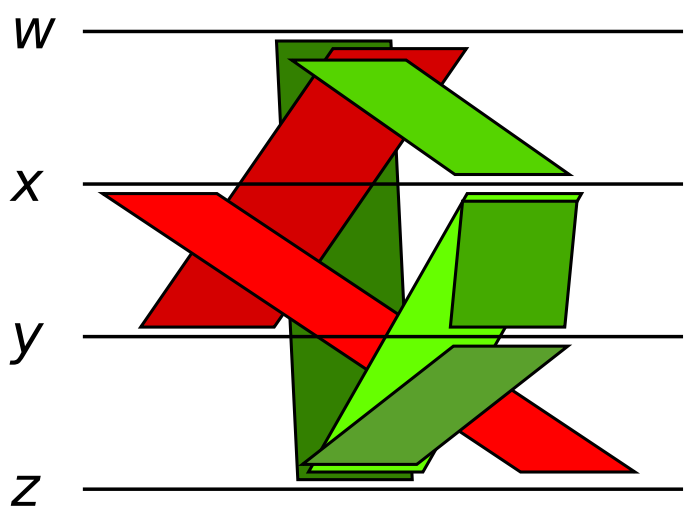
Complexity of MCASP

- MCASP $\in$ NP (reducible to 'Multiple Alignment Problem' in P)
- optimal solution: check each subset of $\mathcal{A}$ for consistency
- exponential growth
  - 7 alignments: 128 subsets



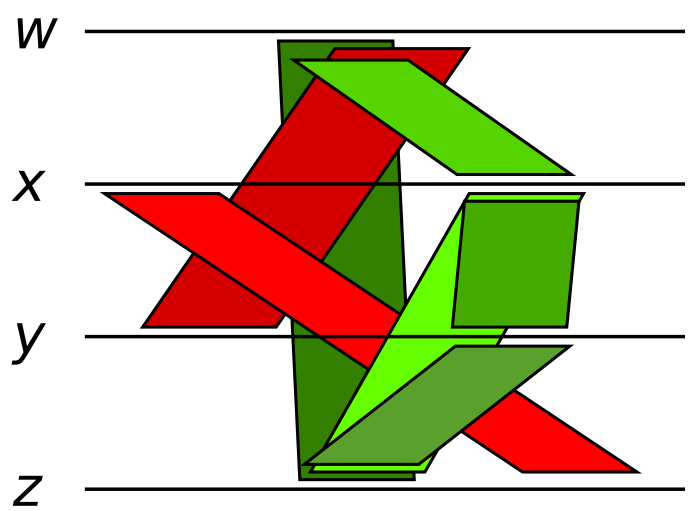
Complexity of MCASP

- MCASP $\in$ NP (reducible to 'Multiple Alignment Problem' in P)
- optimal solution: check each subset of $\mathcal{A}$ for consistency
- exponential growth
  - 7 alignments: 128 subsets
  - 250 alignments: $\sim 10^{75}$ subsets



Complexity of MCASP

- MCASP $\in$ NP (reducible to 'Multiple Alignment Problem' in P)
- optimal solution: check each subset of $\mathcal{A}$ for consistency
- exponential growth
  - 7 alignments: 128 subsets
  - 250 alignments: $\sim 10^{75}$ subsets
- biological data sets contain of millions of alignments $\Rightarrow$ need for heuristic approach



Heuristic: Algorithmic Sketch

- input: arbitrary set $\mathcal{A}$ of local pairwise alignments

Heuristic: Algorithmic Sketch

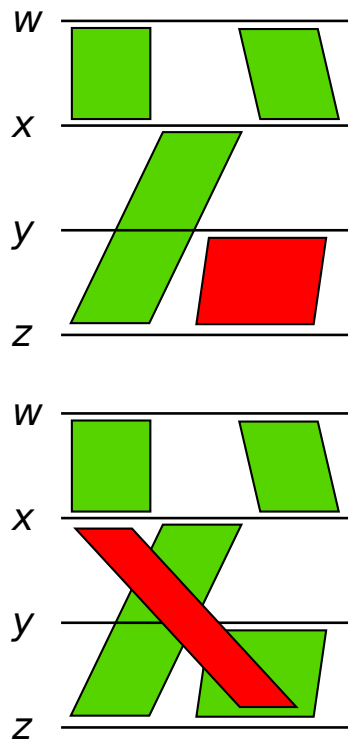
- input: arbitrary set $\mathcal{A}$ of local pairwise alignments
- assemble multiple alignment M, starting with $M = \emptyset$

Heuristic: Algorithmic Sketch

- input: arbitrary set $\mathcal{A}$ of local pairwise alignments
- assemble multiple alignment M, starting with $M = \emptyset$
- checking iteratively all alignments $A \in \mathcal{A}$

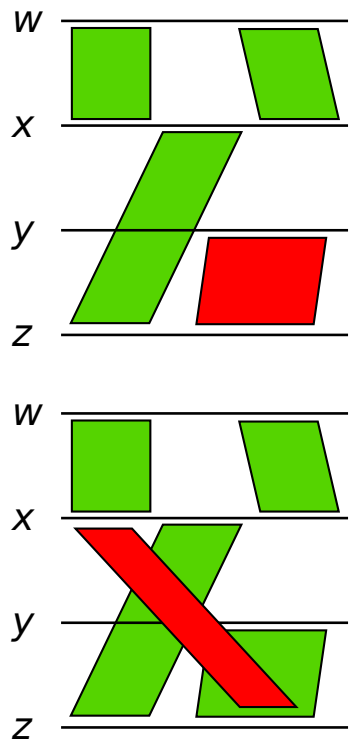
Heuristic: Algorithmic Sketch

- input: arbitrary set $\mathcal{A}$ of local pairwise alignments
- assemble multiple alignment M, starting with $M = \emptyset$
- checking iteratively all alignments $A \in \mathcal{A}$
- consistent alignments are inserted, inconsistent are rejected



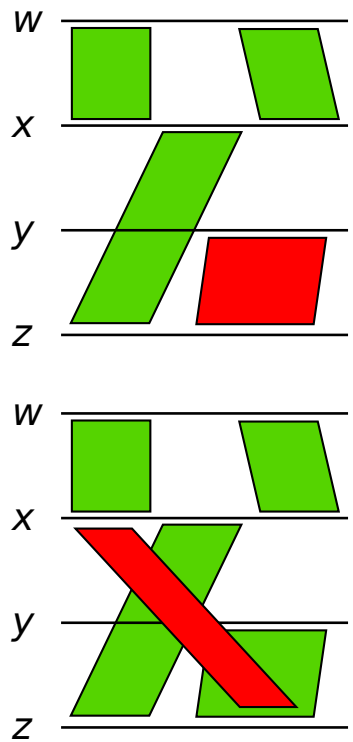
Heuristic: Algorithmic Sketch

- input: arbitrary set $\mathcal{A}$ of local pairwise alignments
- assemble multiple alignment M, starting with $M = \emptyset$
- checking iteratively all alignments $A \in \mathcal{A}$
- consistent alignments are inserted, inconsistent are rejected
- alignments in M are consistent subset of $\mathcal{A}$



Heuristic: Algorithmic Sketch

- input: arbitrary set $\mathcal{A}$ of local pairwise alignments
- assemble multiple alignment M, starting with $M = \emptyset$
- checking iteratively all alignments $A \in \mathcal{A}$
- consistent alignments are inserted, inconsistent are rejected
- alignments in M are consistent subset of $\mathcal{A}$
- problem: inserted alignments cannot be removed or corrected $\Rightarrow$ insertion order is crucial

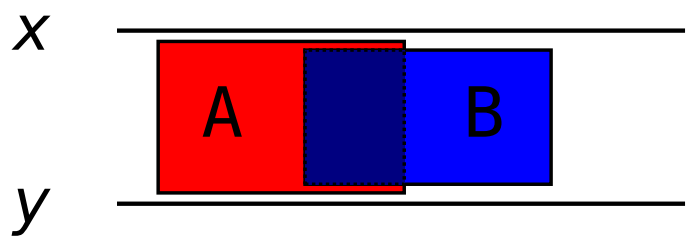


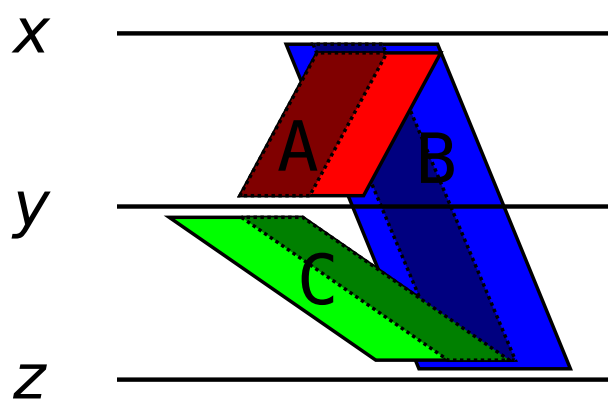
Heuristic: Extended Scores

- start with alignments that are most supported by other alignments

Heuristic: Extended Scores

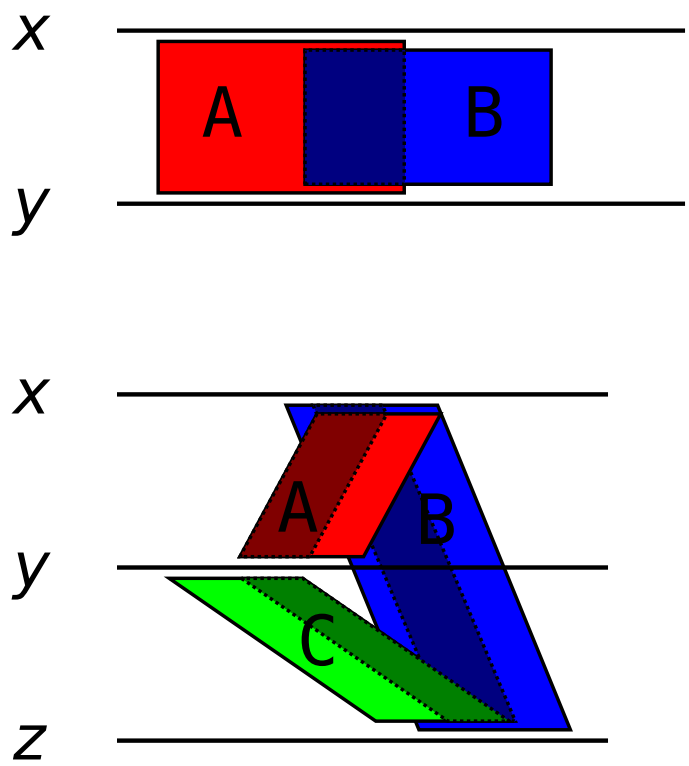
- start with alignments that are most supported by other alignments





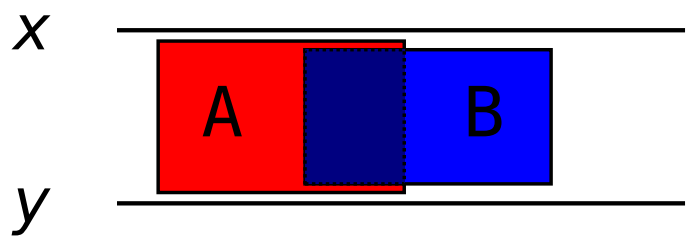
Heuristic: Extended Scores

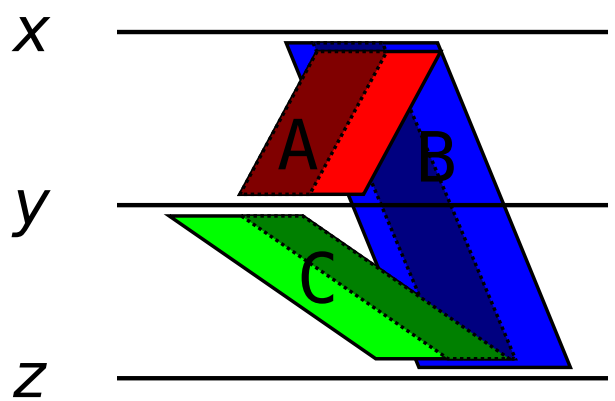
- start with alignments that are most supported by other alignments
- express support by score ⇒ extended scores



Heuristic: Extended Scores

- start with alignments that are most supported by other alignments
- express support by score ⇒ extended scores
- similar to `T-Coffee`[a]

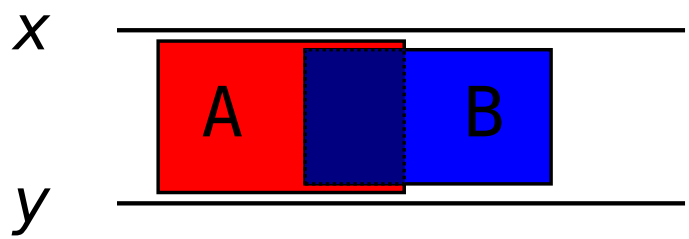


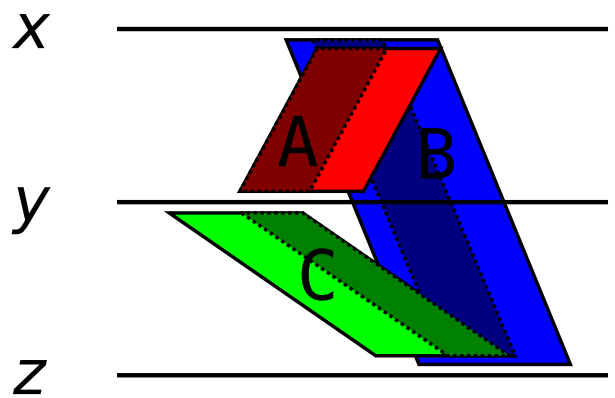


[a] Notredame *et al.*:T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**(1), 205–217

Heuristic: Extended Scores

- start with alignments that are most supported by other alignments
- express support by score ⇒ extended scores
- similar to `T-Coffee`[a]
- basic score plus bonus for each direct / indirect support

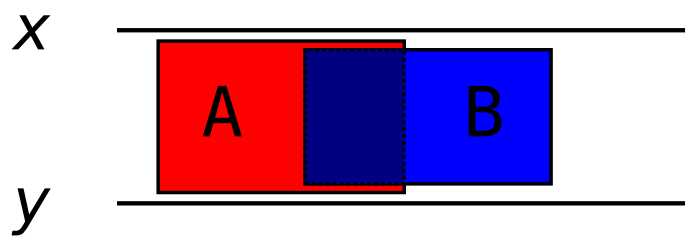


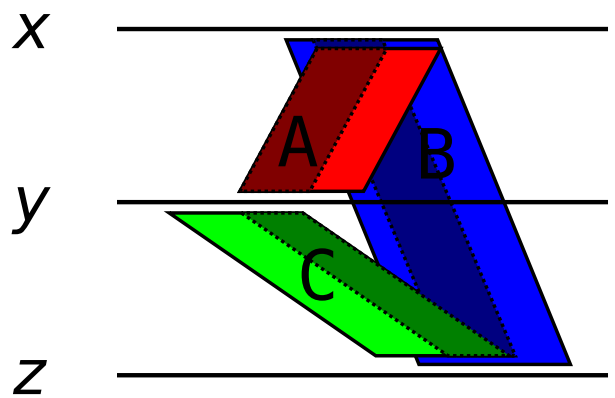


[a] Notredame *et al.*:T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**(1), 205–217

Heuristic: Extended Scores

- start with alignments that are most supported by other alignments
- express support by score ⇒ extended scores
- similar to `T-Coffee`[a]
- basic score plus bonus for each direct / indirect support
- insert alignments in order based on extended score





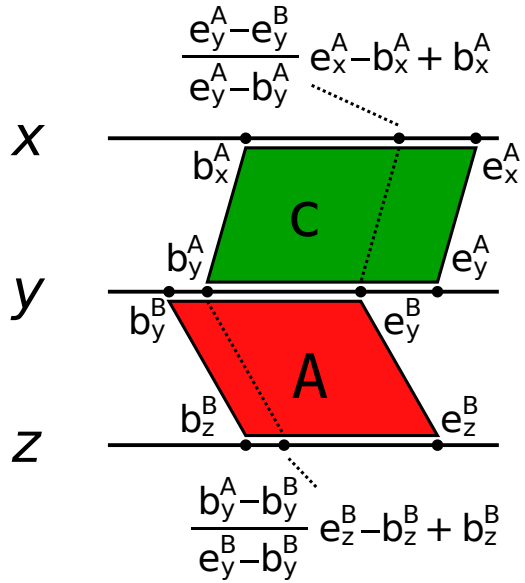
[a] Notredame *et al.*:T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**(1), 205–217

Heuristic: Algorithmic Tuning

- abstract alignments by intervals
  $A = \{[x, b_x, e_x], [y, b_y, e_y]\}$

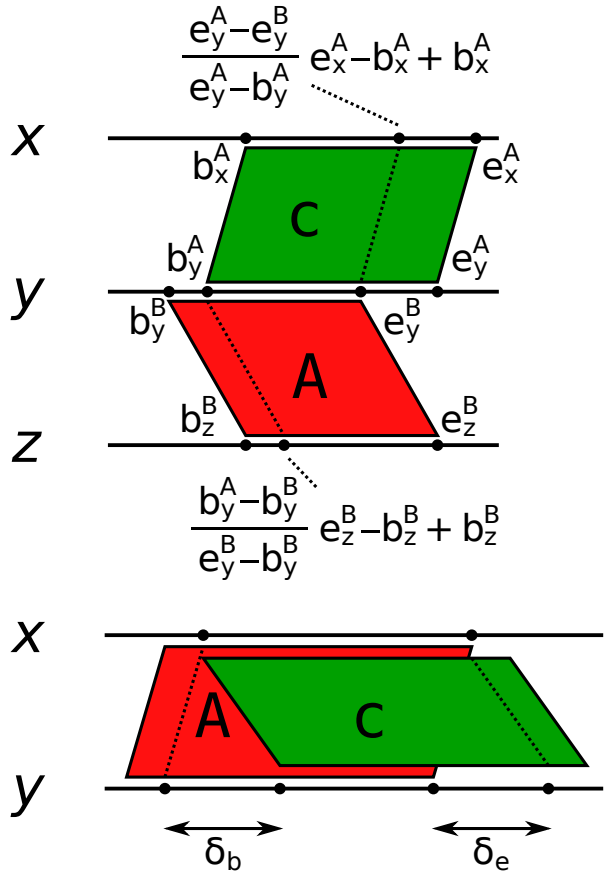
Heuristic: Algorithmic Tuning

- abstract alignments by intervals $A = \{[x, b_x, e_x], [y, b_y, e_y]\}$
- calculate intermediate positions by linear interpolation



Heuristic: Algorithmic Tuning

- abstract alignments by intervals $A = \{[x, b_x, e_x], [y, b_y, e_y]\}$
- calculate intermediate positions by linear interpolation
- allow adjustable error tolerance

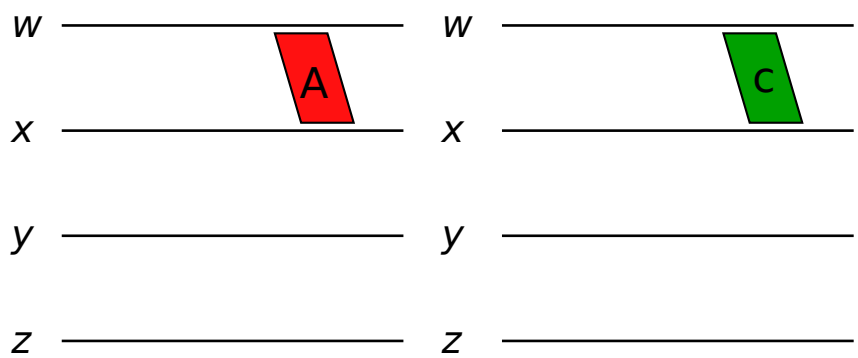


Heuristic: Assembly

- inserted alignments define alignment columns

Heuristic: Assembly

- inserted alignments define alignment columns

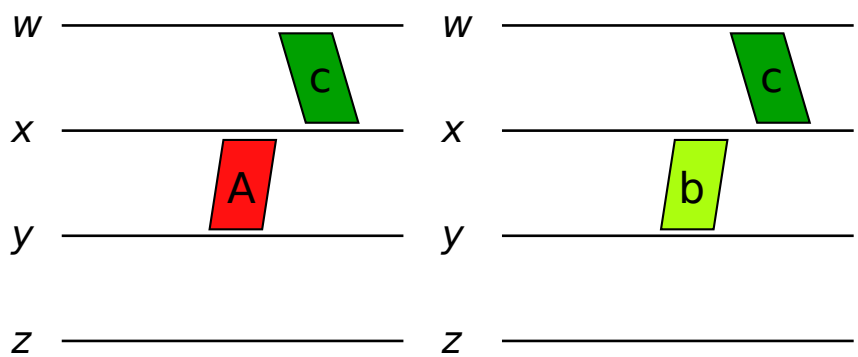


Heuristic: Assembly

- inserted alignments define alignment columns
- save columns as ordered list based on order defined by side space

Heuristic: Assembly

- inserted alignments define alignment columns
- save columns as ordered list based on order defined by side space

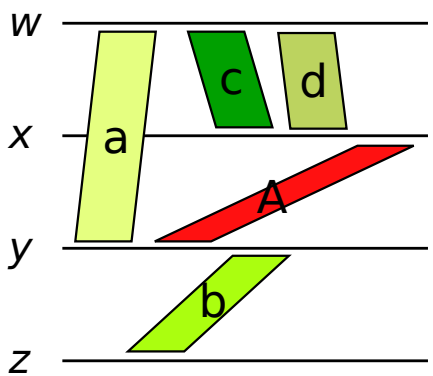


Heuristic: Assembly

- inserted alignments define alignment columns
- save columns as ordered list based on order defined by side space
- new insertion: check columns sequential, remember first suffix and last prefix column for each alignment sequence

Heuristic: Assembly

- inserted alignments define alignment columns
- save columns as ordered list based on order defined by side space
- new insertion: check columns sequential, remember first suffix and last prefix column for each alignment sequence

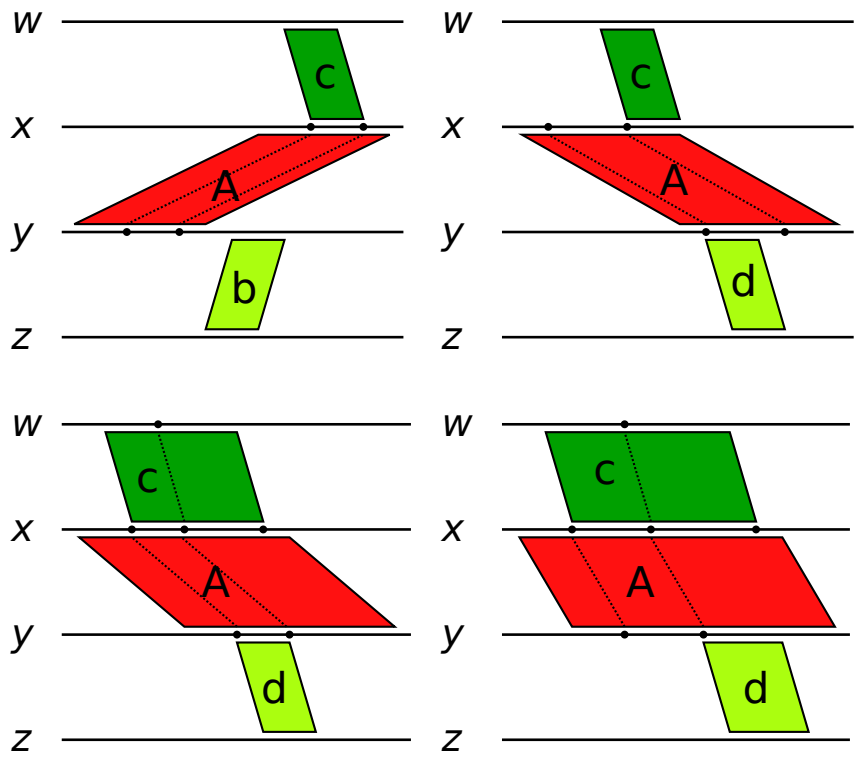


Heuristic: Assembly

- inserted alignments define alignment columns
- save columns as ordered list based on order defined by side space
- new insertion: check columns sequential, remember first suffix and last prefix column for each alignment sequence
- insertion can cause split, switch or merge of columns and alignment

Heuristic: Assembly

- inserted alignments define alignment columns
- save columns as ordered list based on order defined by side space
- new insertion: check columns sequential, remember first suffix and last prefix column for each alignment sequence
- insertion can cause split, switch or merge of columns and alignment



Heuristic: Complexity

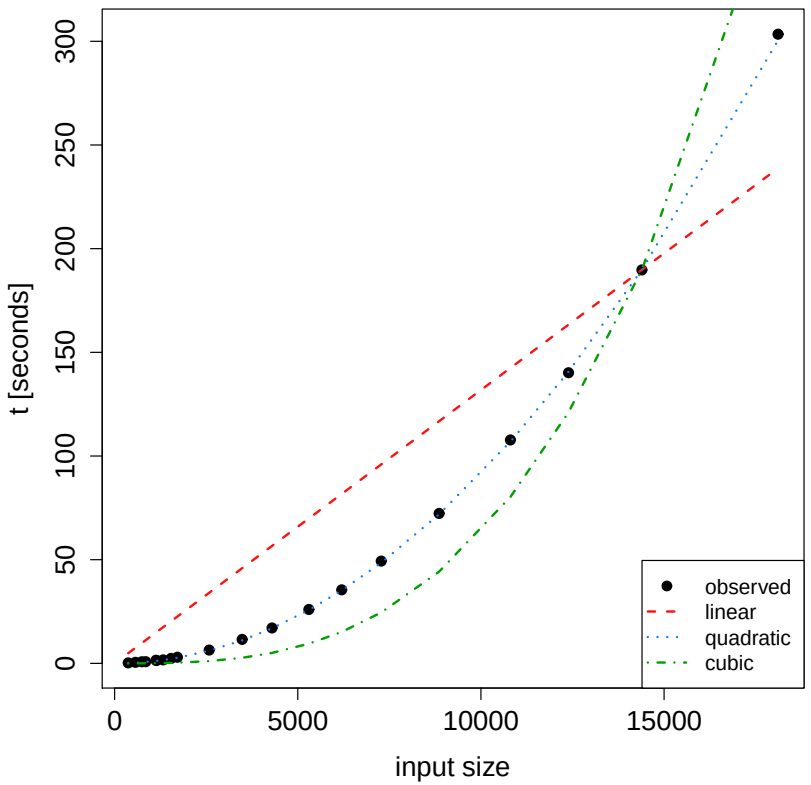
- insertion of n alignments over m sequences with length l is in $O(nlm)$

Heuristic: Complexity

- insertion of n alignments over m sequences with length l is in $O(nlm)$
- calculation of extended scores is in $O(n^3)$

Heuristic: Complexity

- insertion of n alignments over m sequences with length l is in $O(nlm)$
- calculation of extended scores is in $O(n^3)$
- artificial data sets with different set sizes: mean runtime in $O(n^2)$



Results: Maximal Consistent Subsets

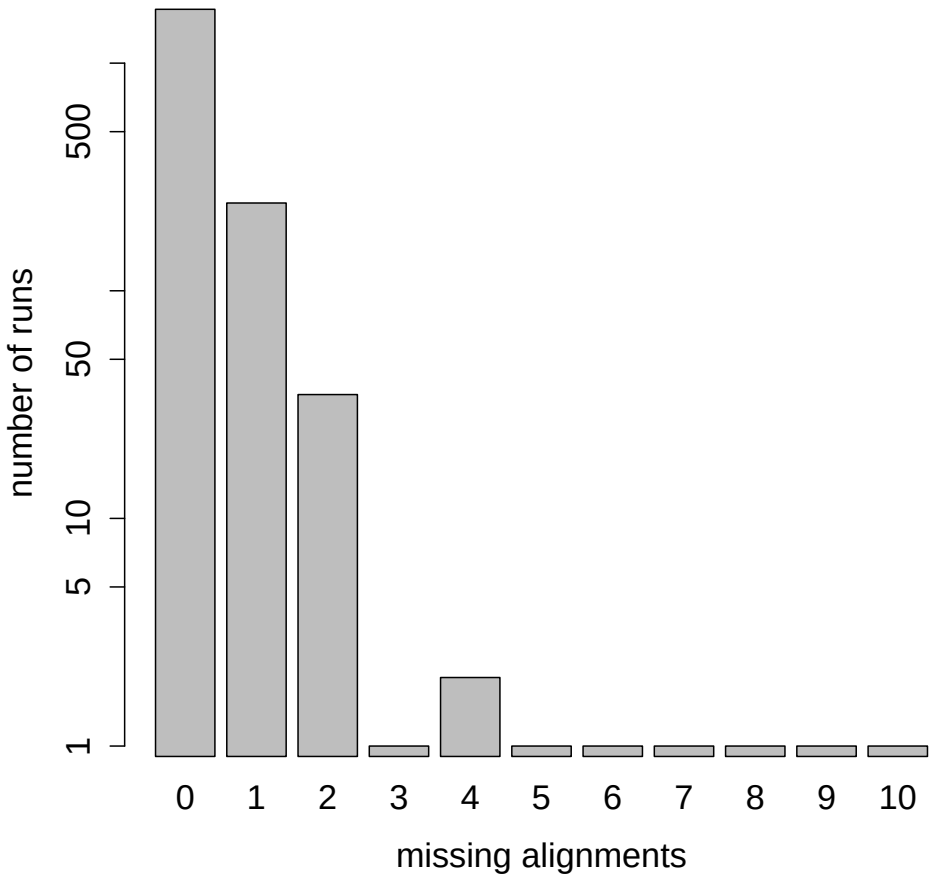
- artificial data sets $\mathcal{A}$ with up to 30 alignments

Results: Maximal Consistent Subsets

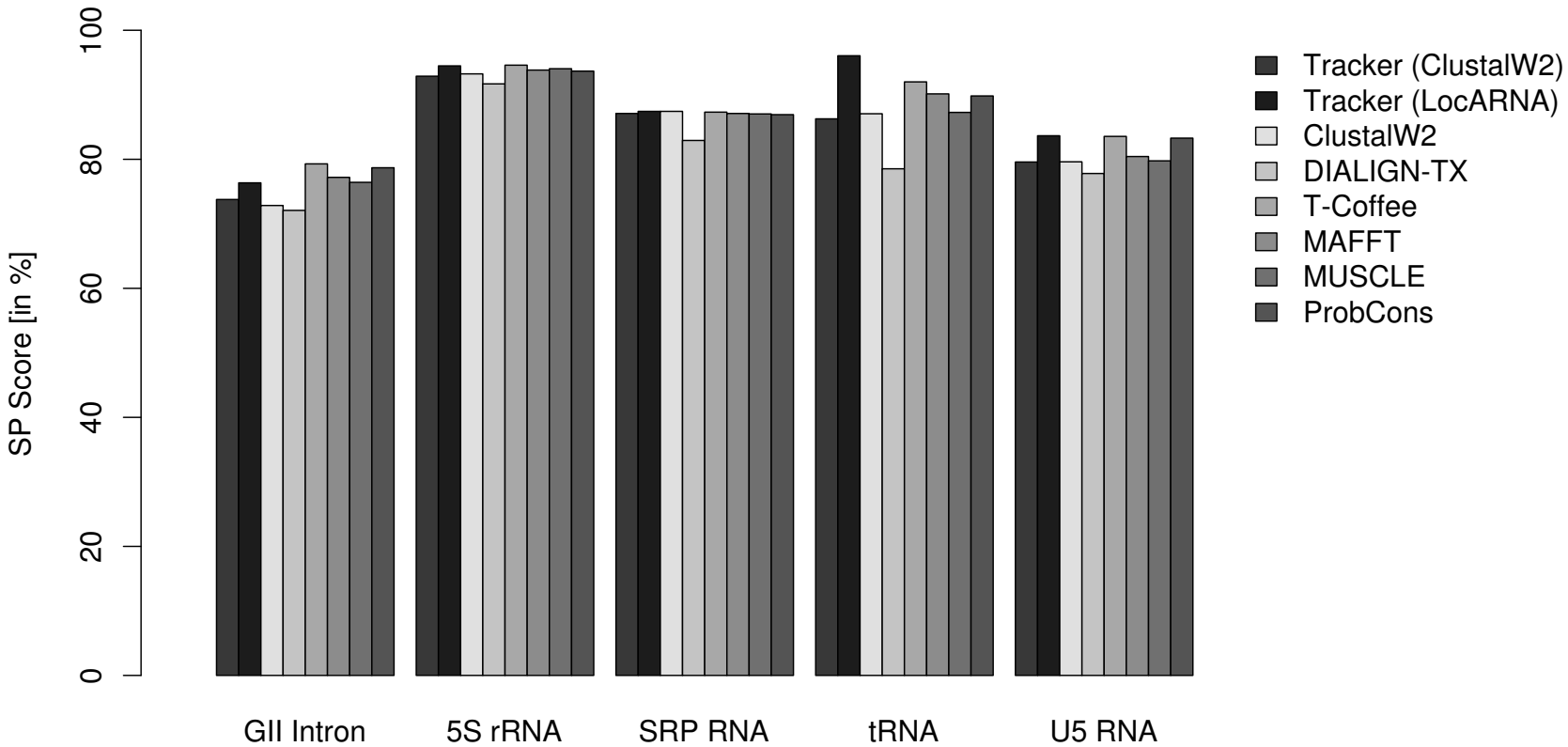
- artificial data sets $\mathcal{A}$ with up to 30 alignments
- comparison of heuristic and optimal (checking all subsets) solutions

Results: Maximal Consistent Subsets

- artificial data sets $\mathcal{A}$ with up to 30 alignments
- comparison of heuristic and optimal (checking all subsets) solutions
- optimal result found in most cases, number of missing alignments relative to optimal solution is low



Results: Alignment Calculation on BRaliBase II



Footprint Detection with `Tracker`

- calculate local pairwise alignments with low stringency between all input sequences

Footprint Detection with `Tracker`

- calculate local pairwise alignments with low stringency between all input sequences
- remove repetitive areas based on entropy and mutual information content

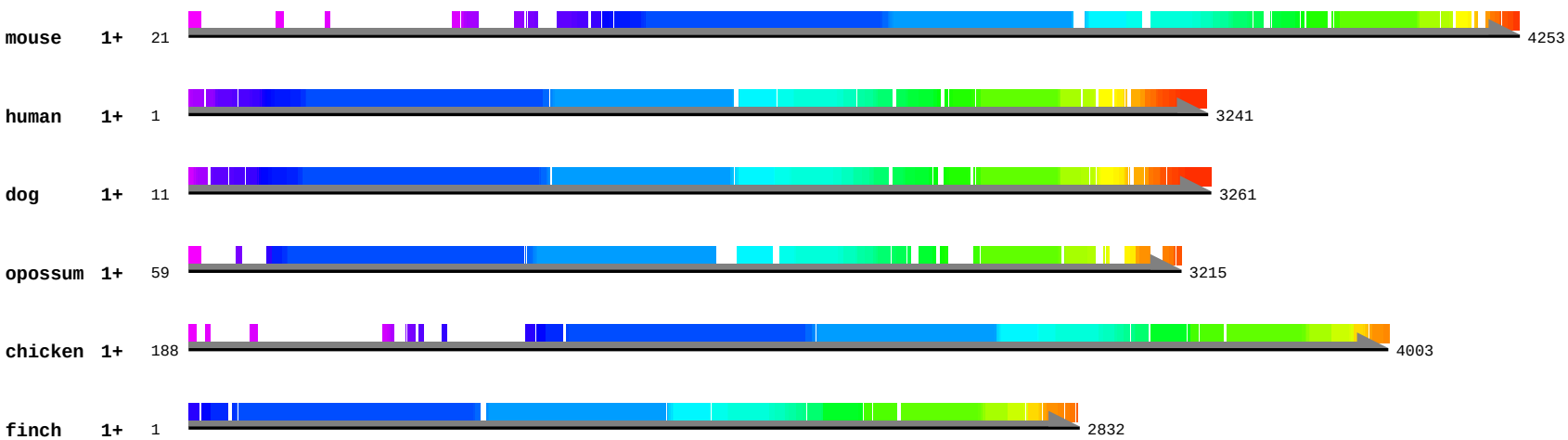
Footprint Detection with `Tracker`

- calculate local pairwise alignments with low stringency between all input sequences
- remove repetitive areas based on entropy and mutual information content
- calculate maximal consitent subset of alignment set

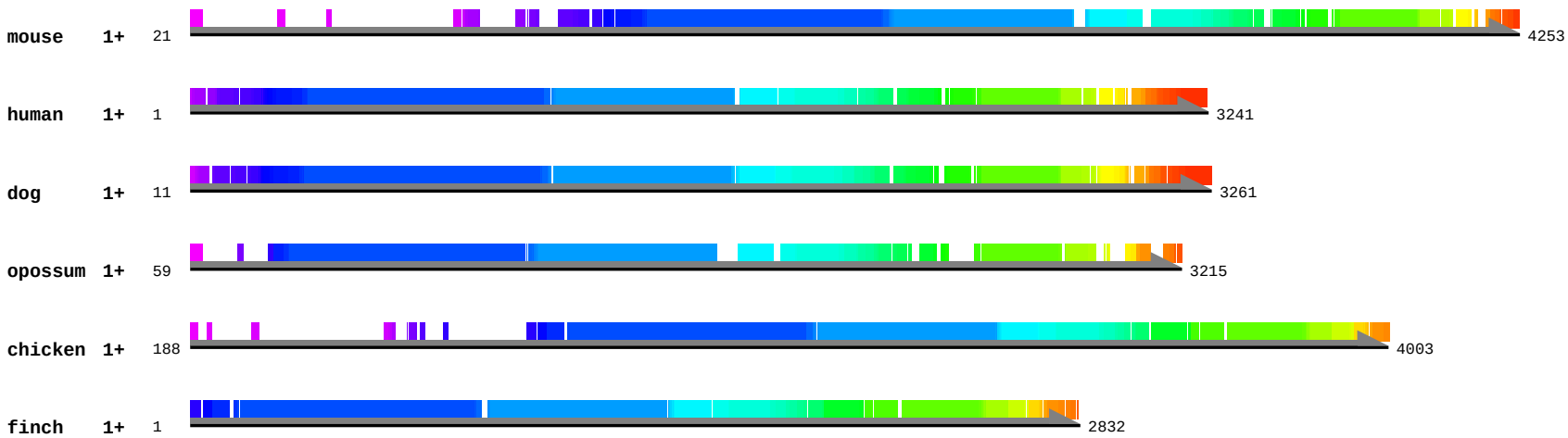
Footprint Detection with `Tracker`

- calculate local pairwise alignments with low stringency between all input sequences
- remove repetitive areas based on entropy and mutual information content
- calculate maximal consitent subset of alignment set
- correct column transition errors caused by inconsistency-tolerance

Tracker: CSB and Digit Indentity in Birds



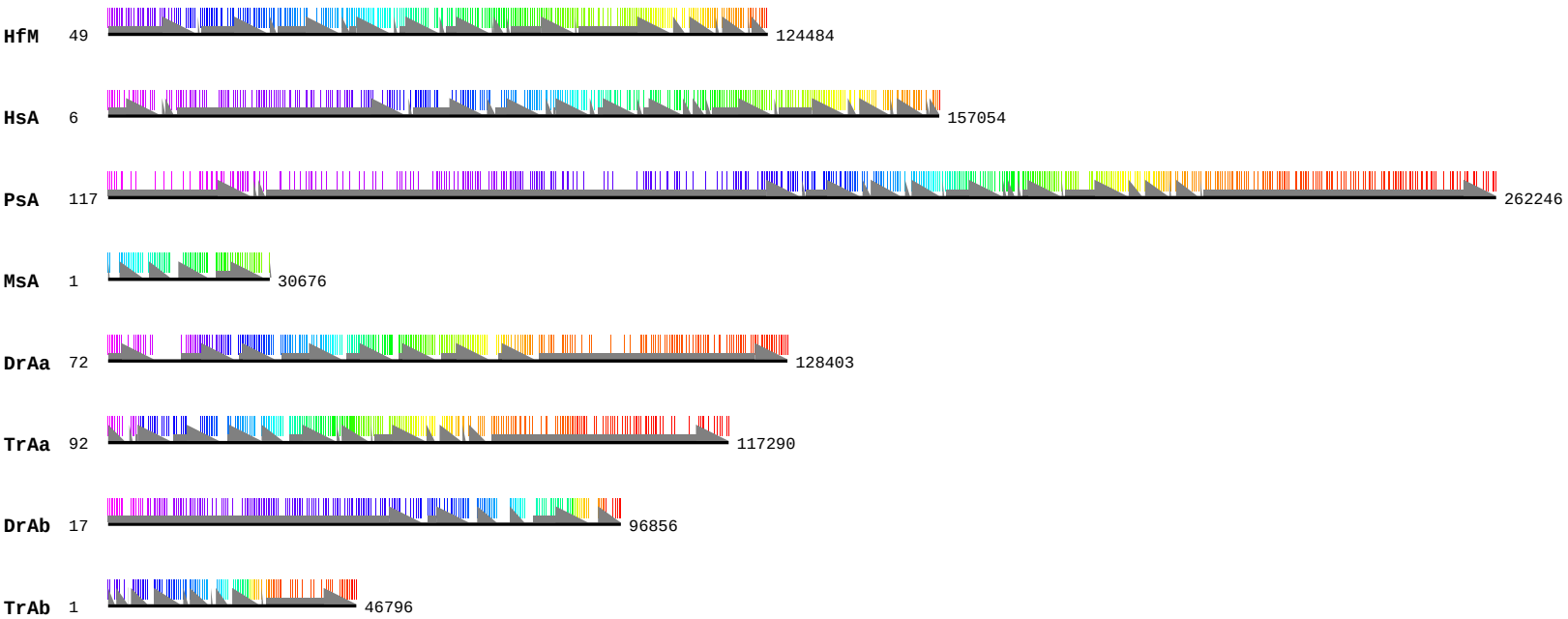
Tracker: CSB and Digit Indentity in Birds




```
mouse:    0|+|3404-3435|32  74 (0)  TCATTACCT-TTTTGGAAAAACACTTCTCTCCC  (27) 77
human:    0|+|2251-2282|32  72 (16) TCATTACTT-TTTCAGAAAAGCACTTTTTTCCC  (0)  76
dog:      0|+|2251-2282|32  72 (16) TCATTACCT-TTTCGGAAAAGCACTTTTTTGCC  (0)  76
opossum:  0|+|2312-2342|31  74 (0)  TCATTGCCTCTTTTGAAAGAGCA--TGTTTCCC  (0)  76
chicken:                            --------------------------------
finch:                              --------------------------------
                                    ***** * * ***   ** * **  * * * **

mouse:                                   -------------------------
human:                                   -------------------------
dog:                                     -------------------------
opossum:                                 -------------------------
chicken:  0|+|3376-3400|25  82 (0)  AAAAGGAGGTAATACTTAAAGGAAA  (1) 93
finch:    0|+|2145-2169|25  82 (0)  ATAAGGAGGCAATACTTAAAGTGAA  (7) 93
                                    * ******* ***********  **
```

Tracker: HoxA-Clusters in Vertebrates



Summary: Detection of Regulatory Elements

- `tracker` detects phylogenetic footprints by computing a new form of multiple alignments consisting of local motifs that still satisfy sequence order condition

Summary: Detection of Regulatory Elements

- `tracker` detects phylogenetic footprints by computing a new form of multiple alignments consisting of local motifs that still satisfy sequence order condition
- computation of initial alignment sets and other alignment steps are completely generic and can be adopted to new alignment algorithms

Summary: Detection of Regulatory Elements

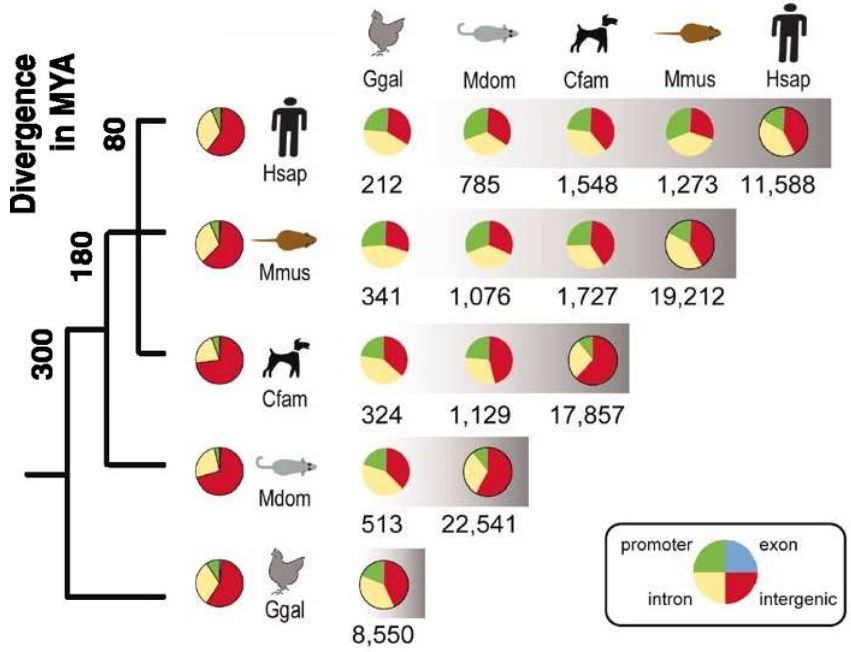
- `tracker` detects phylogenetic footprints by computing a new form of multiple alignments consisting of local motifs that still satisfy sequence order condition
- computation of initial alignment sets and other alignment steps are completely generic and can be adopted to new alignment algorithms
- todo: usage of phylogenetic information, check for motif overrepresentation in footprints

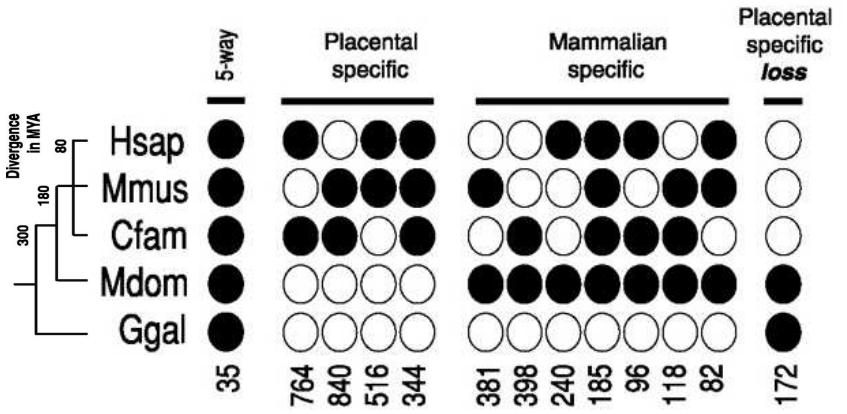
Evolutionary Analysis of Regulatory Elements

Evolutionary Analysis of Regulatory Elements

Schmidt *et al.*, Science, May 2010:

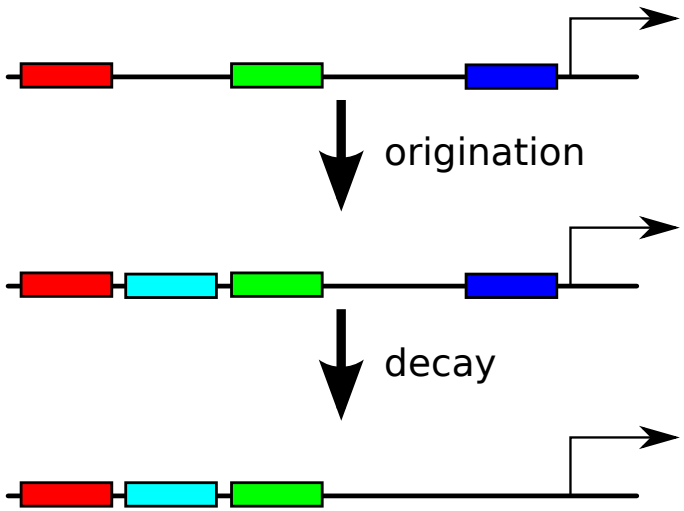
- determination of genome-wide occupancy for CCAAT/enhancer-binding protein alpha (CEBPA) in five vertebrates
- CEBPA: regulation of leptin and growth arrest in cultured cells





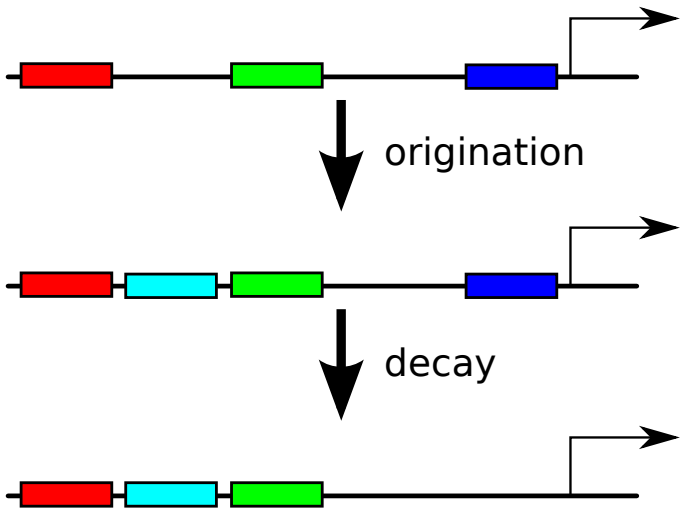
Binding Site Turnover

- TFBS turnover (loss and generation of TFBS) is common event even when transcriptional output remains conserved



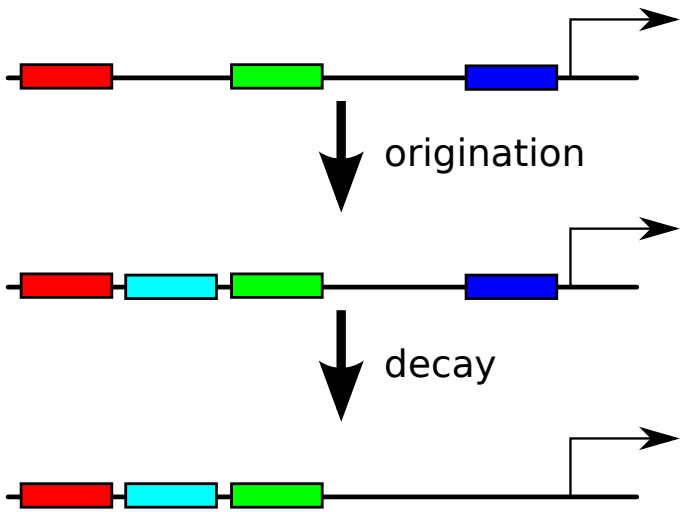
Binding Site Turnover

- TFBS turnover (loss and generation of TFBS) is common event even when transcriptional output remains conserved
- existence of TFBS is more important than exact location



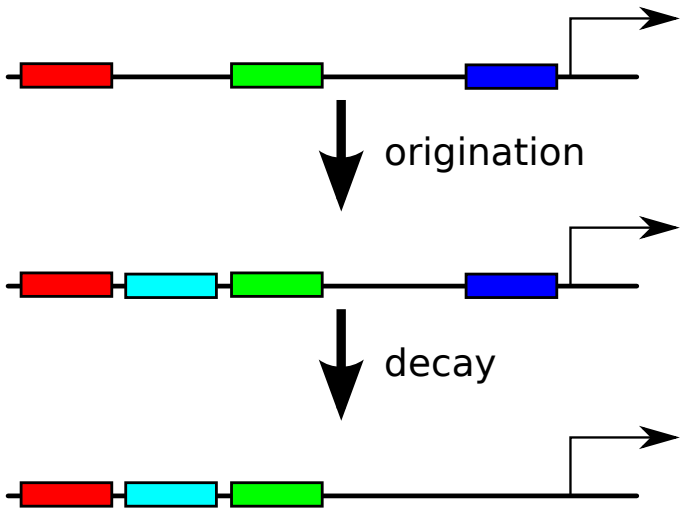
Binding Site Turnover

- TFBS turnover (loss and generation of TFBS) is common event even when transcriptional output remains conserved
- existence of TFBS is more important than exact location
- supported by variability of TF binding



Binding Site Turnover

- TFBS turnover (loss and generation of TFBS) is common event even when transcriptional output remains conserved
- existence of TFBS is more important than exact location
- supported by variability of TF binding
- arrival of new TFBS with origination rate and retention with decay rate



Turnover in Evolutionary Context

- evolutionary events are likely to cause lineage specific differences

Turnover in Evolutionary Context

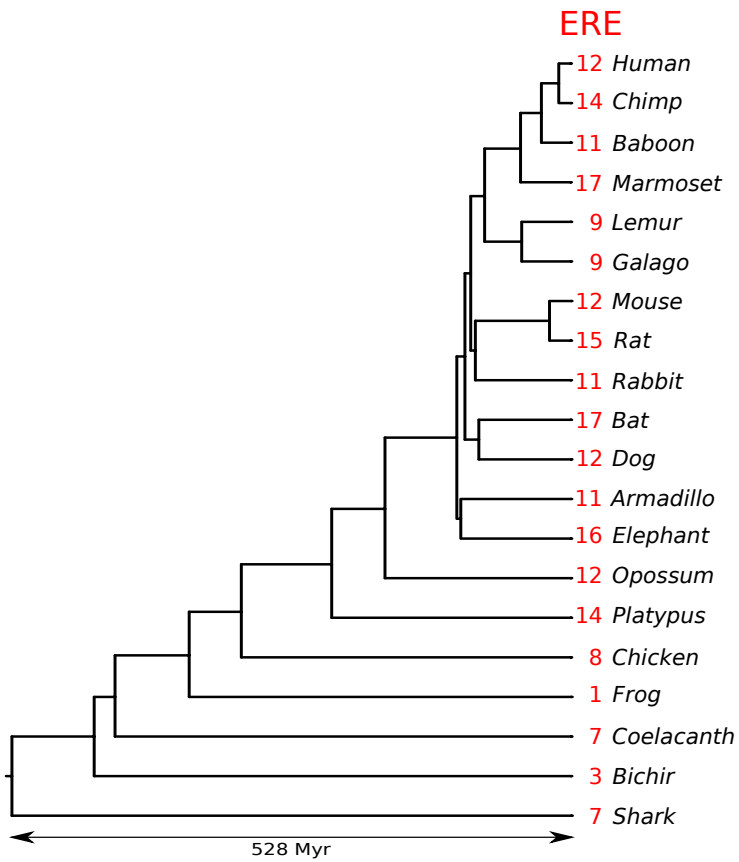
- evolutionary events are likely to cause lineage specific differences
- different rates can indicate timing and kind of evolutionary changes that affect gene regulation

Turnover in Evolutionary Context

- evolutionary events are likely to cause lineage specific differences
- different rates can indicate timing and kind of evolutionary changes that affect gene regulation
- problem: evolutionary rates are unknown

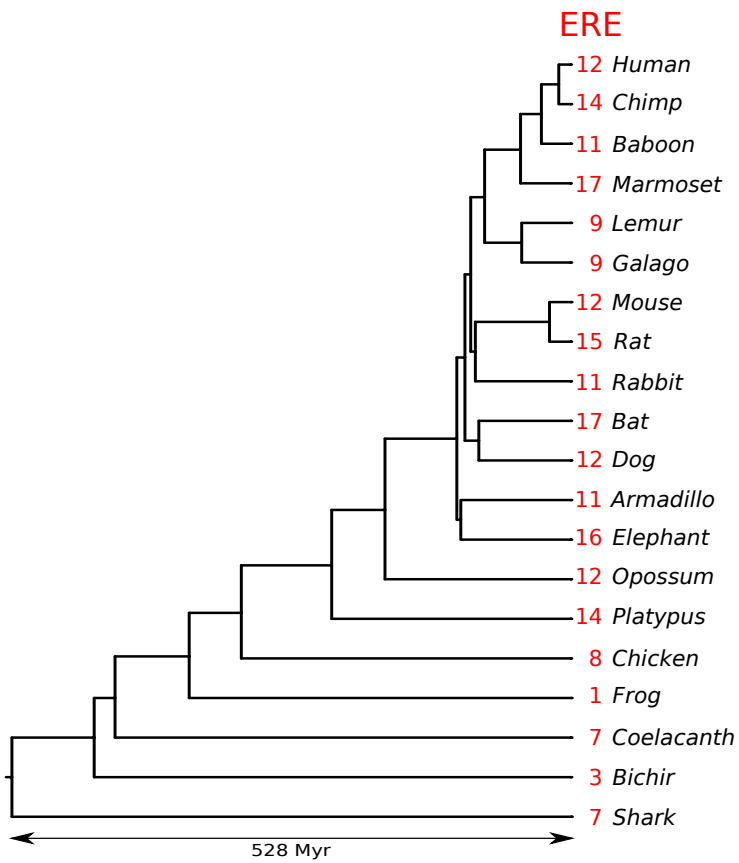
Turnover in Evolutionary Context

- evolutionary events are likely to cause lineage specific differences
- different rates can indicate timing and kind of evolutionary changes that affect gene regulation
- problem: evolutionary rates are unknown
- know phylogenetic relationships and binding site numbers for terminal nodes



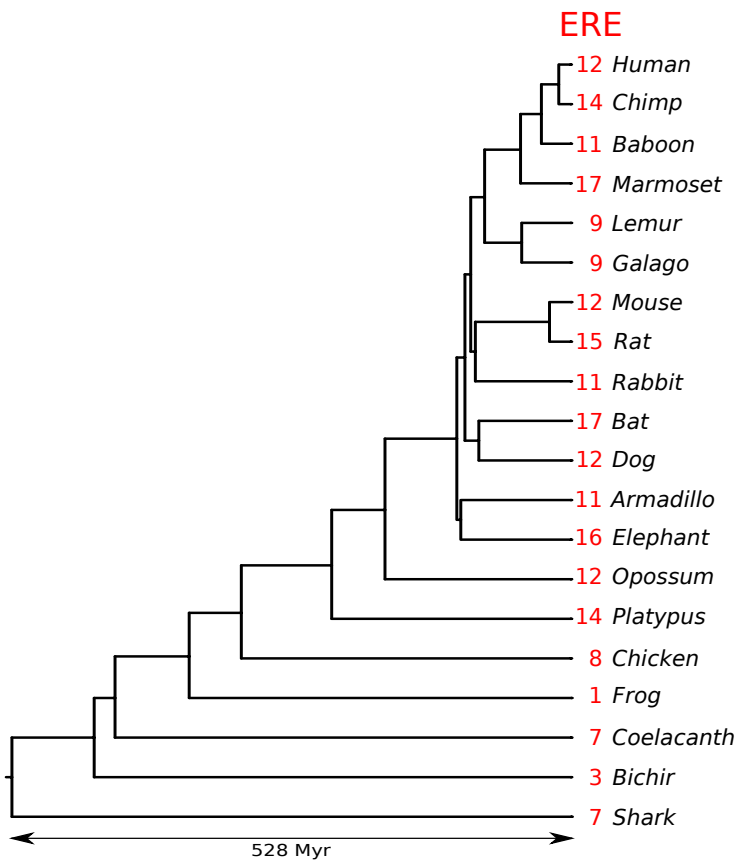
Model for Binding Site Turnover

- what are the most likely rates, how likely is given tree in respect to these rates?



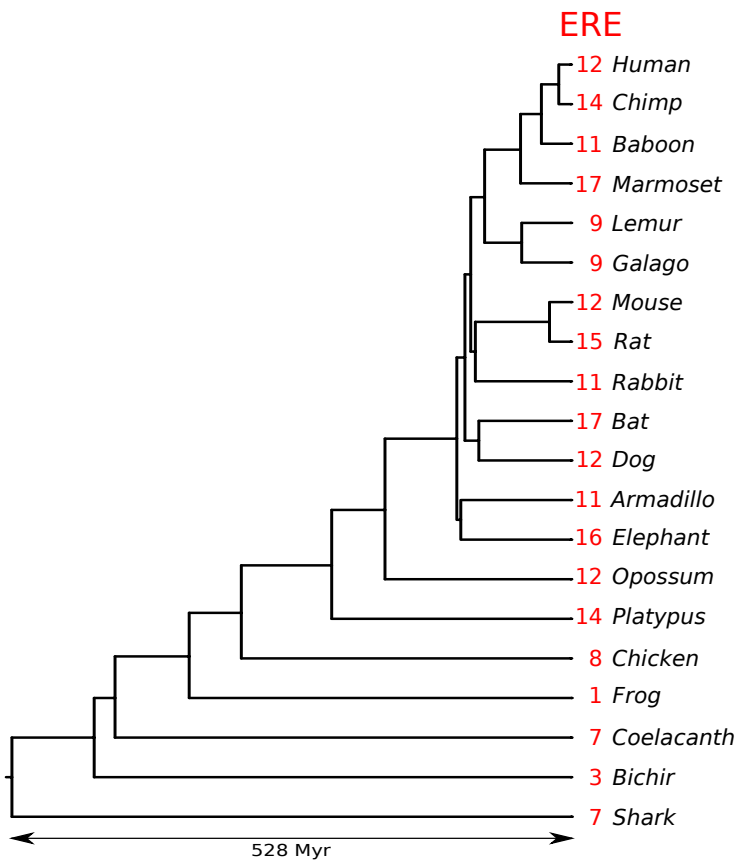
Model for Binding Site Turnover

- what are the most likely rates, how likely is given tree in respect to these rates?
- calculation of tree likelihood is easy $\Rightarrow$ need probability distribution of TFBS on branches



Model for Binding Site Turnover

- what are the most likely rates, how likely is given tree in respect to these rates?
- calculation of tree likelihood is easy $\Rightarrow$ need probability distribution of TFBS on branches
- what is the probability of having a specific binding site number, given start number, time and evolutionary rates?

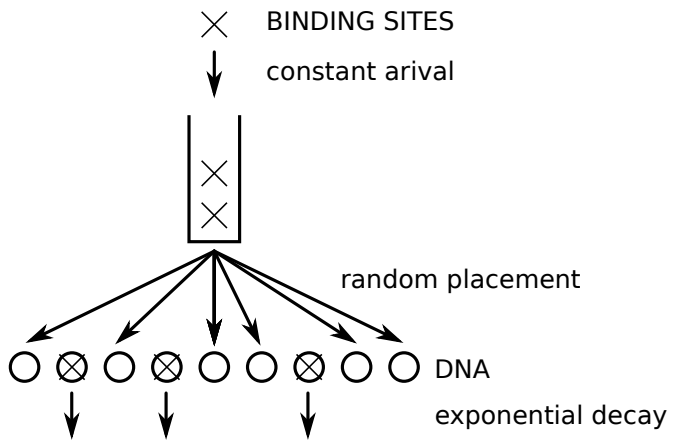


Model: Assumptions

- arrival of new TFBS is not influenced by number of existing TFBS

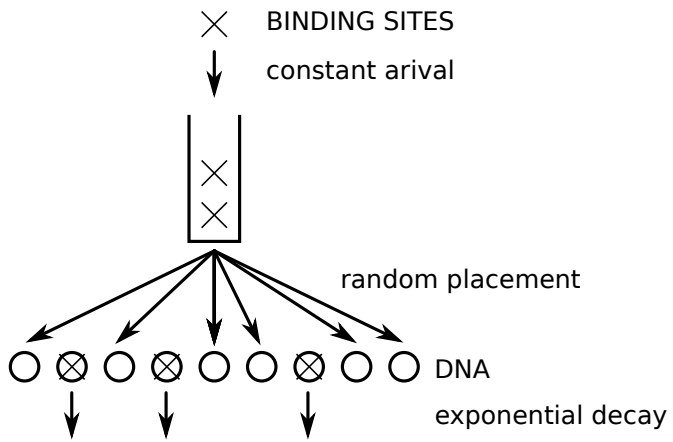
Model: Assumptions

- arrival of new TFBS is not influenced by number of existing TFBS
- TFBS arise with constant rate λ and decay with constant rate μ



Model: Assumptions

- arrival of new TFBS is not influenced by number of existing TFBS
- TFBS arise with constant rate λ and decay with constant rate μ



$\Rightarrow$ Kolmogorov forward equation:

$$\dot{p}_0(t) = -\lambda p_0(t) + \mu p_1(t) \quad (1)$$

$$\dot{p}_n(t) = \lambda p_{n-1}(t) - (\lambda + \mu n) p_n(t) + (n+1)\mu p_{n+1}(t) \quad (2)$$

Model: Sketch of Derivation

- replace probability distribution by generating function

Model: Sketch of Derivation

- replace probability distribution by generating function
- partial differential equation (PDE):

$$\frac{\partial P(z,t)}{\partial t} = \lambda(z-1)P(z,t) + \mu(1-z)\frac{\partial P(z,t)}{\partial z} \qquad (3)$$

Model: Sketch of Derivation

- replace probability distribution by generating function
- partial differential equation (PDE):

$$\frac{\partial P(z,t)}{\partial t} = \lambda(z-1)P(z,t) + \mu(1-z)\frac{\partial P(z,t)}{\partial z} \qquad (3)$$

- use characteristic equations to solve PDE

Model: Sketch of Derivation

- replace probability distribution by generating function
- partial differential equation (PDE):

$$\frac{\partial P(z,t)}{\partial t} = \lambda(z-1)P(z,t) + \mu(1-z)\frac{\partial P(z,t)}{\partial z} \qquad (3)$$

- use characteristic equations to solve PDE
- expected binding site number:

$$E[n(t)] = \frac{\lambda}{\mu}(1-e^{-\mu t}) + E[n(t=0)]e^{-\mu t} \qquad (4)$$

Model: Transient probability distribution

- solution:

$$p_n(t) = \frac{1}{n!} e^{-(\lambda/\mu)(1-e^{-\mu t})} \sum_{k=0}^{min(n_0,n)} k! \binom{n}{k} \binom{n_0}{k} \left(\frac{\lambda}{\mu}\right)^{n-k} \times (e^{-\mu t})^k (1-e^{-\mu t})^{n+n_0-2k} \quad (5)$$

Model: Transient probability distribution

- solution:

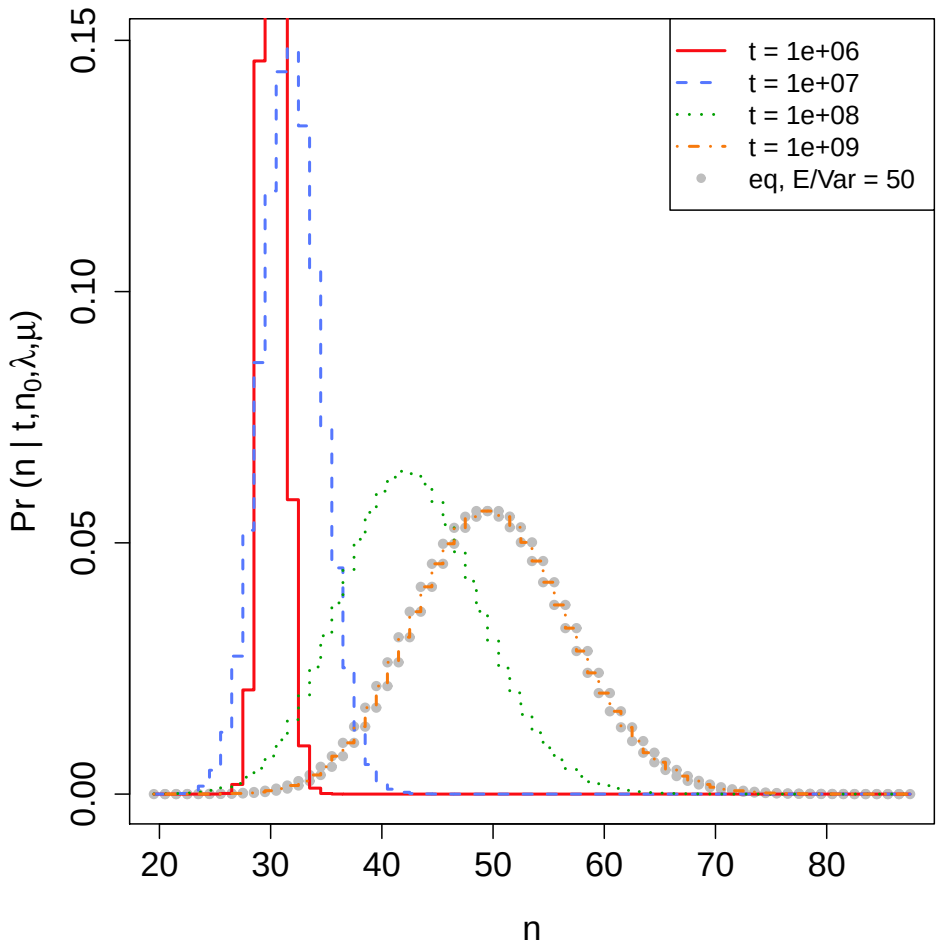
$$p_n(t) = \frac{1}{n!} e^{-(\lambda/\mu)(1-e^{-\mu t})} \sum_{k=0}^{min(n_0,n)} k! \binom{n}{k} \binom{n_0}{k} \left(\frac{\lambda}{\mu}\right)^{n-k} \times (e^{-\mu t})^k (1-e^{-\mu t})^{n+n_0-2k} \quad (5)$$

- for $t \to \infty$ follows stationary Poisson distribution:

$$\hat{p}_n = \left(\frac{1}{n!}\right) \left(\frac{\lambda}{\mu}\right)^n e^{-\lambda/\mu} \quad (6)$$

Characteristics

- probability for binding site numbers at different times
  - start binding site number for $t = 0$: $n_0 = 30$
  - origination rate: $\lambda = 5 \times 10^{-7}$
  - decay rate: $\mu = 1 \times 10^{-8}$
  - ratio: $\lambda/\mu = 50$.



Validation by Sequence Evolution

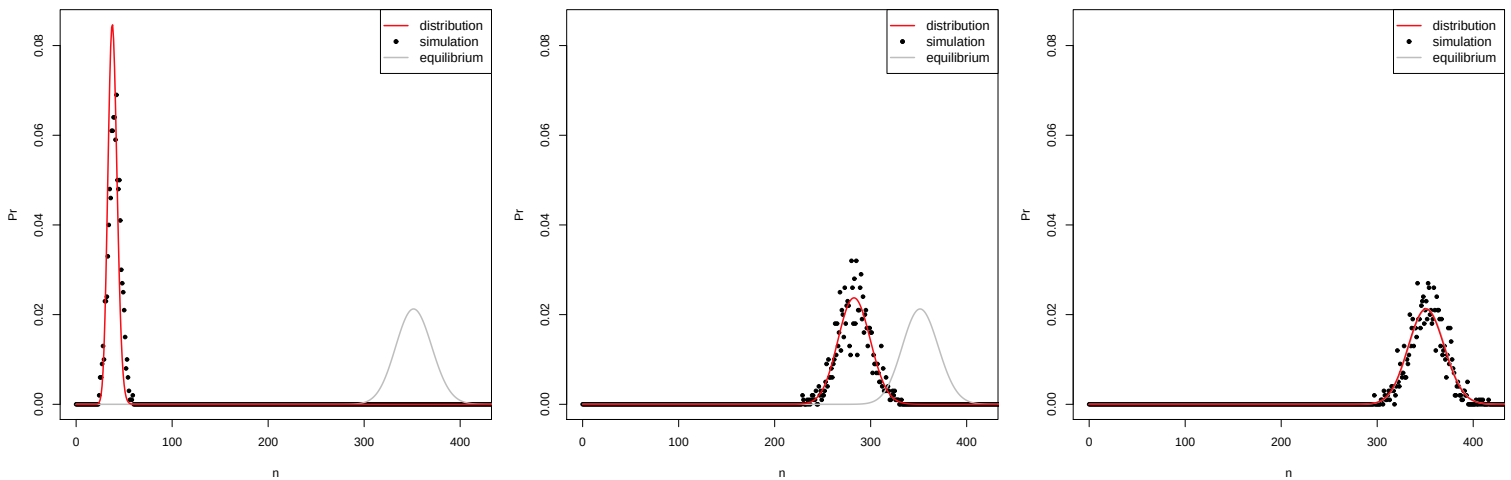
- simulate sequence evolution with specific mutation and fixation rate

Validation by Sequence Evolution

- simulate sequence evolution with specific mutation and fixation rate
- determine distribution by counting of TFBS at certain time points

Validation by Sequence Evolution

- simulate sequence evolution with specific mutation and fixation rate
- determine distribution by counting of TFBS at certain time points



Tree Likelihood

- tree likelihood:

$$L = \sum_{n=n_{min}}^{n_{max}} \pi(n) L_r(n) \qquad (7)$$

Tree Likelihood

- tree likelihood:

$$L = \sum_{n=n_{min}}^{n_{max}} \pi(n) L_r(n) \qquad (7)$$

- likelihoods of subtree defined by node i:

$$L_i(n) = \prod_{j \in children(i)} \sum_{m=n_{min}}^{n_{max}} Pr(m|n, t_j)\, L_j(m) \qquad (8)$$

Tree Likelihood

- tree likelihood:

$$L = \sum_{n=n_{min}}^{n_{max}} \pi(n) L_r(n) \tag{7}$$

- likelihoods of subtree defined by node i:

$$L_i(n) = \prod_{j \in children(i)} \sum_{m=n_{min}}^{n_{max}} Pr(m|n, t_j)\, L_j(m) \tag{8}$$

- breakup criteria for leaves:

$$L_i(n) = \begin{cases} 1 & : \quad n = bs(i) \\ 0 & : \quad else \end{cases} \tag{9}$$

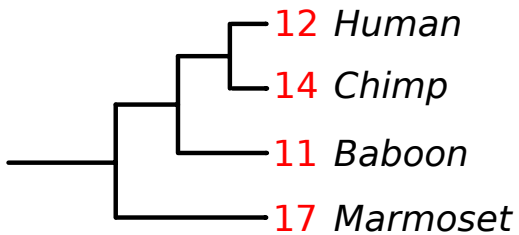
Tree Likelihood

- tree likelihood:

$$L = \sum_{n=n_{min}}^{n_{max}} \pi(n) L_r(n) \qquad (7)$$

- likelihoods of subtree defined by node i:

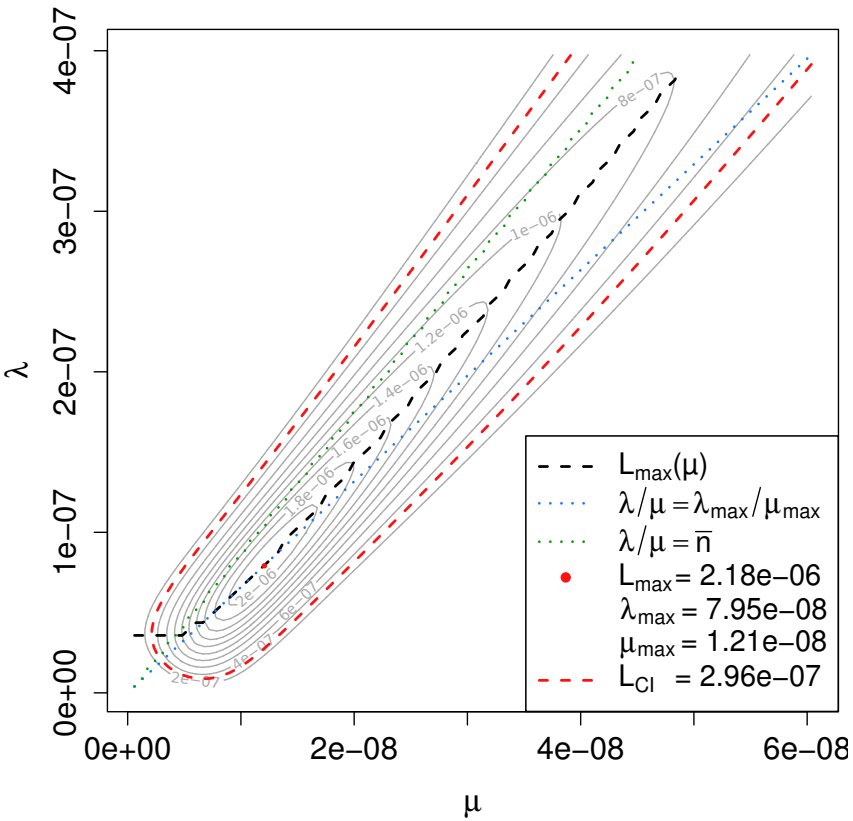
$$L_i(n) = \prod_{j \in children(i)} \sum_{m=n_{min}}^{n_{max}} Pr(m|n, t_j)\, L_j(m) \qquad (8)$$

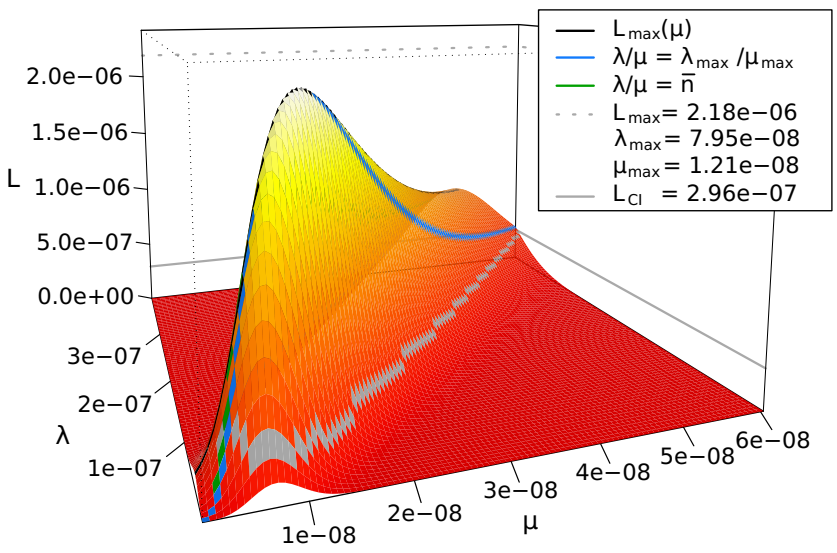


- breakup criteria for leaves:

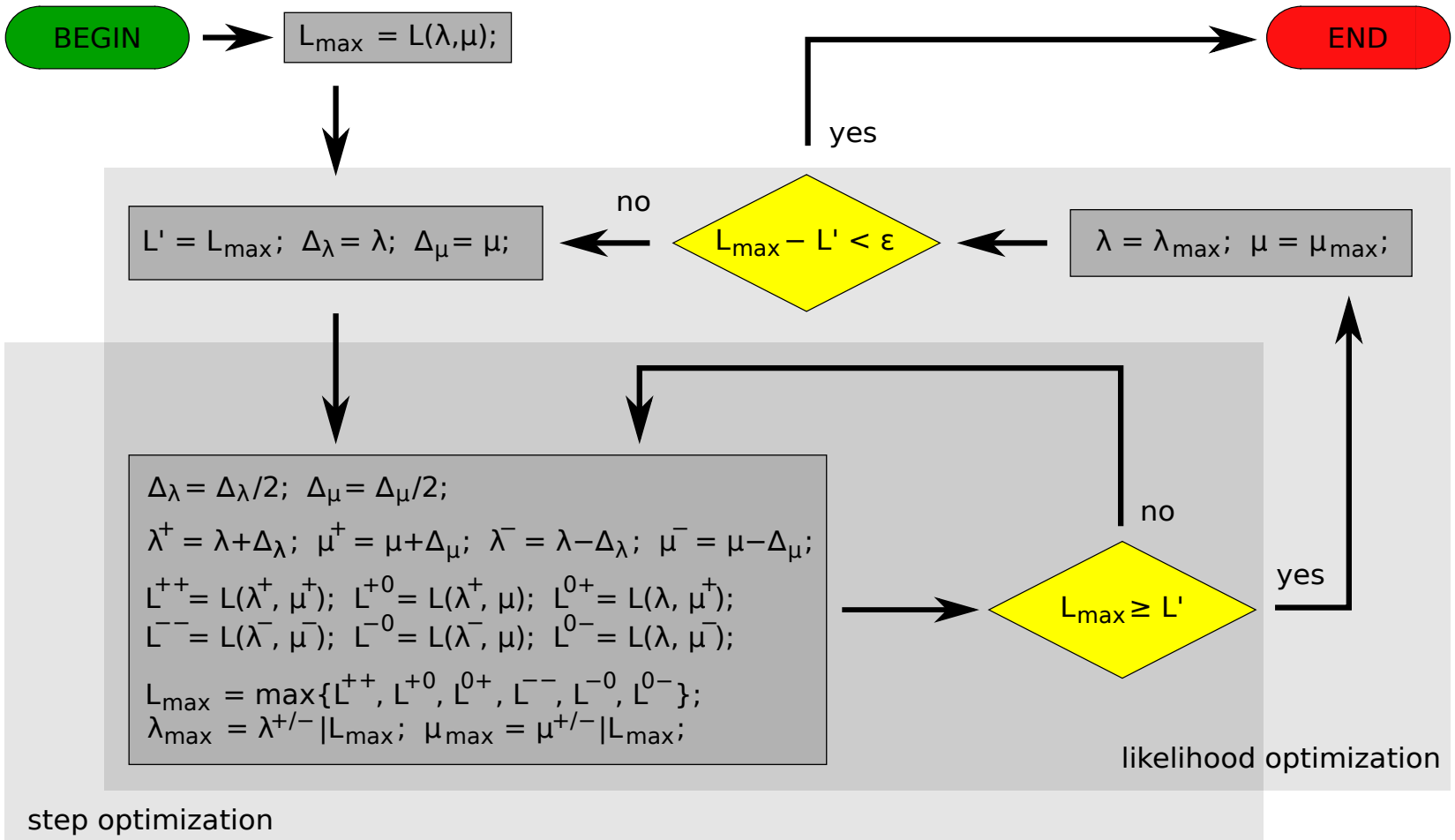
$$L_i(n) = \begin{cases} 1 & : \quad n = bs(i) \\ 0 & : \quad else \end{cases} \qquad (9)$$

Likelihood Landscape



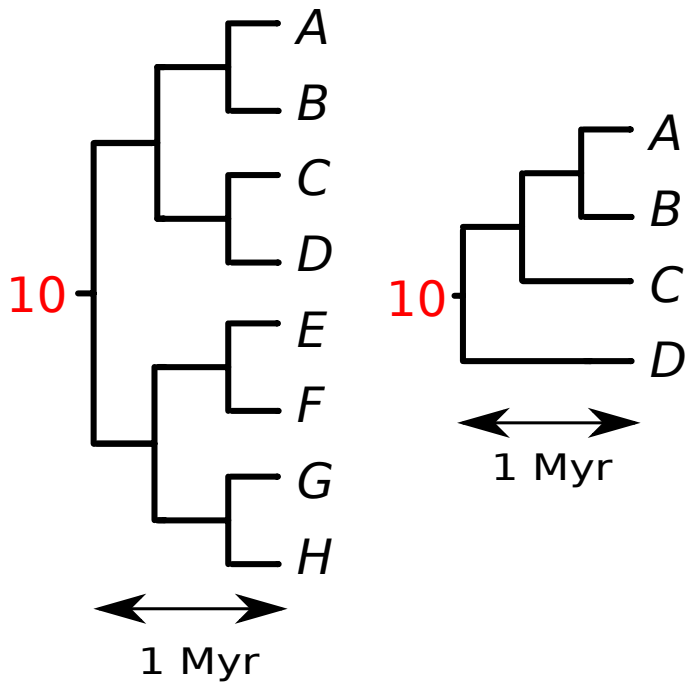


Optimization: Hill Climbing



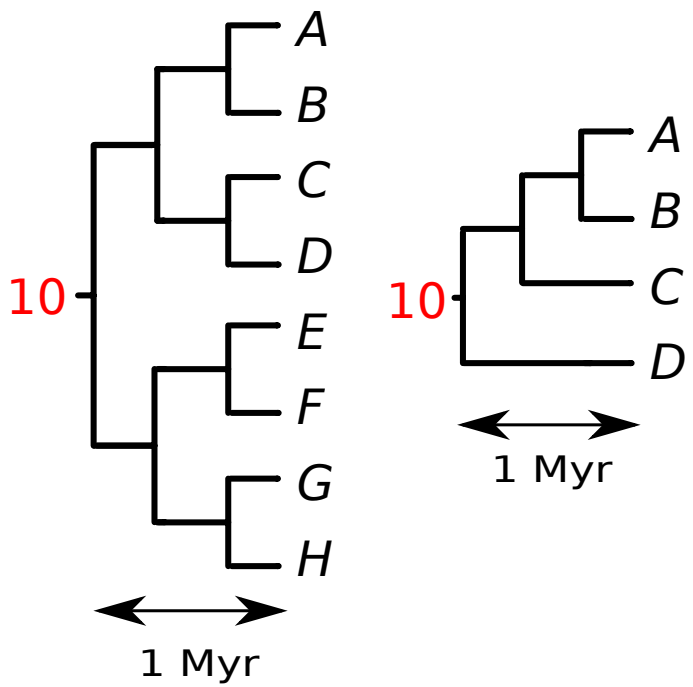
Results: Simulation of Evolution

- simulation on linear and binary trees with different taxa number



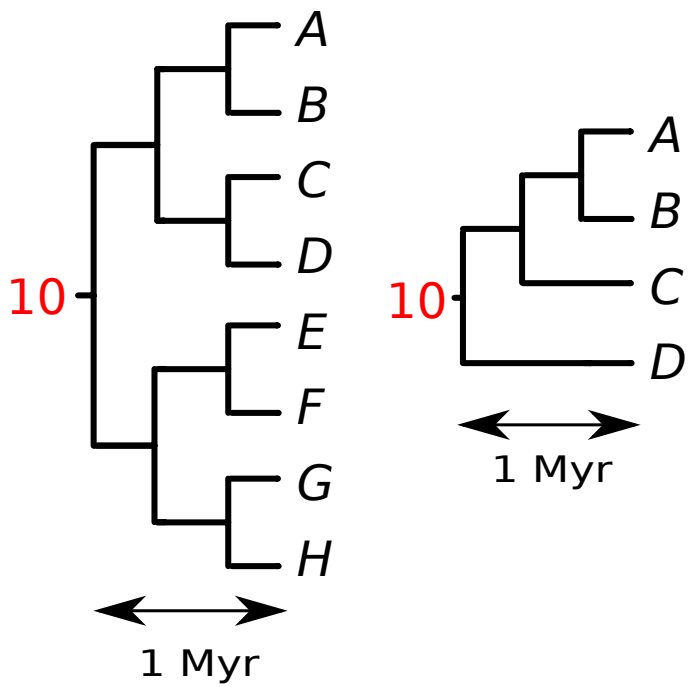
Results: Simulation of Evolution

- simulation on linear and binary trees with different taxa number
- number at root: 10



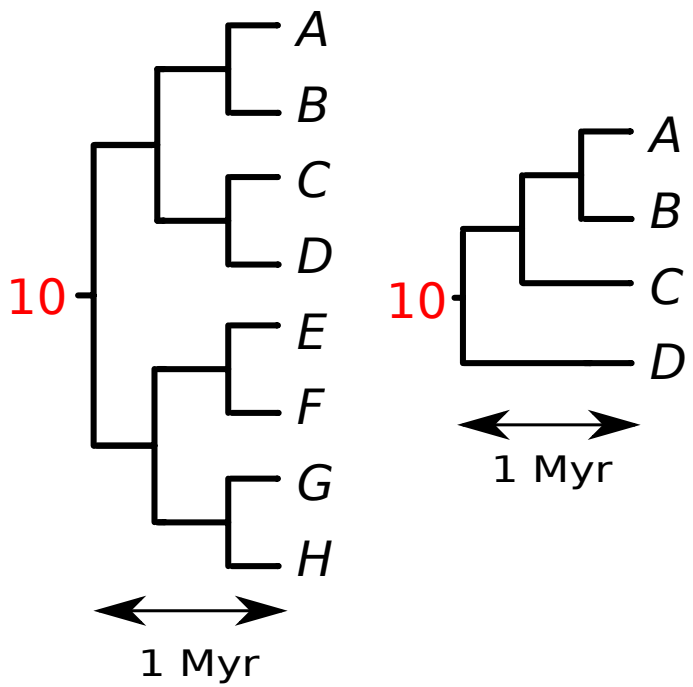
Results: Simulation of Evolution

- simulation on linear and binary trees with different taxa number
- number at root: 10
- simulation for different realtive clade ages (RCA = age of root / half-life time of TFBS)



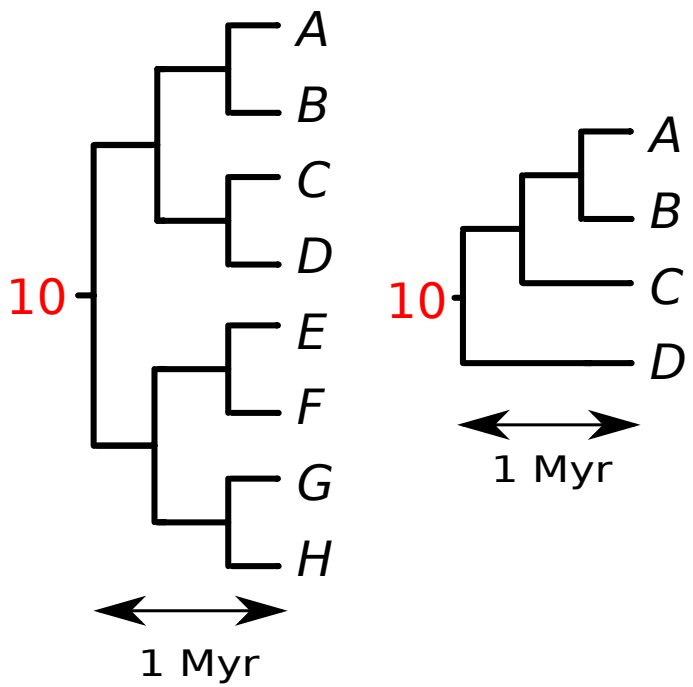
Results: Simulation of Evolution

- simulation on linear and binary trees with different taxa number
- number at root: 10
- simulation for different realtive clade ages (RCA = age of root / half-life time of TFBS)
- age of root node: 10^6, adjust age by μ and λ

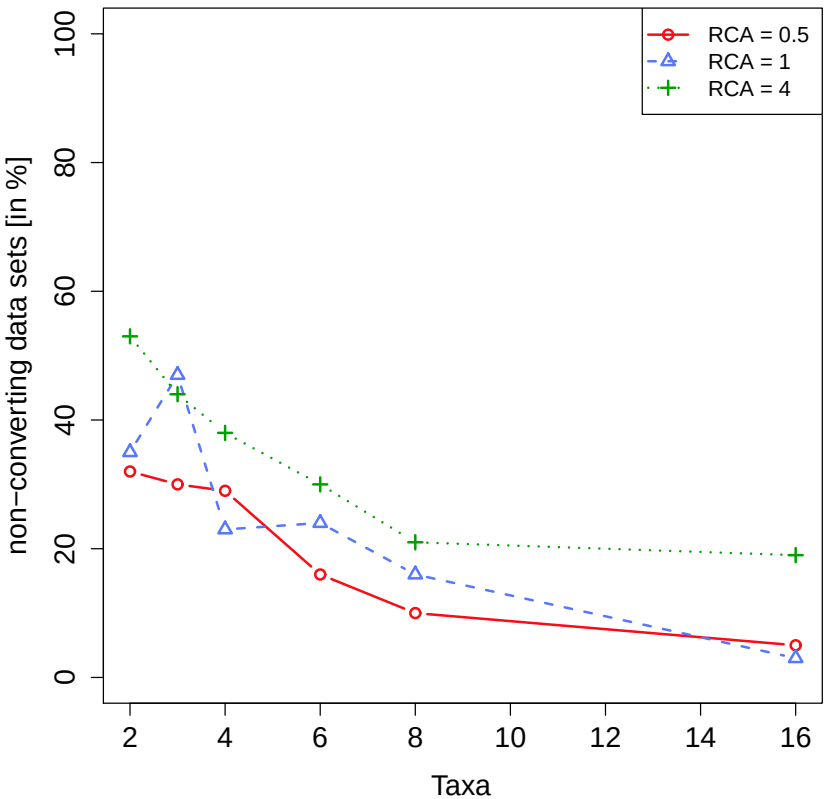


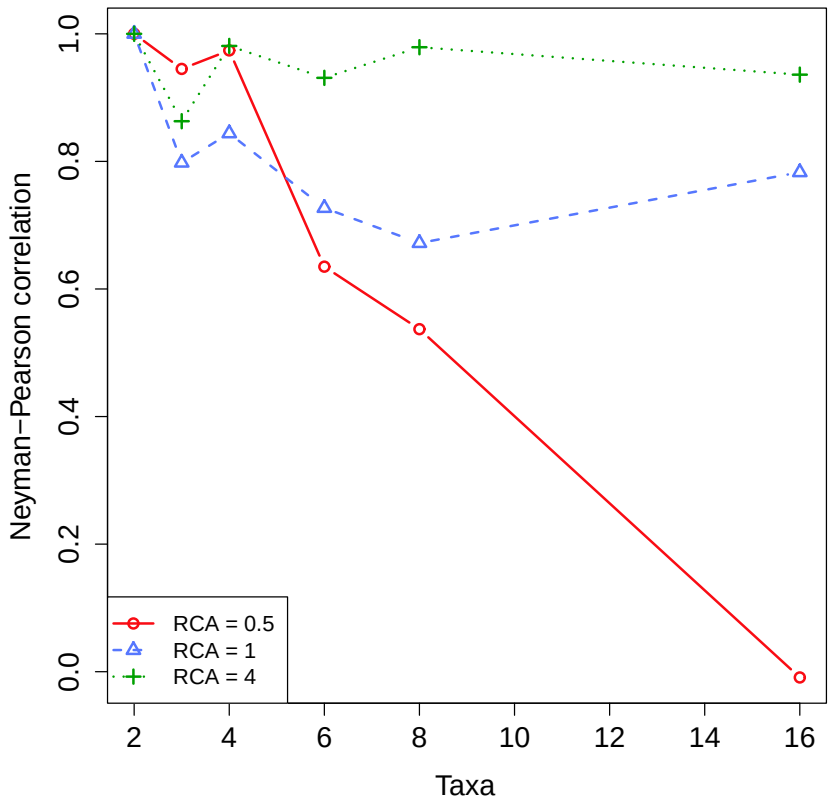
Results: Simulation of Evolution

- simulation on linear and binary trees with different taxa number
- number at root: 10
- simulation for different realtive clade ages (RCA = age of root / half-life time of TFBS)
- age of root node: 10^6, adjust age by μ and λ
- draw TFBS number at inner nodes randomly based on distribution

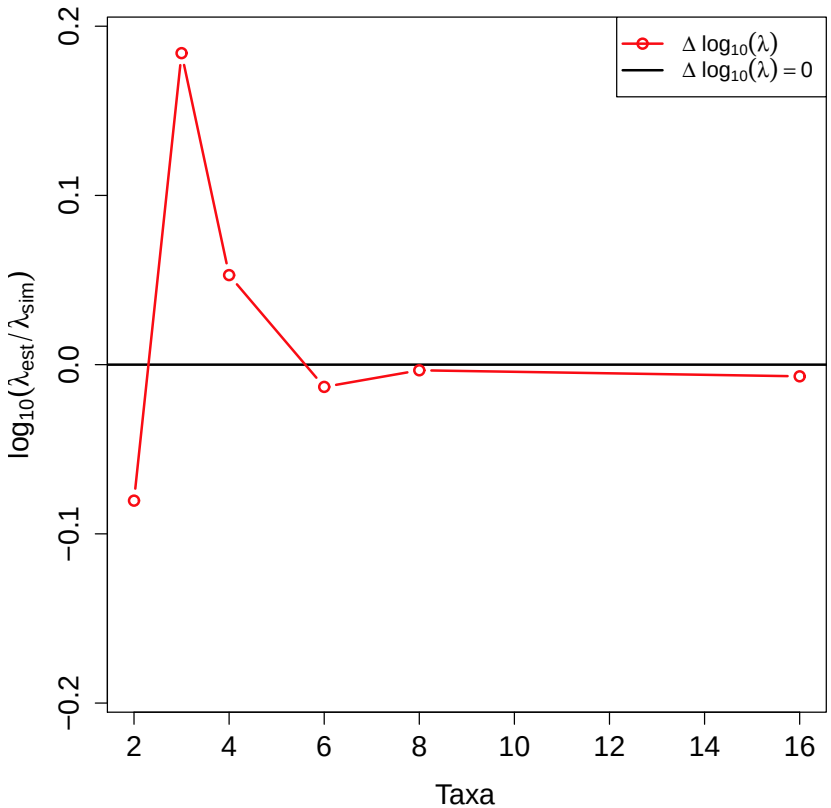


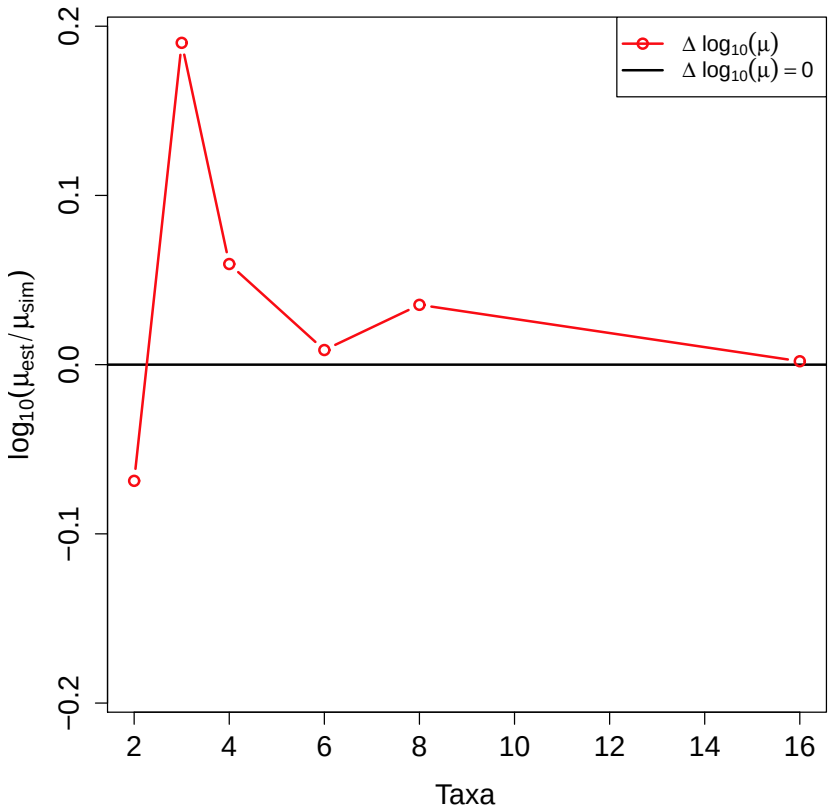
Results: Conversion and λ/μ–$\bar{n}$–Correlation



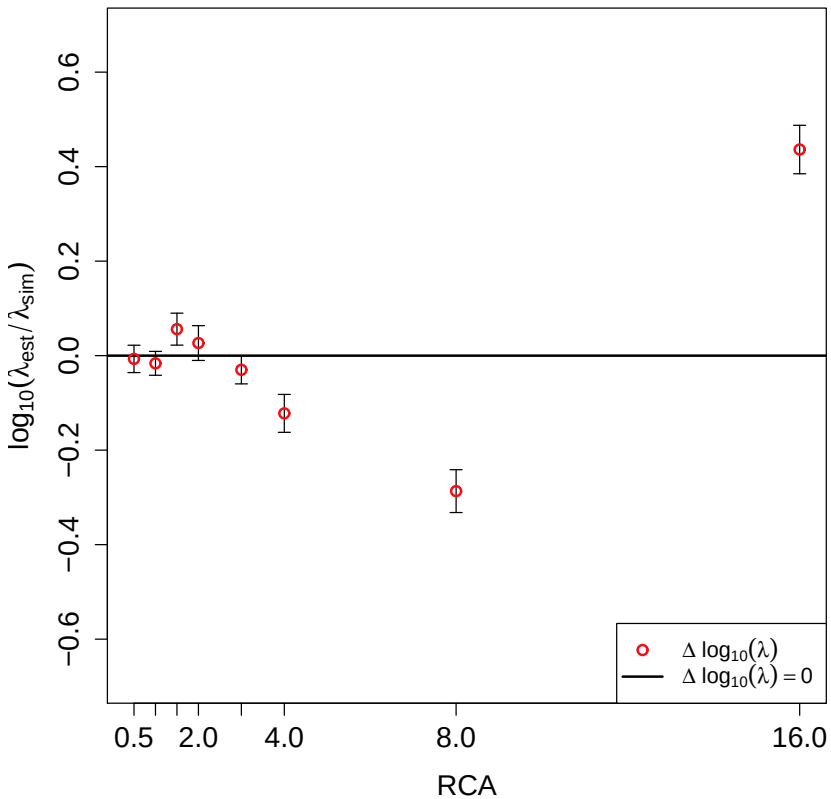


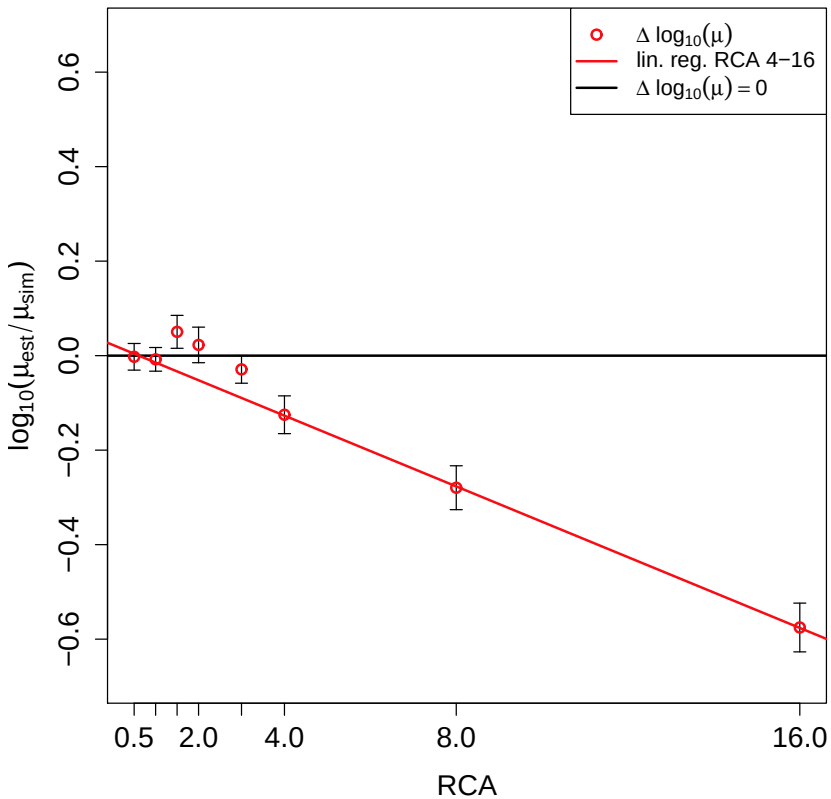
Results: Accuracy in Dependence of Taxa



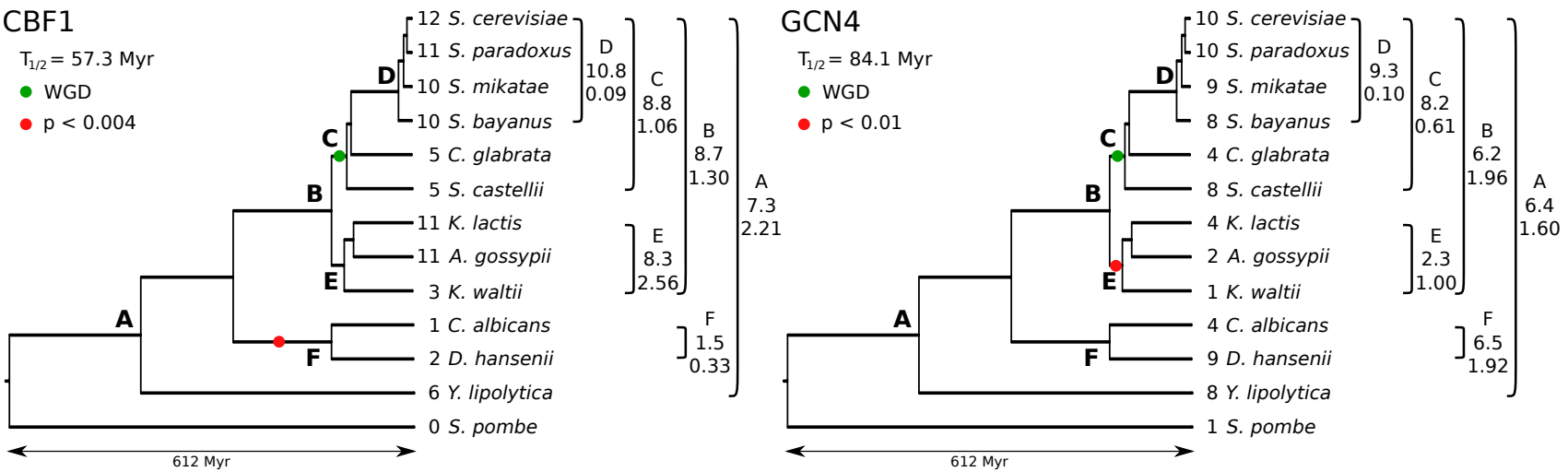


Results: Accuracy in Dependence of RCA

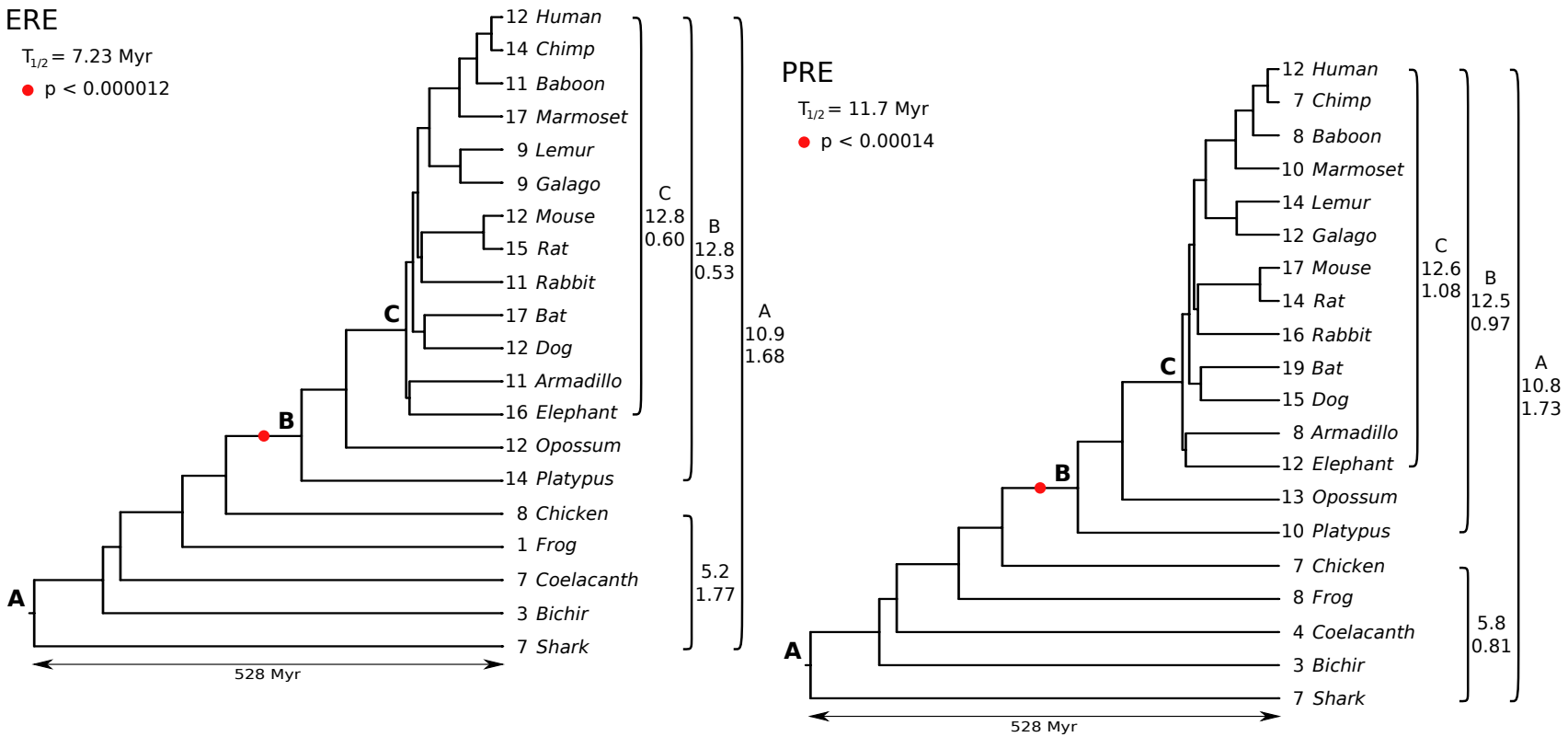




Results: Evolution of Methionine Pathway in Yeast



Results: Evolution of Vertebratee HoxA Clusters



Summary: Evolutionary Analysis of Regulatory Elements

- completely new approach independent of conserved regulatory sequences (works even with data where turnover changed location and arrangement of binding sites)

Summary: Evolutionary Analysis of Regulatory Elements

- completely new approach independent of conserved regulatory sequences (works even with data where turnover changed location and arrangement of binding sites)
- simple, mathematically non-trivial, phenomenological model for binding site number evolution at a genomic locus

Summary: Evolutionary Analysis of Regulatory Elements

- completely new approach independent of conserved regulatory sequences (works even with data where turnover changed location and arrangement of binding sites)
- simple, mathematically non-trivial, phenomenological model for binding site number evolution at a genomic locus
- allows detection of heterogeneity in rate of origination/decay between different lineages/clades $\Rightarrow$ hints for functionally important changes in the evolution of regulation

Final Summary

- `tracker` and `creto` present complete new approaches to well known problems concerning detection and evolutionary analysis of regulatory elements

Final Summary

- `tracker` and `creto` present complete new approaches to well known problems concerning detection and evolutionary analysis of regulatory elements
- both programs have been tested on artificial and biological data and are available for download via homepage of institute

Acknowledgments

- I want to thank:
  - Peter F. Stadler and Sonja Prohaska
  - Gunter P. Wagner, Charlie and Vinny
  - all colleagues and friends, especially Linda

Acknowledgments

- I want to thank:
  - Peter F. Stadler and Sonja Prohaska
  - Gunter P. Wagner, Charlie and Vinny
  - all colleagues and friends, especially Linda
- My PhD-time was supported by:
  - International Max Planck Research School 'Mathematics in the Sciences'
  - Konrad–Adenauer–Foundation

Thank You

Then the Dean repeated the mantra that has had such a marked effect on the progress of knowledge throughout the ages.

"Why don't we just mix up absolutely everything and see what happens?" he said.

And Ridcully responded with the traditional response.

"It's got to be worth a try." he said.

Terry Pratchett, Hogfather