Phylogenetic Footprints and Consitent Sets of Local Alignments

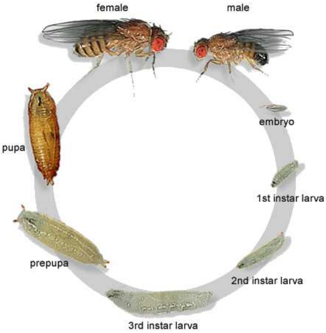
Wolfgang Otto

Bioinformatics Leipzig, IMPRS–MIS, UFZ

22nd Annual Symposium on Combinatorial Pattern Matching
June 2011

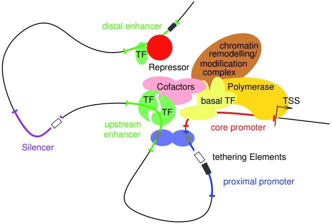
Gene Expression and Regulation

- life depends on ability of cells to synthesize information from genes into corresponding products
- control the timing, the location, and the amount of gene expression is crucial
- regulation is basis for differentiation, morphogenesis and versatility and adaptability of any organism
- understanding of process is one of the main targets in life sciences



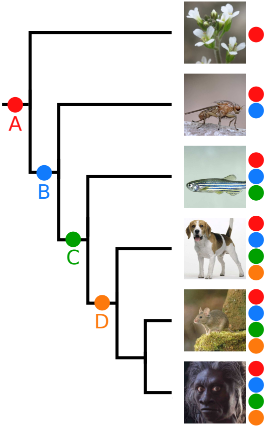
Regulation of Transcription

- complicated process
- molecules bind to *regulatory elements* on DNA and modify production rate of functional element
  - wide variety of mechanism exists
  - enhancer directly increase rate of transcription
  - silencers prevent transcription of genes.
- how to find regulatory elements?



Detection of Regulatory Elements

- regulatory elements are crucial for all 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 ⇒ phylogenetic footprinting

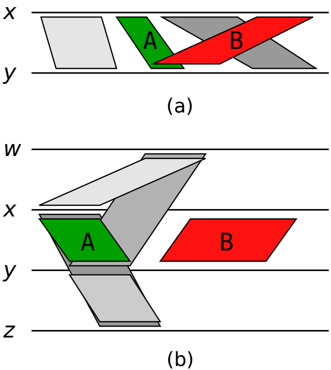


Bioinformatic Challenge

- search for short motifs (down to 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

Phylogenetic Footprinting

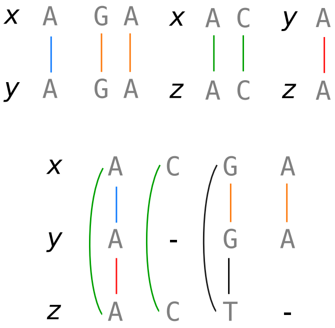
- use evolutionary information:
  - (a) order of motifs defines windows for new motifs ⇒ consistence
  - (b) function motifs are widely conserved ⇒ support
- existing approaches: multiple alignments disregard segments, local alignments disregard order information
- idea: calculate pairwise local alignments with low stringency, determine maximal consistent subsets based on support



Consistent Alignments

Definition (Consistency)

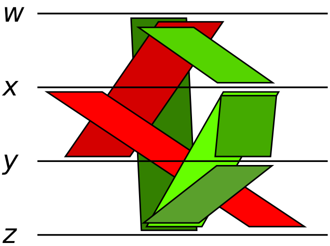
An alignment collection $\mathcal{A} = \{A_1, \ldots, A_n\}$ over sequences $\mathcal{S} = \{S_1, \ldots, S_m\}$ is *consistent* ⇔ it exists a multiple alignment M over $\mathcal{S}$ so that all pairs of nucleotides aligned by alignments in $\mathcal{A}$ are also aligned in M.



Optimization Problem

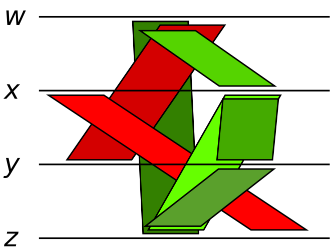
Definition (Maximal Consistent Alignment Subset Problem)

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.



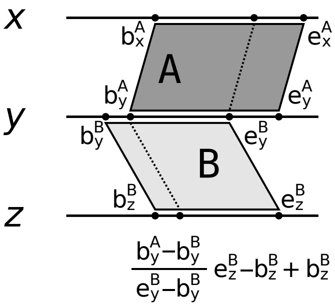
Complexity of MCASP

- MCASP is NP-complete (contains Multiple Alignment Problem)
- optimal solution: check each subset for consistency
- exponential growth
  - 7 alignments: 128 subsets
  - 250 alignments: 10^{75} subsets
- need for heuristic approach



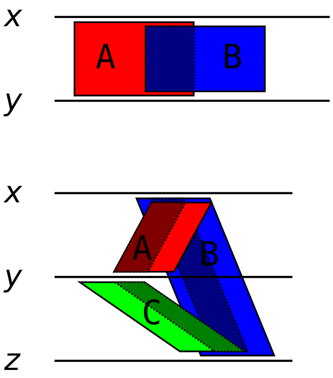
Algorithmic Sketch

- abstract alignments by intervals $A = \{[x, b_x, e_x], [y, b_y, e_y]\}$
- calculate intermediate positions by linear interpolation
- construct M by iteratively checking all alignments $A \in \mathcal{A}$
- consistent alignments are inserted, inconsistent are rejected
- inserted alignments cannot be removed or corrected ⇒ insertion order is crucial



Extended Scores

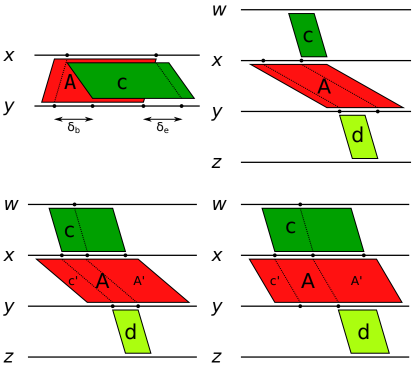
- start with alignments that are most supported by other alignments
- express support by score ⇒ extended scores
- similar to `T-Coffee`[a]
- basic score plus bonus for each direct / indirect support



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

Assembly

- inserted alignments define alignment columns
- accept small contradictions with error rate δ
- insertion can cause switch, merge or split of columns and alignment
- insertion of n alignments over m sequences with length l is in $O(nlm)$
- calculation of extended scores is in $O(n^3)$



Maximal Consistent Subsets

- artificial data sets $\mathcal{A}$
    - m sequences, each with l motifs
    - prob. for ith motif to be k is

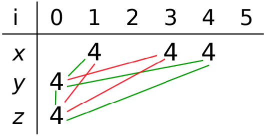
$$p_i(k) = \frac{i^k}{k!}e^{-i}$$

(⇒ conflicts, diff. support)

- permut. of motifs (⇒ crossings)
- alignments between equal motifs inserted with prob. $e/(m-1)$, $1 \leq e < m$ (⇒ evol. distance)
- variation of m, l, e / 250 sim.
- comparison with optimal solutions (NP-complete algorithm)

i	0	1	2	3	4	5
x	0	1	3	4	4	5
y	0	1	2	3	3	4
z	0	1	3	4	5	5

i	0	1	2	3	4	5
x	0	4	1	4	4	3
y	4	3	0	1	3	2
z	4	1	5	3	0	5

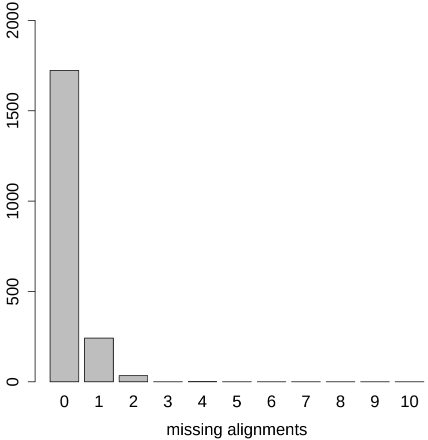


Maximal Consistent Subsets

Model (m/l/e)	$\|\mathcal{A}\|$	$\|\mathcal{A}'\|$	#Opt.	#Heur.	Correct (in %)	Optimal (in %)
3/8/1	10.65	5.36	21.62	3.64	100.00	86.80
4/3/1	5.96	4.16	6.74	2.08	100.00	96.40
4/3/2	11.60	6.84	29.11	3.06	100.00	87.20
4/3/3	15.72	8.66	51.40	3.91	100.00	96.40
4/4/1	7.58	5.01	12.21	2.36	100.00	96.00
4/6/1	10.90	6.54	32.65	2.95	100.00	87.60
4/8/1	14.09	8.02	63.80	3.38	100.00	72.80
5/8/1	16.22	10.00	140.58	3.05	100.00	65.20

Maximal Consistent Subsets

- all heuristic results are consistent subsets that are maximal
- optimal result found in most cases
- number of missing alignments relative to optimal solution is low



Alignment Calculation

- BRaliBase II database with RNA families and reference alignments
- calculation of pairwise alignments with ClustalW2 (seq) and LocARNA(seq-struc)
- splitting of alignment columns in single edges and insertions of all edges in $\mathcal{A}$
- calculation of consistent subsets and multiple alignment M
- correctness: percent of reference edges in M
- comparison with other alignment programs

Alignment Calculation

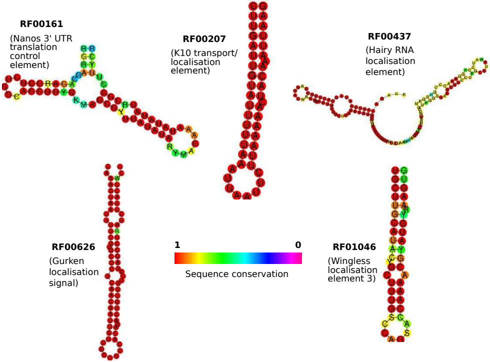
Program	GII Intron	5S rRNA	SRP RNA	tRNA	U5 RNA
Heur. (ClustalW2)	73.77	92.88	87.10	86.27	79.58
Heur. (LocARNA)	76.35	94.48	87.43	96.05	83.65
ClustalW2	72.84	93.24	87.43	87.06	79.61
DIALIGN-TX	72.08	91.69	82.92	78.53	77.80
T-Coffee	79.29	94.59	87.31	92.00	83.55
MAFFT	77.20	93.83	87.10	90.14	80.43
MUSCLE	76.43	94.04	87.03	87.27	79.76
ProbCons	78.69	93.67	86.92	89.82	83.28

Alignment Calculation (True Positives)

Program	GII Intron	5S rRNA	SRP RNA	tRNA	U5 RNA
Heur. (ClustalW2)	76.57	93.18	86.97	87.52	81.05
Heur. (LocARNA)	78.18	94.13	87.28	96.08	84.59
ClustalW2	71.89	92.59	86.32	87.04	79.06
DIALIGN-TX	79.46	92.46	84.51	81.18	81.39
T-Coffee	79.02	93.91	86.65	92.01	83.69
MAFFT	78.17	93.23	86.10	90.35	81.00
MUSCLE	77.62	93.70	86.22	87.79	80.32
ProbCons	78.63	93.17	86.29	90.16	83.41

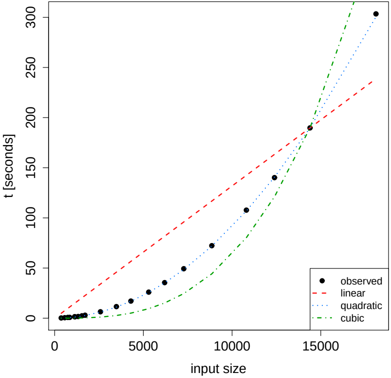
Pattern recognition

- merge 15 sequences of 5 different RNA families (Rfam)
- calculate local, pairwise alignments with LocARNA (seq-struc)
- check multiple alignment for patterns that are characteristic for RNA family


RF00161
(Nanos 3' UTR translation control element)
RF00207
(K10 transport/ localisation element)
RF00437
(Hairy RNA localisation element)
RF00626
(Gurken localisation signal)
1
0
Sequence conservation
RF01046
(Wingless localisation element 3)

Complexity

- artificial data sets with different set sizes
- measure amount of time *t* for calculation of all solutions,
- comparison with linear (red), quadratic (blue) and cubic function (green)
- scaled by a linear factor, all curves go through penultimate data point



Acknowledgments

- I thank:
  - Peter F. Stadler
  - Sonja Prohaska
  - Linda Gerlach
- This project was supported by:
  - International Max Planck Research School for Math in Sciences
  - Helmholtz Centre for Environmental Research