NcDNAlign - a flexible and efficient package for the generation of non- protein coding multiple alignments from genomic sequences

Dominic Rose[1], Jana Hertel[2], Kristin Reiche[1], Peter F. Stadler[1,2,3,4], Jörg Hackermüller[2,4]

[1]Bioinformatics Group, Department of Computer Science, University of Leipzig, Leipzig, Germany; [2]Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig, Germany; [3]The Santa Fe Institute, Santa Fe, New Mexico,USA; [4] Fraunhofer Institute for Cell Therapy and Immunology, Leipzig, Germany

At a glance…

Overall goal: Efficiently and customizably generate multiple sequence alignments of non-coding genomic sequences .

Approach:

- Pipeline concept, implemented in PERL, combining custom written and standard bioinformatic software tools
- Only sequence data required as input
- Sequence regions not of interest can be discarded beforehand
- Follows a progressive paradigm starting form pairwise BLAST alignments
- Uses a configurable heuristics for alignment “beautification”
- Proof of principle applications to identification of ultra-conserved elements in nematodes and fishes and ncRNA finding using RNAz in *Gammaproteobacteria*

Results/Conclusions:

NcDNAlign…

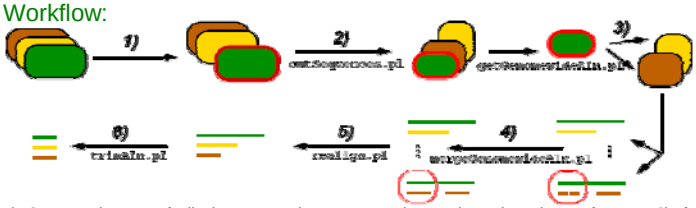
- Displays comparable sensitivity and specificty as TBA in RNA gene finding using RNAz
- Aligns less DNA than TBA, albeit at higher alignment quality
- Significantly outperforms TBA in computational effort
- Appears to be a reasonable alternative to TBA for pilotstudies, when infrastructure is limited, or when analysis pipelines should be rerun frequently to incorporate newly available data

Availability:

- NcDNAlign is available under the GNU Public License from http://www.bioinf.uni-leipzig.de/Software/NcDNAlign/.

1) NcDNAlign

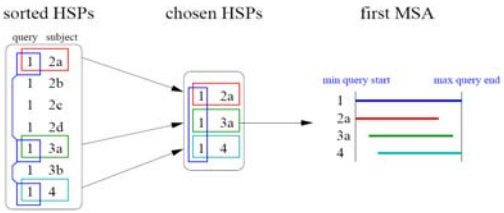
Workflow:



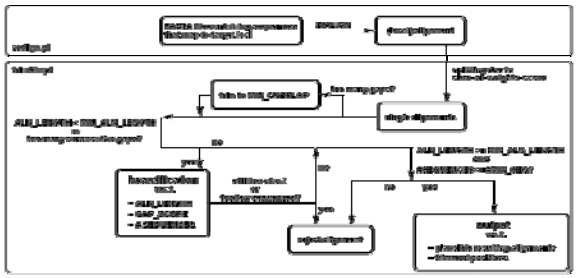
1) One species out of all given genomic sequences has to be selected as reference. 2) If selected, sequence data are ridded of potentially interfering or uninteresting sequence stretches, reducing the data set to genomic sub sequences. 3) All subsequences of the reference are compared to all subsequences of all other species and local alignments are calculated heuristically (BLAST). 4) Adjacent compatible hits are combined. 5) The best hits (E-value) in each organism for each subsequence of the reference are aligned (DIALIGN). 6) Finally, the alignments are pruned, poorly aligned sequences are removed and the remaining sequences are optionally realigned to obtain an optimal alignment.

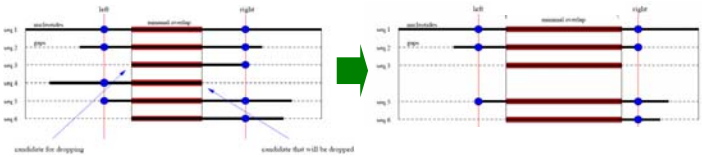
Generating multiple alignments:

MSAs are set up by grouping heuristic pairwise alignments. Starting from tabular BLAST output, HSPs are sorted by query position and one “best” representative is chosen.



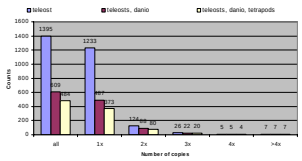
Beautifying Alignments:

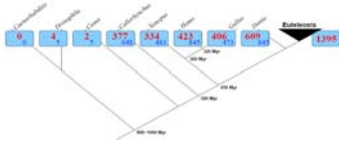




2) Results: Teleost UCRs

- **Identification of ultra conserved elements (UCE) in teleost genomes using NcDNAlign**
- **UCE defined as 100% sequence identity among at leas three sequences in alignment and min. 50 nt in length**
- **Study effect of teleost specific duplication on UCRs**





3) Results: Structured ncRNAs in *Gammaproteobacteria*

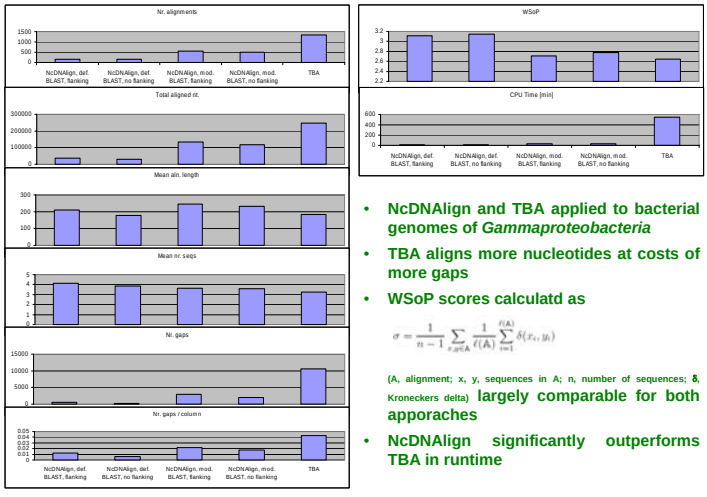
Flanking regions	NcDNAlign default Blast + (1)	NcDNAlign default Blast - (2)	NcDNAlign modified Blast + (3)	NcDNAlign modified Blast - (4)	TBA n.a. (5)
(a)			**CPU time**		
Total [min]	15.68	14.35	29.27	27.52	548.36
(b)			**Alignments**		
Nr. alignments	169	169	542	499	1347
Nr. overlapping aln.	155		483		479 [a]
Total align. nucleotides	35 153	29 596	125 907	113 829	235 986
% of *E. coli* genome	0.76 %	0.64 %	2.71 %	2.45 %	5.09 %
(c)			***RNAz***		
Nr. hits	126	122	339	300	658
Overlap	99		280		393 [b]
Overall length of hits	25 100	20 618	94 995	80 680	92 888
Mean length of hits	199	169	280	269	141
FDR	0.32	0.33	0.25	0.24	0.25
Nr. annotatable hits	102 (.80)	98 (.80)	212 (.62)	189 (.63)	469 (.71)
Nr. non-annot. hits	24 (.19)	24 (.19)	127 (.37)	111 (.37)	189 (.29)
(d)			**Sensitivity**		
rRNA, 22 annot.	10 (.46)	9 (.41)	14 (.64)	14 (.64)	14 (.64)
tRNA, 86 annot.	55 (.64)	61 (.71)	62 (.72)	62 (.72)	56 (.65)
Misc. RNA, 49 annot.	5 (.10)	6 (.12)	18 (.37)	16 (.33)	23 (.47)
Overall, 157 annot.	70 (.46)	76 (.50)	94 (.60)	92 (.59)	93 (.59)

(a) Consuming approximately the 20 to 30-fold amount of CPU time, TBA produces roughly 2.5 fold more alignments than NcDNAlign

(b) RNAz applied to these alignments yields half the number of hits in NcDNAlign as compared to TBA at comparable false discovery rates (c).

(d) Despite the discrepancy in the number of alignments, the sensitivity of the RNAz-based detection of annotated ncRNAs is equal for both approaches.

4) Comparison of NcDNAlign and TBA



- **NcDNAlign and TBA applied to bacterial genomes of *Gammaproteobacteria***
- **TBA aligns more nucleotides at costs of more gaps**
- **WSoP scores calculatd as**

$$\sigma = \frac{1}{n-1} \sum_{x,y \in A} \frac{1}{\ell(A)} \sum_{i=1}^{\ell(A)} \delta(x_i, y_i)$$

(A, alignment; x, y, sequences in A; n, number of sequences; δ, Kroneckers delta) largely comparable for both apporaches

- **NcDNAlign significantly outperforms TBA in runtime**

Correspondence:
Jörg Hackermüller
Phone: +49 (0) 341 3550 842
Fax: +49 (0) 341 3550 855
email: joerg.hackermueller@izi.fraunhofer.de

Funding: German DFG Bioinformatics Initiative BIZ-6/1-2 and 6th Framework Programme of the European Union (SYNLET)

Institute Address:
Fraunhofer Institut für Zelltherapie und Immunologie IZI
Deutscher Platz 5e
D-04103 Leipzig, Germany
Phone: +49 (0) 3 41 / 355 36 - 0
Fax: +49 (0) 3 41 / 355 36 - 109
www.izi.fraunhofer.de