NcDNAlign: Plausible Multiple Alignments of Non-Protein-Coding Genomic Sequences

- SUPPLEMENT -

Dominic Rose [a], Jana Hertel [b], Kristin Reiche [a],
Peter F. Stadler [a,b,c,d], Jörg Hackermüller [e,*]

[a] *Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*

[b] *Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*

[c] *Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*

[d] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

[e] *Fraunhofer Institute for Cell Therapy and Immunology — IZI Deutscher Platz 5e, D-04103 Leipzig, Germany*

---

* Corresponding author. Fax: +49 341 3550 855
Email addresses: dominic@bioinf.uni-leipzig.de (Dominic Rose), jana@bioinf.uni-leipzig.de (Jana Hertel), kristin@bioinf.uni-leipzig.de (Kristin Reiche), stadler@bioinf.uni-leipzig.de (Peter F. Stadler), joerg.hackermueller@izi.fraunhofer.de (Jörg Hackermüller).

Genomics *7 February 2008*

Table 1
Overview of applied nematode genomes.

organism	source	assembly (date)	size [Mb]
Caenorhabditis elegans	UCSC	ce4, Jan 2007	100
Caenorhabditis briggsae	UCSC	cb3, Jan 2007	109
Caenorhabditis remanei	UCSC	caeRem2, Mar 2006	162
Caenorhabditis brenneri	UCSC	caePb110, Jan 2007	207
Pristionchus pacificus	UCSC	priPac1, Feb 2007	175

Table 2
Overview of required CPU-time for aligning gammaproteobacteria.

	`NcDNAlign`, default Blast		`NcDNAlign`, modified Blast		TBA
Flanking regions	+	-	+	-	n.a.
	(1)	(2)	(3)	(4)	(5)
CPU time					
cutSequences.pl	1m44	1m43	1m42	1m45	n.a.
getGwAln.pl	3m28	3m24	4m90	4m40	n.a.
merge	5m60	5m00	6m56	6m59	n.a.
realign	4m25	3m57	11m39	10m48	n.a.
trimAln	1m11	1m11	5m00	4m06	n.a.
Total	15.68	14.35	29.27	26.98	548.36

Table 3
Overlap to CNEs provided by the CONDOR [1] database.
We performed `BLAST` searches against four sets of vertebrate CNEs given by the CONDOR project. We list the number of recovered UCRs that have a significant `BLAST` hit (`E-value` <= 1e-3) and the amount of hits with 100 % sequence identity. CONDOR CNEs only require > 65 % sequence identity over at least 40 nt yielding a plenty of sequences conserved between fugu and other mammals. However, we notice 810 UCRs conserved between the genomes of fugu and human (see Figure 3 of the main paper) but only 612 resp. 491 UCRs have sequence similarity with fugu resp. human CONDOR CNEs.

species	# CONDOR CNEs	recovered UCRs	100 % identity
fugu	6 794	612	597
human	6 771	491	83
mouse	6 489	454	87
rat	5 601	402	75
all	-	411	396

Fig. 1. A simple example of consistency checks validating three global alignments A, B, C.
The validation process sets up the four graphs G_S, G_C, G_I and G_F.
G_S: An edge between two vertices (local alignments) is inserted if they have a distance $\leq 30\,nt$ to find all combinations of consistent global alignments.
G_C: Edges occur between consistent pairs of local alignments.
G_I: Edges occur between inconsistent pairs of local alignments.
G_F: An edge is inserted between x and y if there exists at least one path from x to y in G_C which does not contain vertices connected in G_I
Cliques in G_F are local alignments which can be combined to a consistent global alignment. In the example there is one single trivial clique, thus A and C are consistent and can be combined.

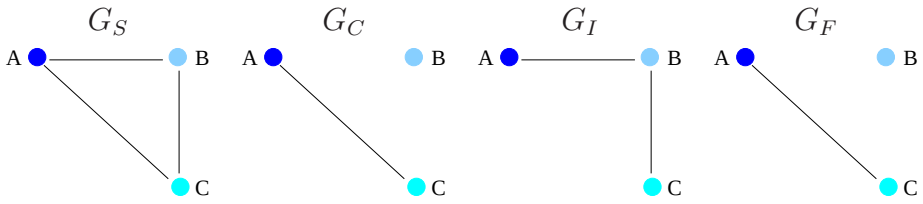

Table 4
Sensitivity analysis of the gammaproteobacteria RNAz screen, part A.

Flanking regions	NcDNAlign, default Blast + (1)	NcDNAlign, default Blast - (2)	NcDNAlign, modified Blast + (3)	NcDNAlign, modified Blast - (4)	TBA n.a. (5)
RNAz					
Nr. RNAz hits	126	122	339	300	658
Overlap	99		280		260 [a]
Overall length of hits	25 100	20 618	94 995	80 680	92 888
Nr. hits per 100k aligned target	358	412	296	372	279
Mean length of hits	199	169	280	269	141
Nr. hits in random. aln.	41	41	87	74	166
FDR	0.32	0.33	0.25	0.24	0.25
Overall length of rand. hits	7 220	6 164	33 283	27 683	19 475
Nr. false discovered nucleotides per 100k aln	117	139	69	65	70
Mean length of random RNAz hits	62	44	383	374	117
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)
	Hits overlapping given annotation				
rRNA	25	21	98	84	65
tRNA	52	57	61	56	12
misc. RNA	5	6	18	16	9
CDS	0	0	0	0	0
Gene	82	84	174	153	57
Repeat region	0	0	0	0	0
Rep. origin	1	0	1	1	1

[a] (3)+(5)

Table 5
Sensitivity analysis of the gammaproteobacteria RNAz screen, part B.

Flanking regions	NcDNAlign, default Blast + (1)	NcDNAlign, default Blast - (2)	NcDNAlign, modified Blast + (3)	NcDNAlign, modified Blast - (4)	TBA n.a. (5)
	Known RNAs detected				
Initial MSAs					
rRNA	22	22	22	22	n.a.
tRNA	84	84	86	86	n.a.
miscRNA	12	12	23	23	n.a.
After beautification					
rRNA	16	16	21	21	20
tRNA	72	74	69	70	81
miscRNA	5	6	21	19	33
Found by RNAz					
rRNA	10	9	14	14	14
tRNA	55	61	62	62	56
miscRNA	5	6	18	16	23
	Sensitivity [a]				
rRNAs/detectable	.62 (10/16)	.56 (9/16)	.66 (14/21)	.66 (14/21)	.70 (14/20)
rRNAs/all known	.45 (10/22)	.40 (9/22)	.63 (14/22)	.63 (14/22)	.64 (14/22)
tRNAs/detectable	.76 (55/72)	.82 (61/74)	.89 (62/69)	.88 (62/70)	.69 (56/81)
tRNA/all known	.63 (55/86)	.70 (61/86)	.72 (62/86)	.72 (62/86)	.65 (56/86)
Misc. RNAs/detectable	1.00 (5/5)	1.00 (6/6)	.85 (18/21)	.84 (16/19)	.70 (23/33)
Misc. RNAs/all known	.10 (5/49)	.12 (6/49)	.36 (18/49)	.32 (16/49)	.47 (23/49)
Genes/detectable	.75 (70/93)	.79 (76/96)	.84 (95/112)	.83 (92/110)	.69 (94/137)
Genes/all known	.01 (70/4437)	.01 (76/4437)	.02 (95/4437)	.02 (92/4437)	.02 (94/4437)
	Number of RNAz hits annotatable by public ncRNA databases				
Rfam	84	85	151	145	127
Noncode	7	9	16	16	17
ncRNAdb	10	9	9	28	30

[a] Based on known RNAs detected by RNAz; Obviously, only those RNAs which are present in the RNAz input alignments are detectable.

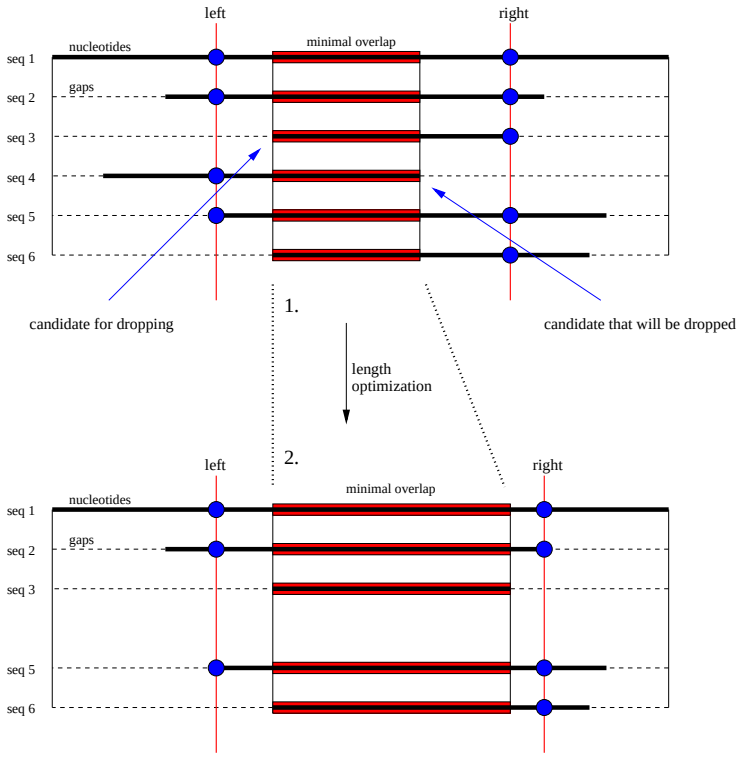

Fig. 2. Exemplary illustration of the alignment beautification procedure.
Consider an initial six-way alignment. Taking into account to minimize the number of sequence losses, the two most restricting sequences of the alignment length are determined. Herein, the optimization algorithm decides to drop sequence four because there are simply more sequences involved (blue dots) at the right than at the left side (5>4) of the minimal overlapping region (red). As a consequence, the length of the minimal overlapping region increases. Repeating the beautification could enlarge the overlap again until a certain cutoff-length is reached or the alignment contains no more dispensable sequences.

References

[1] A. Woolfe, D. K. Goode, J. Cooke, H. Callaway, S. Smith, P. Snell, G. K. McEwen, G. Elgar, CONDOR: a database resource of developmentally associated conserved non-coding elements., BMC Dev Biol 7 (2007) 100.