Use of Near-Neighbor PCR to Close Scaffold Gaps in Microbial Genomes

A. Christine Munk, Yan Xu, Avinash Kewalramani, Riley Arnaudville, Roxanne Tapia, Thomas S. Brettin, C. S. Han
Los Alamos National Laboratory, Los Alamos, N. M.

LAUR-07-1997

ABSTRACT #FF006

One of the challenges of finishing microbial genomes is large numbers of uncloned regions (scaffold gaps). Gaps with no clone links require many expensive PCR reactions to connect scaffolds. Ordering and orienting scaffolds with no clone links is difficult unless a closely-related genome is available to identify possible links. This abstract describes near-neighbor PCR (nnPCR), a method of closing these 'scaffold' gaps with a minimum of PCR reactions using closely related finished genomes.

In the manual process of near-neighbor PCR, scaffold ends are identified visually using the Consed program's Assembly view and map of scaffolds. Blast is performed using as query a fasta file containing alternatively 1) the sequence of all unordered contigs <2kb and >10 reads, or 2) the final 10kb of each scaffold. A closely-related-genome downloaded from Genbank is used as subject. The blast output is parsed and a tab-delimited file is produced which can be opened as a spreadsheet. Links between contigs are identified and PCR primers are chosen and paired according to the observed links. PCR is performed and products are end-sequenced and assembled to close scaffold gaps. If PCR products are >1500 bp, they are shattered, subcloned and subclones are end-sequenced and assembled to make a consensus sequence to close scaffold gaps.

In one microbial finishing project with 14 uncaptured gaps considered, 9 out of 14 PCR reactions were successful, and 9 gaps were closed. In another project with 23 uncaptured gaps, 8 out of 19 PCR reactions were successful and 5 gaps were closed. This can be compared to combinatorial PCR, which would require 378 and 703 PCR reactions respectively.

Software to automate and improve this procedure has been developed and is currently being implemented in production. Improvements are currently being tested.
This nnPCR software uses the Consed autoreport function to generate a file which contains a map of contigs in scaffolds, and creates a fasta file for the blast query with sequence from the contigs at scaffold ends. nnPCR uses megablast to search all finished genomes currently in Genbank for hits within the same organism within a 20 kb range. The software chooses primers and pairs them, and submits a work order directly to the lab.

For the project with 14 uncaptured gaps considered, 12 out of 14 PCR reactions chosen by the nnPCR software were successful. For the project with 23 uncaptured gaps, 3 of 6 PCR reactions were successful.

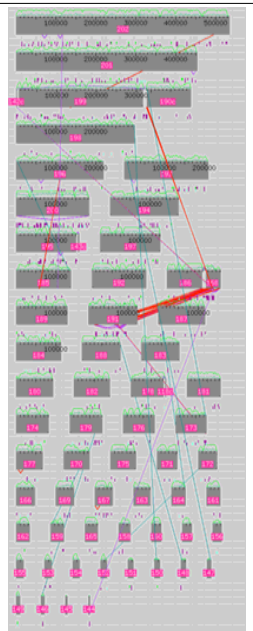
Several microbial genomes have been significantly improved using near-neighbor PCR. This method has significantly reduced the number of PCR reactions that would be required if combinatorial PCR were necessary. Countless hours of finishers' time have been spent on more productive efforts.

US Department of Energy's Office of Science, Biological and Environmental Research Program and Los Alamos National Laboratory under contract No. W-7405-ENG-36

Acknowledgement:

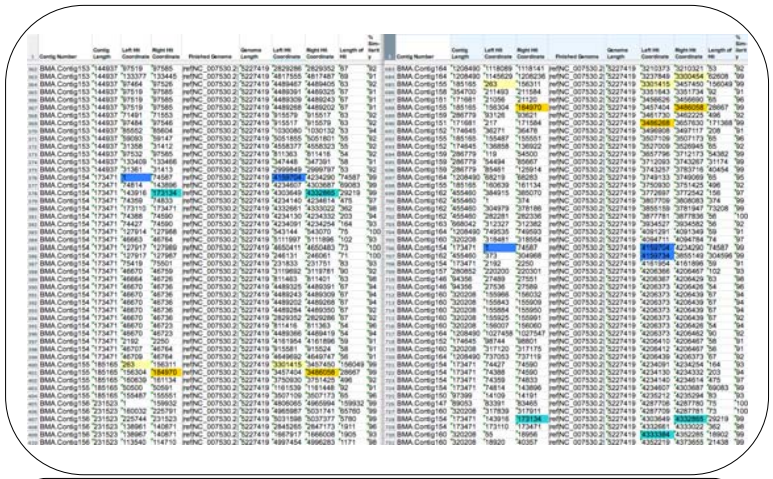
We would like to thank the Intelligence Technology Innovation Center for funding to sequence the genomes used in this analysis.

Consed Assembly View picture of contigs and scaffolds showing draft sequence plus 2 rounds of automatic finishing of a *Bacillus anthracis* genome. Each gray bar is a contig. Intercontig lines represent clones with end sequences in different contigs. Adjacent contigs with clones connecting them make a scaffold. Scaffolds with no clone links are not ordered or oriented.



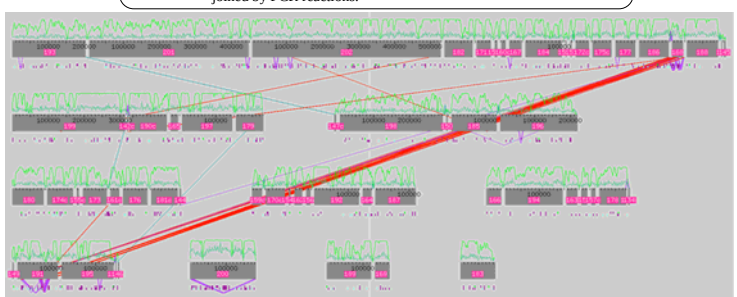
PCR Results for nnPCR and Manual Process

Paired contigs	manual product	manual size	nnPCR product	nnPCR size	both good	manual good	both failed
b202-e201	4SMC	1400	WZX,WZY	1600;1500	x		
a202-b182	4SMD	3200	XAB,XAC	3800;3800	x		
e182-e171	4SME	700	XAT,XAU	1000;900	x		
b171-e153	4SMF	1500	none chosen			x	
b150-e198	4SMG	1500	none chosen			x	
b153-e160	4SMH	2000	WZD	2100	x		
b160-b147	4SMI	2800	XBB,XBC failed	f		x	
e147-b184	4SMJ	750	yes		x		
e184-e151	4SMK	1200	XAJ,XAK failed	f		x	
b151-b152	4SML	2700	XAJ,XAK failed	f		x	
e152-e172	4SMM	650	WZA	1400	x		
b172-b137	4SMN	400	XBI	3000	x		
e137-e175	4SMO	650	XBI	3000	same as above		
b187-e164	4SMP	1000	yes		x		
b164-e192	4SMQ	1350	WZB,WZC	1200;1500	x		
b192-e156	4SMR	1100	yes		x		
b156-e162	4SMS	1600	WZF,WZG	400;1800	x		
b162-b154	4SMT	1100	WZN	1750	x		
e154-b170	4SMU	1000	XAL,XAM	800;900	x		
e170-b159	4SMV	800	yes		x		
e159-b189	4SMW	f	fail				x
e189-b169	4SMX	1200	yes		x		
e169-b199	4SMY	f	failed	f			x
b190-b165	4SMZ	600	WZV,WZW	800;800	x		
e165-b197	4SNA	4200	yes		x		
e197-b179	4SNB	1600	WZJ	1600	x		
e179-e146	4SNC	f	fail				x
b146-b143	4SND	6300	none chosen			x	
b195-e149	4SNE	1100	XAP,XAQ	1100;1000	x		
e199-e192	4SNF	6200	none chosen			x	
e187-e144	4SNH	f	failed	f			x
b191-e149	4SNI	3000	yes		x		
b144-b181	4SNJ	1400	none chosen			x	
e181-e176	4SNK	3800	WZT,WZU	6000;5000	x		
b201-e193	4SNL	550	yes		x		
b193-e147	4SNM	f	fail				x
b147-b198	4SNN	2800	yes		x		
b196-e150	4SNO	500	yes		x		
b185-e150	4SNP	2000	yes		x		
e196-b166	4SNQ	f	fail				x
e166-b194	4SNR	3000	yes		x		
e194-b163	4SNS	1100	XBF,XBG	900;900	x		
e163-b158	4SNT	1250	WZE,XBA	1500;1500	x		
e158-e157	4SNU	1250	none chosen			x	
b157-b178	4SNV	1300	yes		x		
e178-e138	4SNW	550	none chosen			x	
b138-b134	4SNX	2300	none chosen			x	
e134-b180	4SNY	f	fail				x
e180-e174	4SNZ	1200	XAF	750	x		
b174-e155	4SOA	2500	WZL,WZH	4500;4000	x		
b155-b173	4SOB	1000	XBD,XBE	900;900	x		
e173-e183	4SOC	7000	yes		x		
b161-b176	4SOD	1250	yes		x		
b188-e168	4SOF	7000	yes		x		
b168-e186	4SOG	4000	none chosen			x	
b186-e177	4SOH	900	WYX,WYY	2800;2800	x		
b177-b175	4SOI	500	none chosen			x	
e188-b148	4SOJ	4000	yes		x		
b183-e163	4SOK	f	fail				x
e148-b145	4SOL	2000	none chosen			x	
				TOTALS	36	15	8

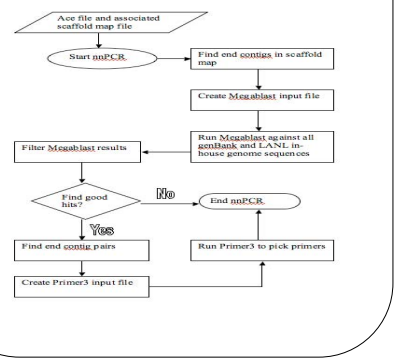


Manual near-neighbor PCR blastParser results. Command-line Blast was performed using a query file containing the sequence of all unordered contigs less than 2kb in length with greater than 10 reads. The finished genome sequence for *Bacillis anthracis* Ames strain was downloaded from Genbank and formatted to use as subject file. Blast output was parsed to output the tab-delimited file shown on the left, ordered by query contigs. The file was copied and re-sorted to represent finished genome order. These files are manually searched to identify links between scaffolds. PCR primers are chosen based on these searches using Consed's primer-picking program. A request for PCR reactions is typed and submitted to the informatics system and forwarded to the lab team.

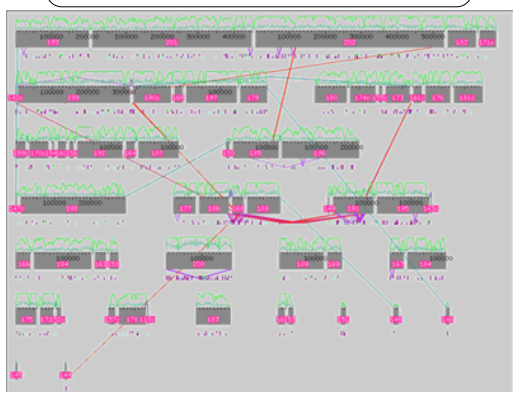
Consed Assembly View picture of contigs and scaffolds showing genome with results of manual near-neighbor PCR. Fifty-one scaffolds have been joined by PCR reactions.



Flow Diagram of nnPCR process



Consed Assembly View picture of contigs and scaffolds showing same genome with results of nnPCR software. Thirty-six scaffolds have been joined by PCR reactions.



Observations and Conclusions

The more closely related microbes are phylogenetically, the more help near neighbors are in connecting sequence scaffolds.

Although the above figures show that additional scaffold links may be found using the manual process, nnPCR software has been recently optimized to pick primer pairs for more scaffold gaps.

Use of near-neighbor microbial genomes to order and orient scaffolds can significantly reduce the number of PCR reactions necessary to close a genome.

The nnPCR program also reduces the amount of the human finisher's time required to close scaffold gaps by automating the process of identifying scaffold links and submitting work lists.

NnPCR can be used with good results as early in the finishing process as after automatic duplication resolution.

nnPCR Results for Various Microbes

PROJECT	# un-captured gaps	#pcr reactions chosen/2	# of gaps closed by nnPCR
Yersinia pestis YPC	17	10	6
Bacillus megaterium BMA	61	46	36
Bacillus anthracis BAA	46	28	23
Clostridium botulinum CLJ	26	6	4
Clostridium botulinum CLM	25	6	4
Escherichia coli ECD	38	23	21
Burkholderia pseudomallei BUC	20	18	17
Vibrio furnissii VFA	52	0	0
Clostridium botulinum CLG	33	0	0

NnPCR was performed at various points early in the finishing process for several projects. The number of primer pairs chosen when nnPCR is run immediately following 2 rounds of dupFinisher* does not vary significantly from the number of pairs chosen after 1 or 2 cycles of autoFinish, or autoFinish plus some manual finishing. *automated duplication resolution

PROJECT	Min # scaffolds	nnPCR pairs chosen	Max # scaffolds	nnPCR pairs chosen
Bacillus anthracis BAA	50	46	63	47
Bacillus megaterium BMA	55	46	66	42
Yersinia pestis YPA	17	11	27	18
Yersinia pestis YPK	11	8	15	10
Francisella tularensis FTB	8	6	11	11
Vibrio cholerae VCA	37	19	49	22
Burkholderia pseudomallei BUD	31	18	36	15