LAGAN and Multi-LAGAN: Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA

Michael Brudno, Chuong B. Do, Gregory M. Cooper, et al.

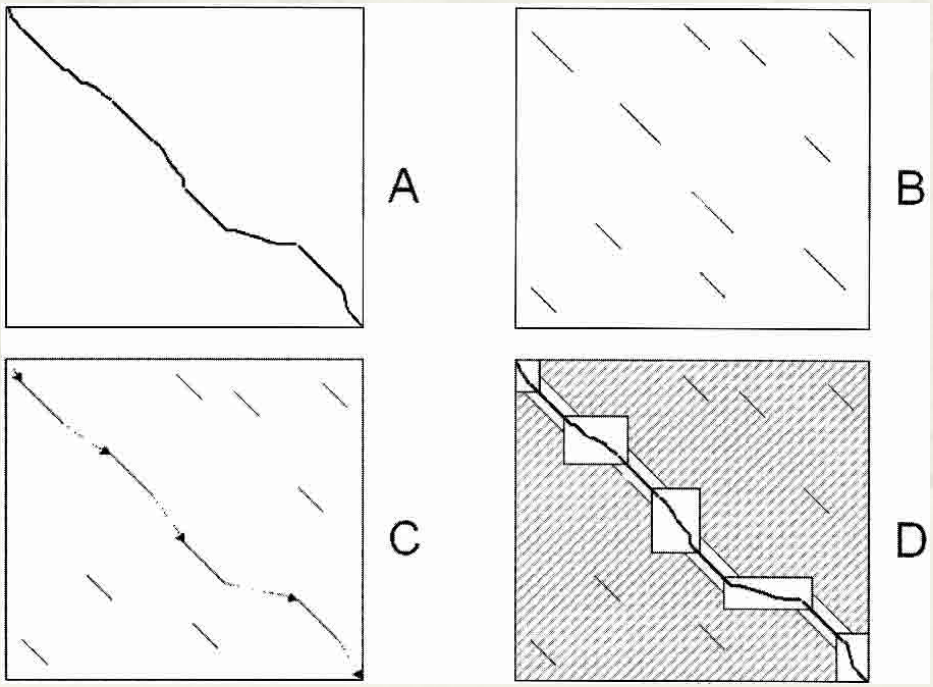
Presented by Xuebei Yang

About Alignments

- Pairwise Alignments
  - Needleman-Wunsch
  - GLASS, WABA, AVID: haven't been tested in alignments between distant relatives
- Multiple Alignments
  - Much harder (NP-Complete)
  - Heuristic approaches
    - Progressive Alignments
- In this article
  - LAGAN (Limited Area Global Alignment of Nucleotides)
  - MLAGAN (Multiple LAGAN)

LAGAN

1. Generation of local alignments
2. Construction of rough global map
3. Computation of global alignment



LAGAN

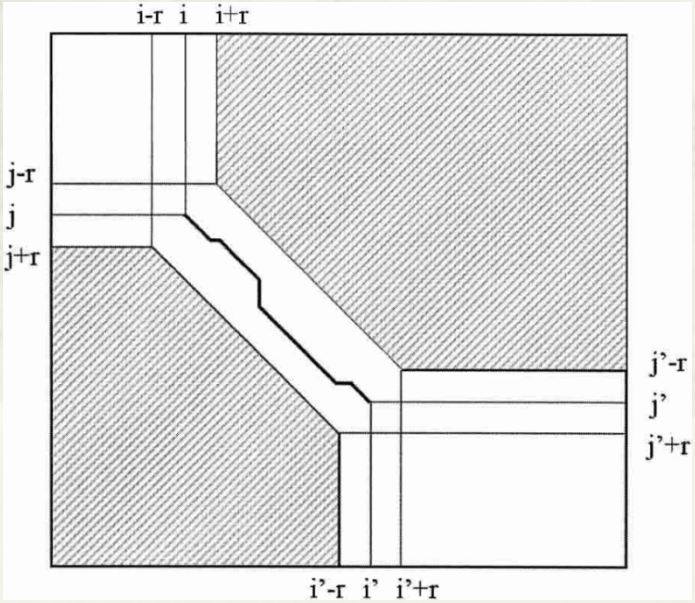
- Generation of Local Alignments
  - CHAOS
    - work well on distant, as well as close organisms
    - (k,c) seed: k-mers that match with at most c differences between the two sequences
    - Chain:
      - x<d, y<d, |x-y|<s
    - Scoring a chain
      - Gap penalty proportional to |x-y|
      - Match & Mismatch scores for letters in each seed
      - Chains below a threshold score are discarded, and a remaining chain is extended ungappedly to find the optimal |x-y| gap position.

LAGAN

- Construction of a Rough Global Map
  - A local alignment (anchor) is a vector (b, e, b', e', s), representing begin and end positions of the alignment in each sequence, and the score of the alignment
  - $A1(b_1, e_1, b_1', e_1', s_1) < A2(b_2, e_2, b_2', e_2', s_2)$, if $e_1 < b_2$, $e_1' < b_2'$
  - A chain is a set of local alignments: $A1<A2<\ldots<Ak$, scoring $s_1+s_2+\ldots s_k$
  - A Rough Global Map is the chain with the highest score
  - O(n log n), n is the total number of local alignments

LAGAN

- Computation of global alignment
  - Computation Area is limited to:
    1. (0, 0) to (i+r, j+r)
    2. (i'-r, j'-r) to (M, N)
    3. the band enclosed by the two diagonals



LAGAN

- Running time
  - k and c
    - k↑ c↓ ➡ running time↓ sensitivity↓
  - Total running time
    - Dominated by those rectangles
    - In most cases, linear with sequence length

MLAGAN

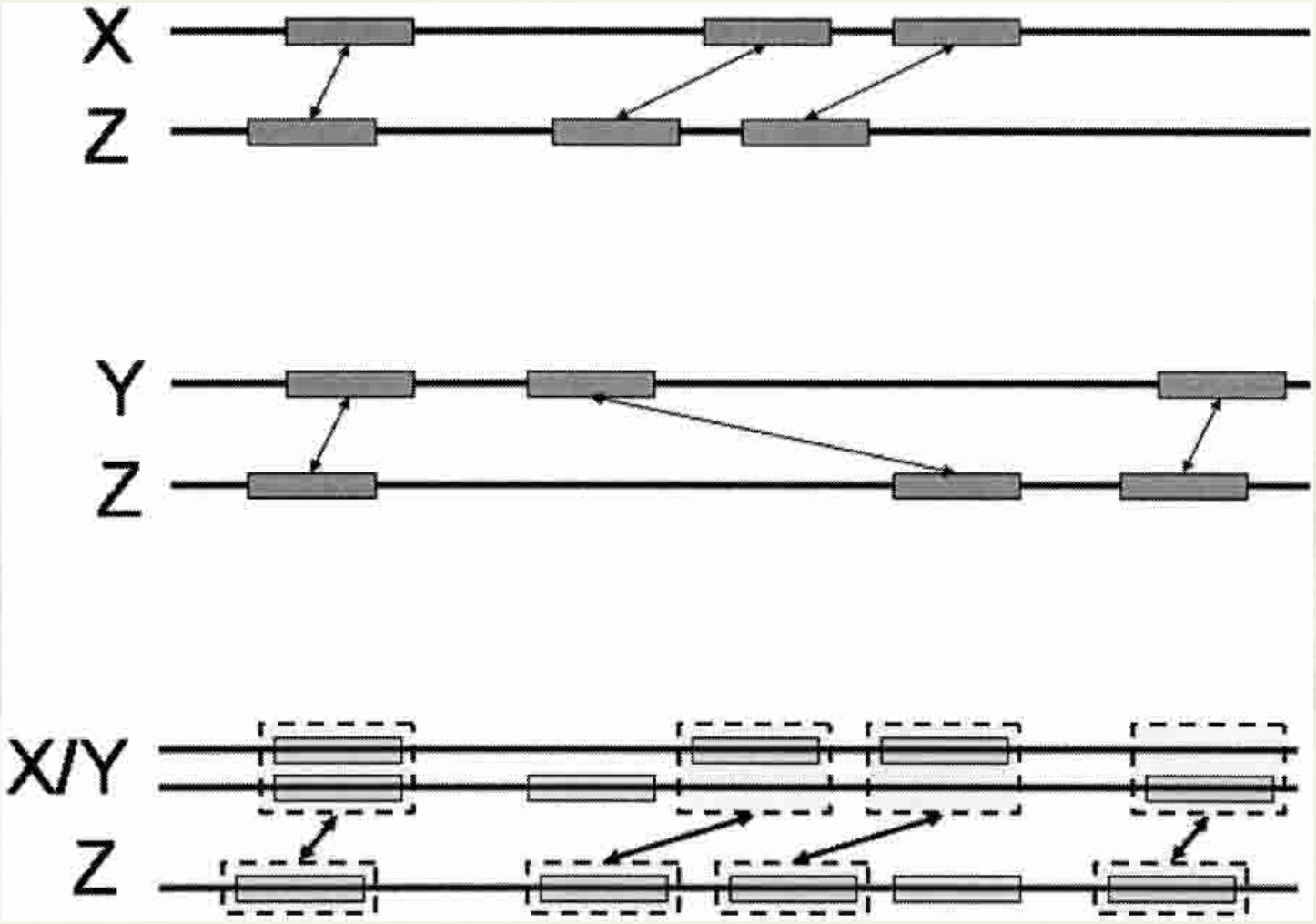
- Progressive Multiple Alignment
  - Assume Phylogenic tree is known
  1. Select the closest two(multiple) sequences
  2. Get rough global map
  3. Align two(multiple) sequences
- Optional iterative improvement phase

MLGAN

- Key point: How to get rough global map?
  1. All anchors in the rough global maps between X and Z, and between Y and Z, become anchors between X/Y and Z, with score equal to their original score.

  2. Anchor between X and Z that overlaps an anchor between Y and Z, is reweighed with score equal to $(s_1 + s_2)^*$ I/U, where s_1, s_2 are the scores of the (X, Z) and (Y, Z) anchors, respectively, I is the length of intersection, and U is the length of union of the anchors.

MLGAN



MLGAN

- Scoring
  - Sum-of-pairs
    - Used for scoring substitution
  - Consensus
    - Used for scoring gaps
    - Gap opening penalty
    - Gap continue penalty
    - Gap ending penalty (to prevent gap opening stack)

-
-
A
A
A
A

MLGAN

- Example of gap opening stack

Seq. 1	AT-------CAG		A----T---CAG	
Seq. 2	AT-------CAG		A----T---CAG	
Seq. 3	ATCTGT---CAG		ATCTGT---CAG	
Seq. 4	ATCTGT---CAG		ATCTGT---CAG	
Seq. 5	ATCTGT---CAG		ATCTGT---CAG	
Seq. 6	ATCTGTATGCAG		ATCTGTATGCAG	
Seq. 7	ATCTGTATGCAG		ATCTGTATGCAG	
		Total		Total
Gap Open	002000300000	5	020000200000	4
Gap Cont.	002222222000	14	022220222000	14
Gap End	000000000200	2	000020000200	4
	A		B	

MLGAN

- Optional iterative improvement phase
  - Each sequence is removed iteratively
  - Each position i of this sequence is scored, and the score is cumulated:
    - $C_i = C_{i-1}$ + the score of this position
    - If $C_i < 0$, set $C_i = 0$
    - If $C_i >$ threshold, set Ci = 0 and create a anchor here
  - Realign this sequence according to new anchors

MLGAN

- Evaluation of iterative refinement

Before Iterative Refinement	After Iterative Refinement
GTGTAT----TTTACCTTATCACAGTTTTAT	GTGTAT----TTTACCTTATCACAG-TTTTAT
GTGTAT----TTTACCTTATCACAGTTTTAT	GTGTAT----TTTACCTTATCACAG-TTTTAT
GTGTGT----TTTACCTTATCACAGTTTTAT	GTGTGT----TTTACCTTATCACAG-TTTTAT
GGGCGTCGCGTCTCCCTTCGCGCAGCTCCGG	GGGCGTCGCGTCTCCCTTCGCGCAG-CTCCGG
GTGTGTT----TTACCTTATCACAG-TTTTA	GTGTGT----TTTACCTTATCACAG-TTTTAT
GTGTGTT----TTACCTTATCATAGTTTTTA	GTGTGT----TTTACCTTATCATAGTTTTTAT
GTGGCT----TTTCCCTTATCACAGGCTTCT	GTGGCT----TTTCCCTTATCACAG-GCTTCT
GTGGCT----TTTCCTTTATCACAGGCTTGT	GTGGCT----TTTCCTTTATCACAG-GCTTGT

Evaluation of Performance

- Two datasets are used
  - The ROSETTA set which contains 129 orthologous annotated genes with complete intron sequences between human and mouse of average length 10 Kbp.
  - The CFTR region, which for the studies described here consisted of 12 orthologous sequences from human, chimpanzee, baboon, cat, dog, cow, pig, mouse, rat, chicken, fugu, and zebrafish, with an average of 1 Mbp.

Evaluation of Performance

Table 1. Performance of Aligners on the ROSETTA Dataset of 1160 Total Exons in Human and Mouse

Aligner	100% exons	90% exons	70% exons	Time (sec)
DIALIGN	89	96	98	388
MUMmer	0	1	3	17
GLASS	91	97	98	154
AVID	90	95	97	19
BlastZ	94	97	98	17
LAGAN	94	97	98	48

Columns show the percentage of exons annotated in human that are aligned to the orthologous mouse exon over at least 70%, 90%, and 100% of their length, and the time required to align the 129 sequences.

Evaluation of Performance

Table 2. Performance of Aligners on the *CFTR* Region

	Baboon	Chimpanzee	Mouse	Rat	Cow	Pig	Cat	Dog	Chicken	Zebrafish	Fugu	Overall
Number of exons	232	176	230	230	224	174	176	182	68	48	150	1914
MUMmer												
100%	100	99	6	7	28	32	38	28	0	0	0	36
90%–100%	100	100	8	9	40	44	47	37	0	0	0	41
70%–100%	100	100	14	16	52	55	56	45	0	0	0	47
AVID												
100%	100	100	94	95	98	97	99	93	66	33	19	88
90%–100%	100	100	98	100	99	100	100	98	79	42	21	91
70%–100%	100	100	100	100	100	100	100	99	85	44	29	92
BlastZ												
100%	100	100	97	97	96	97	100	94	96	73	66	94
90%–100%	100	100	100	100	98	100	100	99	97	79	72	97
70%–100%	100	100	100	100	100	100	100	99	97	79	80	98
LAGAN												
100%	100	100	97	97	98	97	100	94	96	83	72	95
90%–100%	100	100	100	100	99	100	100	99	99	88	77	98
70%–100%	100	100	100	100	100	100	100	99	100	92	81	98
MLAGAN												
100%	100	100	97	97	98	97	100	94	99	88	73	96
90%–100%	100	100	100	100	99	100	100	99	100	98	84	98
70%–100%	100	100	100	100	100	100	100	99	100	100	90	99
Time (sec)												
MUMmer	8	6	7	7	8	6	6	6	3	2	2	61
AVID	82	57	215	221	165	111	139	131	83	53	76	1775
BlastZ	31	24	46	43	40	36	34	33	7	5	6	305
LAGAN	56	50	78	82	60	68	62	78	338	158	133	1135
MLAGAN	—	—	—	—	—	—	—	—	—	—	—	4547
Max memory (MB)												
MUMmer	40	39	40	40	40	39	39	39	39	38	38	40
AVID	598	551	581	584	578	498	522	502	387	340	360	598
BlastZ	239	276	202	212	204	200	208	206	188	185	185	276
LAGAN	90	90	90	89	90	87	88	87	88	87	89	90
MLAGAN	—	—	—	—	—	—	—	—	—	—	—	670

The annotated human exons were aligned with TBLASTN to create *pseudo-annotations* of exons in the other organisms. The table reports the percentage of exons covered by alignments over at least 70%, 90%, and 100% of their length, and the time and memory required to obtain the alignments. The last column reports the percentages of exons aligned out of total number of exons, the total time required for the 11 alignments (for the single 12-sequence multiple alignment in the case of MLAGAN), and the maximum memory required.

Evaluation of Performance

- Result
  - Results demonstrate that LAGAN and MLAGAN are capable of efficiently solving difficult multiple sequence alignment problems.

Discussion

- As the number of sequences increase, pairwise alignment may be increasingly difficult to reconcile into an overall picture of conservation.
- Results suggest that multiple alignments are better than pairwise alignments at aligning conserved exons between distant species.

Questions?