ALNシリーズのアラインメントツール

後藤　修　京都大学大学院情報学研究科

【概要】

配列のアラインメントは、ゲノム・遺伝子・機能性RNAなどの塩基配列やタンパク質のアミノ酸配列を解析する上で最も基本となるバイオインフォマティクスの手法である。我々は、**ALN**シリーズと名付けるアラインメントツール群を開発してきた。それらは、様々な状況に応じて使い分けることが可能な複数のプログラムにより構成される（表１）。これらのプログラムの多くは、**www.genome.ist.i.kyoto-u.ac.jp/~aln_user**より利用可能である。また、同サイトにおいてソースコードの公開も行っている。

開発プログラムとその機能

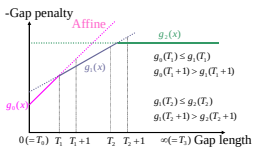
プログラム	機能	文献
Aln	2配列、配列グループ間の大域的アラインメント	1, 2, 3, 4
Swg	2配列間の局所アラインメント	5
Prrn	二重反復改善法による多重配列アラインメント	6, 7, 8
Prime	長いギャップを考慮した多重配列アラインメント	9, 10
Alngg	2ゲノム配列間のアラインメントとエキソン予測	
Cgaln	巨視化によるゲノム配列アラインメントの高速化	11
Spaln	cDNA, EST, アミノ酸配列のゲノムへのマップとスプライスアラインメント	12, 13, 14

1) O. Gotoh, *J. Mol. Biol.* 1982; 2) O. Gotoh, *Bull. Math. Biol.* 1990; 3) O. Gotoh, *CABIOS* 1993; 4) O. Gotoh, *CABIOS* 1994
5) O. Gotoh, *CABIOS* 1987; 6) O. Gotoh, *CABIOS* 1995; 7) O. Gotoh, *J. Mol. Biol.* 1996; 8) O. Gotoh, *Adv. Biophys.* 1999 ;
9) S. Yamada et al. *BMC Bioinf.* 2006; 10) S. Yamada et al. *IPSJ T. Bioinf.* 2008; 11) R. Nakato & O. Gotoh, *APBC* 2008;
12) O. Gotoh, *Bioinformatics* 2000; 13) O. Gotoh, *Nucl. Acids Res.* 2008; 14) O. Gotoh, *Bioinformatics* 2008

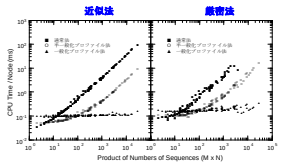
Aln and *Swg*

- *Aln*
  - Needleman-Wunsch-Gotohアルゴリズム
  - アフィン＆ダブルアフィンギャップペナルティ
  - 「厳密な」グループ間アラインメント
  - 省メモリー版、高速版の実装
- *Swg*
  - Smith-Waterman-Gotohアルゴリズム
  - Suboptimalな局所アラインメント検出

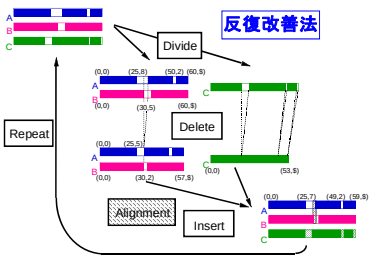
Affine vs. Double Affine Gaps



グループ間アラインメントに要する計算時間



*Prrn*と *Prime*によるマルチプルアラインメン

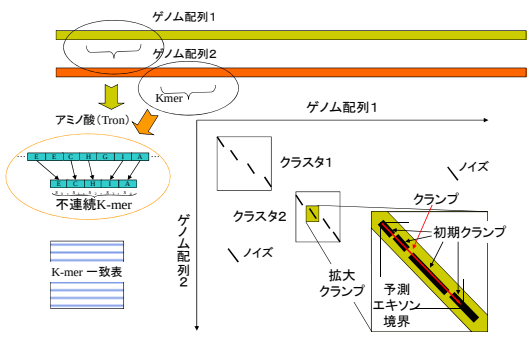


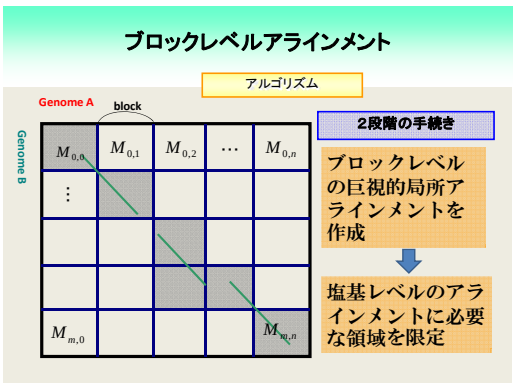
多重アラインメントプログラムの性能比較

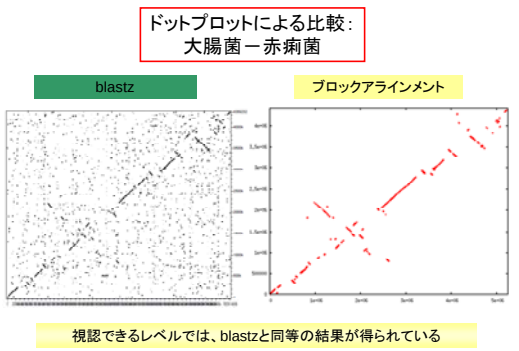
BALiBase3.1 Column Score

プログラム名	Ref. 1.1	Ref. 1.2	Ref. 2	Ref. 3	Ref. 4	Ref. 5	Overall	Ranksum
PRIME p	0.416	0.889	0.446	0.666	0.676	0.662	0.673	846
PRIME s	0.391	0.826	0.413	0.639	0.488	0.496	0.531	958
Prrn	0.334	0.791	0.406	0.469	0.491	0.411	0.499	1080
MAFFT	0.449	0.889	0.436	0.660	0.607	0.644	0.663	789
ProbCons	0.401	0.851	0.374	0.462	0.530	0.509	0.532	847
T-Coffee	0.324	0.882	0.384	0.459	0.563	0.534	0.535	1017
MUSCLE	0.313	0.795	0.343	0.380	0.460	0.408	0.465	1246
DIALIGNT	0.346	0.733	0.390	0.347	0.462	0.389	0.433	1554
POA	0.234	0.676	0.266	0.343	0.413	0.333	0.389	1690
ClustalW	0.221	0.707	0.219	0.371	0.404	0.337	0.368	1497

Alngg と*Cgaln* によるゲノム配列アラインメント

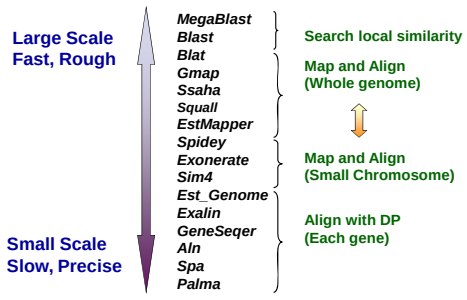






Spaln によるゲノムへのマップとスプライスアラインメント

先行研究

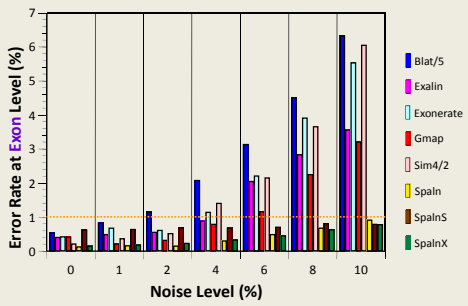


マッピングの性能

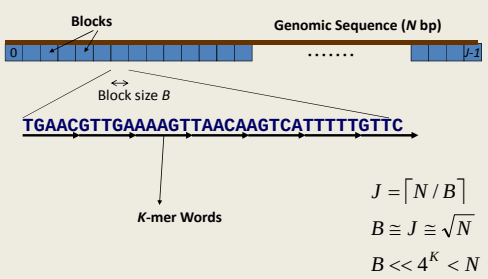
Program	Mapped	Records	Sensitivity (%)	Selectivity (%)	CPU (h)
Blat	124,103	3,774,716	99.80	3.29	37.37
Gmap	123,373	162,700	99.21	75.83	11.00
Megablast	118,643	304,138	95.41	39.01	63.05
SpalnXL	124,112	142,316	99.80	87.21	5.58

Total No. of Queries = 124,355 Unigene Sequences

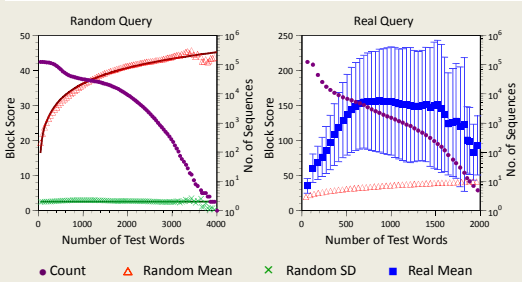
cDNAを鋳型としたときのエキソンレベルのエラー率

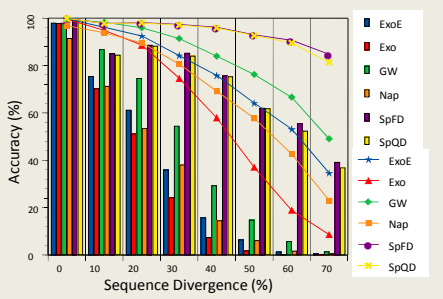


Words and Blocks



ブロックスコアの分布



Whole Human Genome × Unigene Unique Set (124,355 seqs.)

アミノ酸配列を鋳型としたスプライスアラインメントの精度