



完全長cDNAプロジェクト

Last Update 2003.07.02

SpVaPJでは解析配列と対象配列の比較が必須。

従 来

<ローカルアライメント>

- 正確なアライメント情報が得られる。
- × 部分的なアライメントのために、GAP領域の位置など、全体像が把握しにくい。

<グローバルアライメント>

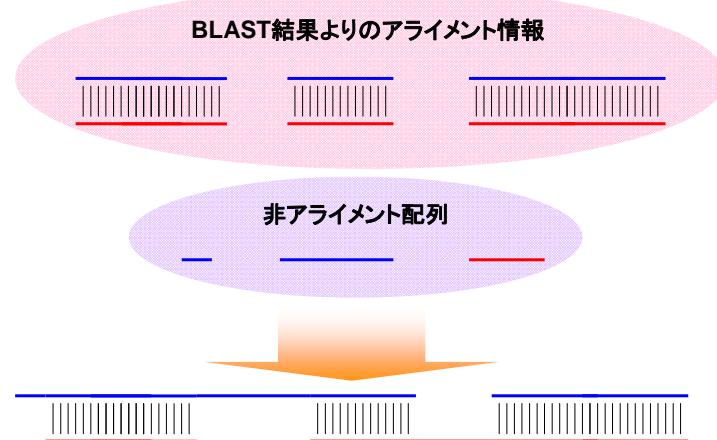
- 全体像が把握し易い。
- × 配列全体にアライメントさせるため、本来アライメント可能にも関わらず、バラバラに貼り付く事がある。
ロングGAPに弱い。



Long GAPを考慮して、配列全体のアライメントを
簡単に把握できるアライメントツールの開発が必要。

アライメントツールのアルゴリズム

BLASTのアルゴリズムを利用して、ローカルアライメントをとり、
その結果を非アライメント領域の配列で繋ぎ一本にしてグローバル化する。



GAPを考慮した高精度のアライメントが可能となる

インストール方法-1

STEP1:アライメントツール実行環境の確認

- 動作確認済みOS : Windows 2000 / XP

- ALVISIONの実行には、Perlのインストールが必要になります。

Active Perl 5.6.1 Build 631 以上のバージョンで動作確認済みです。

使用するパソコンにPerlがインストールされていない場合は、CD-Rに格納されている
インストーラ(ActivePerl-5.8.0.804-MSWin32-x86.msi)でインストールを行って下さい。

- PerlTkのインストール

通常、Perlインストール時にインストールされているものです。

ALVISIONが起動しない場合のみ、インストールの必要があります。

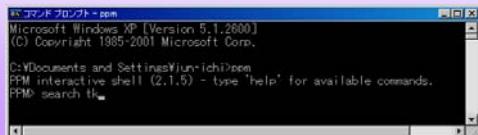
- ・コマンドプロンプトを起動 スタート→プログラム→アクセサリー→コマンドプロンプト

- ・ppmと入力

- ・search tkと入力

- ・install tkと入力

これでインストール完了です。



インストール方法-2

STEP1のつづき

- PDFファイルを閲覧するために、Adobe Acrobat Readerが必要です。

Acrobat Reader 5.0以上のバージョンで動作確認済みです。

(Acrobat Reader 4.X以前のバージョンでは閲覧できません)

使用するパソコンにAcrobat Readerがインストールされていない場合は、CD-Rに格納されているインストーラ(AcroReader51_JPN.exe)でインストールを行って下さい。

※)以前のバージョンが既にインストール済みの場合は、古いAcrobat Readerをアンインストールしてから、インストールして下さい。



STEP2-1：実行プログラムのコピー(iTextFront版)

- CD-Rに格納されている「iTextFront版」フォルダに移動し、「Align」フォルダを“Cドライブ直下”にコピーする。

CD-Rからドラッグアンドドロップで、“Cドライブ直下”にコピーして下さい。



インストール方法-3

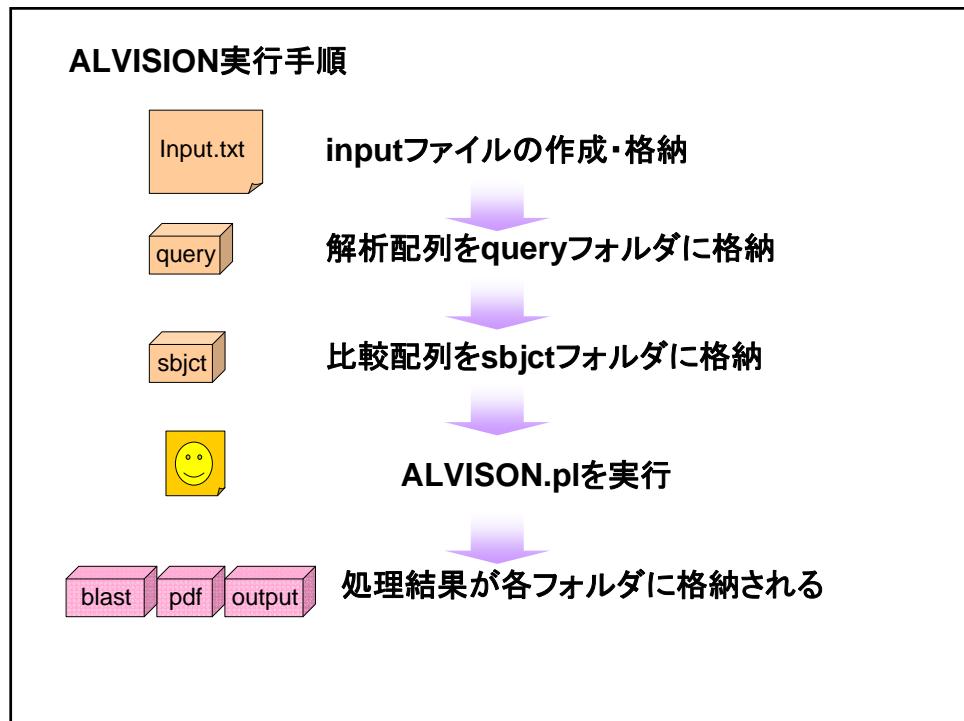
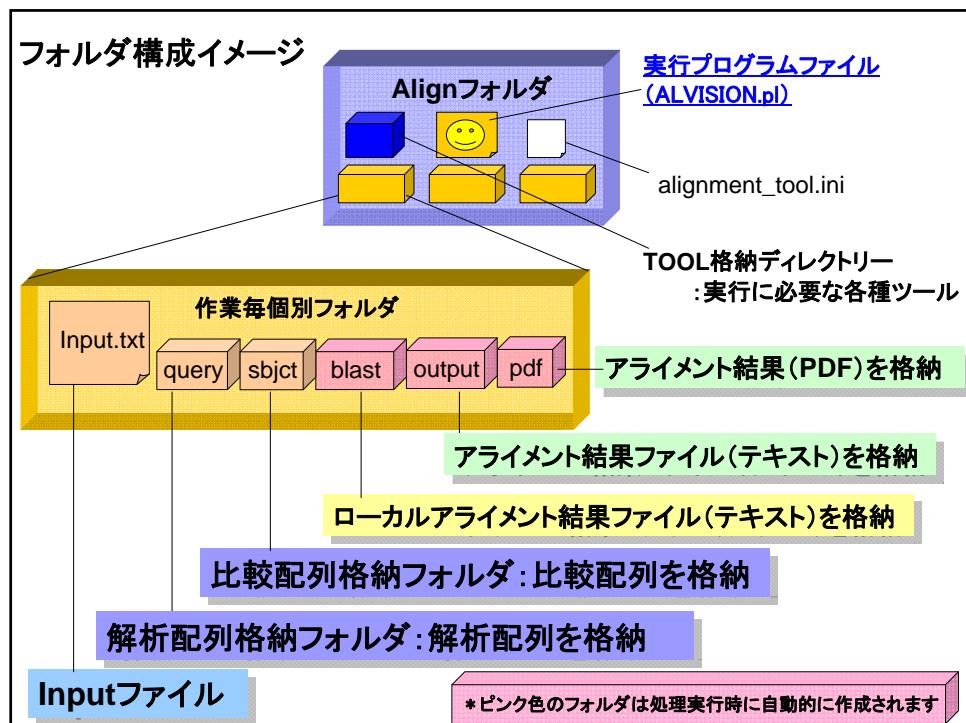
STEP2-2：実行プログラムのコピー(PDFJ版)

(※)このプログラムはiTextFront版で実行した結果、白紙のPDFファイルが出力される場合にご利用ください。既に「iTextFront版」がインストール済みの場合、コピー済みの「Align」フォルダのフォルダ名称は別名に変更しておいて下さい。尚、PDFJ版のALVISIONは、iTextFront版に比べ、処理速度が大幅に低下します。

- CD-Rに格納されている「PDFJ版」フォルダに移動して下さい。
- 「Align」フォルダ、「PDFJ」フォルダ、「PDFJ.pm」ファイルが格納されている事を確認して下さい。
- 「Align」フォルダを“Cドライブ直下”にコピーして下さい。
- 「PDFJ」フォルダ、「PDFJ.pm」ファイルを、ActivePerlをインストールしたフォルダにコピーします。
(注) ActivePerlを“Cドライブ直下”にインストールした場合、「PDFJ」フォルダ、「PDFJ.pm」ファイルのコピー先は、“C:\Perl\Site\lib”的になります。



作業完了



Inputファイル形式-1

解析配列情報				比較配列情報			
解析配列名	解析配列	ORF開始位置	ORF終止位置	比較配列名	比較配列	ORF開始位置	ORF終止位置
Z-ASTRO2007289-01				C-ERLT2002178			
Z-BRACE2006690-01				NM_021160.1			
Z-BRACE3002532-01				NM_032773.1			
Z-BRAMY2014483-01				AL117482.1			
Z-BRAMY2034995-01				AB028968.1			
Z-BRAMY3013305-01				NM_017712.1			
Z-BRAWH-2018633-01				AB020702.1			
Z-BRAWH-3004005-01				NM_144579.1			
Z-BRAWH-3013633-01				AB037784.1			
Z-BRAWH-3023732-01				BC028035.1			
Z-BRAWH-3028717-01				C-FEBRA2000805			
Z-BRAWH-3039770-01				NM_152904.1			
Z-BRHIP2000006-01				NM_052926.1			
Z-BRHIP2006833-01				NM_001606.2			

A列 B列 C列 D列 E列 F列 G列 H列

解析配列名(A列)・比較配列名(E列)を元に、指定フォルダを検索し、一致するファイルから配列情報を抜き出し、B列・F列に記載する。



B列とF列の配列をアライメントさせる

注)実際のInputファイルはエクセル形式ではなく、タブ区切りのテキスト形式になります。

Inputファイル形式-2

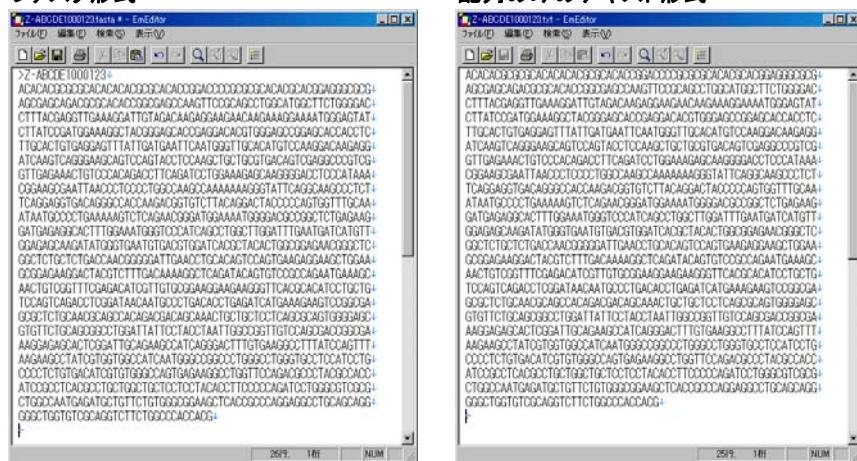
- Inputファイルは以下の形式で作成
ファイル形式 : タブ区切りのテキスト
ファイル名 : input.txt

- A列に解析配列名、E列に比較配列名を記載する。
B・C・D・F・G・H列は現在のツールでは使用しないので、空白にする。
- 前ページでは説明用に1行目に「解析配列名」「解析配列」…があるが
実際のファイルではこの行は削除する

実際のInputファイルは以下のようになります。

左側 (A列)	右側 (E列)
Z-ASTRO2007289-01	NM_021160.1
Z-BRACE2006690-01	NM_032773.1
Z-BRACE3002532-01	AL117482.1
Z-BRAMY2014483-01	AB028968.1
Z-BRAMY2034995-01	NM_017712.1
Z-BRAMY3013305-01	AB020702.1
Z-BRAWH-2018633-01	NM_144579.1
Z-BRAWH-3004005-01	AB037784.1
Z-BRAWH-3013633-01	BC028035.1
Z-BRAWH-3023732-01	C-FEBRA2000805
Z-BRAWH-3028717-01	NM_152904.1
Z-BRAWH-3039770-01	NM_052926.1
Z-BRHIP2000006-01	NM_001606.2
Z-BRHIP2006833-01	

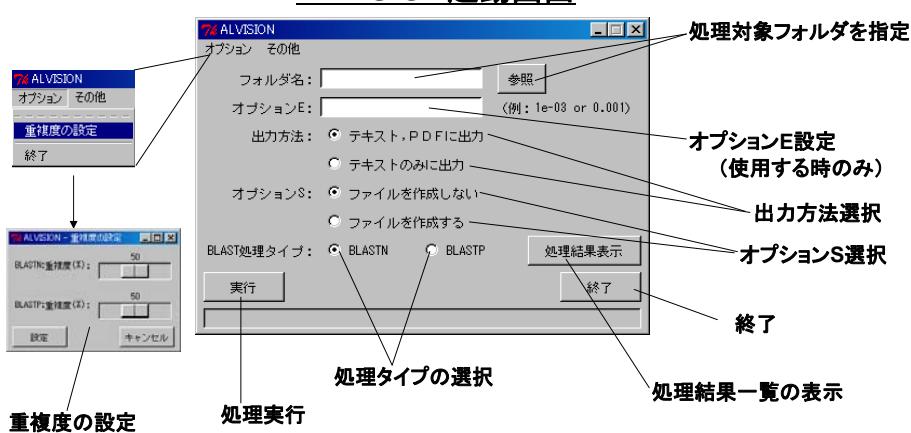
配列ファイル形式 ファスタ形式



queryフォルダ・sbjctフォルダには、上記のどちらかの形式のファイルを格納可能。両方の形式が混在していても構いません。
尚、マルチファスタには対応していません。
Inputファイルに記載されたファイルは必ず格納して下さい。

ALVISION実行方法-1

Alignフォルダ内のALVISION.plをダブルクリックしALVISIONを起動
(Alignフォルダ内にはalignin.plもありますので、ご注意下さい)
ALVISION.plのショートカットをデスクトップ等に作成する事をお勧めします。



* 各機能の説明は次ページ以降を参照

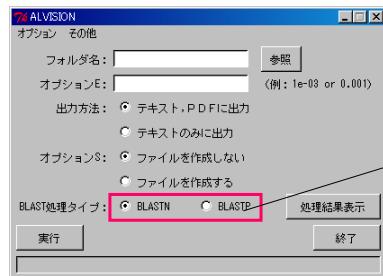
ALVISION実行方法-2

処理タイプの選択について

ALVISIONでは2タイプのアライメント処理を行う事ができます。

1) 核酸 vs 核酸 (BLASTN)
queryフォルダ・sbjctフォルダには核酸配列のファイルを格納

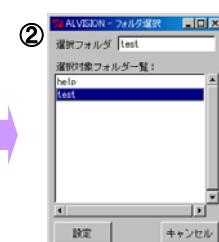
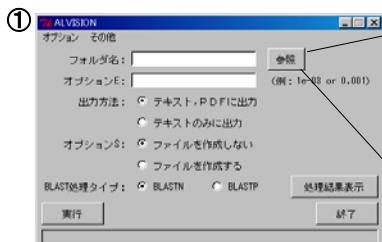
2) アミノ酸 vs アミノ酸 (BLASTP)
queryフォルダ・sbjctフォルダにはアミノ酸配列のファイルを格納



処理実行時にどちらのタイプか選択
(処理タイプを誤って選択すると全てNGで終わります)

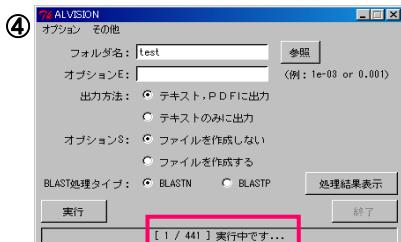
BLASTN実行時には"blastn"フォルダに、BLASTP実行時には"blastp"フォルダに
BLAST処理結果が保存されます。これらのフォルダは実行時に自動的に作成されます。

ALVISION実行方法-3



“参照”ボタンを押してフォルダ選択画面を開く

処理するフォルダを選択

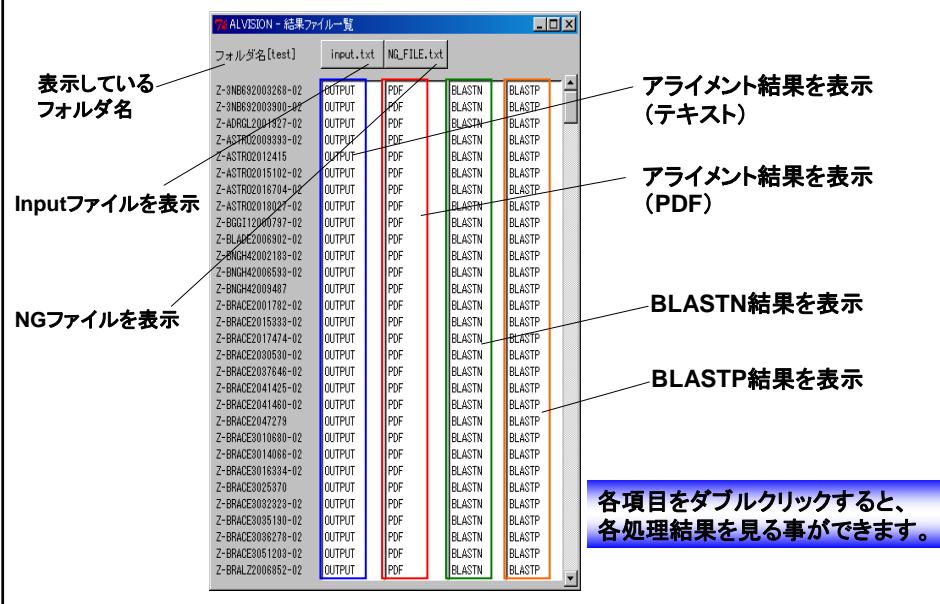


実行タイプを選択し、“実行”ボタンをクリック

処理開始

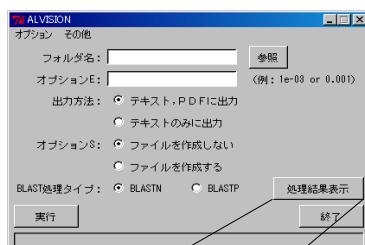
アライメントツール実行方法-4

処理終了後、自動的に結果一覧が表示されます。



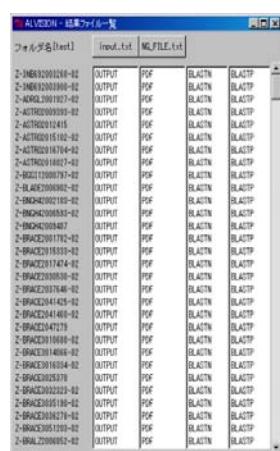
その他の使用方法

①



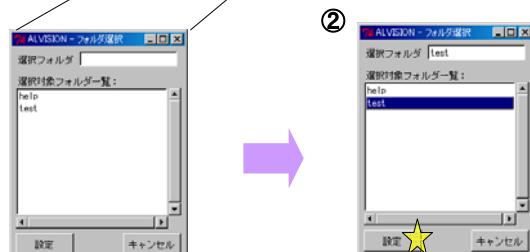
処理済みの結果を結果一覧から
見る事もできます

③



“処理結果”ボタンを押してフォルダ選択画面を開く

②

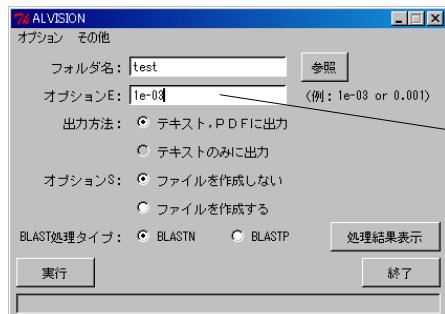


処理するフォルダを選択

結果一覧が表示される

オプションEについて

オプションEを用いると、設定値以上のE-valueのローカルアライメントはグローバル化の際に排除されます。
このオプションを用いる事により、偶発的に貼り付く短いアライメントを排除することが可能となります。



*オプションEについては、次ページの使用上の注意も必ずお読み下さい

オプションEの使用上の注意

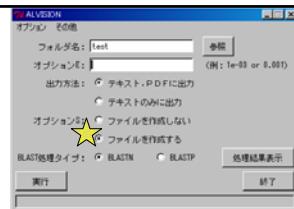
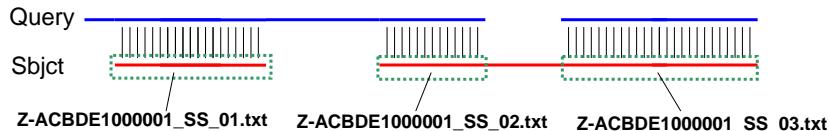
ALVISIONでは、BLASTによるローカルアライメントの結果からスコアの高いアライメントを基準として、整合性が取れるようにアライメント結果の真偽を判定し、真のアライメントと判断されたものを整列化しています。
しかし、あくまでBLASTのアルゴリズムを利用しているため、偶然に短い配列が整合性の取れる位置にアライメントされることがあります。
この様なアライメントについての真偽の判定はゲノム配列等と比較しないとできません。現行のバージョンではそのような機能は付加されていません。

短いアライメントはローカルアライメントのE-valueが大きな値になります。
明らかに偽と判定されるアライメントである場合には、ツール実行時にオプションEを利用して下さい。
これにより、指定したE値以上のローカルアライメントは偽と判定されます。
ただし、短くても真と判定されるべきアライメントも多く存在します。
よって、通常の実行時にはオプションEは使用せず、明らかに問題があると判定される配列に対してのみ、オプションEを使用するようにして下さい。

オプションSについて

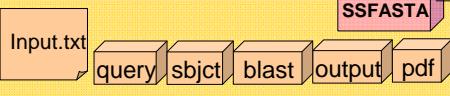
オプションSとは、比較配列(Sbjct)のアライメント領域配列の配列ファイルをアライメント領域毎にOutputする機能です。例えば、下記の様な場合は……

Query Z-ACBDE1000001
Sbjct NM_000016



ファイル名称はクエリーネーム+「_SS_XX」(XXは上流から01・02…と付与します)
本来、比較配列(Sbjct)の名称を基準に名前を付与すべきなのですが、同じ比較配列(Sbjct)を有するクローンが存在する為、重複がない解析配列(Query)の名称をもとに、名前を付与しています。

作業毎個別フォルダ



作成されたファイルは作業毎フォルダ内に自動的に作成される「SSFASTA」フォルダ内に格納されます。
Sオプションを指定しない場合、このフォルダは作成されません。

* オプションSはBLASTN実行時のみ有効で、BLASTP実行時には使用できません。

パラメーター<重複度>計算方法

重複度とは、クエリー配列・サブジェクト配列において同一領域のアライメントが重複して複数箇所に存在する時、その重複領域がローカルアライメントにおいて占める割合を計算したものです。

A Query: 181 ggaaaactggactgaagtggcatccagataaaaatccctgagaataaagaad 231
Sbjct: 226 ggaaaactggactgaagtggcatccagataaaaatccctgagaataaagaad 276

B Query: 225 aaagaaggattaaatggtgaggaggatggaaagtcatttgacagtccatttgaattt 284
Sbjct: 363 aaagaaggattaaatggtgaggaggatggaaagtcatttgacagtccatttgaattt 422

重複度は重複領域を持つアライメントの中で最もアライメントスコアが高いアライメントを基準配列とし、その基準配列に対する重複度を計算します。例えば、上図の様なアライメント結果で、AがBよりアライメントスコアが高いとした時、AとBの配列におけるクエリー配列の重複領域がBのアライメントにおいて占める割合を計算します。

上図の場合、重複領域は225-231の7塩基、
Bのアライメント長は225-284の60塩基。
よって重複度は $7 \div 60 \times 100 = 11.7\%$ となります。

パラメーター＜重複度＞について-1

ALVISIONでは、ローカルアライメントの結果からアライメントスコアの高いものを基準とし、位置情報の整合性が取れるかどうかローカルアライメントの真偽を判定し、眞のアライメントと判断されたものを整列化しています。

しかし、ローカルアライメントの結果からは、アライメント領域の末端が他のアライメント領域の末端配列と同じ配列を有しているために、整合性が取れない場合があります。

例1

A Query: 181 ggaaaactggcactgaagtggcatccagataaaaatccctgagaataaagaad 231
Sbjct: 226 ggaaaactggcactgaagtggcatccagataaaaatccctgagaataaagaad 276

B

Query: 225 aaagaaggattaaatggtgaggaggagggtggaaagtcatttgacagtccatttgaattt 284
Sbjct: 363 aaagaaggattaaatggtgaggaggagggtggaaagtcatttgacagtccatttgaattt 422

上図のAとBではQueryの225～231がA・B両方のアライメントに存在し、位置的な矛盾が起こっています。
しかし、225～231は偶然一致したものであり、実際にはA・B両アライメントが眞であり、下図の様なアライメントを取るのが理想的です。この様な場合、ほとんどのアライメントにおいて重複度は10%以下になります。

重複配列削除

Query : ctcgtaaaaatggtgattactatggaaatgttaggcgtgggggacatgcctaccggggatattaaaaaggcataccggaaactggcaactggaaatgg
Sbjct : ctcgtaaaaatggtgattactatggaaatgttaggcgtgggggacatgcctaccggggatattaaaaaggcataccggaaactggcaactggaaatgg
155 165 175 185 195 205 215 225 235 245
210 220 230
Query : catccagataaaaatccctgagaataaagaag
Sbjct : catccagataaaaatccctgagaataaagaag 255 265 275 285 295 305 315 325 335 345
255 265 275 285 295 305 315 325 335 345
Query : attaaatggtgaggaggagggtggaaagtcatttgacagtccatttgaatttgcacattccgttaaccacgg
Sbjct : TCTATGACAAAATATGGGAAGAAAGattaaatggtgaggaggagggtggaaagtcatttgacagtccatttgaatttgcacattccgttaaccacgg
355 365... 375 385 395 405 415 425 435 445

パラメーター＜重複度＞について-2

例2：同塩基の連続した配列の場合

A

Query: 977 aaaaaaaaaaa; 987
Sbjct: 2114 aaaaaaaaaaa; 2124

B

Query: 977 aaaaaaaaaaaa; 987
Sbjct: 2116 aaaaaaaaaaaa; 2126

重複度 $11 \div 11 \times 100 = 100\%$

例2は連続する同塩基の配列が、少しずつずれた状態でアライメントされています。
実際にアライメントに使用するべき配列はAのみで、Bのアライメントは排除すべきものです。
理想的なアライメントは下図の様になります。

Query: 977 aaaaaaaaaaa 987
Sbjct: 2114 aaaaaaaaaaaa 2126

例2・3のように重複領域が存在し、一方のアライメントを排除するような場合、
ほとんどのアライメントにおいて、非常に高い重複度を示します。

例3: 同一配列が複数箇所にある場合

A

Query: 1551 atgcatcatcatgtat; 1563
Sbjct: 2264 atgcatcatcatgtat; 2276

B

Query: 1551 atgcatcatcatgtat; 1563
Sbjct: 2534 atgcatcatcatgtat; 2546

重複度 $13 \div 13 \times 100 = 100\%$

例3は同一配列が複数箇所に存在するために、複数箇所でローカルアライメントがとれます。
しかし、実際に使用するのはAのみでBのアライメントは排除すべきものです。

パラメーター＜重複度＞について-3

重複度を用いる事により、ローカルアライメント同士で位置的な矛盾があった場合、例1のようにローカルアライメントを採用すべきケースと、例2・3のように排除すべきケースの分類を行う事ができます。

現在は50%に設定されています。

アライメントツールでは、設定値以上の重複度のアライメントを排除されます。例1のようなパターンは多くのケースにおいて、アライメント長が長く重複度は低くなります。それに対して、例2・3のようなパターンはアライメント長が短く重複度は高くなります。

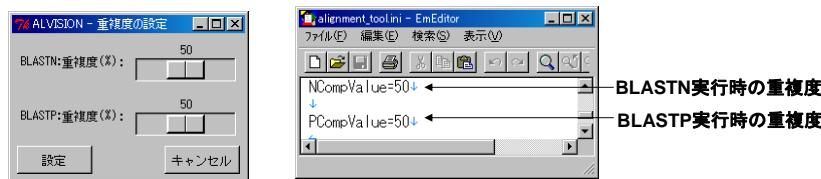
パラメーター設定方法

重複度の設定はオプションの“重複度の設定”で行います。

ALVISON起動時にはデフォルト値が設定されています。

ここで設定した値は起動中のみ有効であり、ALVISON終了後は再度設定する必要があります。

起動時のデフォルト値はalignment_tool.iniファイルにて変更可能です。



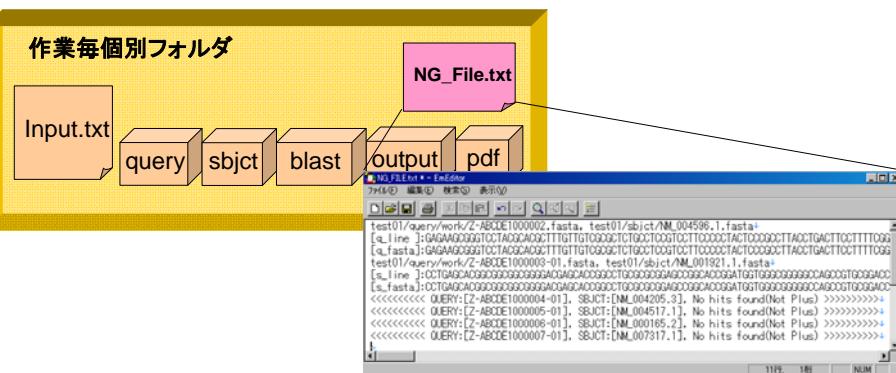
NG_FILEについて

実行時にエラーがあった場合、各エラー内容がNG_FILE.txtに出力されます。

このファイルは作業毎フォルダ内に自動的に作成されます。

ただし、エラーがない場合は作成されません。

- アライメントに用いている配列と、形成されたアライメントの配列が一致しない時。
- アライメントの取れる領域がない場合



実行結果表示例

アライメントが取れるところは小文字の塩基と「|」で表示。
ミスマッチのところは大文字の塩基と、「*」で表示。



に関するお問い合わせは下記まで

完全長cDNAプロジェクト
磯貝 隆夫 flj-cdna@nifty.com