phred によるベースコール

オートシーケンサから得られるデータは塩基配列そのものではなく、多くの場合各塩基に対応した波長光のシグナル強度の時間的変化を記録した波形ファイルとなっている。シーケンサの制御ソフトが標準でそのデータから塩基配列を読み出す (この操作をベースコールと呼ぶ) ものもあるが、エラーの有無を確認したり、信頼性を評価するには直に波形ファイルを扱った方がよい。

ここでは、優秀なベースコールソフトウェアである phred を利用して日立製スラブゲル型オートシーケンサ SQ-5500 から得られる SCF ファイルと、AppliedBiosystems 製キャピラリ型オートシーケンサ Prism 310 および 3730xl から得られる AB1 ファイルからベースコールを行い、phred が出力するベースコールの信頼性と波形を同時に参照しながら配列を確認・編集するまでを説明する。SCF・AB1 共にシーケンサ制御マシンが MacOS 8/9 機の場合、拡張子が無いかもしれませんが Windows 機の phred でも問題無く扱えます。ただし Mac バイナリ (Mac では Windows のように拡張子と対応ソフトを関連付けるのではなく、ファイル本体にそのファイルを作成したソフトなどの情報を付加します。それが Mac バイナリで、Windows/UNIX 機上ではファイル本体にくっついたゴミでしかありません。) を除去する必要があるかもしれません。Mac バイナリカッターとか macutils などというソフトで除去できるようです。

最近の ABI 製マシンでは KB Basecaller という優秀なベースコールソフトウェアが付属しており、phred 同様に信頼性を評価することができます。ABI から無料で配布されている Windows 用の Sequence Scanner という閲覧･編集用ソフトウェアを用いてこれらのファイルを扱うことができます。

必要なソフトウェア

・phred

当然これが無くては始まらない。学術研究利用なら無料で入手できます。リンク先から「How to get」の項を読んで入手して下さい。私が入手したときと変わっていなければ、指定のメールアドレスに必要事項を記入してメールを送ると返信メールの添付ファイルとしてソースコードが送られてくるはずです。コンパイルとインストールに関しては後述。

・phred の出力する信頼性を表示可能なソフト

無料で使えるものに以下のようなものがあります。

・FinchTV (Windows/Linux/MacOS X)
・4Peaks (MacOS X)
・TraceViewer (Java)
・Trev (Staden Package に収録) (Windows/Linux/MacOS X/ その他 UNIX)

他にもありますが、系統推定ではシンプルなこれらが使いやすいと思います。

phred のコンパイルとインストール

既に述べたように、phred はソースコードの形で送られてきますので、コンパイルする必要があります。UNIX、MacOS X の場合はコンパイラが標準で入っているか、簡単に導入できるはずです。Windows の場合は MinGW + MSYS か Services for UNIX(WinXP Home では使用不可) などを導入すればいいでしょう。導入方法はここでは説明しません。Cygwin や coLinux でも構いません。

コンパイル

とりあえずターミナルが立ち上げられるようになっていることを前提に説明します。ソースコードの入った圧縮ファイル (phred-dist-xxxxxx.c-acd.tar.Z) のあるディレクトリで以下のコマンドを実行します。

```
% mkdir phred
% cd phred
% tar xvzf ../phred-dist-xxxxxx.c-acd.tar.Z
% make
```

これでコンパイルができるはずですが、もし cc が無いというエラーが出た場合 (MinGW とか) は最後のコマンドを

```
% make CC=gcc
```

として下さい。これで実行ファイルができあがります。

インストール

phred は環境変数 PHRED_PARAMETER_FILE に設定されている設定ファイルを読み込んで動作します。配布ファイルを展開すると出てくる phredpar.dat が標準の設定ファイルです。やや旧式のメジャーなシーケンサなら編集せずに使用しても問題無いと思いますが、ここで扱う SQ-5500 と Prism 310 および 3730xl では設定が必要です。UNIX 改行コードが扱えるエディタで以下の 3 行を末尾に加えて下さい。

```
"DT310POP6{BDv3}v1.mob" terminator big-dye ABI_3100
"KB_3730_POP7_BDTv3.mob" terminator big-dye ABI_3700
"" terminator big-dye Beckman_CEQ_2000
```

1 行目は ABI Prism 310 + BigDye Terminator v3.x Cycle Sequencing Kit + POP6 泳動したファイルの場合、BigDye を用いて Dye Terminator 法で標識したサンプルを ABI 3100 で泳動したものとしてベースコールを行うという設定です。2 行目が ABI 3730xl + BigDye Terminator v3.x Cycle Sequencing Kit + POP7 泳動の場合、BigDye を用いて Dye Terminator 法で標識したサンプルを ABI 3700 で泳動したものとして、3 行目は設定ファイル内のどの条件にも一致しないサンプルは、BigDye を用いて Dye Terminator 法で標識したサンプルを Beckman CEQ 2000 で泳動したものとしてベースコールを行うという設定です。SQ-5500 + Amersham Thermo Sequenase Primer Cycle Sequencing Kit + LongRanger ゲル泳動の SCF ファイルはこの設定でベースコールするのが経験上最も良い (というか、マシ) のでこのようにしていますが、適当に変えていただいても構いません。

DT310POP6{BDv3}v1.mob や KB_3730_POP7_BDTv3.mob などに当たるものが分からない場合はとりあえず phred でベースコールしてみればエラーメッセージに出ますので、それを見ればいいでしょう。例えば上記の1行目を phredpar.dat に加えずに phred で ABI Prism 310 の出力ファイルを処理しようとすると以下のようなメッセージが出ます。

```
ファイル名
ファイル名 : unable to match primer ID string: skipping chromatogram
  unknown chemistry (DT310POP6{BDv3}v1.mob) in chromat ファイル名
  add a line of the form
  "DT310POP6{BDv3}v1.mob"    <chemistry>      <dye type>      <machine type>
  to the file C:\Windows\system32\phredpar.dat
  type `phred -doc' for more information
```

phredpar.dat を編集したら、適当な場所に置いておきます。Windows の場合はコンパイルしてできる phred.exe と共に C:\Windows\System32 にでも置いて

```
set PHRED_PARAMETER_FILE=C:\Windows\system32\phredpar.dat
```

をコマンドプロンプトで実行して環境変数を設定します。「コントロールパネル→システム→詳細設定→環境変数」から設定しても構いません。Win9x/Me 機では AUTOEXEC.BAT に上記の行を追記して下さい。

各種 UNIX、MacOS X の場合は

```
% su
# mkdir /usr/local/etc/phred
# cp phred phredpar.dat /usr/local/etc/phred
# echo -e "#!/bin/sh\nPHRED_PARAMETER_FILE=\
/usr/local/etc/phred/phredpar.dat \
/usr/local/etc/phred/phred $*" > /usr/local/bin/phred
# chmod 755 /usr/local/bin/phred
# exit
```

とでもすればよいだろう (インストール先は適宜変更)。この例では /usr/local/etc/phred に phred と phredpar.dat を置いておき、環境変数と引数を与えて phred を起動する sh スクリプト (/usr/local/bin/phred) を実行ファイルとして作成している。

phred でベースコールする

ベースコールの対象となるファイルのある場所で以下のコマンドを実行する。

```
phred 入力ファイル名 -c 出力ファイル名
```

これで、Standard Chromatgram Format(SCF) の出力ファイルが得られる。入力ファイルは SCF でも AB1 でも問題無い。出力ファイルにはベースコールに用いた波形とベースコール後の塩基配列、各塩基配列の信頼性の情報が全て含まれている。塩基配列のみ、信頼性のみをそれぞれ出力することもできるが、ファイルが別々になり管理が面倒なので筆者は SCF ファイルで出力するようにしている。

ファイルが大量にある場合、上記のように 1 ファイルずつ処理するのは面倒である。そこで、入力ファイル名の代わりに「-if 入力ファイルリストの記述してあるテキストファイル」や「-id 入力

ファイルのあるディレクトリ」を使うことで一気に大量のファイルを処理することができる。例えば、

```
mkdir basecalled
phred -id 入力ディレクトリ -cd basecalled
```

とすれば、入力ディレクトリ内の処理可能なファイルが全て処理され、同じファイル名で basecalled ディレクトリに SCF ファイルが出力される。ただし残念ながらサブディレクトリまでは処理されない。Windows の場合、

```
mkdir %1\basecalled
phred -id %1 -cd %1\basecalled
```

と書いたバッチファイル (またはそのショートカット) を「送る」(SendTo) に入れておくと、エクスプローラ上でディレクトリを右クリックから一発でディレクトリ内のファイルを処理できる。

その他のオプションについては

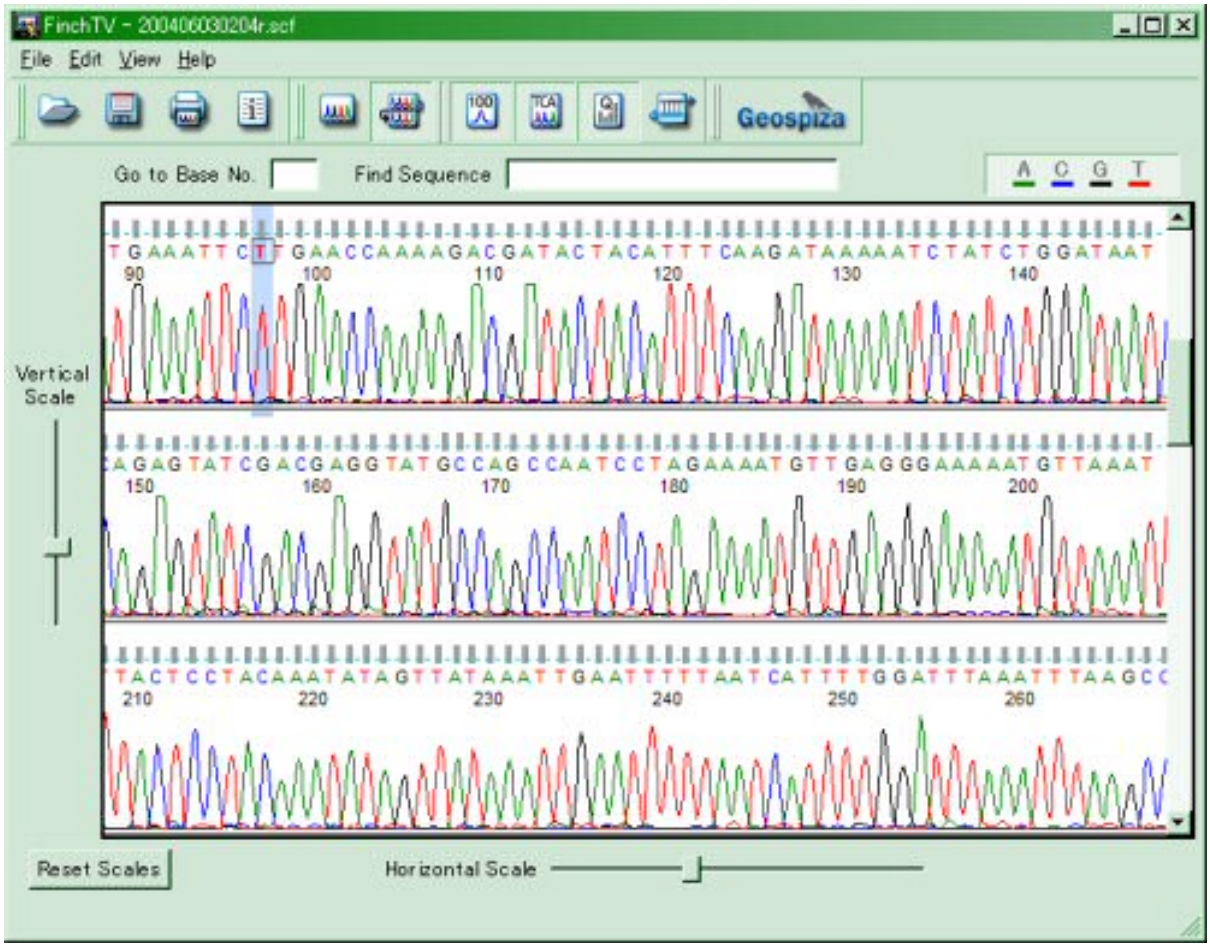
```
phred -doc
```

の出力を見て下さい。

ベースコール結果を確認・編集

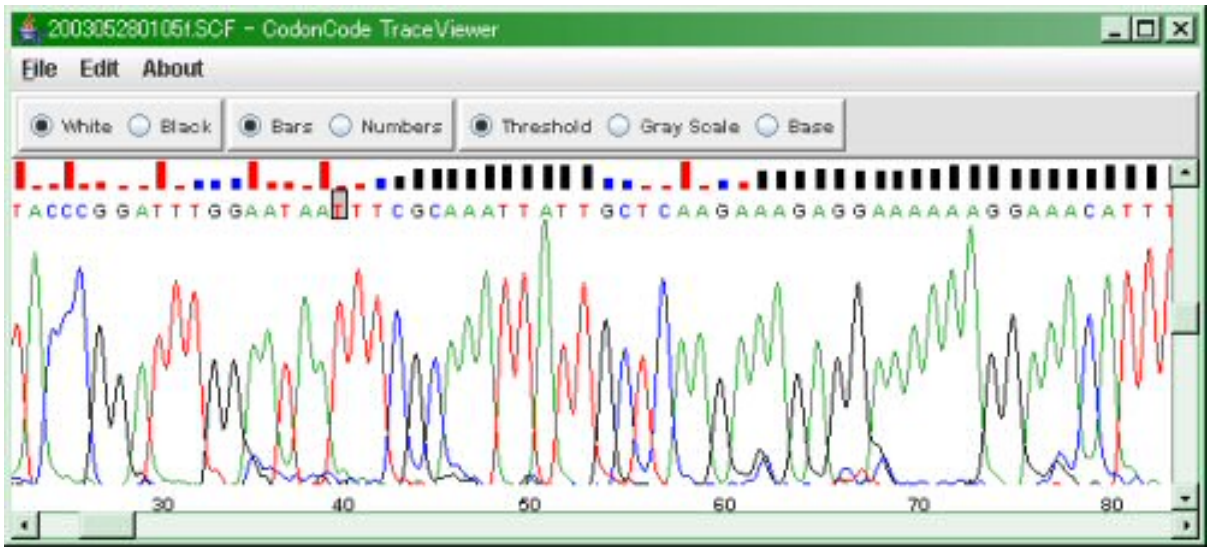
あとは出力されたファイルを前述のソフトで開けばベースコールの結果を確認し、おかしいところがあれば編集することができます。4Peaks は知りませんが、FinchTV と TraceViewer には FASTA 形式で塩基配列を出力する機能がありますので、様々なソフトに読み込むことができます。

結果を FinchTV で表示



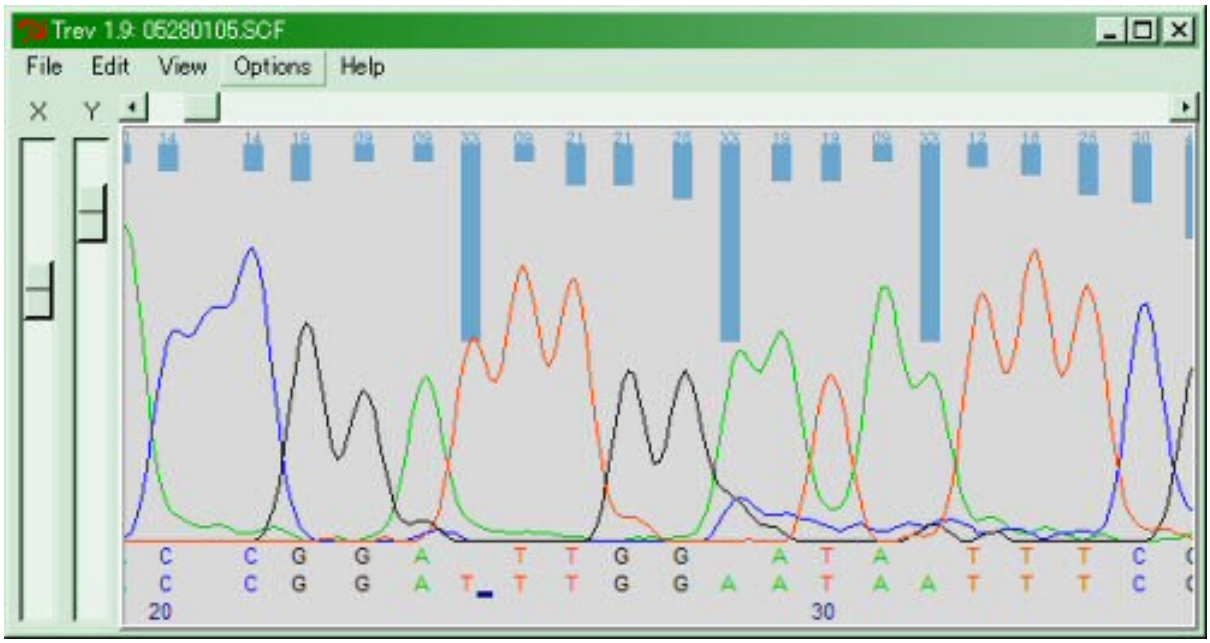
波形の上にベースコールの結果の塩基、その上に信頼性を表す棒グラフが表示されている。縮小しているため見えなくなっているが、信頼性が十分か不十分かの基準となる線も表示されている。その線よりも棒グラフが高いデータはほぼ見直す必要が無い。FinchTV では水平方向・垂直方向のスケールを下と左のバーで独立に変更することができる。手動で修正した塩基にはアンダーラインが付く。標準では 1 段表示だが「View → Wrapped View」で折り返し多段表示に変更できる。

結果を TraceViewer で表示



波形の上にベースコールの結果の塩基、その上に信頼性を表す棒グラフが表示されている。信頼性の高いデータでは棒グラフは黒、ほぼ問題無いデータでは青、見直す必要のあるデータでは赤となる。赤く表示されているのに高い棒グラフのデータは手動で編集したもの。垂直方向のスケールは調整可能だが水平方向は固定。

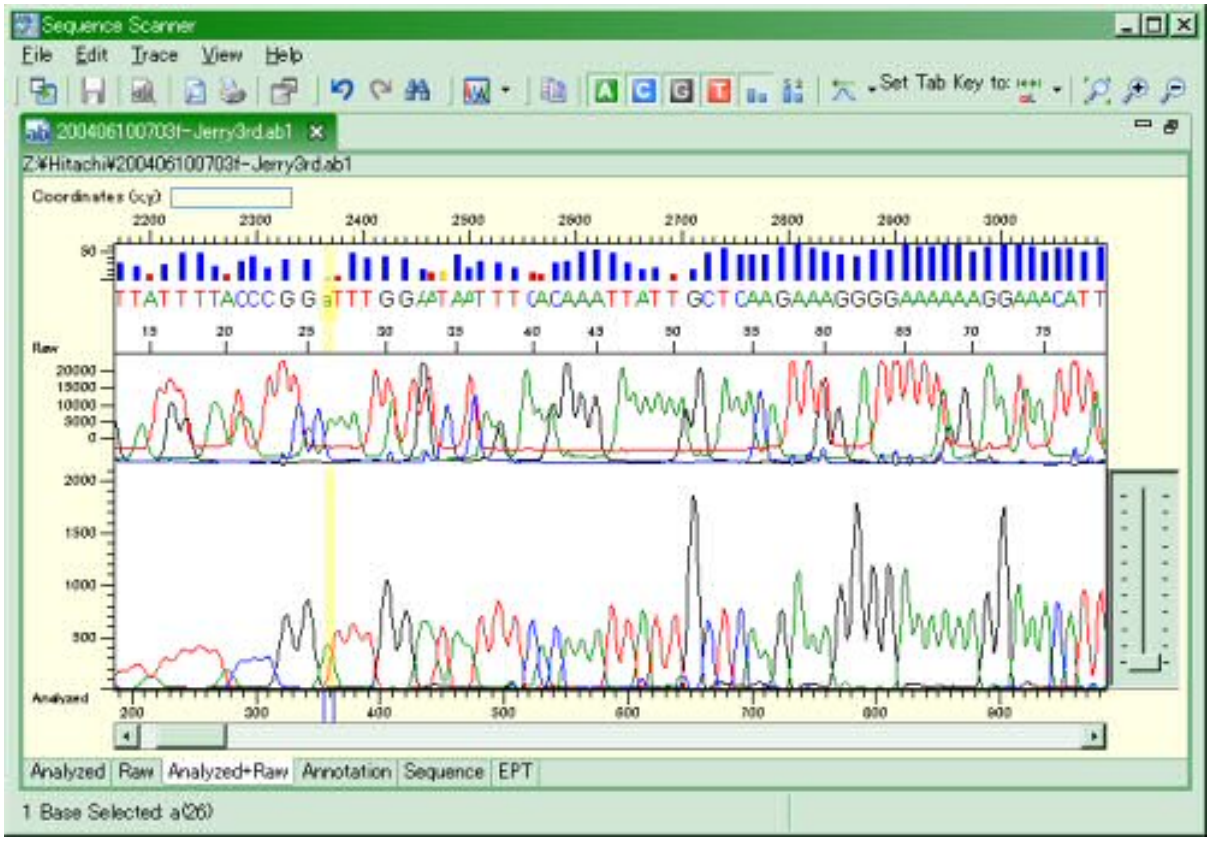
結果を Trev で表示



下から編集後の塩基配列、編集前の塩基配列、波形データ、信頼性の棒グラフ。信頼性が MAX なのは手動で編集したもの。標準では編集後の配列と信頼性は表示されないが、「View」メニューで「Display edits」と「Display confidence」にチェックを入れると表示されるようになる。配列編集時には「Edit」メニューで「Sequence」をチェックしてから行う。垂直・水平方向のスケールは左側のバーで調整可能。

結果を Sequence Scanner で表示

Sequence Scanner は phred が吐き出す SCF ファイルを扱うことができませんので、信頼性値を表示させたければ KB Basecaller でベースコールした AB1 ファイルを読み込ませる必要があります。SCF2ABI というソフトで SCF ファイルから ABI 形式へ変換することもできますが、この際に信頼性値は失われますので意味がありません。



ベースコール元の波形の上にシーケンサが吐いた Raw データの波形、さらにベースコール結果の塩基と信頼性を表す棒グラフが表示されている。水平方向・垂直方向のスケールを任意に変更可能。信頼性の低い塩基に Tab キーで飛ぶことができるのが便利。