前面的话

免费软件的一大特点,就是帮助文件一般较为简单,这个软件包尤为突出,不信选择程序的 HELP,你会发现虽然程序功能强大,但帮助文件只有非常简单的一些使用信息,实在令人失望。但你可以在解压出的文件中,找到 doc_ant.exe 文件,解压在另一个目录,是 HTML格式的介绍文件,对软件进行了一些介绍。

我花费了大量时间,写出了这篇使用介绍,希望对你的使用有所帮助,更高深的使用只能由自己去挖掘。希望谁有心得,能写出来提供给大家。如果你觉得我介绍的软件与写出的帮助对你的工作非常有帮助,请您寄给我 10 元钱或者更多,这不仅是对我经济上的资助,这对我也是精神上的鼓励,鼓励我将这个网站继续做下去。毕竟,我的收入是非常有限的,不一定能保证我一直将这个网站继续下去。谢谢。

地址:石家庄市 华药四区 5-3-401

谈杰

050031

生物的化学,就是蛋白质与核酸的化学。对蛋白质的研究是生物化学领域一个非常重要的部分。近年来,通过实施基因组计划,得到了大量的蛋白序列数据,但是,面对如此众多的蛋白质序列数据,其分析工作是一个非常困难的工作。用人工的方法是不可能完成如此大量的分析工作的。运用计算机,利用一定的运算规则,进行蛋白序列

分析是唯一的方法。蛋白序列分析软件包 ANTHEPROT 4.3 正是这样的一个程序。

蛋白序列分析软件包 ANTHEPROT 4.3 是位于法国的蛋白质生物与化学研究院(Institute of Biology and Chemistry of Proteins)用十多年时间开发出的蛋白质研究软件包。软件包包括了蛋白质研究领域所包括的大多数内容,功能非常强大。应用此软件包,使用个人电脑,便能进行各种蛋白序列分析与特性预测。感谢软件的编写者,他们将此软件包免费提供给此领域的同行使用,这对于我国的科学技术人员来说,实在是一件好事。因为一个功能再多再好的专业应用软件,如果作为商业软件进行销售的话,往往使大多数需要使用此软件的科技人员因为科研经费的问题,无缘使用,这点在我国尤为突出。

软件分为三个版本,分别用于 Unix 系统,DOS 系统与 WINDOWS95 或 98 系统,让我以用于 WINDOWS 系统的版本为例,介绍一下这个软件的基本用法。

软件包为一个自解压执行文件,文件名 Anthe4_3c.exe,大小为 1.9 兆。执行此文件,输入解压后存放的目录名,便可将所有文件解压在此目录下。主程序名为 Anthewin,在桌面上建立它的快捷图标,双击快捷图标便打开了 ANTHEPROT 4.3 主窗口。

通过主程序,我们可以载入蛋白序列,对序列进行编辑、打印、 拷贝、改变设定等操作,更重要的是,我们可以在此调用各种所需的 分析工具,对蛋白序列进行分析。

一. 序列编辑功能

ANTHEPROT 4.3 的主窗口便是其序列编辑窗口:

可以使用 按键或菜单中的 File/Open 命令打开各种序列文件,程序识别的文件类型包括:

单序列文件格式:*.SEQ , (ANTHEPROT 3.3 支持以下格式的单序列文件:DNA/Strider 格式、EMBL 格式、NBRF 格式、Pearson/Fasta格式、PIR 格式与 IG/Stanford 格式);

含有多个蛋白序列的蛋白数据库文件:*.BAS,(以 Pearson/Fasta 格式添加序列,上限为30条序列或32K文件大小);

ClustalV 或 ClustalW 文件: *.ALN 含有 ClustalW 格式的多队列;

含有多队列的 Multalin 4.1 格式文件: *.MUL;

在 Prosite 蛋白数据库中查寻出 Site 结果的文件: *.SIT,

蛋白质原子空间结构的 PDB 格式文件:*.PDB;

使用 IBCP (http://wwww.ibcp.fr/predict.html) 服务器预测蛋白二级结构所获得的结果文件格式:*.CNS;

序列可以从 ANTHEPROT 编辑窗口以键盘输入,或者以*.SEQ 等文件格式载入,也可以从其它文件复制粘贴到编辑窗口。最简单的序列格式为 Pearson/Fasta 格式,为文本格式文件,带有>开始的为序列蛋白的名称,其后便是蛋白序列。下面为一个此格式的例子:

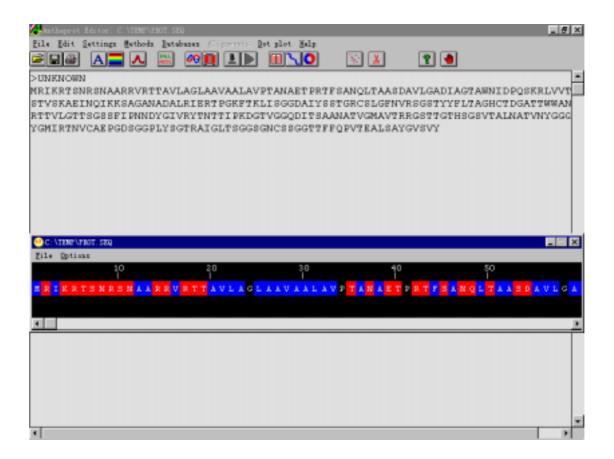
>MBN

 $VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDL\\ KKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGD\\ FGADAQGAMNKALELFRKDIAAKYKELGYQG$

程序支持的一些编辑功能如下:

- 取消,恢复(Undo): Ctrl+Z。
- 查找 Ctrl+T。
- 支持复制与粘贴。
- 可以自定义字符颜色,背景色与打印字体。
- 可打印序列文件。

输入或载入序列后,便同时打开了一个图像窗口,以彩图的形式显示出蛋白序列。不同的蛋白残基以不同的颜色表示。下面便是打开一个序列后的主程序窗口的例子:



二.工具栏与基本功能

当序列格式识别后,Methods 菜单便被激活,相应的快捷按键也出现在工具栏中。



这些按键依次为:

打开文件按键 📴 🛚

存储文件按键 📳 ;

打印按键 🞒

更改字体与格式按键 🔼 ;

更改设定颜色按键 🔲 🦼

```
进行蛋白序列二级结构预测 《 ;
在蛋白序列中查找符合 PROSITES 数据库的特征序列 绘制出蛋白序列的所有理化特性曲线 ;
在 Internet 或本地蛋白序列数据库中查找类似序列 》 ;
计算蛋白序列分子量,比重与各蛋白残基百分组成  ;
计算蛋白序列滴定曲线与等电点  ;
选定一个片段后,绘制 Helical Wheel 图 ;
进行点阵图(Dot Plot)分析 ;
计算信号肽潜在的断裂位点 。 ;
打开一个简单的帮助文件 ?
退出程序 。 。
```

三.基本分析工具介绍

除了通过单击快捷按键的方式进行各种蛋白序列分析,通过菜单进行各种操作也非常方便。下面让我将一些主要分析工具的功能介绍一下:

1. 点阵图 (Dot Plot)

Dot Plot 分析方法对以下分析工作非常有用:

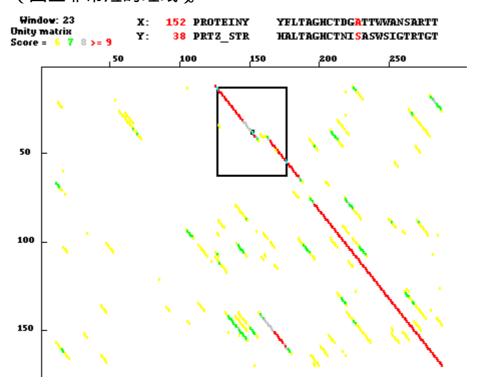
- 在一条序列内查找重复序列;
- 在 2 条序列中不同位置查找类似片段;
- 在不同序列中查找同源性。

只有当程序载入一条序列时,才可以激活 Dot Plot 程序,选择菜单命令 Dot Plot 或单击按键 运行 Dot Plot 程序,可再选择要进行比较的第二条序列。然后,再确定参数与限制条件,包括:

窗口尺寸:移动片段的长度

矩阵:选择替代矩阵

下图为含有两个蛋白酶序列的点阵图,两个酶具有同源性,因为由点阵图可以看到,基本上有非常好的斜线。但由于此斜线的位置开始于位置 115,很容易知道,水平轴的蛋白序列含有前体(由1到第114 个残基组成),也清楚地看到,它们具有一个小的内部重复片段(图上非常短的红线)。



2. 类似性查找 (Similarity Search)

此程序用来在一个蛋白数据库(可以是本地数据库,也可是网上的数据库)中,找出所有与给定序列具有同源性的序列。

选择菜单命令 Methods/Similarity Search/WWW or Local Fasta,或者 Methods/Similarity Search/WWW-Blast at PBIL NPS@ server 便实现在蛋白数据库中对类似性蛋白序列的查找。需注意,如果在本地蛋白数据库中查找,需要先下载 SwissProt 蛋白质数据库。

3. 查找 Si te/si gnature

在 ANTHEPROT 主窗口单击 链 按键或选取菜单命令 Methods / Site detection (Prosite),便进行位点与特征序列的查找。程序查找给定序列是否存在 PROSITE 数据库中已定义的特征序列,缺省设置为100%匹配,也可设定为 Mismatch 1 or 2(一到两个不匹配)或者Similarity level(类似水平)。当前 PROSITE 数据库是 95 年 2 月发布的,共 1054 个不同位点与特征序列,含有两个文件,PROSITE.DAT与 PROSITE.DOC。此两个文件必需均放在 ANTHEPROT 程序的目录下,程序才能进行。

下图为查找结束后程序输出的结果:

```
Site
          424
                    429
                         GOKVIE
                                   Observed frequency
                                                        1.14E-005
Site
          465
               to
                    470
                         GSFEAA
                                   Observed frequency:
                                                        3.92E-005
          502 to
Site
                   507 GILDSF
                                  Observed frequency: 1.02E-005
ATP/GTP-binding site motif A (P-loop) PS00017
 MG]-\times(4)-G-K-[ST]
Theoretical frequency:
                        1.09E-004
          169 to
                    176 GDRQTGKT Observed frequency: 2.42E-005
ATP synthase alpha and beta subunits signature P-[SAP]-[IV]-[DNH]-x(3)-S-x-S
                                                        PS00152
Theoretical frequency:
                       5.70E-007
          366 to
                   375 PAVNPGISVS
                                       Observed frequency: 5.20E-008
Elasped searching time: 8.12862458409424 Sec
```

绿色的文字可连接到 PROSITE 文件中相应的位置。

需要知道的是,在你的蛋白序列中找到的特征序列并不一定具有生物学意义,只能说,查找的结果对寻找潜在的有意义片段具有很大帮助。

4. 在序列数据库中查找一个自定义的特征序列: Pattern search

选择菜单命令 Databases/Pattern search ,便开始在网上的蛋白数据库或本地蛋白数据库中查找含有自定义特征序列的蛋白。 所定义的特征序列必需符合 PROSITE 规定的语法。 语法规则如下:

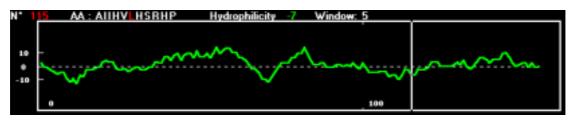
- 位置分隔符;
- -[]允许此位置为括号内的任何一个残基;
- { } 允许此位置为除了括号内所包括的任何一个残基;
- -x 代表任何残基;
- x(3) 代表任何3个氨基酸残基,

根据以上语法,如果要查找的特征序列为

N-[PT]-{GM}-x(2)-[ILVM],则序列 N-P-K-G-H-V, N-T-L-K-G-M 将能够找到,而序列 N-L-K-G-H-V, N-T-G-K-H-V 将不能找到。查找结束后,所有找到的序列将以 Pearson 格式,以 SEQ_x.BAS 为文件名存储在硬盘中。

5. 预测物理化学特性曲线

Antheprot 可以以图示的方法,显示出预测的蛋白质的物理化学特性曲线,下图为一个亲水性特性曲线的例子。在图中,一个可移动的光标,显示了目前在序列中的位置为 115,此位置的氨基酸残基(L)显示为红色,同时显示前 5个与后五个残基(为 AIIHVLHSRHP),特性值(-7)与参数(5)也显示在图中。



光标可以如下方式移动:

鼠标左键在选定位置单击;

方向键左或右为每次移动一个位置;

Shift+方向键左或右为每次移动十个位置;

方向键上或下为放大缩小。

可显示的物理化学特性曲线包括:

亲水曲线(Hydrophobicity profile)

疏水曲线(Hydrophilicity profile)
易溶性曲线(Solvent accessibility profile)
抗原性曲线(Antigenicity profile)
跨膜区域曲线(transmembranous regions profile)
结合抗原性曲线(Combined antigenicity profile)

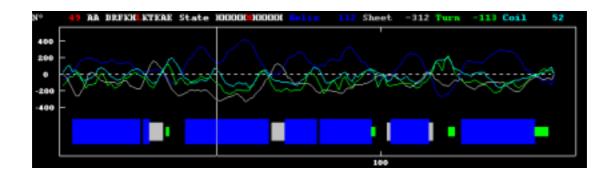
6. 二级结构预测

Antheprot 提供了几种方法进行蛋白质二级结构预测,包括:

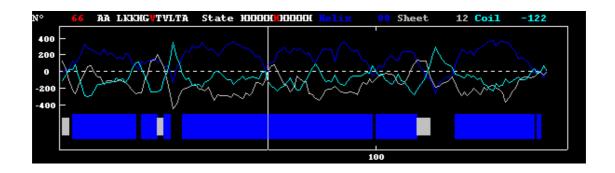
- GOR I 方法:在菜单中选择 Method/Secondary structure prediction/Garnier et al
- Gibrat 方法:在菜单中选择 Method/Secondary structure prediction/Gibrat et al
- Double Prediction Method DPM 方法:在菜单中选择 Method/Secondary structure prediction/DPM
- Homologue 方法:在菜单中选择 Method/Secondary structure prediction/Levin
- Predator 方法:在菜单中选择 Method/Secondary structure prediction/ Predator

五种方法的输出形式分别举例如下:

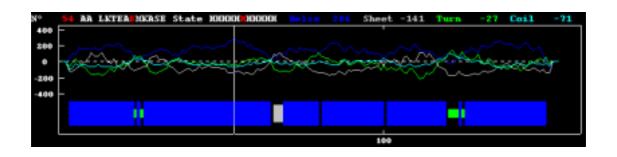
GOR I 方法



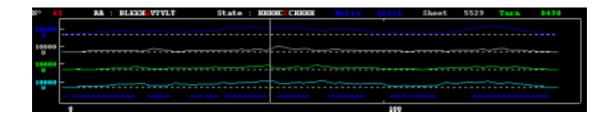
Gibrat 方法



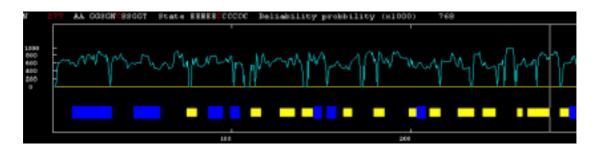
Double Prediction Method – DPM 方法



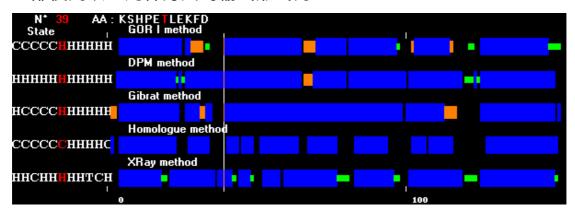
Homologue 方法



Predator 方法



全部使用以上几种方法的输出形式为



上图中加入了自定义的由 X 射线衍射法确定的二级结构图,可以显示在一个窗口,与其它预测的二级结构相比较。自定义的二级结构由以*.XRA 为扩展名的文件输入。其格式为 Pearson/Fasta 文件格

式,下面是一个*.xra 文件的例子,

>Myoglobin

СССНИННИННИННИНТТЅИНИННИННИННИННИННИНТСИННИТССЯНИ НИНИСИНИННИН

ННИННИННИНТТТТССИНИННИННИННИНТТСССИНИННИННИННИННИН СТТТТЅИНИННИ

НННННННННННННННТССС

- H 代表螺旋,图中表示为蓝色;
- E 代表折叠,图中表示为橙色;
- T 代表转角,图中表示为绿色;
- C 代表其它松散结构,图中表示为黑色。
- 7. 计算蛋白序列分子量,比容与各蛋白残基百分组成

程序可以非常方便地对蛋白序列进行分子量 ,比容与各蛋白残基百分组成的计算。单击 b键或选取 Methods/AA Composition, specific volume, M. Weight 便进行计算。快捷键为 Ctrl+M。

输出形式如下:

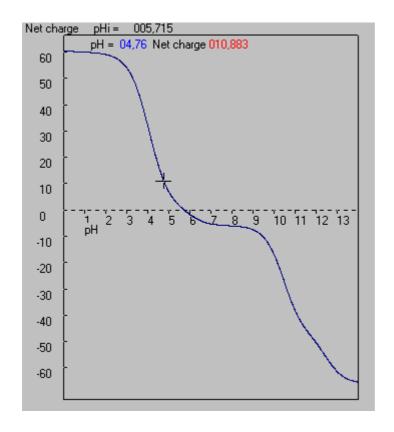
```
_ (# X
Eile Edit Settings
Selection in protein UNKNOWN
AA 1 to AA 300
MRIKETSNESNAAREVETTAVLAGLAAVAALAVPTANAETPETESAMQLT
AABDAVLGADIAGTAWNIDPQBKBLVVTVDBTVSKAEINQIKKBAGANAD
ALRIERTPGKFTKLISGGDAIYSSTGRCSLGFNVRSGSTYYFLTAGHCTD
GATTWWANSARTTVLGTTSGSSFIPNNDYGIVRYTNTTIPKDGTVGGQDI
TBAANATVOKAVTRROBTTOTHBOSVTALNATVNYOGODVVYOKIRTNVC
AEPGDSGGPLYSGTRAIGLTSGGSGNCSSGGTTFFQPUTEALSAYGVSVY
Amino acid Number
Ala(A): 38
           38
                       12.66
Asx(B):
Cys (C):
App (D) :
Glu(E):
Phe (F):
                        2.33
G1y(G):
              37
                       12.33
                       .66
4.66
His (E):
Ile(I):
Lys(K):
                        2.66
Leu(L):
              15
Het (M):
Asn (N):
                        5.66
Pro(P) I
               9
                        1.66
Gln(Q):
Arg(R):
              18
30
Ser (8):
                      10
The (T):
              40
                      13.33
Val (V):
                        7.66
              23
Top (W) :
Unit (X) :
                        0
Tyr(T):
G1x(Z):
              10
                        3.33
              0
```

8. 计算蛋白序列滴定曲线与等电点

ANTHEPRO4.3 还可根据一定算法,对输入的蛋白序列计算其滴 定曲线与等电点。

单击 Y键或选取 Methods/Titration curve 便可进行计算,快捷键为 Ctrl+U。

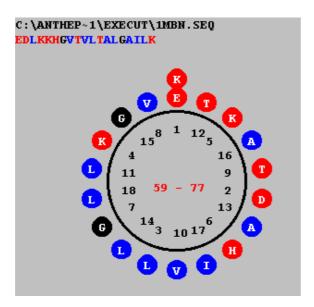
输出图形如下:



9. 选定一个片段后,绘制 Helical Wheel图

ANTHEPRO4.3 还可先选定一个蛋白片段,绘制其 Helical Wheel 图。单击 按键或选取 Methods/ Helical Wheel 便可进行计算与绘制。在二级结构预测输出窗口,用鼠标左键定义起始点,用鼠标右键定义结束点,选择菜单命令 Helical Wheel,也可很方便的绘制此定义片段的 Helical Wheel 图

输出图形的例子如下:

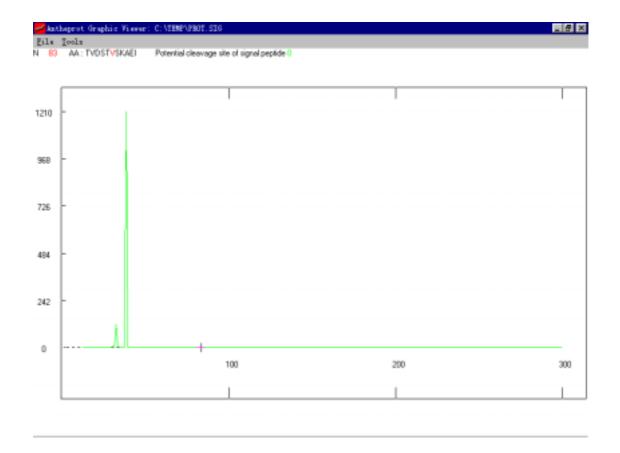


在此窗口,单击鼠标左键放大图形,单击鼠标右键缩小图形。

10. 进行潜在的信号肽与断裂位点预测

ANTHEPRO4.3 还对输入的蛋白片段,进行潜在的信号肽与断裂位点预测。单击 送键或选取 Methods/Potential cleavage site of signal pept.,再选择是原核序列或或者真核序列便可进行预测。

输出结果举例如下:



四. 结果的输出以及与其它应用程序进行数据交换

ANTHEPRO4.3 所输出的所有文字、结果图形与曲线,均可以打印,也可以拷贝到剪贴板,粘贴到其它应用软件如 WORD 中,或者输出为*.BMP 格式的文件,使用其它软件进行再编辑。程序的缺省设置将输出图形背景色设为黑色,在所有图形输出窗口,均可以以 Invert Mode (即黑白反色模式)输出,或者可以在 Options 菜单由用户定义背景与前景色颜色。这样,可以将黑色背景反转为白色,符合我们的习惯,打印时也节约些墨水。

ANTHEPRO4.3 蛋白质分析软件包的简单介绍到这里就结束了,本文只是对此软件进行了一个粗略的介绍,许多细节的用法还有待大家去发掘。总的来说,ANTHEPRO4.3 蛋白质分析软件包可以极大地加快研究蛋白序列的速度,并且能够获得具有发表质量的图表与数据,它对生物化学工作者来说,是一个不可多得的好软件。可以毫不夸张地说,ANTHEPRO4.3 蛋白质分析软件包是蛋白分析领域的"重磅炸弹",威力巨大,每一个生物化学工作者都应熟悉它,用好它。

ANTHEPRO4.3 蛋白质分析软件包与相应蛋白质数据库可以在ftp://ftp.ibcp.fr/pub/ANTHEPROT/WINDOWS 下载,也可以到我的个人主页:"免费生物化学软件天地"(http:://biosoft.yeah.net)下载,那里还有许多其它的生物化学相关软件可供下载。我的 Email 地址为:tanjie@earthling.net,希望大家常与我联系。