

# 运用生物信息学的方法对胶质纤维酸性蛋白的初步分析

贺金堂

学号: 10311034

**摘要:** 胶质纤维酸性蛋白是星形胶质细胞中的一种标志蛋白,它在神经元内环境的维持和血脑屏障中起着重要作用。本文主要运用生物信息学的方法,对呈现在双向电泳凝胶上、并用质谱鉴定出来的胶质纤维酸性蛋白进行了初步分析,从而对其性质有一个全面的了解。

胶质纤维酸性蛋白(glial fibrillary acidic protein, GFAP) 是星形胶质细胞中的一种标志蛋白,它参与神经元内环境的维持和血脑屏障。在脑损伤的炎症中会出现以胶质纤维酸性蛋白为特征的纤维化。这种纤维化形成“疤痕”,有隔离和控制细胞外液的离子和分子成分(如  $K^+$ , 神经递质等)的作用。

蛋白质组学的方法是系统化地研究蛋白质性质的一把利器。双向电泳、质谱和生物信息学是蛋白质组学的三个基本技术平台。从人脑组织中提取蛋白,然后进行双向电泳,切蛋白点,胶内酶切后进行质谱鉴定,发现在双向电泳图谱上,胶质纤维酸性蛋白并不是与单个点相对应,而是令人吃惊的对应着 50 多个点,这主要是由于基因的拼接和翻译后修饰造成的。正是由于胶质纤维酸性蛋白以多种形式存在,从而使其能够非常灵活的参与生理活动的调节。本文主要运用生物信息学的方法对胶质纤维酸性蛋白进行了初步分析,得到了蛋白质组学研究中感兴趣的一些蛋白质基本参数,包括分子量、等电点、氨基酸残基的组成、疏水性、跨膜区、三维结构、亚细胞定位、基本功能等基本信息,并通过 blast 找到其同源序列,建立了进化树。

首先选择 Swissprot,进行 SRS Quick Research,搜索 GFAP,种属选择 Human,部分信息如下:

```
ID  GFAP_HUMAN      STANDARD;      PRT;   432 AA.
AC  P14136;
DT  01-JAN-1990 (Rel. 13, Created)
DT  01-JAN-1990 (Rel. 13, Last sequence update)
DT  10-OCT-2003 (Rel. 42, Last annotation update)
DE  Glial fibrillary acidic protein, astrocyte (GFAP).
GN  GFAP.
OS  Homo sapiens (Human).
```

OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
 OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.  
 OX NCBI\_TaxID=[9606](#);  
 DR EMBL; [J04569](#); [AAA52528](#).1; -.  
 DR EMBL; [S40719](#); [AAB22581](#).1; -.  
 DR EMBL; [M26638](#); [AAA52529](#).1; -.  
 DR EMBL; [BC013596](#); [AAH13596](#).1; -.  
 DR EMBL; [BC041765](#); [AAH41765](#).1; -.  
 DR PIR; [A32936](#); A32936.  
 DR HSC-2DPAGE; [P14136](#); HUMAN.  
 DR Genew; HGNC:4235; GFAP.  
 DR MIM; [137780](#); -.  
 DR GO; GO:0005882; C:intermediate filament; TAS.  
 DR GO; GO:0005200; F:structural constituent of cytoskeleton; TAS.  
 DR InterPro; [IPR006821](#); Filament\_head.  
 DR InterPro; [IPR001664](#); IF.  
 DR InterPro; [IPR002957](#); Keratin\_I.  
 DR Pfam; [PF00038](#); filament; 1.  
 DR Pfam; [PF04732](#); filament\_head; 1.  
 DR PRINTS; [PR01248](#); TYPE1KERATIN.  
 DR PROSITE; [PS00226](#); IF; 1.  
 KW Intermediate filament; Coiled coil.  
 FT [DOMAIN](#) 1 72 HEAD.  
 FT [DOMAIN](#) 73 377 ROD.  
 FT [DOMAIN](#) 378 432 TAIL.  
 FT [DOMAIN](#) 73 104 COIL 1A.  
 FT [DOMAIN](#) 105 115 LINKER 1.  
 FT [DOMAIN](#) 116 214 COIL 1B.  
 FT [DOMAIN](#) 215 230 LINKER 12.  
 FT [DOMAIN](#) 231 252 COIL 2A.  
 FT [DOMAIN](#) 253 256 LINKER 2.  
 FT [DOMAIN](#) 257 377 COIL 2B.  
 SQ SEQUENCE 432 AA; 49880 MW; E6C3B3454C3F1250 CRC64;  
 MERRRITSAA RRSYVSSGEM MVGGLAPGRR LGPGTRLSLA RMPPPLPTRV DFSLAGALNA  
 GFKETRASER AEMMELNDRF ASYIEKVRFL EQQNKALAAE LNQLRAKEPT KLADVYQAEI  
 RELRLRLDQL TANSARLEVE RDNLAQDLAT VRQKLQDET N LRLEAENLA AYRQEAD EAT  
 LARLDLERKI ESLEEEIRFL RKIHEEEVRE LQEQLARQQV HV ELDVAKPD LTAALKEIRT  
 QYEAMASSNM HEAEWYRSK FADLTDA AAR NAELLRQAKH EANDYRRQLQ SLTCDLESLR  
 GTNESLERQM REQEERHVRE AASYQEALAR LEEEGQSLKD EMARHLQEYQ DLLNVKLALD  
 IEIATYRKLL EGEENRITIP VQTFSNLQIR ETSLDTKSVS EGHLKRNIVV KTVEMRDGEV  
 IKESKQEHKD VM//

进入 SSH, 建立文件 GFAP.fasta, 将 GFAP 的序列 copy 到此文件中。然后  
 用 EMBOSS 命令结合一些生物信息学工具进行分析:

## 1. 统计蛋白质序列“GFAP.fasta”的参数

(1) 用 pepstats 命令统计蛋白质序列“GFAP.fasta”中各种氨基酸的含量，并得到分子量、等电点、带电荷数以及 280nm 光吸收值等有用信息。

```
pepstats GFAP.fasta
```

输出结果如下：

Molecular weight = 49880.06                      Residues = 432

Average Residue Weight = 115.463              Charge = -9.0

Isoelectric Point = 5.2047

A280 Molar Extinction Coefficient = 18490

A280 Extinction Coefficient 1mg/ml = 0.37

Probability of expression in inclusion bodies = 0.781

Residue	Number	Mole%	DayhoffStat
A = Ala	47	10.880	1.265
B = Asx	0	0.000	0.000
C = Cys	1	0.231	0.080
D = Asp	21	4.861	0.884
E = Glu	60	13.889	2.315
F = Phe	7	1.620	0.450
G = Gly	13	3.009	0.358
H = His	8	1.852	0.926
I = Ile	13	3.009	0.669
K = Lys	21	4.861	0.737
L = Leu	56	12.963	1.752
M = Met	12	2.778	1.634
N = Asn	17	3.935	0.915
P = Pro	9	2.083	0.401
Q = Gln	26	6.019	1.543
R = Arg	47	10.880	2.220
S = Ser	23	5.324	0.761
T = Thr	20	4.630	0.759

V = Val	20	4.630	0.701
W = Trp	1	0.231	0.178
X = Xaa	0	0.000	0.000
Y = Tyr	10	2.315	0.681
Z = Glx	0	0.000	0.000

Property	Residues	Number	Mole%
Tiny	(A+C+G+S+T)	104	24.074
Small	(A+B+C+D+G+N+P+S+T+V)	171	39.583
Aliphatic	(I+L+V)	89	20.602
Aromatic	(F+H+W+Y)	26	6.019
Non-polar	(A+C+F+G+I+L+M+P+V+W+Y)	189	43.750
Polar	(D+E+H+K+N+Q+R+S+T+Z)	243	56.250
Charged	(B+D+E+H+K+R+Z)	157	36.343
Basic	(H+K+R)	76	17.593
Acidic	(B+D+E+Z)	81	18.750

(2) 为了更加全面地统计各参数，用 expasy 中的工具 ProtParam (<http://cn.expasy.org/tools/protparam.html>)，结果如下：

Number of amino acids: 432

Molecular weight: 49880.2

Theoretical pI: 5.42

Amino acid composition:

Ala (A)	47	10.9%
Arg (R)	47	10.9%
Asn (N)	17	3.9%
Asp (D)	21	4.9%
Cys (C)	1	0.2%
Gln (Q)	26	6.0%
Glu (E)	60	13.9%
Gly (G)	13	3.0%
His (H)	8	1.9%
Ile (I)	13	3.0%
Leu (L)	56	13.0%
Lys (K)	21	4.9%
Met (M)	12	2.8%
Phe (F)	7	1.6%
Pro (P)	9	2.1%

Ser (S)	23	5.3%
Thr (T)	20	4.6%
Trp (W)	1	0.2%
Tyr (Y)	10	2.3%
Val (V)	20	4.6%
Asx (B)	0	0.0%
Glx (Z)	0	0.0%
Xaa (X)	0	0.0%

Total number of negatively charged residues (Asp + Glu): 81

Total number of positively charged residues (Arg + Lys): 68

**Atomic composition:**

Carbon	C	2140
Hydrogen	H	3516
Nitrogen	N	654
Oxygen	O	691
Sulfur	S	13

**Formula:** C<sub>2140</sub>H<sub>3516</sub>N<sub>654</sub>O<sub>691</sub>S<sub>13</sub>

**Total number of atoms:** 7014

**Extinction coefficients:**

Conditions: 6.0 M guanidium hydrochloride

0.02 M phosphate buffer

pH 6.5

Extinction coefficients are in units of M<sup>-1</sup> cm<sup>-1</sup>.

The first table lists values computed assuming ALL Cys residues appear as half cystines, whereas the second table assumes that NONE do.

	276	278	279	280	282
	nm	nm	nm	nm	nm
Ext. coefficient	19900	19600	19110	18490	17600
Abs 0.1% (=1 g/l)	0.399	0.393	0.383	0.371	0.353

	276	278	279	280	282
	nm	nm	nm	nm	nm
Ext. coefficient	19900	19600	19110	18490	17600
Abs 0.1% (=1 g/l)	0.399	0.393	0.383	0.371	0.353

**Estimated half-life:**

The N-terminal of the sequence considered is M (Met).

The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro).

>20 hours (yeast, in vivo).

>10 hours (Escherichia coli, in vivo).

### Instability index:

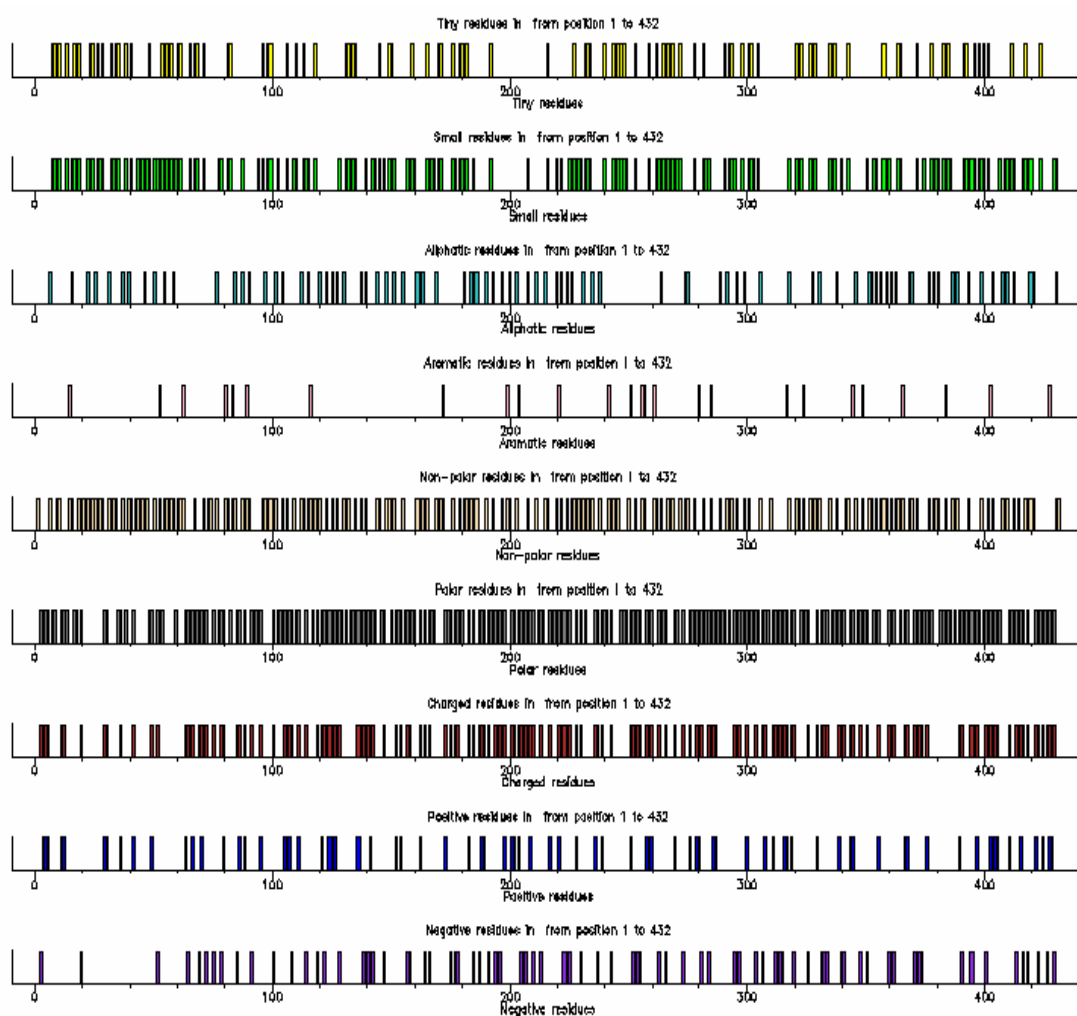
The instability index (II) is computed to be 52.74

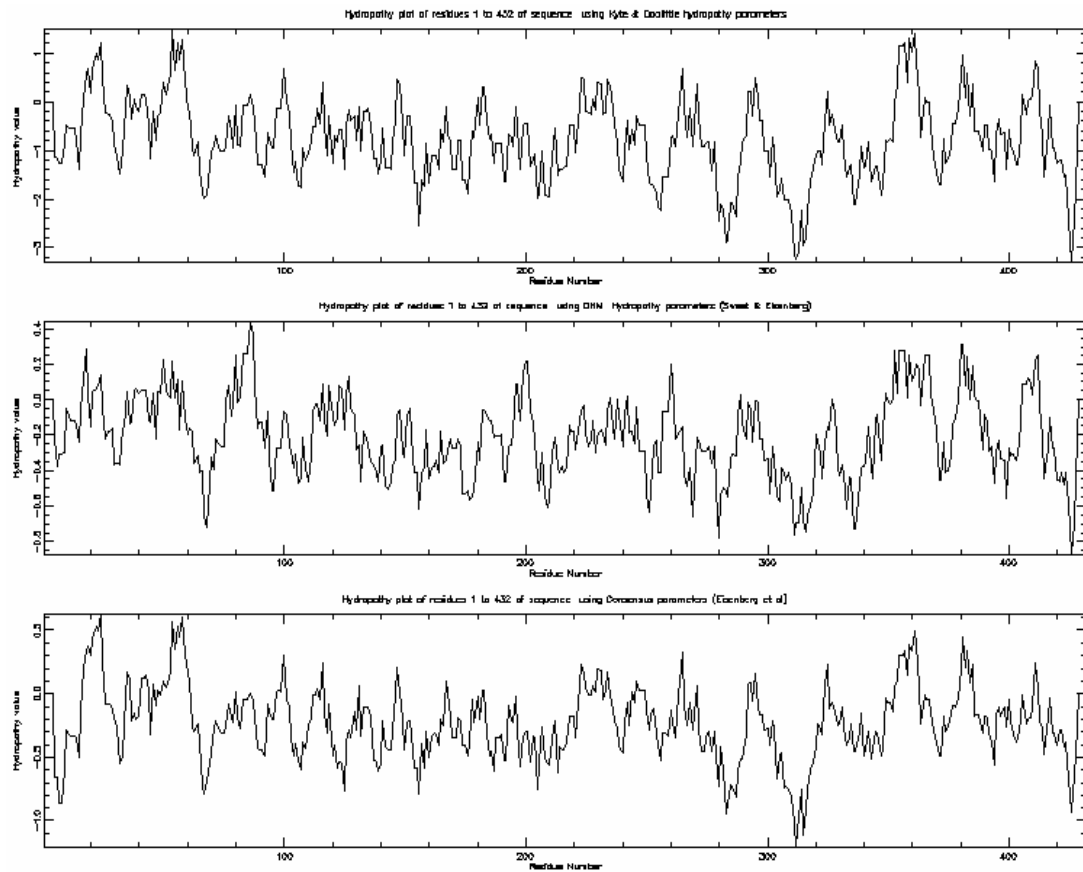
This classifies the protein as unstable.

Aliphatic index: 86.60

Grand average of hydropathicity (GRAVY): -0.773

2. 用 `pepinfo` 命令以图形方式显示蛋白质序列“GFAP.fasta”中各种不同性质氨基酸的含量。 输出结果如下:

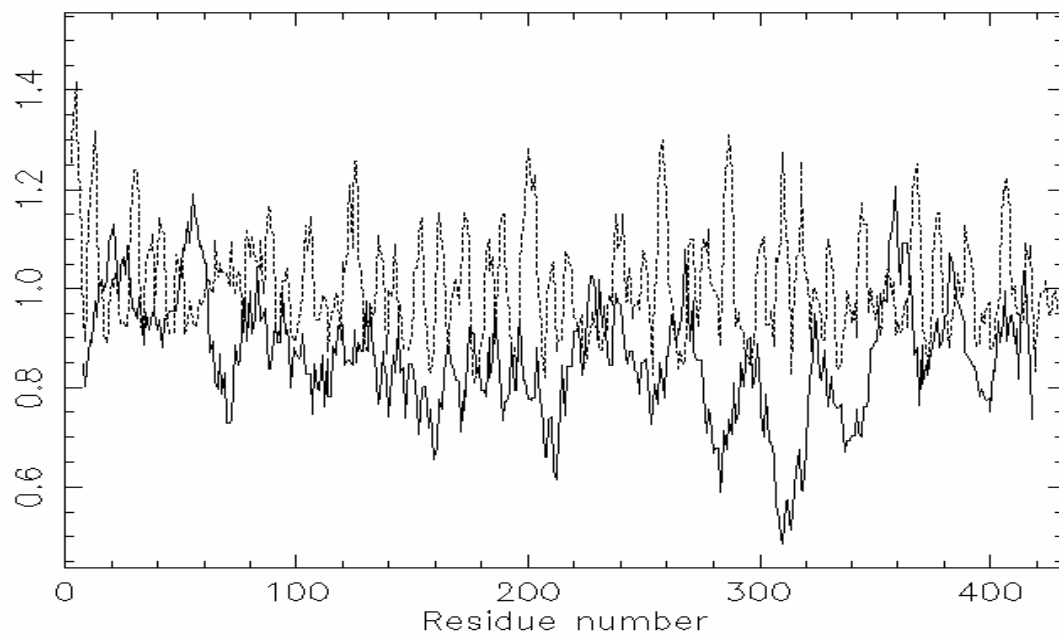




### 3. 预测蛋白质序列“GFAP.fasta”的跨膜区。

(1) 用 tmap 命令预测如下：

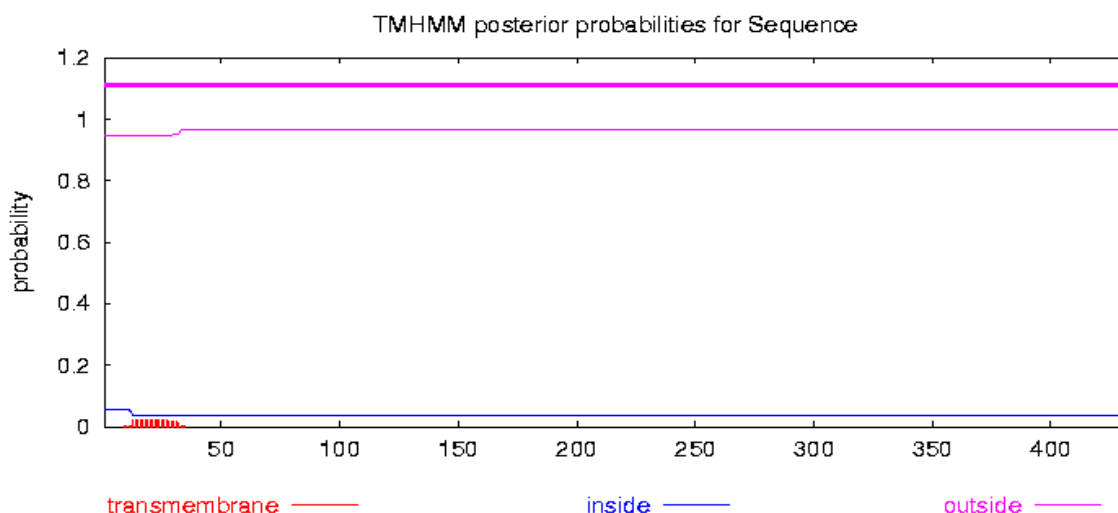
Tmap



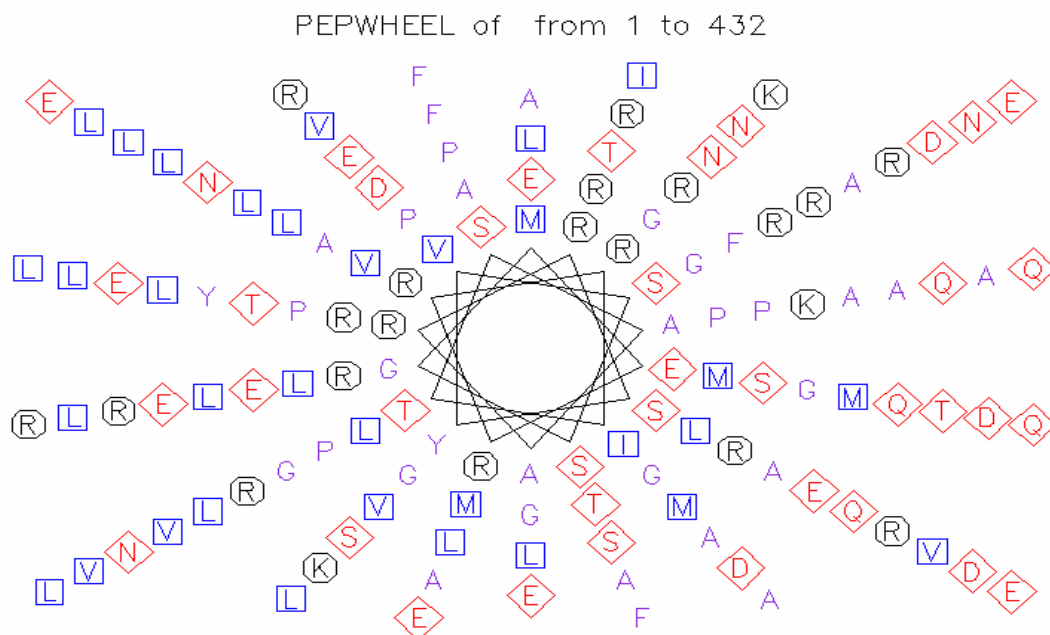
由上图可知此蛋白不是膜蛋白，无跨膜区。

(2) 为了更直观的显示跨膜区，可以借助 `expasy`(<http://cn.expasy.org/>) 中的 `TMHMM` 工具 (<http://www.cbs.dtu.dk/services/TMHMM-2.0/>)，预测结果如下：

```
# Sequence Length: 432
# Sequence Number of predicted TMHs: 0
# Sequence Exp number of AAs in TMHs: 0.38312
# Sequence Exp number, first 60 AAs: 0.3831
# Sequence Total prob of N-in: 0.05392
Sequence      TMHMM2.0      outside      1      432
```



4. 用 `pepwheel` 命令以图形方式显示蛋白质序列 “`GFAP.fasta`” 的所有 432 个残基的螺旋轮。输出结果如下：



## 5. GFAP 疏水性的预测:

蛋白质疏水性的预测在蛋白质组学研究中是必不可少的,通常根据蛋白质的 GRAVY 值来预测。GRAVY 值的范围在 2 与-2 之间,正值表明此蛋白为疏水蛋白,负值表明为亲水蛋白。

从通过 protparam 工具统计的结果中可知, **Grand average of hydropathicity (GRAVY)**为 -0.773,这说明人源的 GFAP 为亲水蛋白。

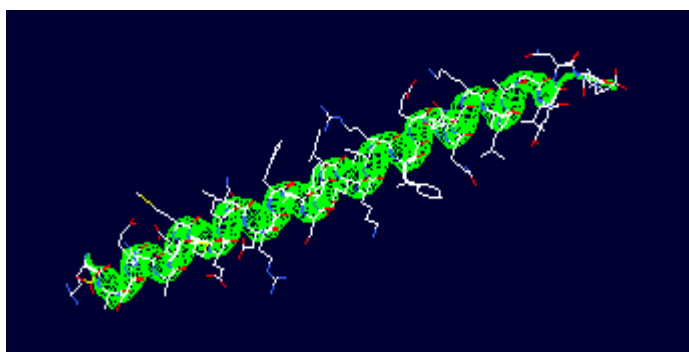
## 6. 预测 GFAP 的亚细胞定位:

通过工具 PSORT (<http://psort.nibb.ac.jp/form2.html>) 进行预测,结果如下:

47.8 %: nuclear  
39.1 %: mitochondrial  
13.0 %: cytoplasmic

## 7. 预测 GFAP 的三维结构:

通过 SRS 搜索,没有得到 GFAP 的 PDB 文件,说明此蛋白的三维结构还没有解析出来。用 expasy 中的 SWISS-MODEL 工具 (<http://swissmodel.expasy.org/>) 预测如下:



## 8. 预测 GFAP 的功能:

用 Gene ontology(GO)工具 (<http://www.ebi.ac.uk/ego/>) 预测 GFAP 的功能,结果如下:

All annotation for the protein [GFAP HUMAN](#) (P14136).

Select	Name	GO ID	Source	Evidence	Reference	With
<b>function (2)</b>						
<input type="checkbox"/>	structural constituent of cytoskeleton	<a href="#">GO:0005200</a>	<a href="#">Proteome Inc</a>	<a href="#">TAS</a>	<a href="#">PubMed: 2740350</a>	

<input type="checkbox"/>	structural molecule activity	<a href="#">GO:0005198</a>	<a href="#">InterPro</a>	<a href="#">IEA</a>	<a href="#">InterPro: IPR001664</a>	
<b>component (3)</b>						
<input type="checkbox"/>	intermediate filament	<a href="#">GO:0005882</a>	<a href="#">Proteome Inc</a>	<a href="#">TAS</a>	<a href="#">PubMed: 2740350</a>	
<input type="checkbox"/>	intermediate filament	<a href="#">GO:0005882</a>	<a href="#">InterPro</a>	<a href="#">IEA</a>	<a href="#">InterPro: IPR006821</a>	
<input type="checkbox"/>	intermediate filament	<a href="#">GO:0005882</a>	<a href="#">Keyword</a>	<a href="#">IEA</a>		

由上表可知，GFAP 是一个骨架蛋白，是中间纤维的重要组成部分。这与实验所观察到的现象是一致的，在脑损伤的炎症中出现的疤痕正是由于骨架蛋白 GFAP 的纤维化形成的。

### 9. 多序列比对，并建进化树

首先在 NCBI 中 BLAST 从而得到许多 GFAP 的同源蛋白，然后将同源性较高的 5 个蛋白的序列下载为 fasta 格式，并把这些序列合并到一个文件中，5 个序列：

```
>gi|4503979|ref|NP_002046.1| glial fibrillary acidic protein [Homo sapiens]
>gi|16265836|gb|AAL16662.1|AF419299_1 glial fibrillary acidic protein [Homo sapiens]
>gi|38566198|gb|AAH62609.1| GFAP protein [Homo sapiens]
>gi|28849921|ref|NP_776490.1| glial fibrillary acidic protein [Bos taurus]
>gi|417050|sp|P03995|GFAP_MOUSE Glial fibrillary acidic protein, astrocyte (GFAP)
```

然后用 clustalw 进行多序列比对，得到 aln 文件，结果如下：

```
gi|4503979      MERRRITSAARRSYVSSGEMMVGGLAPGRRLLGPGTRLSLARMPPPLPTRV
gi|16265836    MERRRITSAARRSYVSSGEMMVGGLAPGRRLLGPGTRLSLARMPPPLPTRV
gi|38566198    -ERRRITSAARRSYVSSGEMMVGGLAPGRRLLGPGTRLSLARMPPPLPTRV
gi|28849921    MERRRVTSATRRSYVSSSEMVG---GRRLGPGTRLSLARMPPPLPARV
gi|417050      MERRRITS-ARRSYAS--ETVVRGLGPSRQLGTMPRFSLSRMTPPLPARV
                ***:* ** :***.* * * :*      .*:**. .*:**:**.***:**

gi|4503979      DFSLAGALNAGFKETRASERAEMMELNDRFASYIEKVRFLFLEQQNKALAAE
gi|16265836    DFSLAGALNAGFKETRASERAEMMELNDRFASYIEKVRFLFLEQQNKALAAE
gi|38566198    DFSLAGALNAGFKETRASERAEMMELNDRFASYIEKVRFLFLEQQNKALAAE
gi|28849921    DFSLAGALNSGFKETRASERAEMMELNDRFASYIEKVRFLFLEQQNKALAAE
gi|417050      DFSLAGALNAGFKETRASERAEMMELNDRFASYIEKVRFLFLEQQNKALAAE
                *****:*****
```

gi | 4503979 LNQLRAKEPTKLADVYQAE LRELRLRLDQLTANSARLEVERDNLAQDLAT  
gi | 16265836 LNQLRAKEPTKLADVYQAE LRELRLRLDQLTANSARLEVERDNLAQDLAT  
gi | 38566198 LNQLRAKEPTKLADVYQAE LRELRLRLDQLTANSARLEVERDNLAQDLAT  
gi | 28849921 LNQLRAKEPTKLADVYQAE LRELRLRLDQLTANSARLEVERDNLAQDLGT  
gi | 417050 LNQLRAKEPTKLADVYQAE LRELRLRLDQLTANSARLEVERDNFAQDLGT  
\*\*\*\*\*:\*\*\*\*.\*

gi | 4503979 VRQKLQDETNRLEAENNLAA YRQEADEATLARLDLERKIESLEEEIRFL  
gi | 16265836 VRQKLQDGTNRLEAENNLAA YRQEADEATLARLDLERKIESLEEEIRFL  
gi | 38566198 VRQKLQDETNRLEAENNLAA YRQEADEATLARLDLERKIESLEEEIRFL  
gi | 28849921 LRQKLQDETNRLEAENNLAA YRQEADEATLARLDLERKIESLEEEIRFL  
gi | 417050 LRQKLQDETNRLEAENNLAA YRQEAHEATLARVDLERKVESLEEEIQFL  
:\*\*\*\*\* \*\* \*\*\*\*\*:\*\*\*\*\*:\*\*\*\*\*:\*\*\*\*\*:\*\*\*\*.\*

gi | 4503979 RKIHHEEVRELQEQLARQQVHVELDVAKPDLTAALKEIRTQYEAMASSNM  
gi | 16265836 RKIHHEEVRELQEQLARQQVHVELDVAKPDLTAALKEIRTQYEAMASSNM  
gi | 38566198 RKIHHEEVRELQEQLARQQVHVELDVAKPDLTAALKEIRTQYEAMASSNM  
gi | 28849921 RKIHHEEVRELQEQLAQQQVHVEMDVAKPDLTAALREIRTQYEAVASSNM  
gi | 417050 RKIYEEVVDLREQLAQQQVHVEMDVAKPDLTAALREIRTQYEAVATSNM  
\*\*\*:\*\*\*\*\*:\*.\*\*\*\*:\*\*\*\*\*:\*\*\*\*\*:\*\*\*\*\*:\*\*\*\*\*:\*.\*\*\*

gi | 4503979 HEAEWYRSKFADLTDAARNAELLRQAKHEANDYRRQLQSLTCDLESLR  
gi | 16265836 HEAEWYRSKFADLTDAARNAELLRQAKHEANDYRRQLQSLTCDLESLR  
gi | 38566198 HEAEWYRSKFADLTDAARNAELLRQAKHEANDYRRQLQSLTCDLESLR  
gi | 28849921 HEAEWYRSKFADLNDAAARRNAELVRQAKHEANDYRRQLQALTCDESLR  
gi | 417050 QETEEWYRSKFADLTDAASRNAELLRQAKHEANDYRRQLQALTCDESLR  
:\*.\*\*\*\*\*.\*\*\* \*\*\*\*\*:\*\*\*\*\*:\*\*\*\*\*:\*\*\*\*\*:\*\*\*\*\*

gi | 4503979 GTNESLERQMREQEERHVREAA SYQEALARLEEEGQSLKDEMARHLQEYQ  
gi | 16265836 GTNESLERQMREQEERHVREAA SYQEALARLEEEGQSLKDEMARHLQEYQ  
gi | 38566198 GTNESLERQMREQEERHVREAA SYQEALARLEEEGQSLKDEMARHLQEYQ  
gi | 28849921 GTNESLERQMREQEDAHAREAA SYQEALARLEEEGQSLKDEMARHLQEYQ  
gi | 417050 GTNESLERQMREQEERHARESASYQEALARLEEEGQSLKEEMARHLQEYH  
\*\*\*\*\*: \*.\*\* \*\*\*\*\*:\*\*\*\*\*:\*\*\*\*\*:\*\*\*\*\*:\*\*\*\*\*:

gi | 4503979 DLLNVKLALDIEIATYRKLLEGEENRITIPVQTFSNLQIRETSLDTKSVS  
gi | 16265836 DLLNVKLALDIEIATYRKLLEGEENRITIPVQTFSNLQIRETSLDTKSVS  
gi | 38566198 DLLNVKLALDIEIATYRKLLEGEENRITIPVQTFSNLQIRETSLDTKSVS  
gi | 28849921 DLLNVKLALDIEIATYRKLLEGEENRITIPVQTFSNLQIRETSLDTKSVS  
gi | 417050 DLLNVKLALDIEIATYRKLLEGEENRITIPVQTFSNLQIRETSLDTKSVS  
\*\*\*\*\*

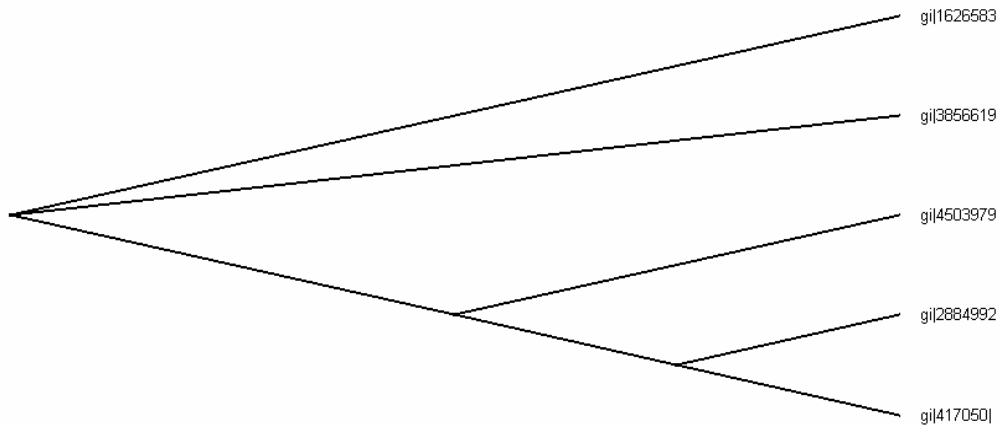
gi | 4503979 EGHKRNIVVKTVEMRDGEV IKESKQEHKDVM  
gi | 16265836 EGHKRNIVVKTVEMRDGEV IKESKQEHKDVM

```

gi|38566198      EGH LKR NIVVKT VEMRDGEV IKESKQE HKDVM
gi|28849921     EGH LKR NIVVKT VEMRDGEV IKESPQE HKDVM
gi|417050       EGH LKR NIVVKT VEMRDGEV IKD-----
                *****:

```

之后，开始建立进化树。主要用到 seqboot、protdist、neighbor 等命令，最终生成 treefile 文件，用 treeview 软件打开此文件，可以得到进化树，见下图：



由此进化树可以看出，人来源的蛋白 gi|16265836 和 gi|38566198 在进化上距离较近，而本文所分析的人来源蛋白 gi|4503979 则距离前两种蛋白较远，牛来源的蛋白 gi|28849921 和鼠来源的蛋白 gi|417050 在进化上距离较近，但距离人来源的三种蛋白较远。

### 结论：

通过以上分析，可以得到如下结论：胶质纤维酸性蛋白(GFAP)为一中等分子量的酸性蛋白，分子中含有较多的 Glu，每个分子大约带 9 个负电荷。该蛋白无跨膜区域，并且亲水，不是膜蛋白。GFAP 是一种结构蛋白，它主要参与中间纤维的构成，在神经元内环境的维持和血脑屏障中起着重要作用。到现在为止，该蛋白的空间结构尚未解析出来。该蛋白还有待于进一步的研究。