Protein sequence analysis

>sp|P01966|HBA_BOVIN

MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG

AKVAAALTKAVEHLDDLPGALSELSDLHAHKLRVDPVNFKLLSHSLLVTLASHLPSDFTP

AVHASLDKFLANVSTVLTSKYR

What can we do with sequence?

Amino acid occurrence

It is the number of amino acids of each type present in a protein.

E.g. THISISAPEPTIDE

A: 1; C: 0; D: 1; E: 2 etc.

```
  1   MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD
 81   NLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
15		2	7	8	8	13	9		11	18	2	6	7	3	3	5	7	18	2	3

Amino acid composition

It is the number of amino acids of each type normalized with the total number of residues.

It is defined as:

$$\text{Comp(i)} = \Sigma\, n_i * 100.0/N$$

where i stands for the 20 amino acid residues;

n_i is the number of residues of each type

N is the total number of residues.

The summation is through all the residues in the considered protein.

E.g. Comp(Ala) = 15*100/147 = 10.2

Algorithm

Algorithm

1. Read 20 amino acid residues: E.g. aa(i), i=1,20

2. Read the sequence: seq(i),i=1,n

3. Normalize number of residues: no(i)=0

4. Compare each residue in the sequence with standard 20 residues

```
do i=1,n
do j = 1,20
if (seq(i).eq.aa(j)) no(j) = no(j)+1
```

5. Count the number with respective residues for each match

6. Result will give the occurrence of each residue

7. Normalize with total number of residues

```
comp(i)=no(i)/n
```

amino.dat

ADCEFGHIKLMNPQRSTVWY

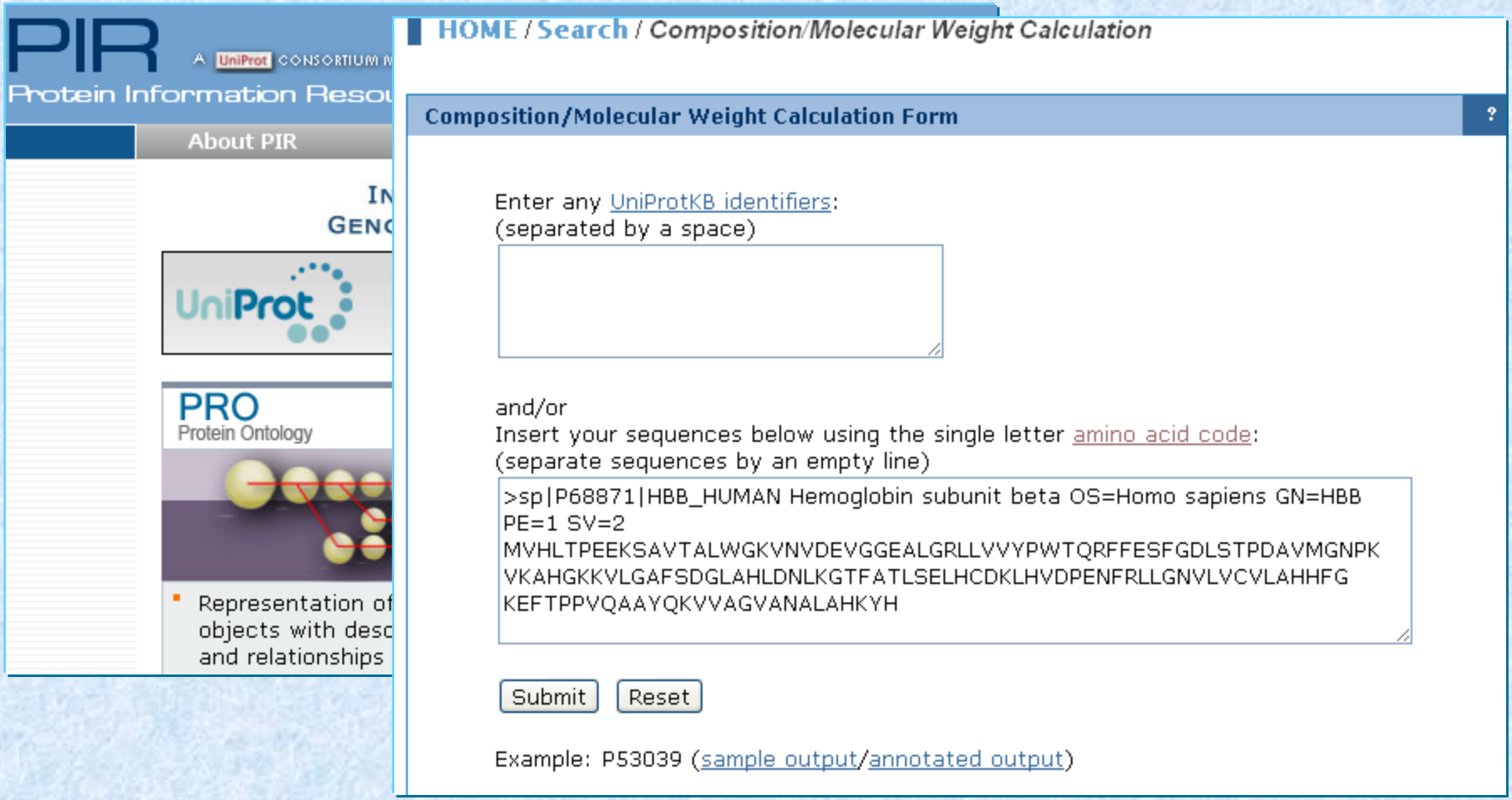
seq.dat

79

MALLPAAPGAPARATPTRWPVGCFNRPWTKWSYDEALDGIKAAGYAWTGLLTASKPSLHH
ATATPEYLAALKQKSRHAA

A	**17**	**21.52**
D	**2**	**2.53**
C	**1**	**1.27**
E	**2**	**2.53**
F	**1**	**1.27**
G	**5**	**6.33**
H	**3**	**3.80**
I	**1**	**1.27**
K	**5**	**6.33**
L	**8**	**10.13**
M	**1**	**1.27**
N	**1**	**1.27**
P	**8**	**10.13**
Q	**1**	**1.27**
R	**4**	**5.06**
S	**4**	**5.06**
T	**7**	**8.86**
V	**1**	**1.27**
W	**4**	**5.06**
Y	**3**	**3.80**

Composition



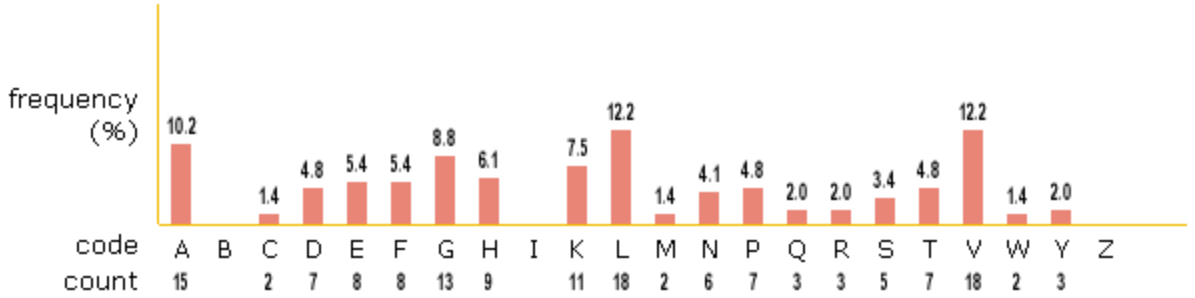
Composition

Molecular Weight & Composition

SEQUENCE:

```
      >SP|P68871|HBB_HUMAN HEMOGLOBIN SUBUNIT BETA OS=HOMO SAPIENS GN=HBB PE=1 SV=2
  1   MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD
 81   NLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

COMPOSITION:



Number of residues = 147 Molecular weight = 15998.34

Molecular weight

Molecular weight = Σ n(i)*w(i) – 18*(N-1)

Number of residues = 147 Molecular weight = 15998.34

CALCULATION NOTES:

1. Molecular weight = sum of individual residues weights - water molecular weight × (number of residues - 1)

where, water molecular weight = 18.015;

2. For each residue, the table gives the molecular weight:

A	Ala	89.09	G	Gly	75.07	N	Asn	132.12	V	Val	117.15
B	Asx	132.61	H	His	155.16	P	Pro	115.13	W	Trp	204.23
C	Cys	121.15	I	Ile	131.17	Q	Gln	146.15	Y	Tyr	181.19
D	Asp	133.10	K	Lys	146.19	R	Arg	174.20	Z	Glx	146.64
E	Glu	147.13	L	Leu	131.17	S	Ser	105.09			
F	Phe	165.19	M	Met	149.21	T	Thr	119.12	X	else	128.16

seq.dat

79

MALLPAAPGAPARATPTRWPVGCFNRPWTKWSYDEALDGIKAAGYAWTGLLTASKPSLHH
ATATPEYLAALKQKSRHAA

A	Ala	89.09	G	Gly	75.07	N	Asn	132.12	V	Val	117.15
B	Asx	132.61	H	His	155.16	P	Pro	115.13	W	Trp	204.23
C	Cys	121.15	I	Ile	131.17	Q	Gln	146.15	Y	Tyr	181.19
D	Asp	133.10	K	Lys	146.19	R	Arg	174.20	Z	Glx	146.64
E	Glu	147.13	L	Leu	131.17	S	Ser	105.09			
F	Phe	165.19	M	Met	149.21	T	Thr	119.12	X	else	128.16

A	**17**	**21.52**
D	**2**	**2.53**
C	**1**	**1.27**
E	**2**	**2.53**
F	**1**	**1.27**
G	**5**	**6.33**
H	**3**	**3.80**
I	**1**	**1.27**
K	**5**	**6.33**
L	**8**	**10.13**

M	**1**	**1.27**
N	**1**	**1.27**
P	**8**	**10.13**
Q	**1**	**1.27**
R	**4**	**5.06**
S	**4**	**5.06**
T	**7**	**8.86**
V	**1**	**1.27**
W	**4**	**5.06**
Y	**3**	**3.80**

79	8129

Why composition is important?

>tr|D9PL53|D9PL53
MALLPAAPGAPARATPTRWPVGCFNRPWTKWSYDEALDGIK
AAGYAWTGLLTASKPSLHHATATPEYLAALKQKSRHAA

>1a91_
AMENLNMDLLY MAAAVMMGLA AIGAAIGIGI LGGKFLEGAA RQPDLIPLLR
TQFFIVMGLV DAIPMIAVGL GLYVMFAVA

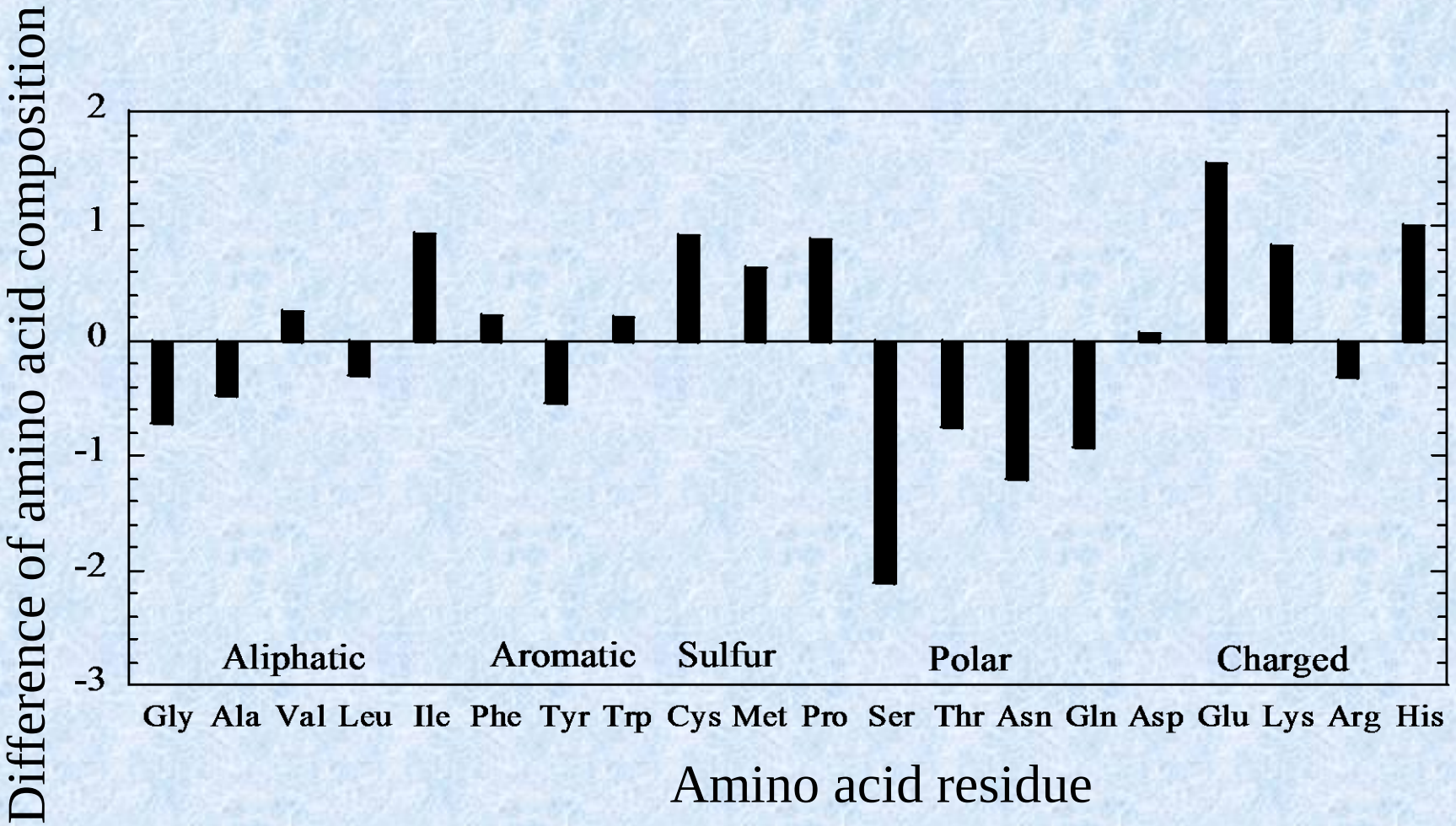
>2erv_
A ADVSAAVGAT GQSGMTYRLG LSWDWDKSWW QTSTGRLTGY WDAGYTYWEG
GDEGAGKHSL SFAPVFVYEF AGDSIKPFIE AGIGVAAFSG TRVGDQNLGS
SLNFEDRIGA GLKFANGQSV GVRAIHYSNA GLKQPNDGIE SYSLFYKIPI

Example

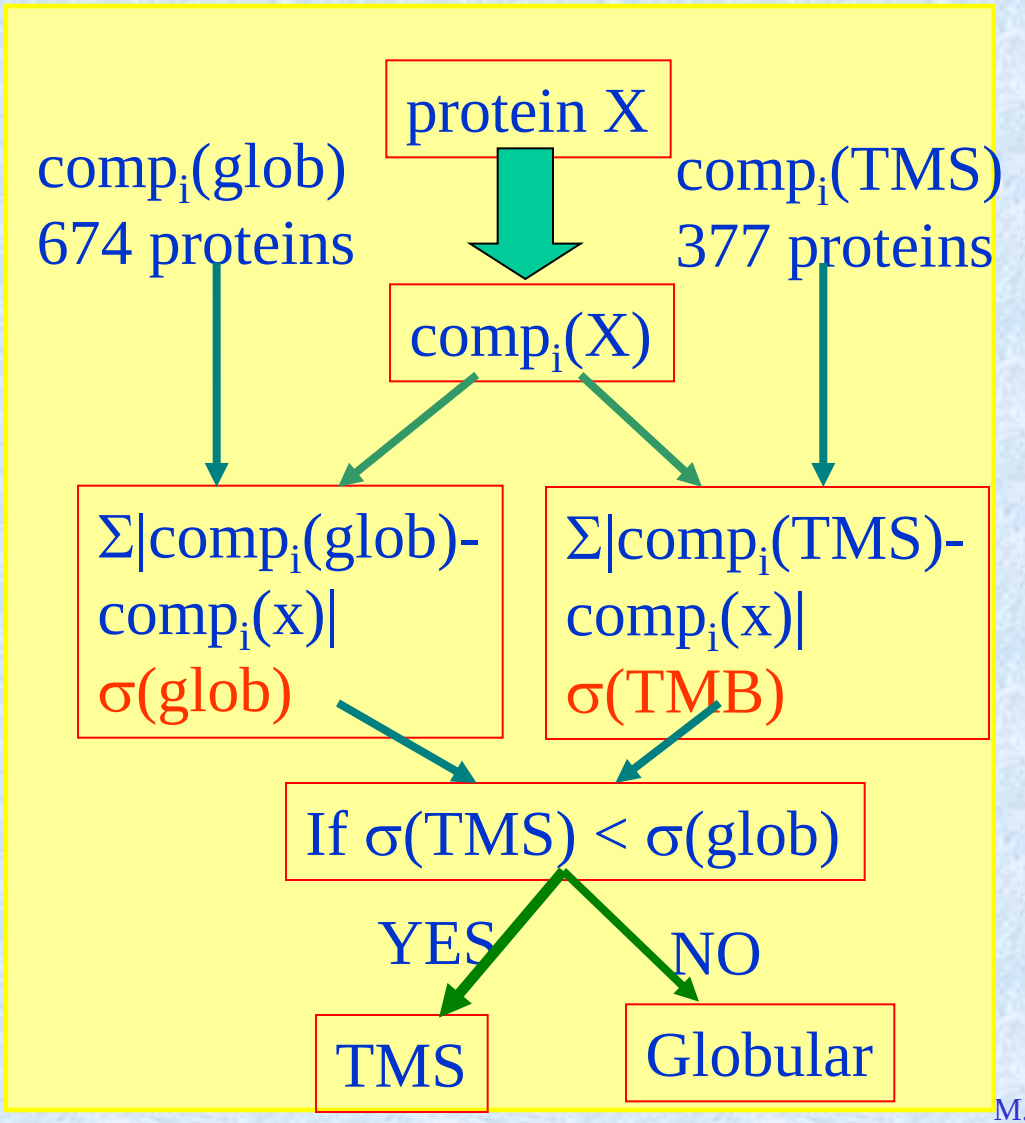
Residue	Composition (%)	
	Globular	OMP
Ala	8.47	8.95
Asp	5.97	5.91
Cys	**1.39**	**0.47**
Glu	**6.32**	**4.78**
Phe	3.91	3.68
Gly	7.82	8.54
His	**2.26**	**1.25**
Ile	**5.71**	**4.77**
Lys	5.76	4.93
Leu	8.48	8.78

Residue	Composition (%)	
	Globular	OMP
Met	2.21	1.56
Asn	**4.54**	**5.74**
Pro	4.63	3.74
Gln	**3.82**	**4.75**
Arg	4.93	5.24
Ser	**5.94**	**8.05**
Thr	5.79	6.54
Val	7.02	6.76
Trp	1.44	1.24
Tyr	3.58	4.13

Example


Difference of amino acid composition
2
1
0
-1
-2
-3
Aliphatic
Aromatic
Sulfur
Polar
Charged
Gly Ala Val Leu Ile Phe Tyr Trp Cys Met Pro Ser Thr Asn Gln Asp Glu Lys Arg His
Amino acid residue

Applications: Discrimination of beta barrel membrane proteins



Example

Table 2. Steps to discriminate globular and outer membrane proteins in two typical proteins

Residue	Adenovirus DNA-Binding Protein (1ADT)				OutD protein			
	N	Comp	σ_{glob}	σ_{OMP}	*N*	Comp	σ_{glob}	σ_{OMP}
Ala	10	11.11	2.64	2.16	54	8.31	0.16	0.64
Asp	4	4.44	1.53	1.47	40	6.15	0.18	0.24
Cys	1	1.11	0.28	0.64	1	0.15	1.24	0.32
Glu	7	7.78	1.46	3.00	31	4.77	1.55	0.01
Phe	5	5.56	1.65	1.88	21	3.23	0.68	0.45
Gly	3	3.33	4.49	5.21	46	7.08	0.74	1.46
His	4	4.44	2.18	3.19	3	0.46	1.80	0.79
Ile	1	1.11	4.60	3.66	35	5.38	0.33	0.61
Lys	7	7.78	2.02	2.85	28	4.31	1.45	0.62
Leu	10	11.11	2.63	2.33	53	8.15	0.33	0.63
Met	4	4.44	2.23	2.88	19	2.92	0.71	1.36
Asn	4	4.44	0.10	1.30	43	6.62	2.08	0.88
Pro	3	3.33	1.30	0.41	21	3.23	1.40	0.51
Gln	4	4.44	0.62	0.31	33	5.08	1.26	0.33
Arg	3	3.33	1.60	1.91	37	5.69	0.76	0.45
Ser	3	3.33	2.61	4.72	55	8.46	2.52	0.41
Thr	6	6.67	0.88	0.13	47	7.23	1.44	0.69
Val	6	6.67	0.35	0.09	64	9.85	2.83	3.09
Trp	2	2.22	0.78	0.98	6	0.92	0.52	0.32
Tyr	3	3.33	0.25	0.80	12	1.85	1.73	2.28
Total			34.18	39.89			23.70	16.09
Discrimination	*Globular protein*				*Outer membrane protein*			

N: number of residues; $\sigma_{glob} = |comp - comp(glob)|$; $\sigma_{OMP} = |comp - comp(OMP)|$.

M.M. Gromiha and M. Suwa (2005)
***Bioinformatics* 21, 961-968.**

http://psfs.cbrc.jp/tmbetadisc-comp/

TMBETA-DISC: Online servers

Discrimination of Beta-Barrel Membrane Proteins from Amino Acid Sequen

TMBETADISC-CO

Discrimination of Outer Membr

Amino Acid Compo

Discriminating outer membrane proteins from other folding types of globular and meml
outer membrane proteins (OMPs) from genomic sequences and for the successful pred

We have systematically analyzed the amino acid composition of globular proteins from
proteins. We found that the residues, Glu, His, Ile, Cys, Gln, Asn and Ser show a signif
proteins. Based on this information, we have devised a statistical method for discrimina
membrane proteins. Our approach correctly picked up the outer membrane proteins wi
proteins. On the other hand, our method has correctly excluded the globular proteins a
674 proteins.

TMBETADISC-COMP discriminates the beta-barrel membrane proteins using
membrane protein enter your sequence in a single letter code in the following bo

```
MAPKDNTWYTGAKLGWSQYHDTGLINNNGPTHENKLGAGAFGGYQVNPYVGFEMGYDWLG
RMPYKGSVENGAYKAQGVQLTAKLGYPITDDLDIYTRLGGMVWRADTYSNVYGKNHDTGV
SPVFAGGVEYAITPEIATRLEYQWTNNIGDAHTIGTRPDNGMLSLGVSYRFG
```

Predict Reset

Residue	Occurrence	Composition(%)	Globular Protein Diff.	OMP Diff.
A	12	6.98	1.49	1.97
D	10	5.81	0.16	0.10
C	0	0.00	1.39	0.47
E	6	3.49	2.83	1.29
F	4	2.33	1.58	1.35
G	26	15.12	7.30	6.58
H	4	2.33	0.07	1.08
I	7	4.07	1.64	0.70
K	7	4.07	1.69	0.86
L	11	6.40	2.08	2.38
M	5	2.91	0.70	1.35
N	12	6.98	2.44	1.24
P	8	4.65	0.02	0.91
Q	5	2.91	0.91	1.84
R	6	3.49	1.44	1.75
S	6	3.49	2.45	4.56
T	14	8.14	2.35	1.60
V	10	5.81	1.21	0.95
W	5	2.91	1.47	1.67
Y	14	8.14	4.56	4.01
Total	172	-	37.78	36.66

Amino acid sequence seems to be an Outer Membrane Protein

Residue pair preference (Dipeptide composition)

It is a measure to quantify the preference of amino acid residue pairs in a sequence.

$$Dipep(i,j) = \Sigma N_{ij} * 100 / (\Sigma N_i + \Sigma N_j)$$

where i,j stands for the distribution of 20 amino acid residues at positions i and i+1.

$N_{i,j}$ is the number of residues of type i followed by the residue j.

ΣN_i and ΣN_j are the total number of residues of type i and j, respectively.

E.g. DIPEPTIDESAREPEPTIDEPAIRS

Dipep(PE) = 2*100/(5+5) = 20

>tr|D9PL53|D9PL53
MALLPAAPGAPARATPTRWPVGCFNRPWTKWSYDEALDGIK
AAGYAWTGLLTASKPSLHHATATPEYLAALKQKSRHAA

AA	4	11.76
AD	0	.00
AC	0	.00
AE	0	.00
AF	0	.00
AG	1	4.55
AH	0	.00
AI	0	.00
AK	0	.00
AL	3	12.00