Idursulfasebeta

Taeyoon Kim

May 12, 2022

CONTENTS

The data in the report is intended to be a resource to guide the development and lead-optimization of a clinical antibody. These data can be used in a preemptive fashion - for example, in the decision to substitute an exposed residue on the antibody surface that may be prone to the kind of chemical modification that might affect the stability of the antibody. They can also be used to assist the troubleshooting of problems that can arise in the course of an antibody's clinical development - for example if an antibody displays stability issues in storage, or unacceptably high levels of immunogenicity in early clinical trials. There is always a great deal of risk involved in the development of any therapeutic molecule but experience has shown that the kind of data presented in this report is an invaluable tool for mitigating that risk - either by helping to identify potential problems before they occur, or by guiding the troubleshooting ofproblems that can occur during the antibody's development and lead-optimization.

Introduction

This report contains a structure and sequence-based analysis of Aflibercept.

> Aflibercept, sold under the brand names Eylea and Zaltrap, is a medication used to treat wet macular degeneration and metastatic colorectal cancer. It was developed by Regeneron Pharmaceuticals and is approved in the United States and the European Union. –wiki

Aflibercept is a recombinant fusion protein consisting of vascular endothelial growth factor (VEGF)-binding portions from the extracellular domains of human VEGF receptors 1 and 2, that are fused to the Fc portion of the human IgG1 immunoglobulin.

History

Regeneron commenced clinical testing of aflibercept in cancer in 2001. In 2003, Regeneron signed a major deal with Aventis to develop aflibercept in the field of cancer. In 2004 Regeneron started testing the compound, locally delivered, in proliferative eye diseases, and in 2006 Regeneron and Bayer signed an agreement to develop the eye indications.

Table of contents

CHAPTER

ONE

PROTEIN PARAMETERS ANALYSIS

The program performs most of the same functions as the Expasy ProtParam tool. Protein parameters for analysing protein sequences.

The program calculates:

1. Molecular weight
2. Chemicalcomposition
3. Amino acid composition
4. pI
5. Extinction coefficient
6. Aromaticity
7. GRAVY and Instability_index

1.1 Input sequences

```
SDTGRPFVEM YSEIPEIIHM TEGRELVIPC RVTSPNITVT LKKFPLDTLI PDGKRIIWDS
RKGFIISNAT YKEIGLLTCE ATVNGHLYKT NYLTHRQTNT IIDVVLSPSH GIELSVGEKL
VLNCTARTEL NVGIDFNWEY PSSKHQHKKL VNRDLKTQSG SEMKKFLSTL TIDGVTRSDQ
GLYTCAASSG LMTKKNSTFV RVHEKDKTHT CPPCPAPELL GGPSVFLFPP KPKDTLMISR
TPEVTCVVVD VSHEDPEVKF NWYVDGVEVH NAKTKPREEQ YNSTYRVVSV LTVLHQDWLN
GKEYKCKVSN KALPAPIEKT ISKAKGQPRE PQVYTLPPSR DELTKNQVSL TCLVKGFYPS
DIAVEWESNG QPENNYKTTP PVLDSDGSFF LYSKLTVDKS RWQQGNVFSC SVMHEALHNH
YTQKSLSLSP G
```

(Known Disulfide bridge: 30-79, 124-185, 246-306, 352-410, 211-211', 214-214')

1.2 Molecular weight

Amino acids are the building blocks that form polypeptides and ultimately proteins. Calculates the molecular weight of a protein.

```
Molecular weigh of Idursulfasebeta is 59,297.10 Da.
Total number of amino acid in Idursulfasebeta is 525.
```

1.3 Chemical composition

A chemical formula is a way of presenting information about the chemical proportions of atoms that constitute a particular chemical compound or molecule, using chemical element symbols, numbers, and sometimes also other symbols, such as parentheses, dashes, brackets, commas and plus (+) and minus (−) signs. These are limited to a single typographic line of symbols, which may include subscripts and superscripts.

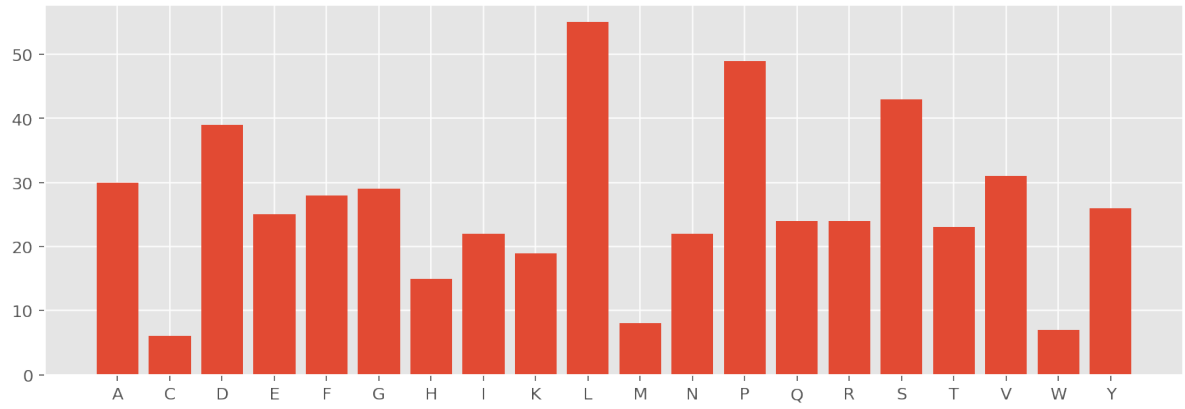
```
Chemical formula of Idursulfasebeta is C2689H4057N699O792S14.
Total number of atom in Idursulfasebeta is 8251.
```

1.4 Amino acid composition

We can easily count the number of each type of amino acid.

1.4.1 Number of each amino acids

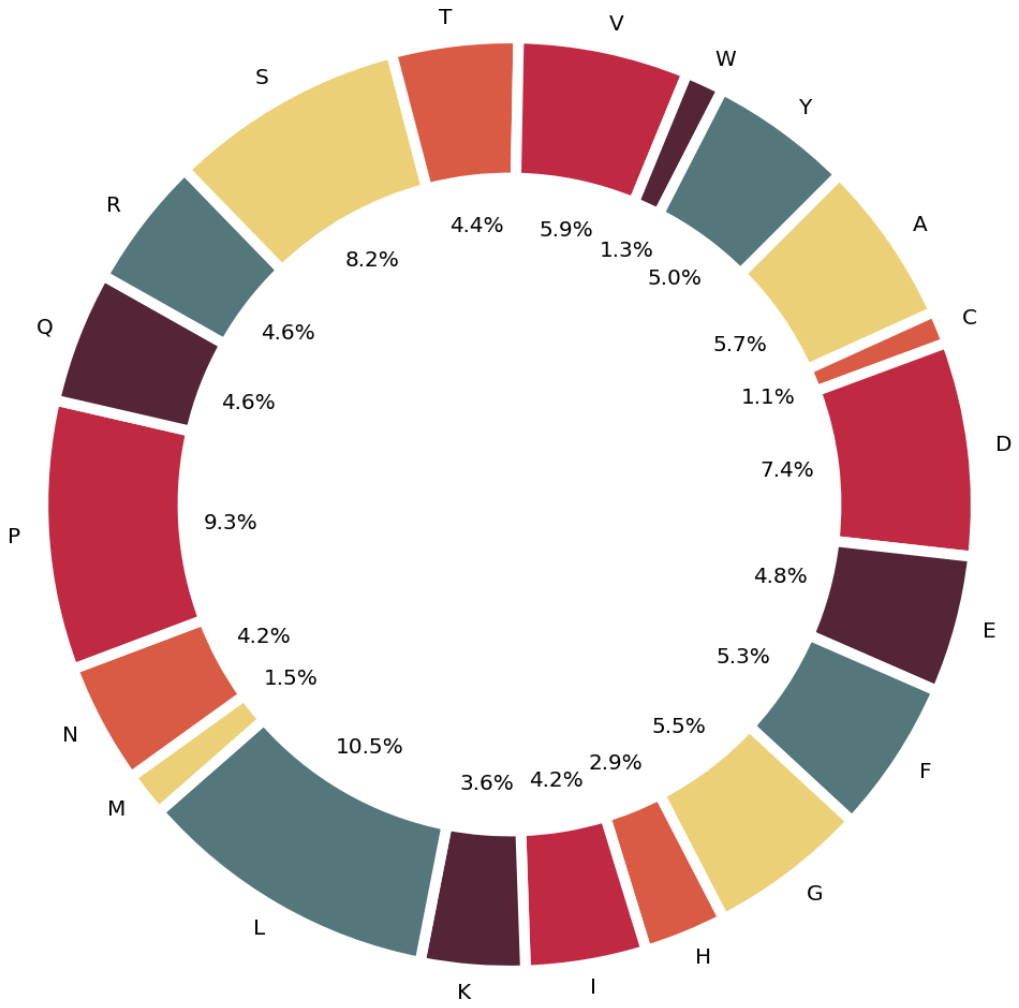
Simply counts the number times an amino acid is repeated in the protein sequence.




```
Total number of positively charged residues (Arg + Lys) of Idursulfasebeta is 43.
Total number of negatively charged residues (Asp + Glu) of Idursulfasebeta is 64.
```

1.4.2 Percent of amino acid contents

The same as number of amino acid, only returns the number in percentage of entire sequence. When the total is taken as 100, the ratio of each amino acid is calculated.



1.5 Theoretical pI

The isoelectric point (pI, pH(I), IEP), is the pH at which a molecule carries no net electrical charge or is electrically neutral in the statistical mean.

The pI value can affect the solubility of a molecule at a given pH. Such molecules have minimum solubility in water or salt solutions at the pH that corresponds to their pI and often precipitate out of solution. Biological amphoteric molecules such as proteins contain both acidic and basic functional groups.

```
Theoretical pI is 5.102.
```

1.5.1 Charge of target protein ins PBS

Phosphate-buffered saline (abbreviated PBS) is a buffer solution (pH ~ 7.4) commonly used in biological research.

Note: PBS has many uses because it is isotonic and non-toxic to most cells

```
Idursulfasebeta is negative charged(-21.412) in PBS
```

1.6 Extinction coefficient

Extinction (or extinction coefficient) is defined as the ratio of maximum to minimum transmission of a beam of light that passes through a polarization optical train.

extinction coefficient in units of M-1 cm-1, at 280 nm measured in water.

1.6.1 with reduced cysteines

```
Extinction coefficient of Idursulfasebeta at reduced condition is 77240.
Abs 0.1%(=1 g/L) is 1.303.
```

1.6.2 with non-reduced cysteines

```
Extinction coefficient of Idursulfasebeta at non-reduced condition is 77615.
Abs 0.1%(=1 g/L) is 1.309.
```

1.7 Aromaticity

Calculate the aromaticity according to Lobry, 1994. Calculates the aromaticity value of a protein according to Lobry, 1994. It is simply the relative frequency of Phe+Trp+Tyr.

```
The Aromaticity of target protein is 11.619%.
```

1.8 GRAVY

Protein GRAVY returns the GRAVY (grand average of hydropathy) value for the protein sequences you enter. The GRAVY value is calculated by adding the hydropathy value for each residue and dividing by the length of the sequence (Kyte and Doolittle; 1982).

A higher value is more hydrophobic. A lower value is more hydrophilic.

```
Idursulfasebeta is kind a more hydrophilic protein. The GRAVY value is -0.371.
```

1.9 Instability_index

Implementation of the method of Guruprasad et al. (1990, Protein Engineering, 4, 155-161). This method tests a protein for stability. Any value above 40 means the protein is unstable (=has a short half life).

```
The instability index of Idursulfasebeta is computed to be 39.375, and it seems␣
↪stable.
```

CHAPTER

TWO

POTENTIAL SITES OF CHEMICAL MODIFICATION

An initial scan of the antibody sequences is presented based purely upon sequence. If a structural analysis was also requested, this section should be used in conjunction with the molecular surface analysis described in a subsequent section. Any of the sites listed below could be candidates for further consideration if the molecular surface analysis shows that they are significantly exposed on the surface of the antibody, increasing their propensity for chemical modification. The canonical sequence analysis is also helpful here, since each of these sites can also be considered in the context of their frequency of occurrence within the canonical library of homologous sequences.

This is an example of a footnote.[1]

2.1 Potential deamidation positions

Asparagine (N) and glutamine (Q) residues are particularly prone to deamidation when they are followed in the sequence by amino acids with smaller side chains, that leave the intervening peptide group more exposed. Deamidation proceeds much more quickly if the susceptible amino acid is followed by a small, flexible residue such as glycine whose low steric hindrance leaves the peptide group open for attack.

2.1.1 Search patterns: ASN/GLN-ALA/GLY/SER/THR

```
4-QA-5
6-NS-7
51-NA-52
56-QA-57
81-NS-82
101-NG-102
142-NT-143
175-QS-176
179-QA-180
252-QA-253
273-QS-274
300-NS-301
371-QS-372
451-NS-452
515-QG-516
```

[1] The definition for referencing footnotes is generally placed at the bottom of the document.

2.2 Potential o-linked glycosylation sites

The O-linked glycosylation of serine and threonine residues seems to be particularly sensitive to the presence of one or more proline residues in their vicinity in the sequence, particularly in the2 -1 and +3 positions.

2.2.1 Search patterns: PRO-SER/THR

```
24-PS-25
61-PS-62
135-PS-136
383-PT-384
400-PS-401
444-PS-445
455-PS-456
```

2.2.2 Search patterns: SER/THR-X-X-PRO

```
92-STIP-95
129-SFPP-132
190-SASP-193
223-TLAP-226
330-THVP-333
380-SLFP-383
439-SQYP-442
445-SDIP-448
452-SDKP-455
```

2.3 Potential n-linked glycosylation sites

2.3.1 Search patterns: ASN-X-SER/THR

```
6-NST-8
90-NFS-92
119-NHT-121
221-NIT-223
255-NIS-257
300-NST-302
488-NFS-490
512-NDS-514
```

CHAPTER

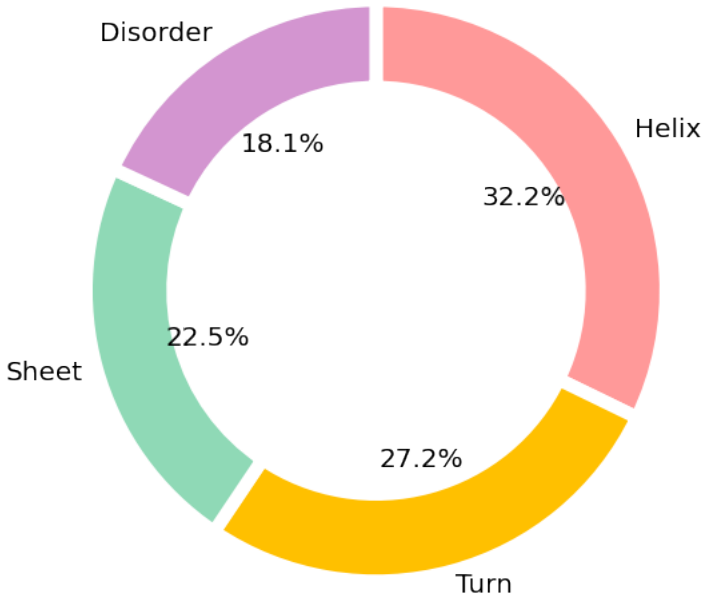
THREE

SECONDARY STRUCTURE FRACTION

A very useful method — .secondary_structure_fraction() — returns the fraction of amino acids that tend to be found in the three classical secondary structures. These are beta sheets, alpha helixes, and turns (where the residues change direction).

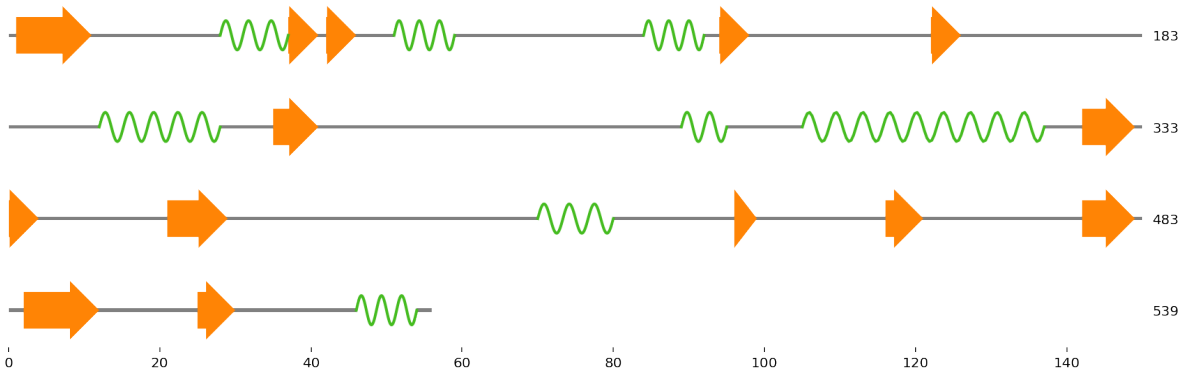
Secondary_structure_fraction: This methods returns a list of the fraction of amino acids which tend to be in helix, turn or sheet.

- Amino acids in helix: V, I, Y, F, W, L.
- Amino acids in turn: N, P, G, S.
- Amino acids in sheet: E, M, A, L.



3.1 Secondary structure prediction

Protein secondary structure prediction is one of the most important and challenging problems in bioinformatics. Here in, the P-SEA algorithm that to predict the secondary structures of proteins sequences based only on knowledge of their primary structure.



CHAPTER

FOUR

PROTEIN SCALES

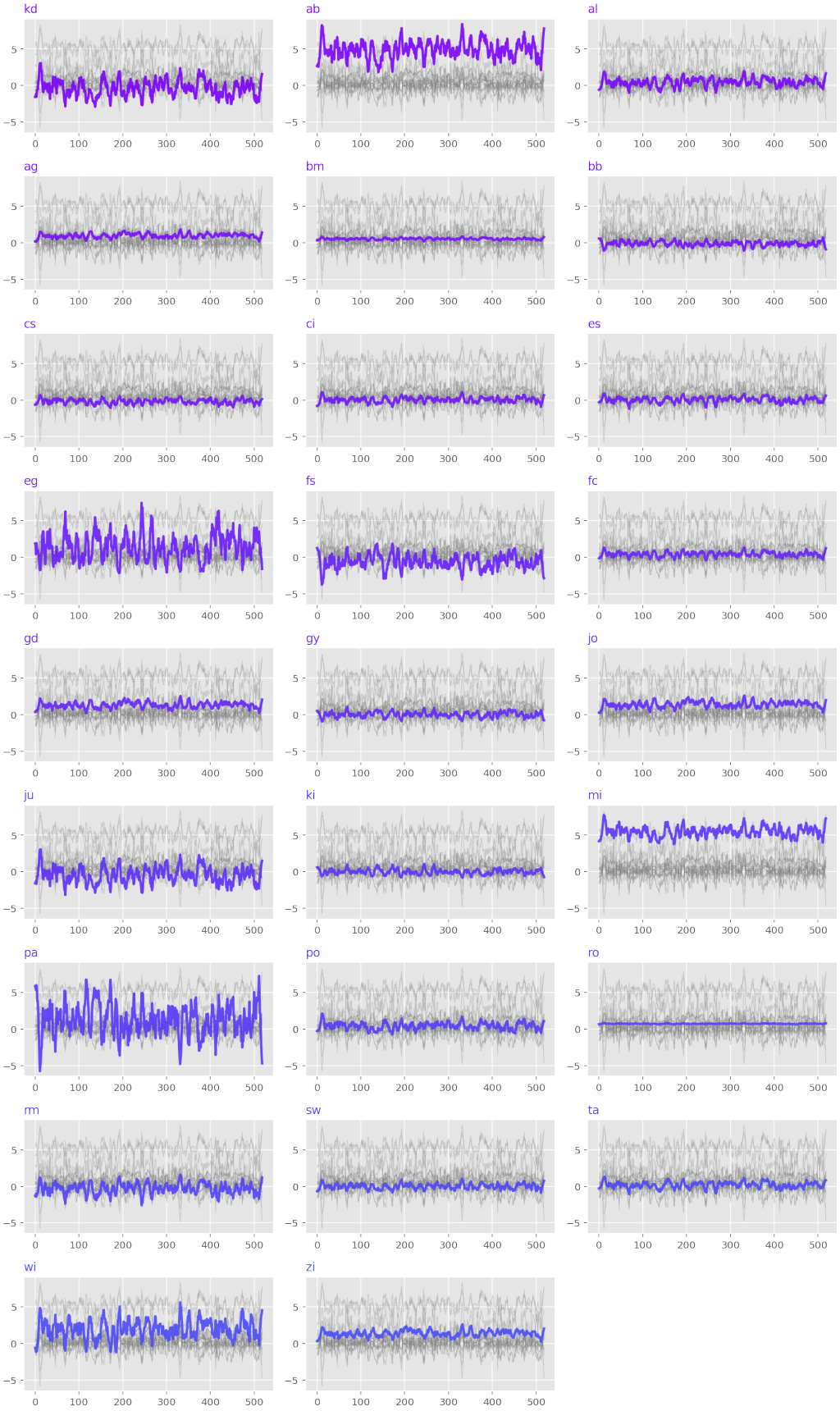
Protein scales are a way of measuring certain attributes of residues over the length of the peptide sequence using a sliding window. Scales are comprised of values for each amino acid based on different physical and chemical properties, such as hydrophobicity, secondary structure tendencies, and surface accessibility. As opposed to some chain-level measures like overall molecule behavior, scales allow a more granular understanding of how smaller sections of the sequence will behave.

Some common scales include:

```
- kd → Kyte & Doolittle Index of Hydrophobicity
- Flex → Normalized average flexibility parameters (B-values)
- hw → Hopp & Wood Index of Hydrophilicity
- em → Emini Surface fractional probability (Surface Accessibility)
```

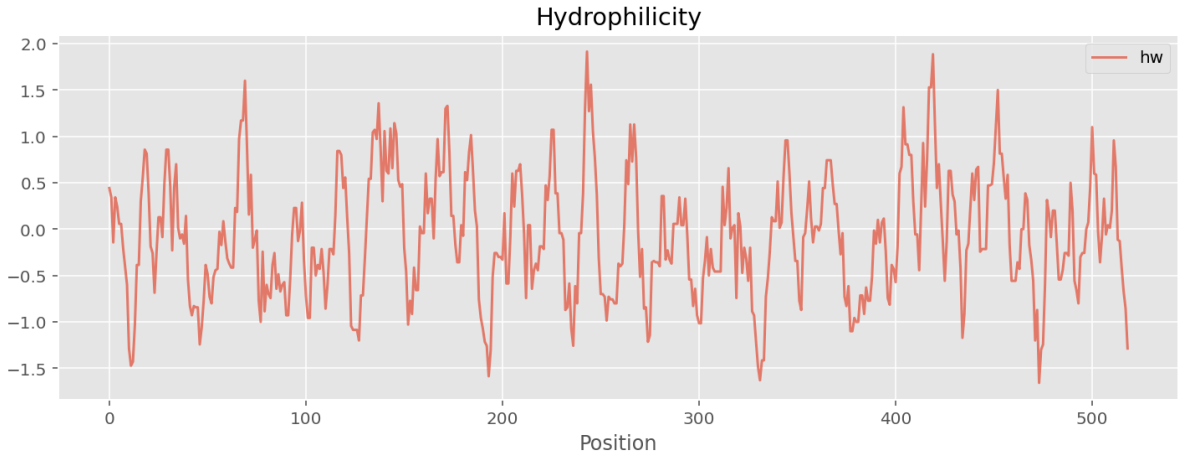
4.1 Hydrophobicity index

hydrophobicity is the physical property of a molecule that is seemingly repelled from a mass of water (known as a hydrophobe).


kd
ab
al
ag
bm
bb
cs
ci
es
eg
fs
fc
gd
gy
jo
ju
ki
mi
pa
po
ro
rm
sw
ta
wi
zi
5
0
−5
0
100
200
300
400
500

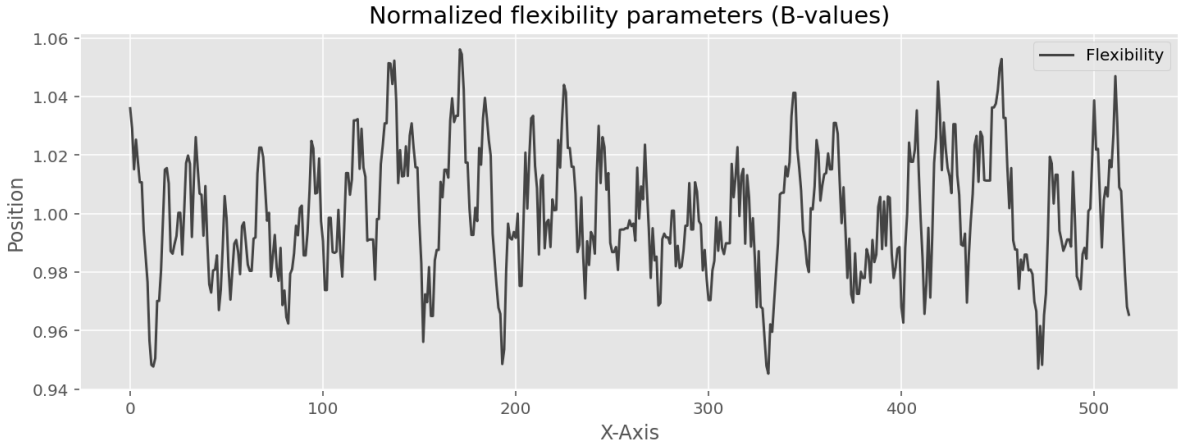
4.2 Hydrophilicity index

Hydrophilicity is the tendency of a molecule to be solvated by water.



4.3 Flexibility index

Proteins are dynamic entities, and they possess an inherent flexibility that allows them to function through molecular interactions within the cell.

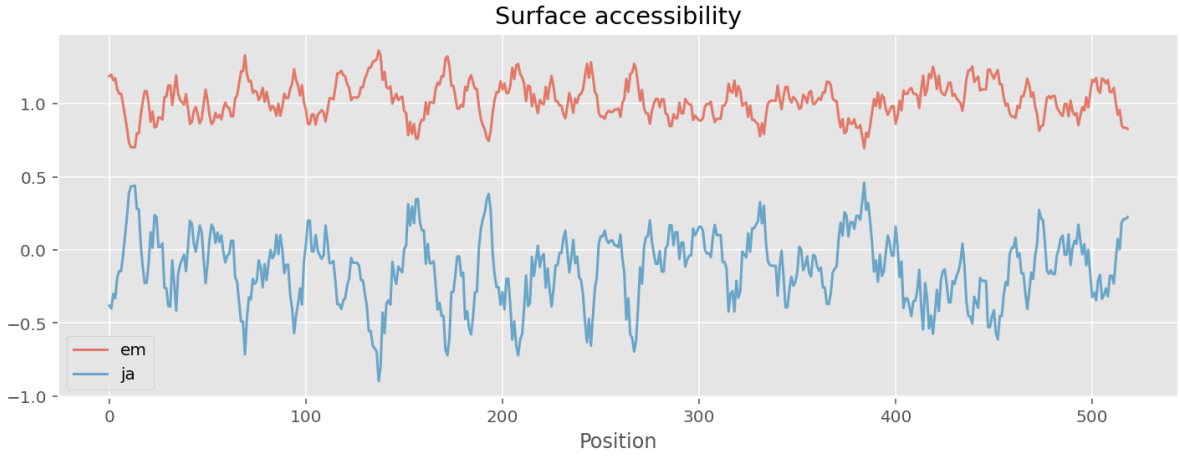


4.4 Surface accessibility

Data describing the solvent-accessible surface of a molecule is of great utility in the development of that molecule as a therapeutic, particularly in the case of antibodies. In the context of this report, the most obvious application of molecular surface data is in combination with the potential sites of chemical modification, described in the previous section. Antibodies (and other proteins) are known to undergo many different chemical modifications as a result of interactions with their aqueous environment. The probability and kinetic rate of such a modification is greatly enhanced by the degree of exposure of the potential modification site to the solvent environment. If a 3D structure was supplied or requested by the client, then the graphs below will show the solvent-accessible surface areas in (Å2) calculated for each residue in the light and heavy chain variable regions of the antibody. These solvent-accessible surfaces are calculated using the “rolling

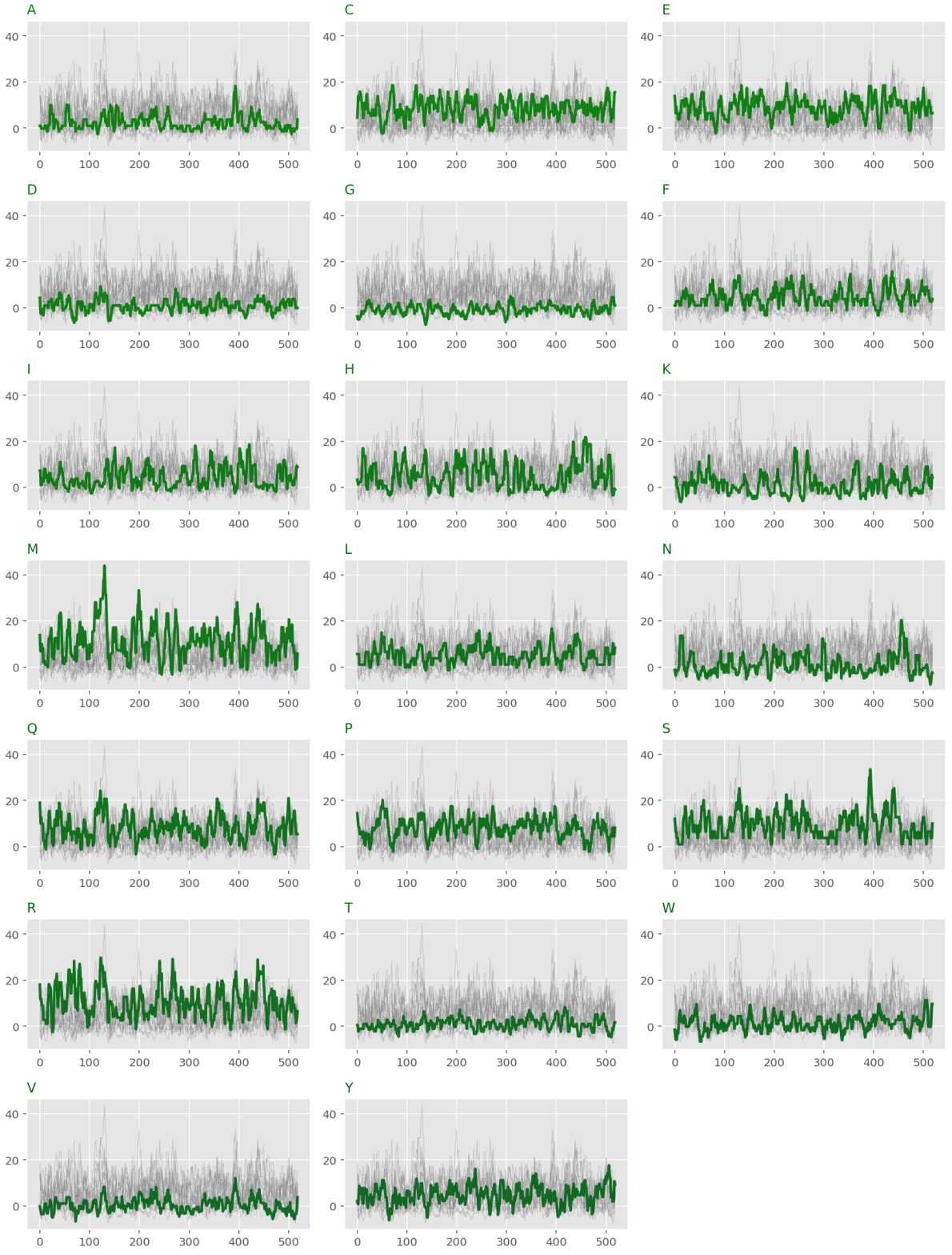
ball" algorithm developed by Shrake & Rupley in which a spherical solvent "probe" is3 used, the accessible surface being generated by tracing the center of the probe sphere as it rolls along the van der Waals surface (as shown below).

The solvent-accessible surface for each residue depends upon the degree of exposure of the residue on the antibody surface, but also on the size of the residue side chain. Residues that are completely buried within the hydrophobic core of the antibody, will have solvent-accessible surface areas at or close to zero. Residues that are highly exposed at the antibody surface will have values that range from around 100 - 200 Å2, depending upon the size of the residue side chain and the degree of exposure.



4.5 Instability index

The instability index provides an estimate of the stability of your protein in a test tube. Statistical analysis of 12 unstable and 32 stable proteins has revealed that there are certain dipeptides, the occurence of which is significantly different in the unstable proteins compared with those in the stable ones.


A
C
E
D
G
F
I
H
K
M
L
N
Q
P
S
R
T
W
V
Y
0
20
40
0
100
200
300
400
500

CHAPTER

FIVE

STRUCTURAL ANALYSIS

5.1 Detection of disulfide bonds

This function detects disulfide bridges in protein structures. Then the detected disulfide bonds are visualized and added to the bonds attribute of the AtomArray.,

The employed criteria for disulfide bonds are quite simple in this case: the atoms of two cystein residues must be in a vicinity of Å and the dihedral angle of must be .

```
    A     171  CYS SG     S        16.601   32.190  143.334
    A     184  CYS SG     S        18.540   32.510  142.815
    A     422  CYS SG     S        26.896   13.750  106.306
    A     432  CYS SG     S        27.418   11.787  106.401
    A     422  CYS SG     S        26.896   13.750  106.306
    A     432  CYS SG     S        27.418   11.787  106.401
```

The found disulfide bonds are visualized with the help of Matplotlib: The amino acid sequence is written on the X-axis and the disulfide bonds are depicted by yellow semi-ellipses.

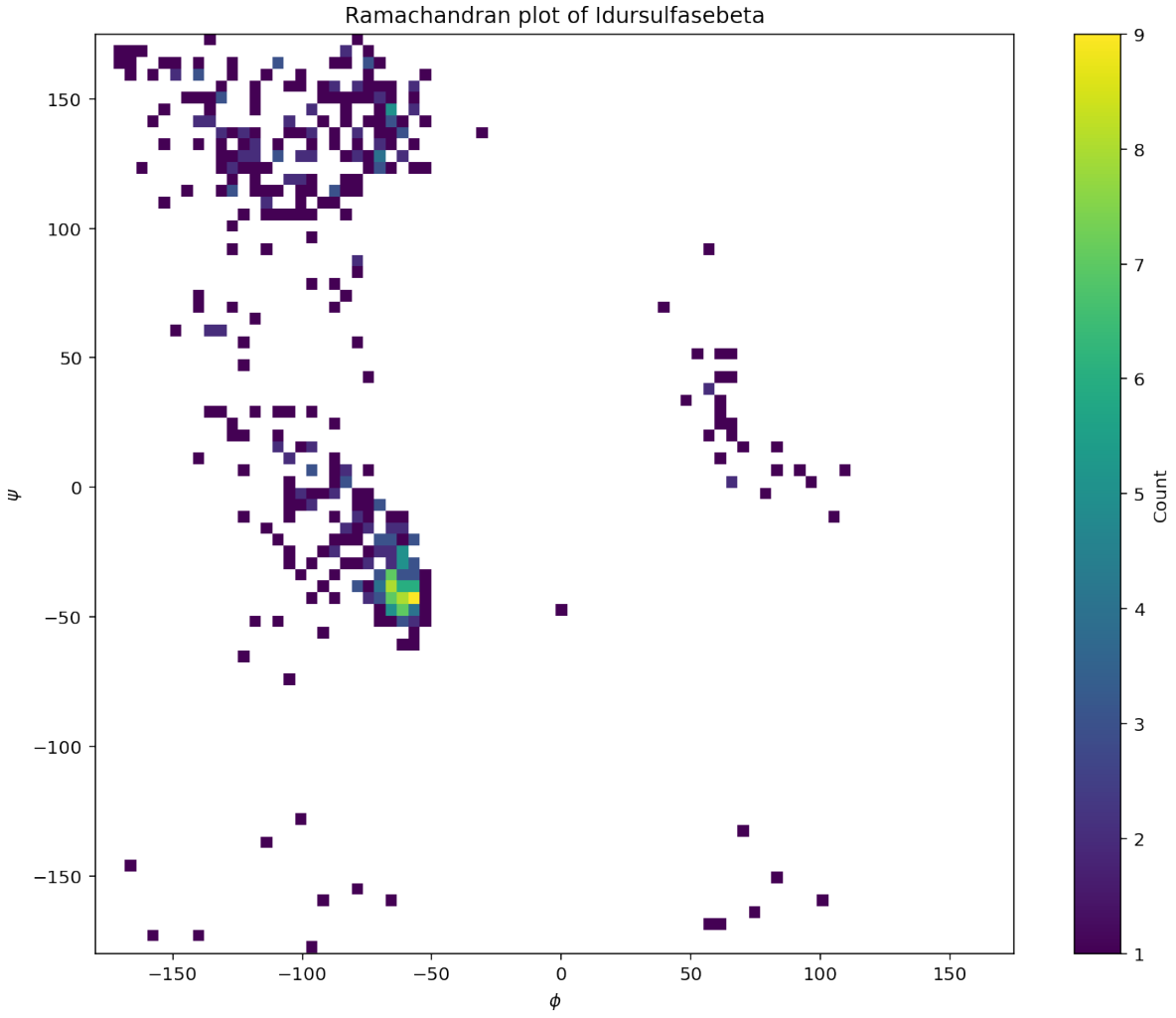
5.2 Calculation of protein diameter

This calculates the diameter of a protein defined as the maximum pairwise atom distance.

```
Diameter of Idursulfasebeta is: 71.274 Angstrom.
```

5.3 Ramachandran plot

This creates a Ramachandran plot of the motor domain of target protein.


Ramachandran plot of Idursulfasebeta
150
100
50
0
−50
−100
−150
ψ
−150
−100
−50
0
50
100
150
ϕ
9
8
7
6
5
4
3
2
1
Count

CHAPTER

SIX

IMMUNOGENICITY ANALYSIS

We uses the method of removing and/or reducing potential T-cell epitopes, as an approach to the management of the immunogenicity of biologics. The protein sequence is scanned in silico, for sequences that have a strong binding signature for a family of 50 MHC Class II receptors , whose alleles cover 96 – 98% of the human population. The presented histograms for each variable region sequence, show the average (for the n positively-testing MHC II alleles) of epitope strength at each position as a percentage for all epitopes above a threshold of 20%. At each position in the sequence, the number of alleles scoring above the threshold is shown above the histogram at that position. The epitopes of most concern for the antibody's immunogenicity are therefore those that have not just the highest average score per allele (as shown by the histogram), but which also score above the threshold across more alleles, since these epitopes are more likely to engender an immune response in a larger fraction of the patient population.

Experience using in silico algorithms of this kind in conjunction with laboratory immunogenicity assays has shown that epitopes below this threshold do not generally contribute significantly to the protein's immunogenicity. The number of alleles, the affected alleles and their individual scores are also listed in the detailed analyses below each histogram figure.

The raw immunogenicity score quoted is the total over all epitopes above the threshold for all affected alleles. The normalized immunogenicity score is this raw score divided by the sequence length, and represents epitope strength per unit sequence to enable comparisons of protein sequences of different lengths.

The absolute magnitudes of these scores are somewhat arbitrary, but they have value as comparative metrics. It has been shown that human serum proteins generally display an immunogenicity potential that is inversely proportional to their abundance in serum . Proteins that are found at very low concentrations in serum, like erythropoietin, can have normalized scores above 80%. By contrast, very abundant human serum proteins like albumin and immunoglobulins typically have normalized scores in the 35 - 50% range.

6.1 MHC class 1

Class I major histocompatibility complex (MHC) molecules bind, and present to T cells, short peptides derived from intracellular processing of proteins. The peptide repertoire of a specific molecule is to a large extent determined by the molecular structure accommodating so-called main anchor positions of the presented peptide.

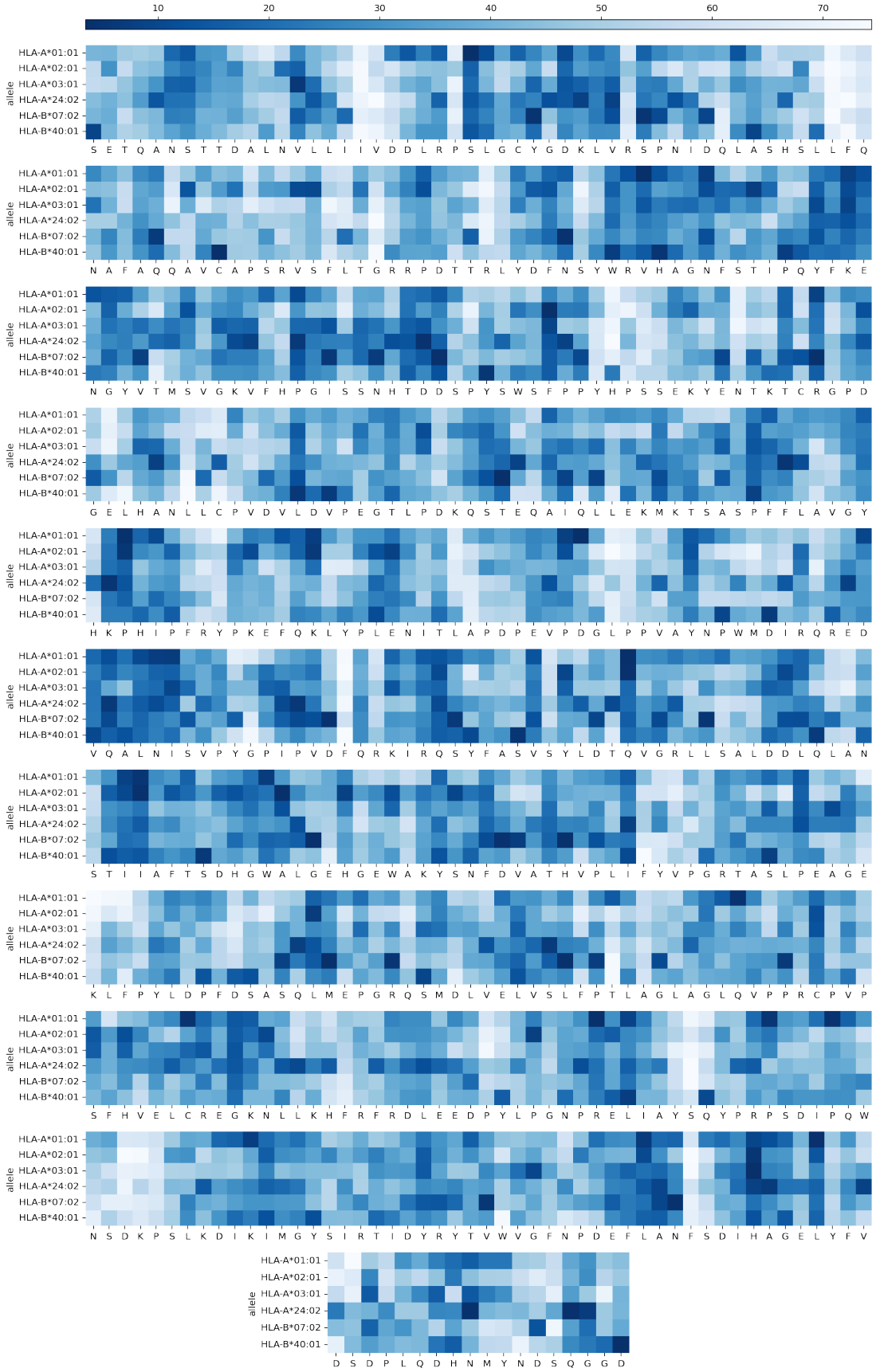
MHC class I molecules are one of two primary classes of major histocompatibility complex (MHC) molecules (the other being MHC class II) and are found on the cell surface of all nucleated cells in the bodies of vertebrates.

Their function is to display peptide fragments of proteins from within the cell to cytotoxic T cells; this will trigger an immediate response from the immune system against a particular non-self antigen displayed with the help of an MHC class I protein. Because MHC class I molecules present peptides derived from cytosolic proteins, the pathway of MHC class I presentation is often called cytosolic or endogenous pathway.[1]

In humans, the HLAs corresponding to MHC class I are HLA-A, HLA-B, and HLA-C. Almost a decade ago we introduced this concept of clustering human leukocyte antigen (HLA) alleles and defined nine different groups, denominated as supertypes, on the basis of their main anchor specificity.

[1] Kimball's Biology Pages, Histocompatibility Molecules

- HLA-A01:01
- HLA-A02:01
- HLA-A03:01
- HLA-A24:02
- HLA-B07:02
- HLA-B40:01


10
20
30
40
50
60
70
allele
HLA-A*01:01
HLA-A*02:01
HLA-A*03:01
HLA-A*24:02
HLA-B*07:02
HLA-B*40:01
S E T Q A N S T T D A L N V L L I I V D D L R P S L G C Y G D K L V R S P N I D Q L A S H S L L F Q
N A F A Q Q A V C A P S R V S F L T G R R P D T T R L Y D F N S Y W R V H A G N F S T I P Q Y F K E
N G Y V T M S V G K V F H P G I S S N H T D D S P Y S W S F P P Y H P S S E K Y E N T K T C R G P D
G E L H A N L L C P V D V L D V P E G T L P D K Q S T E Q A I Q L L E K M K T S A S P F F L A V G Y
H K P H I P F R Y P K E F Q K L Y P L E N I T L A P D P E V P D G L P P V A Y N P W M D I R Q R E D
V Q A L N I S V P Y G P I P V D F Q R K I R Q S Y F A S V S Y L D T Q V G R L L S A L D D L Q L A N
S T I I A F T S D H G W A L G E H G E W A K Y S N F D V A T H V P L I F Y V P G R T A S L P E A G E
K L F P Y L D P F D S A S Q L M E P G R Q S M D L V E L V S L F P T L A G L A G L Q V P P R C P V P
S F H V E L C R E G K N L L K H F R F R D L E E D P Y L P G N P R E L I A Y S Q Y P R P S D I P Q W
N S D K P S L K D I K I M G Y S I R T I D Y R Y T V W V G F N P D E F L A N F S D I H A G E L Y F V
D S D P L Q D H N M Y N D S Q G G D

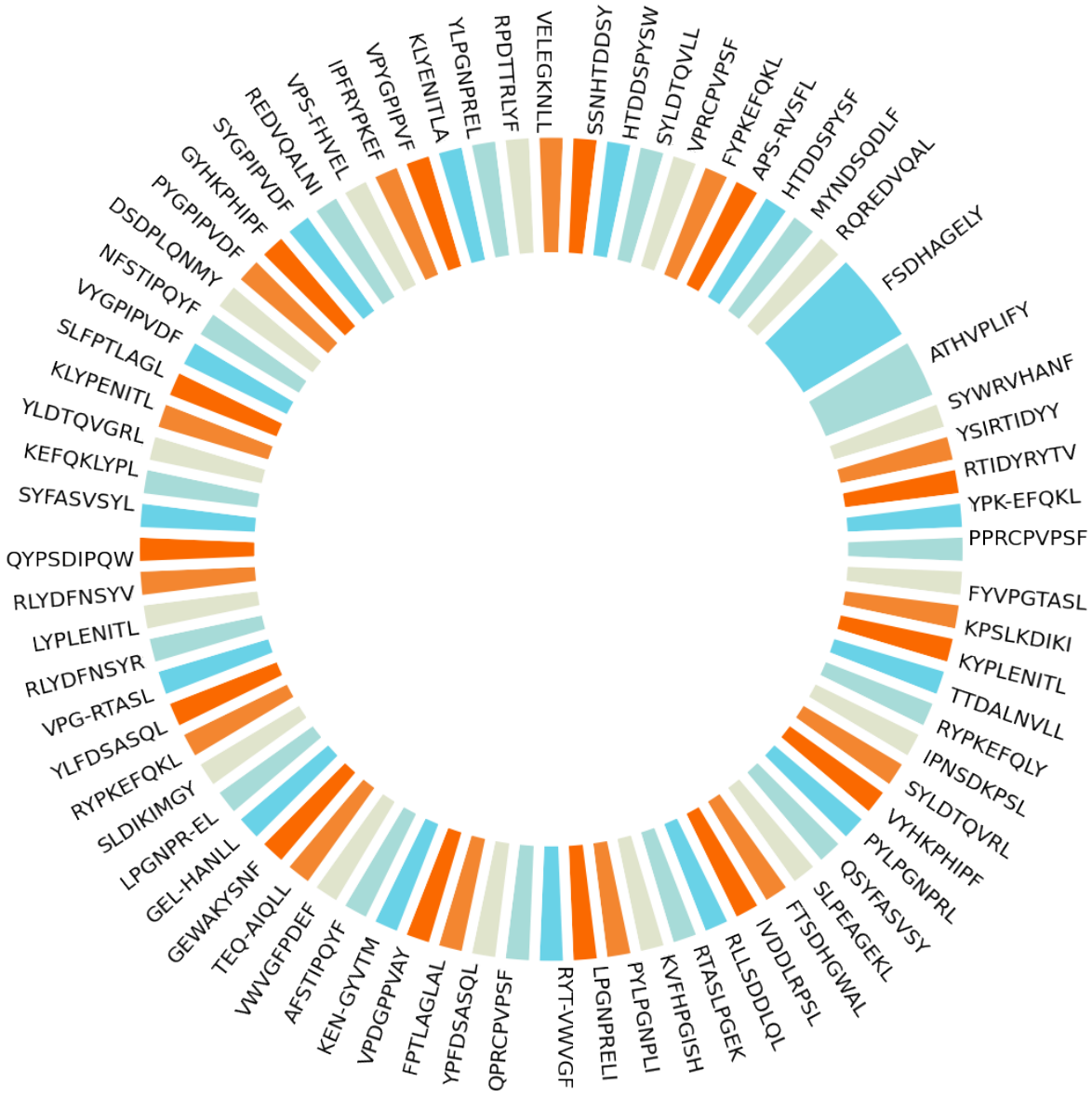
6.1.1 MHC class1 binding peptide

Top10 strong binding peptide

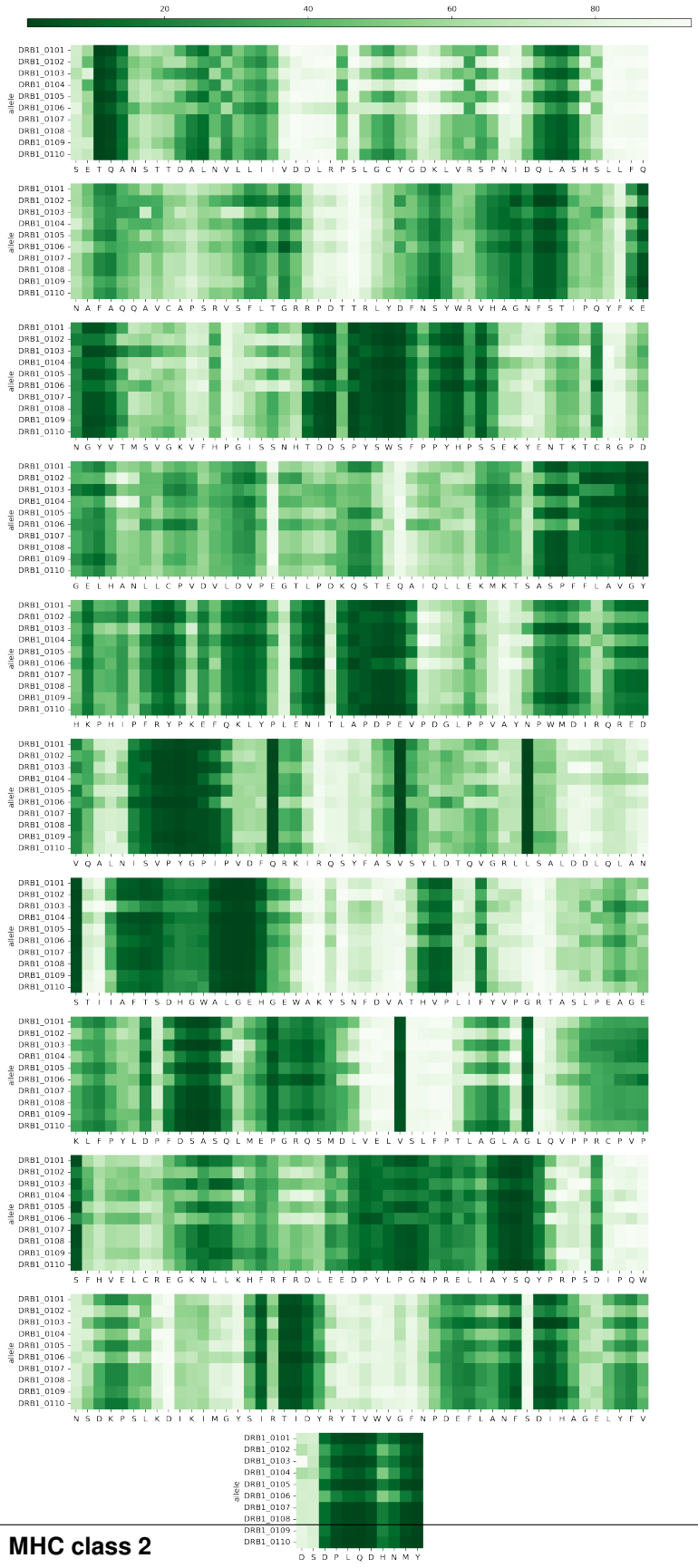
	allele	peptide	Core	Rank
0	HLA-A*01:01	FSDIHAGELY	FSDHAGELY	0.004
1	HLA-A*24:02	RYPKEFQKL	RYPKEFQKL	0.006
2	HLA-A*24:02	PYGPIPVDF	PYGPIPVDF	0.011
3	HLA-A*01:01	DSDPLQDHNMY	DSDPLQNMY	0.026
4	HLA-A*24:02	NFSTIPQYF	NFSTIPQYF	0.035
5	HLA-A*24:02	VPYGPIPVDF	VYGPIPVDF	0.044
6	HLA-A*02:01	SLFPTLAGL	SLFPTLAGL	0.059
7	HLA-A*02:01	KLYPLENITL	KLYPENITL	0.064
8	HLA-B*40:01	KEFQKLYPL	KEFQKLYPL	0.065
9	HLA-A*02:01	YLDPFDSASQL	YLFDSASQL	0.067

Frequency of binding peptide

Pie graphs are drawn to visualize peptides with immunogenicity in various alleles.


VELEGKNLL
SSNHTDSCY
HTDDSPYSW
SYLDTQVLL
VPRCPVPSF
FYPKEFQKL
APS-RVSFL
HTDDSPYSF
MYNDSQDLF
RQREDVQAL
FSDHAGELY
ATHVPLIFY
SYWRVHANF
YSIRTIDYY
RTIDYRYTV
YPK-EFQKL
PPRCPVPSF
FYVPGTASL
KPSLKDIKI
KYPLENITL
TTDALNVLL
RYPKEFQLY
IPNSDKPSL
SYLDTQVRL
VYHKPHIPF
PYLPGNPRL
QSYFASVSY
SLPEAGEKL
FTSDHGWAL
IVDDLRPSL
RLLSDDLQL
RTASLPGEK
KVFHPGISH
PYLPGNPLI
LPGNPRELI
RYT-VWVGF
QPRCPVPSF
YPFDSASQL
FPTLAGLAL
VPDGPPVAY
KEN-GYVTM
AFSTIPQYF
VWVGFPDEF
TEQ-AIQLL
GEWAKYSNF
GEL-HANLL
LPGNPR-EL
SLDIKIMGY
RYPKEFQKL
YLFDSASQL
VPG-RTASL
RLYDFNSYR
LYPLENITL
RLYDFNSYV
QYPSDIPQW
SYFASVSYL
KEFQKLYPL
YLDTQVGRL
KLYPENITL
SLFPTLAGL
VYGPIPVDF
NFSTIPQYF
DSDPLQNMY
PYGPIPVDF
GYHKPHIPF
SYGPIPVDF
REDVQALNI
VPS-FHVEL
IPFRYPKEF
VPYGPIPVF
KLYENITLA
YLPGNPREL
RPDTTRLYF

6.2 MHC class 2



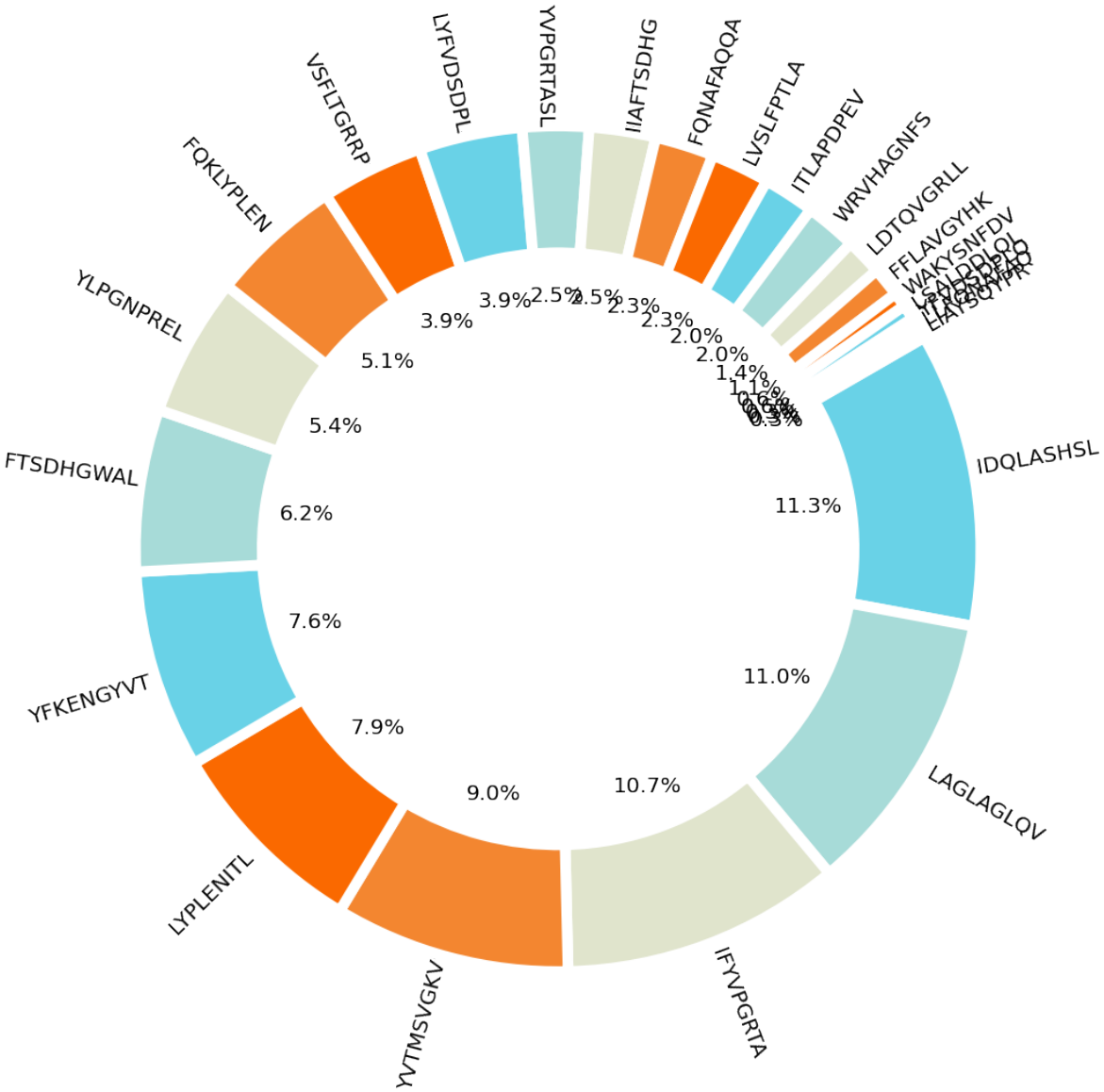
6.2.1 MHC class 2 binder peptide

Top10 peptide

	allele	peptide	Core	Rank
0	DRB1_0102	FPTLAGLAGLQVPPR	LAGLAGLQV	0.07
1	DRB1_0104	FPTLAGLAGLQVPPR	LAGLAGLQV	0.08
2	DRB1_0106	FPTLAGLAGLQVPPR	LAGLAGLQV	0.14
3	DRB1_0102	PTLAGLAGLQVPPRC	LAGLAGLQV	0.18
4	DRB1_0106	VPLIFYVPGRTASLP	IFYVPGRTA	0.22
5	DRB1_0102	LFPTLAGLAGLQVPP	LAGLAGLQV	0.23
6	DRB1_0104	PTLAGLAGLQVPPRC	LAGLAGLQV	0.26
7	DRB1_0101	FPTLAGLAGLQVPPR	LAGLAGLQV	0.28
8	DRB1_0107	FPTLAGLAGLQVPPR	LAGLAGLQV	0.28
9	DRB1_0108	FPTLAGLAGLQVPPR	LAGLAGLQV	0.28

Frequency of binding peptide

Pie graphs are drawn to visualize peptides with immunogenicity in various alleles.


IDQLASHSL
11.3%
LAGLAGLQV
11.0%
IFYVPGRTA
10.7%
YVTMSVGKV
9.0%
LYPLENITL
7.9%
YFKENGYVT
7.6%
FTSDHGWAL
6.2%
YLPGNPREL
5.4%
FQKLYPLEN
5.1%
VSFLTGRRP
3.9%
LYFVDSDPL
3.9%
YVPGRTASL
2.5%
IIAFTSDHG
2.5%
FQNAFAQQA
2.3%
LVSLFPTLA
2.3%
ITLAPDPEV
2.0%
WRVHAGNFS
2.0%
LDTQVGRLL
1.4%
FFLAVGYHK
1.1%
WAKYSNFDV

CHAPTER

SEVEN

APPENDIX