mouseAvastin(B20-4)

Taeyoon kim

May 30, 2022

CONTENTS

1 Protein parameters analysis **3**
- 1.1 Amino acid composition . . . 5
- 1.2 Potential sites of chemical modification . . . 5
- 1.3 Secondary structure fraction . . . 7
- 1.4 Secondary structure prediction . . . 8
- 1.5 Structural analysis . . . 8
- 1.6 Protein Scales . . . 8

2 Immunogenicity analysis **13**
- 2.1 MHC class 1 . . . 13
- 2.2 MHC class 2 . . . 16
- 2.3 Appendix . . . 17

Introduction

The data in the report is intended to be a resource to guide the development and lead-optimization of a therapeutic protein. These data can be used in a preemptive fashion - for example, in the decision to substitute an exposed residue on the surface that may be prone to the kind of chemical modification that might affect the stability of the protein. They can also be used to assist the troubleshooting of problems that can arise in the course of an clinical development - for example if an therapeutic protein 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 of problems that can occur during the therapeutic protein development and lead-optimization.

Background

Bevacizumab, sold under the brand name Avastin, is a medication used to treat a number of types of cancers and a specific eye disease. For cancer, it is given by slow injection into a vein (intravenous) and used for colon cancer, lung cancer, glioblastoma, and renal-cell carcinoma. In many of these diseases it is used as a first-line therapy. For age-related macular degeneration it is given by injection into the eye (intravitreal).

B20-4 is an anti-VEGF antibody, which has previously been used as a surrogate for preclinical modelling of bevacizumab activity.

Protein sequence

```
>mAvastin(B20-4_HC)
EVQLVESGGGLVQPGGSLRLSCAASGFSINGSWIFWVRQAPGKGLEWVGAIWPFGGYTHYADSVKGRFTISADTSKNTAY
LQMNSLRAEDTAVYYCARWGHSTSPWAMDYWGQGTLVTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTV
SWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSCDKTHTCPPCPAPELLG
GPSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREEQYNSTYRVVSVLTVLHQDWLNG
KEYKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPSRDELTKNQVSLTCLVKGFYPSDIAVEWESNGQPENNYKTTPP
VLDSDGSFFLYSKLTVDKSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPG
>mAvastin(B20-4_LC)
DIQMTQSPSSLSASVGDRVTITCRASQVIRRSLAWYQQKPGKAPKLLIYAASNLASGVPSRFSGSGSGTDFTLTISSLQP
EDFATYYCQQSNTSPLTFGQGTKVEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQ
ESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC
```

CHAPTER

ONE

PROTEIN PARAMETERS ANALYSIS

The program performs most of the same functions as the Expasy ProtParam tool.

```
# Name of target protein: ----------------------------------B20_HC
# Molecular weight(Dalton): --------------------------------49,078
# Total number of amino acid: ------------------------------450
# Chemical formula: ----------------------C2198H3377N581O666S15
# Total number of atom is: ---------------------------------6837
# Extinction coefficient(reduced): -------------------------92820
# Reduced Abs 0.1%(=1 g/L): --------------------------------1.891
# Extinction coefficient(non-reduced): ---------------------93445
# Non-reduced Abs 0.1%(=1 g/L): ----------------------------1.904
# Theoretical pI: ------------------------------------------8.210
# Aromaticity: ---------------------------------------------9.78%
# The GRAVY value is ---------------------------------------0.337
  B20_HC is more hydrophilic protein.
# The instability index: -----------------------------------43.002
  B20_HC is seems unstable.
```

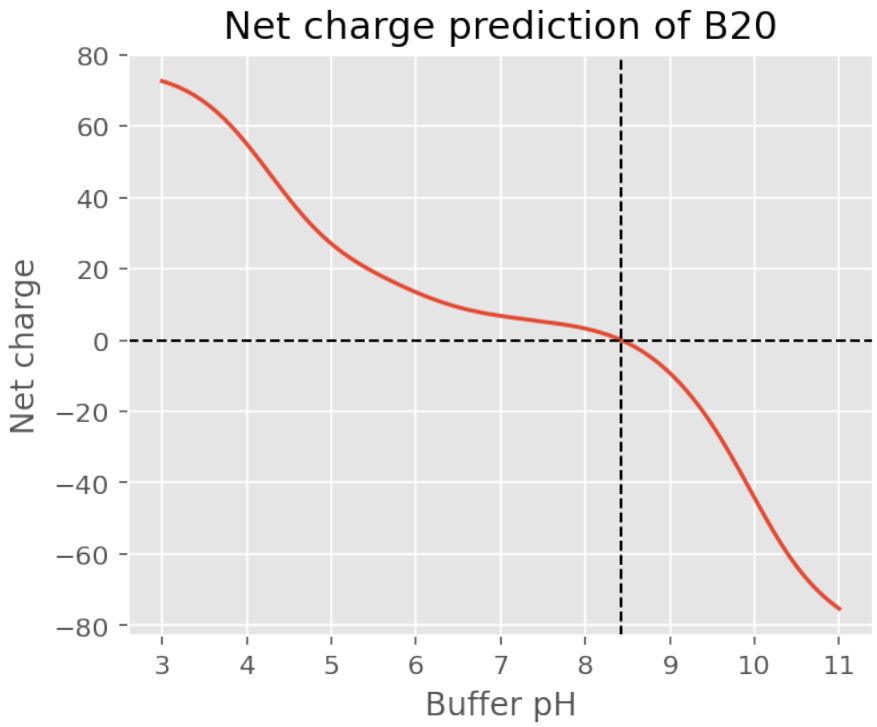
```
# Name of target protein: ----------------------------------B20_LC
# Molecular weight(Dalton): --------------------------------23,231
# Total number of amino acid: ------------------------------214
# Chemical formula: ----------------------C1016H1603N279O332S6
# Total number of atom is: ---------------------------------3236
# Extinction coefficient(reduced): -------------------------22920
# Reduced Abs 0.1%(=1 g/L): --------------------------------0.987
# Extinction coefficient(non-reduced): ---------------------23170
# Non-reduced Abs 0.1%(=1 g/L): ----------------------------0.997
# Theoretical pI: ------------------------------------------8.564
# Aromaticity: ---------------------------------------------8.41%
# The GRAVY value is ---------------------------------------0.409
  B20_LC is more hydrophilic protein.
# The instability index: -----------------------------------57.615
  B20_LC is seems unstable.
```

```
# Name of target protein: ----------------------------------B20
# Molecular weight(Dalton): --------------------------------72,290
# Total number of amino acid: ------------------------------664
# Chemical formula: ----------------------C3214H4978N860O997S21
# Total number of atom is: ---------------------------------10070
# Extinction coefficient(reduced): -------------------------115740
# Reduced Abs 0.1%(=1 g/L): --------------------------------1.601
# Extinction coefficient(non-reduced): ---------------------116740
```

(continues on next page)

(continued from previous page)

```
# Non-reduced Abs 0.1%(=1 g/L): ----------------------------1.615
# Theoretical pI: -------------------------------------------8.425
# Aromaticity: ----------------------------------------------9.34%
# The GRAVY value is ----------------------------------------0.360
  B20 is more hydrophilic protein.
# The instability index: ------------------------------------47.726
  B20 is seems unstable.
```



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

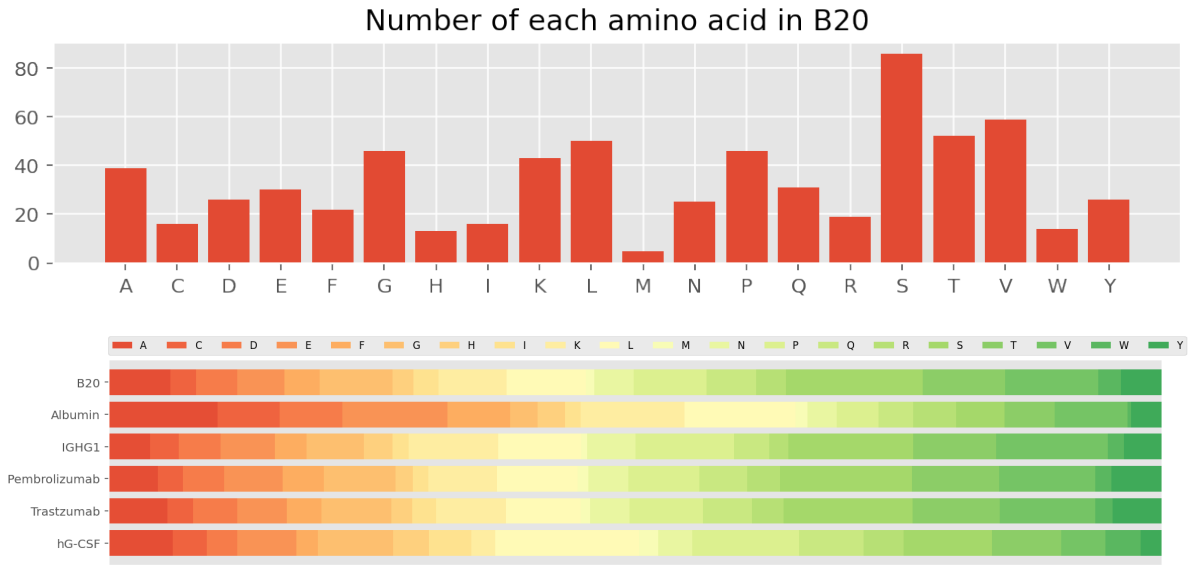
  - Calculate the aromaticity according to Lobry, 1994. Calculates the aromaticity value of a protein according to Lobry, 1994. It is simply the relative frequency of Phe+Trp+Tyr.

- GRAVY
  - 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.
- 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).

1.1 Amino acid composition

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

```
Total number of positively charged residues(Arg + Lys):--------62
Total number of negatively charged residues(Asp + Glu):--------56
```



1.2 Potential sites of chemical modification

An initial scan of the protein 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 protein, 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.

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.

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

```
Deamination pattern found in:--------------B20_HC
30-NG-31
39-QA-40
77-NT-78
84-NS-85
113-QG-114
163-NS-164
179-QS-180
200-QT-201
212-NT-213
290-NA-291
301-NS-302
319-NG-320
388-NG-389
423-QG-424
```

```
Deamination pattern found in:--------------B20_LC
6-QS-7
90-QS-91
92-NT-93
100-QG-101
152-NA-153
155-QS-156
158-NS-159
199-QG-200
```

1.2.2 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 the 2-1 and +3 positions.

- Search patterns: PRO-SER/THR

```
Potential o-linked glycosylation sites:--------------B20_HC
127-PS-128
134-PS-135
193-PS-194
210-PS-211
242-PS-243
357-PS-358
378-PS-379
```

```
Potential o-linked glycosylation sites:--------------B20_LC
8-PS-9
59-PS-60
```

(continues on next page)

(continued from previous page)

```
113-PS-114
120-PS-121
```

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

```
Potential o-linked glycosylation sites:--------------B20_HC
102-STSP-105
124-TKGP-127
128-SVFP-131
229-TCPP-232
258-SRTP-261
354-TLPP-357
397-TTPP-400
446-SLSP-449
```

```
Potential o-linked glycosylation sites:--------------B20_LC
5-TQSP-8
56-SGVP-59
77-SLQP-80
```

1.2.3 Potential n-linked glycosylation sites

- Search patterns: ASN-X-SER/THR

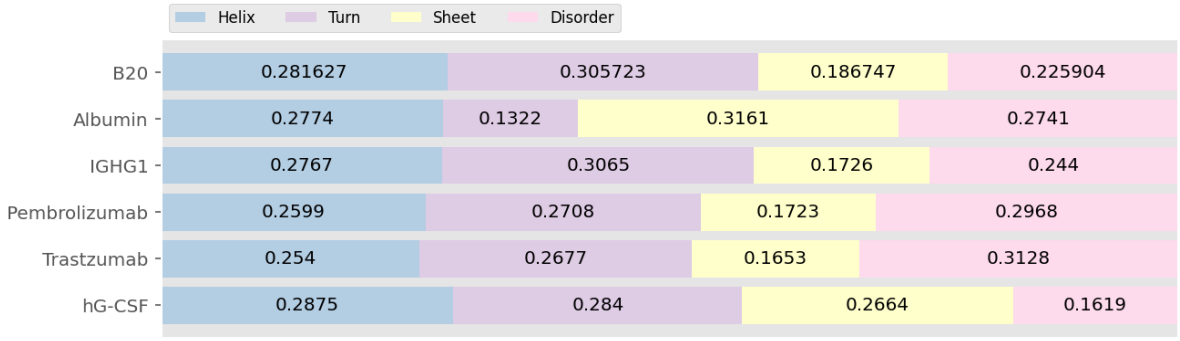
```
Potential n-linked glycosylation sites:--------------B20_HC
30-NGS-32
301-NST-303
```

```
Potential n-linked glycosylation sites:--------------B20_LC
92-NTS-94
```

1.3 Secondary structure fraction

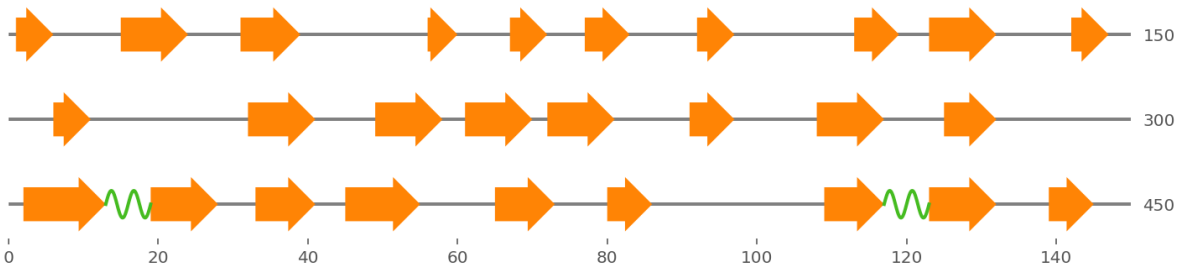
The fraction of amino acids that tend to be found in the three classical secondary structures. These are beta sheets, alpha helixes, and turns (where the residues change direction).

- 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.



1.4 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.



1.5 Structural analysis

1.5.1 Calculation of protein diameter

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

```
# Diameter of B20 is: ----------------------126.996 Angstrong.
```

1.6 Protein Scales

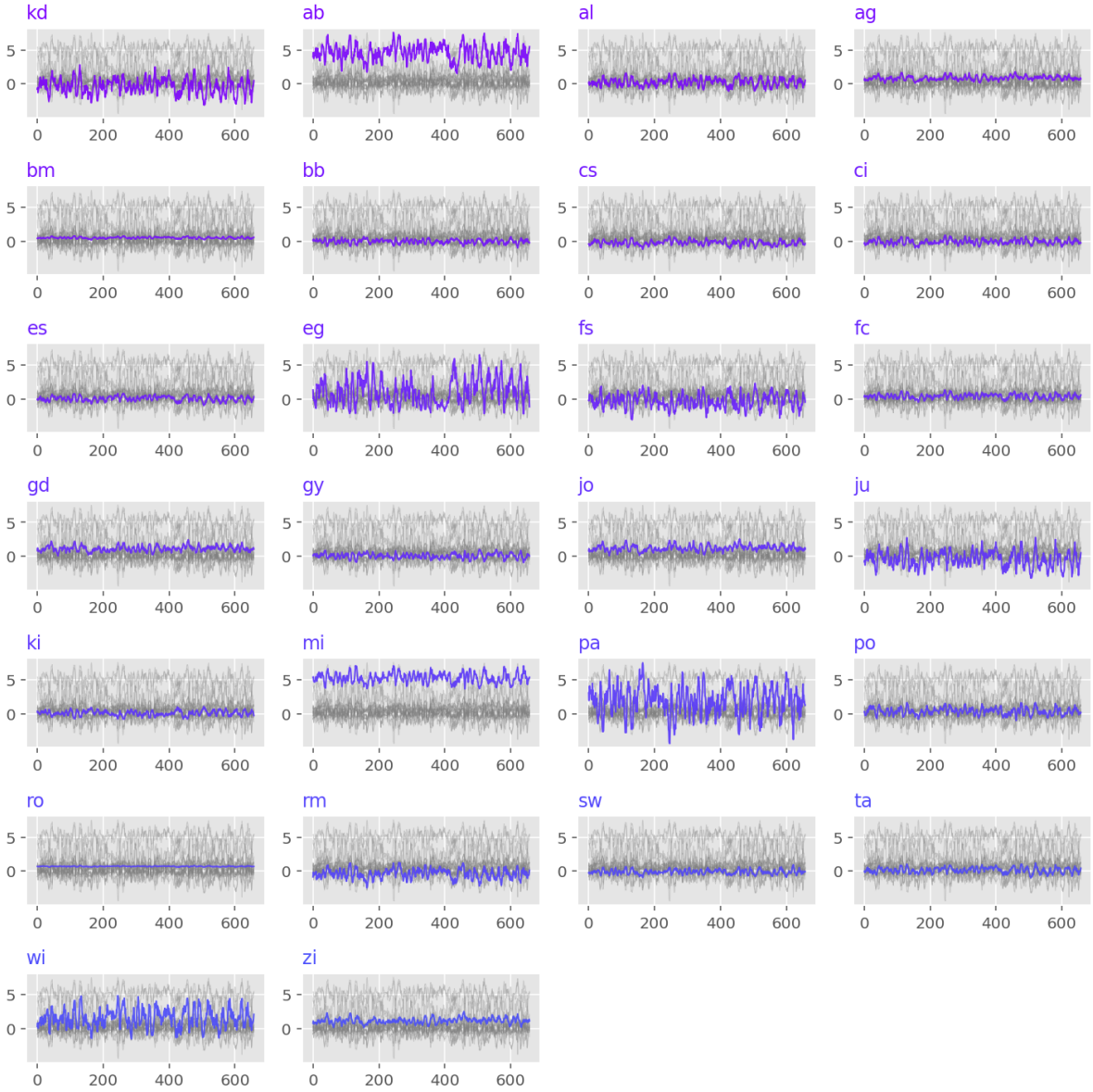
Protein scales are a way of measuring certain attributes of residues over the length of the peptide sequence using a sliding window. Scales are comprised of values for each amino acid based on different physical and chemical properties, such as hydrophobicity, secondary structure tendencies, and surface 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.

- kd → Kyte & Doolittle Index of Hydrophobicity
- hw → Hopp & Wood Index of Hydrophilicity
- em → Emini Surface fractional probability (Surface Accessibility

-

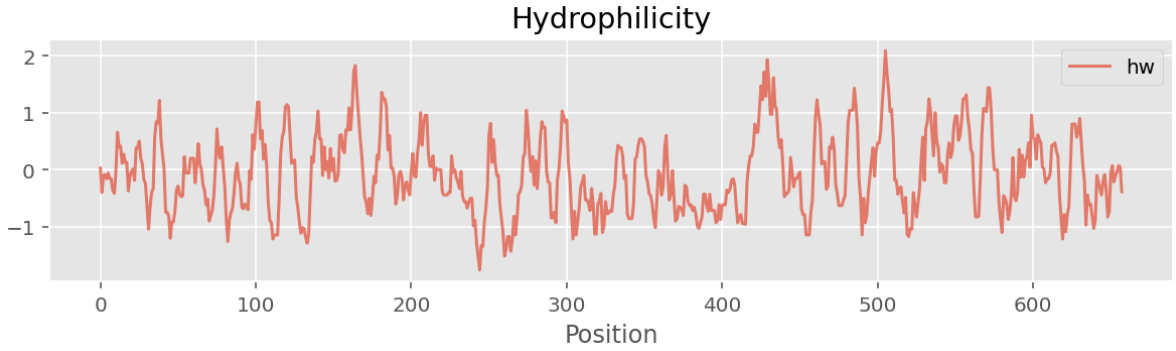
1.6.1 Hydrophobicity index

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



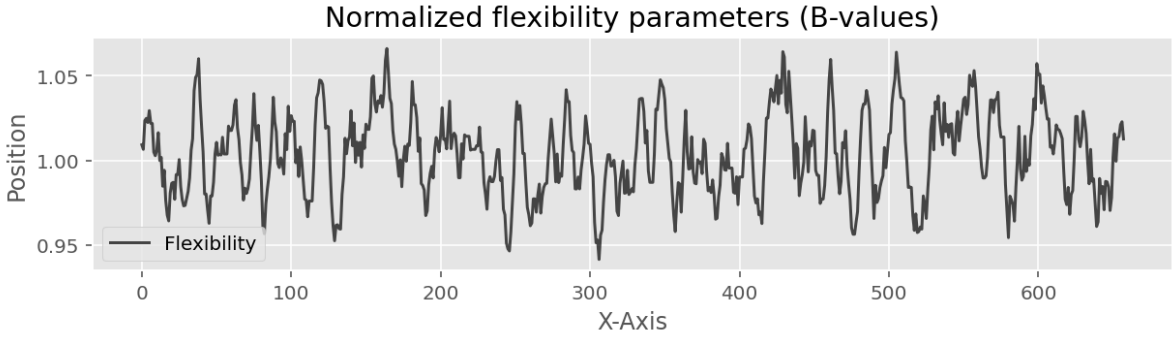
1.6.2 Hydrophilicity index

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



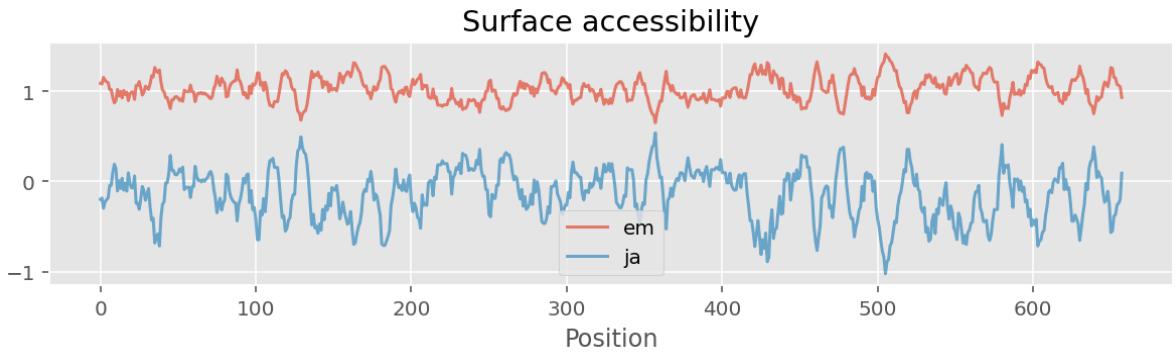
1.6.3 Flexibility index

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



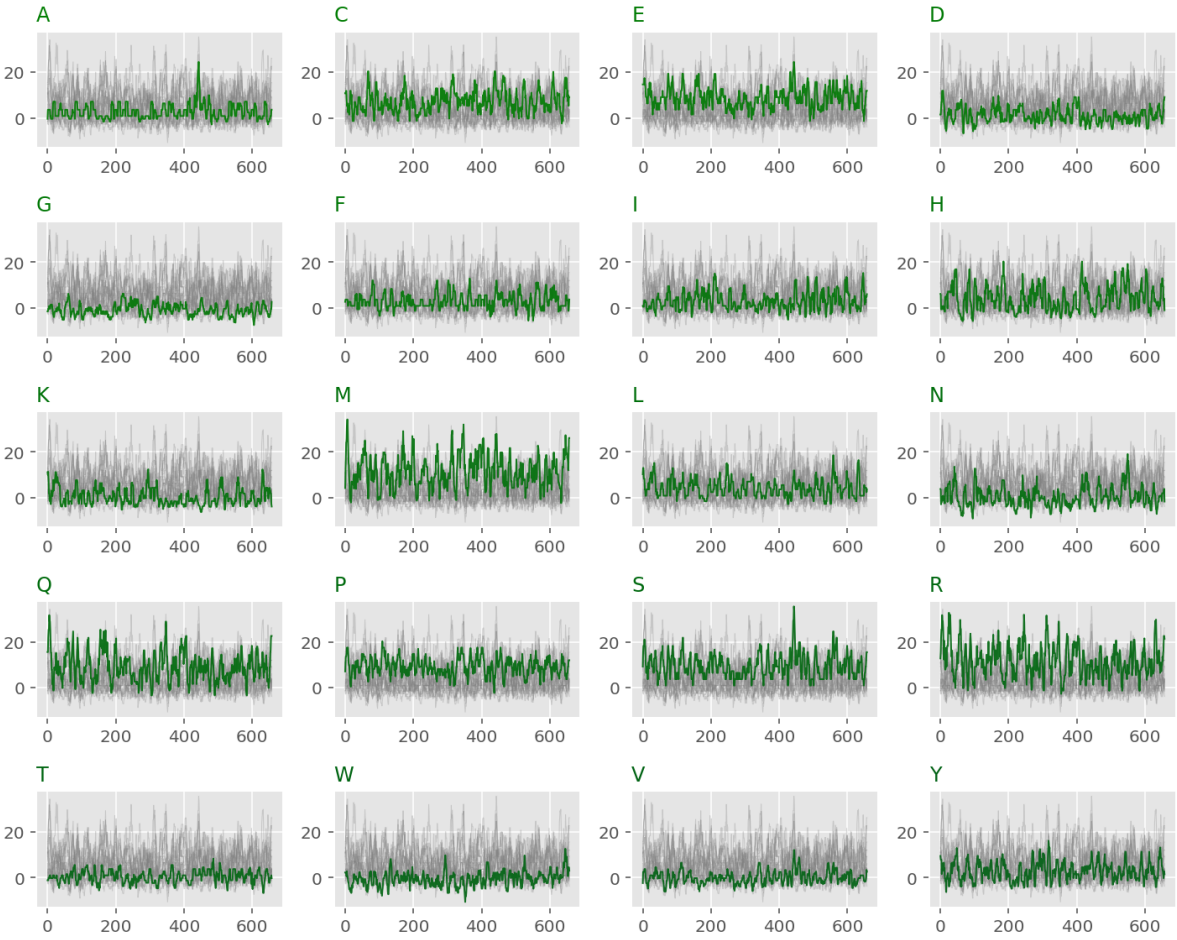
1.6.4 Surface accessibility

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. 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. The solvent-accessible surface for each residue depends upon the degree of exposure of the residue on the surface, but also on the size of the residue side chain.



1.6.5 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.



CHAPTER

TWO

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.

2.1 MHC class 1

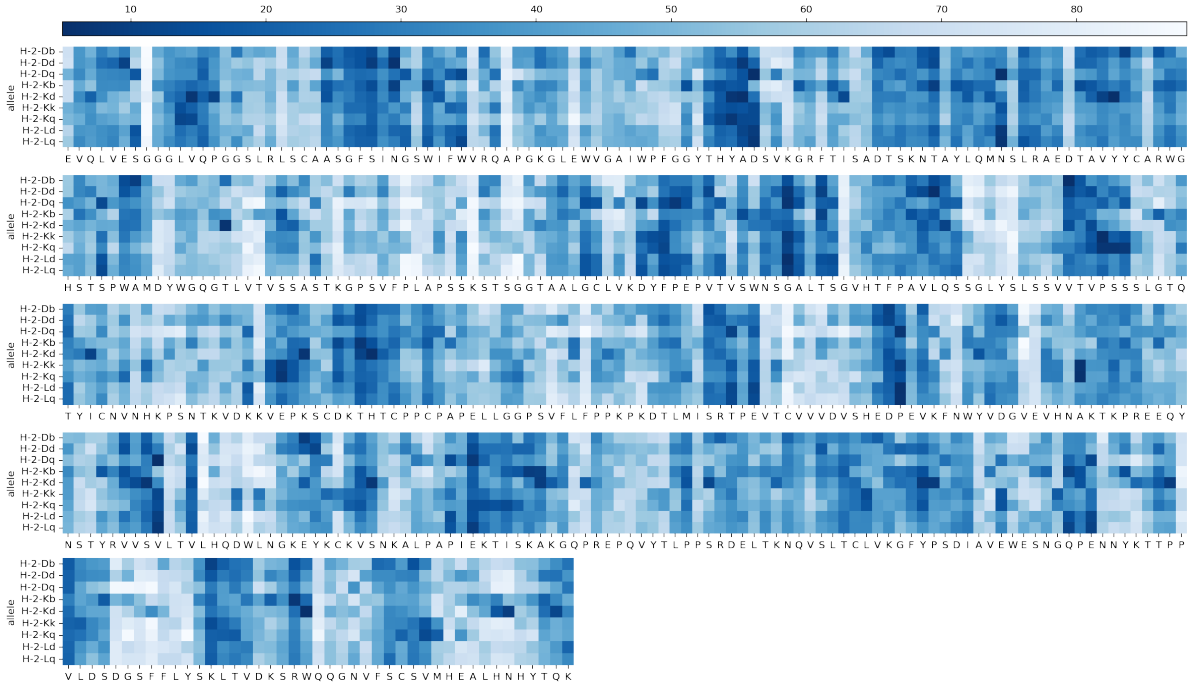
Class I major histocompatibility complex (MHC) molecules bind, and present to T cells, short peptides derived from intracellular processing of proteins. The peptide repertoire of a specific molecule is to a large extent determined by the molecular structure accommodating so-called main anchor positions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

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]

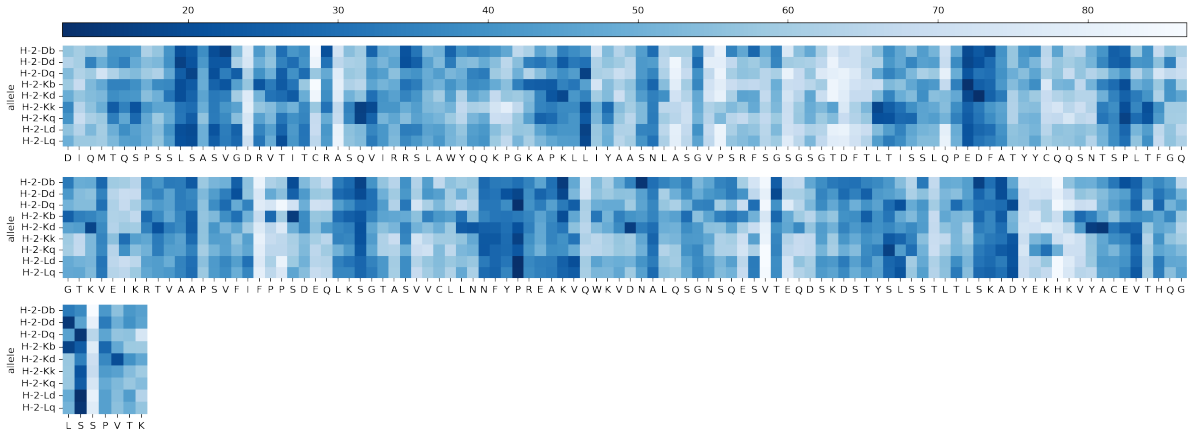
- MHC class 1 superset
  - HLA-A01:01, HLA-A02:01, HLA-A03:01, HLA-A24:02, HLA-B07:02, HLA-B40:01

[1] Kimball's Biology Pages, Histocompatibility Molecules

2.1.1 Predicts binding of peptides to MHC class1




```
make_heatmap(make_table(df_mhc1_lc, second_record, 100),second_record, 11, 2.7,
 ↪'Blues_r',100)
```



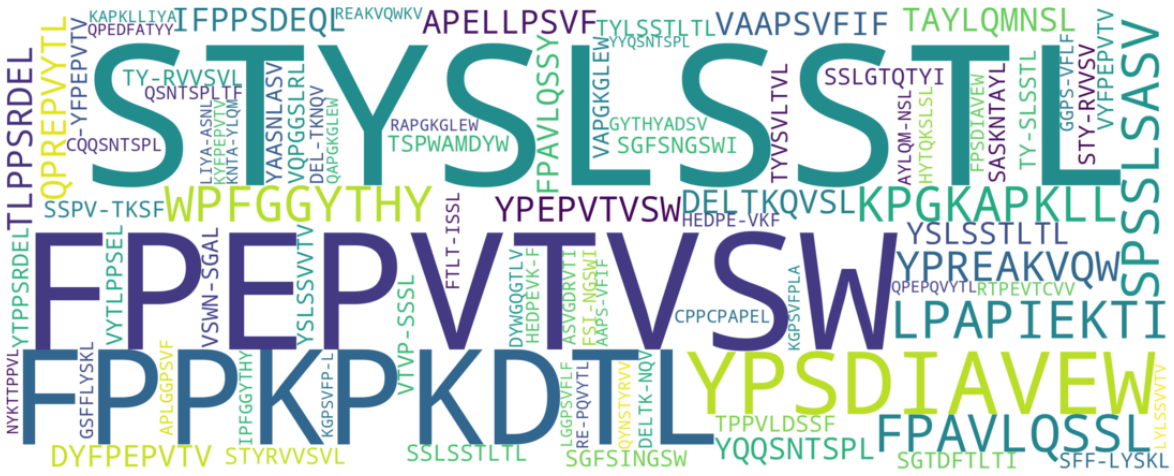

```
len(first_record.seq)
```

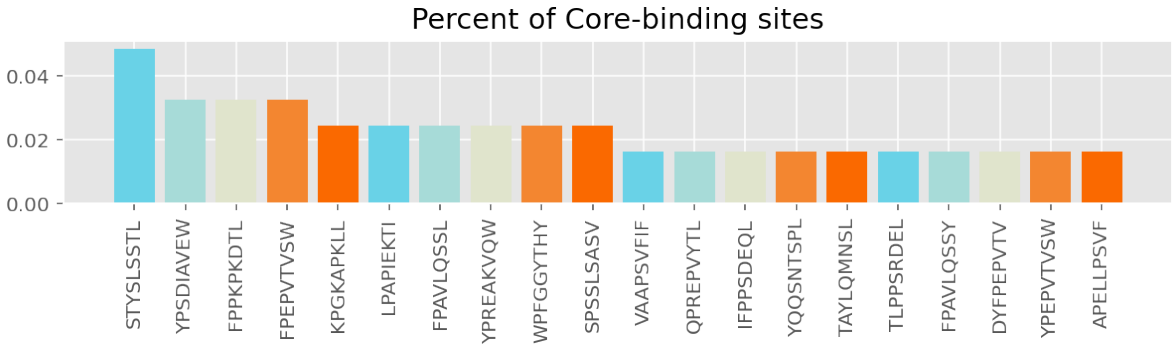
```
450
```

2.1.2 Top10 strong binding peptide

Core	allele	peptide
STYSLSSTL	6	6
FPEPVTVSW	4	4
FPPKPKDTL	4	4
YPSDIAVEW	4	4
FPAVLQSSL	3	3
LPAPIEKTI	3	3
SPSSLSASV	3	3
WPFGGYTHY	3	3
KPGKAPKLL	3	3
YPREAKVQW	3	3

2.1.3 Frequency of binding peptide



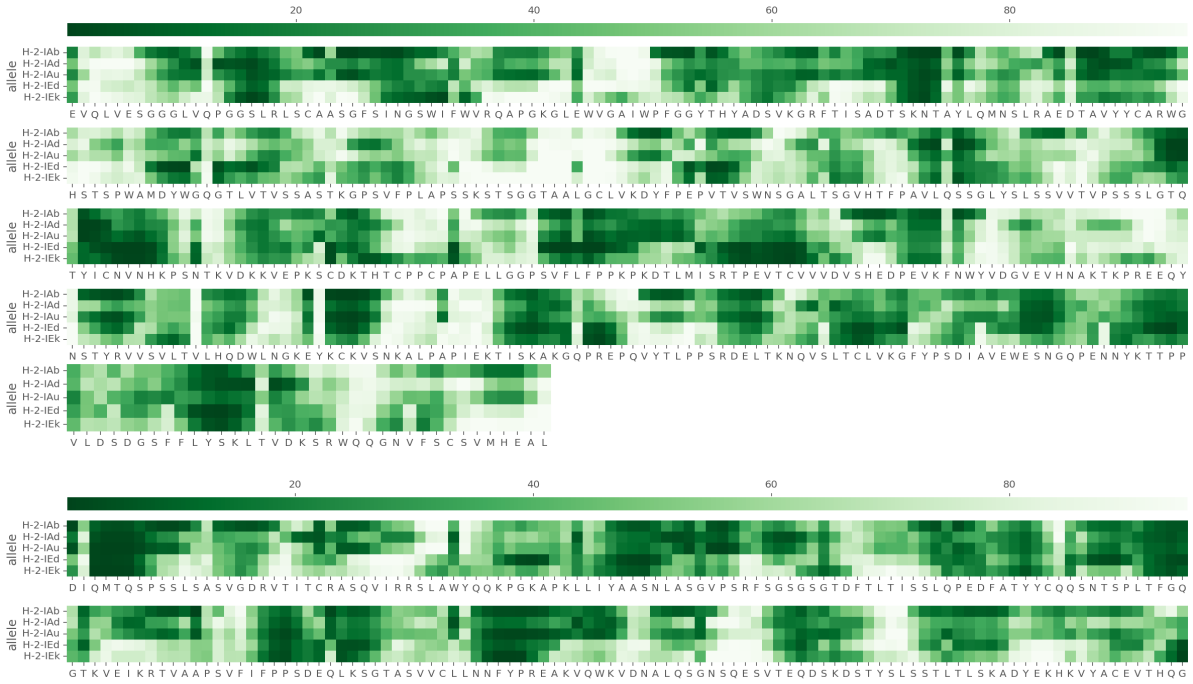


2.2 MHC class 2

MHC Class II molecules are a class of major histocompatibility complex (MHC) molecules normally found only on professional antigen-presenting cells such as dendritic cells, mononuclear phagocytes, some endothelial cells, thymic epithelial cells, and B cells. These cells are important in initiating immune responses.

- MHC class 2 allele superset
  - DRB1_0101,DRB1_0102,DRB1_0103,DRB1_0104,DRB1_0105,DRB1_0106,DRB1_0107,DRB1_0108,DRB1_0109,DRB

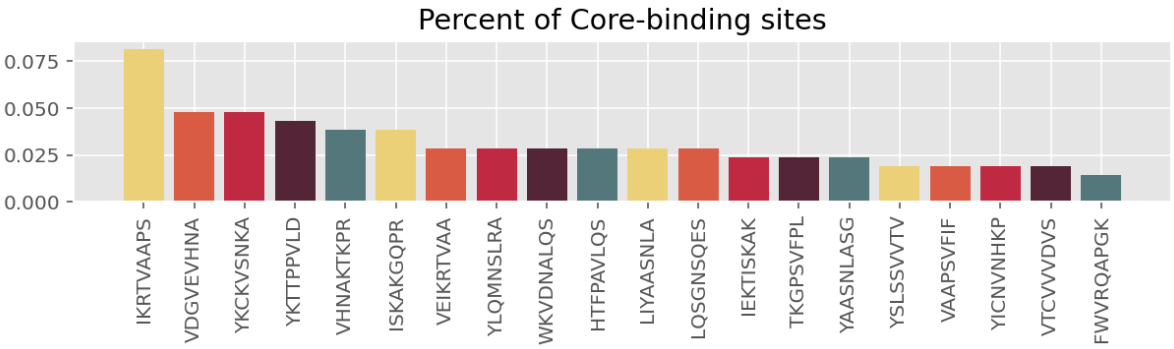
2.2.1 Predicts binding of peptides to MHC class2



2.2.2 Top10 binding peptide

```
<pandas.io.formats.style.Styler at 0x7fd6071566b0>
```

2.2.3 Frequency of binding peptide



2.3 Appendix

nothing yet.