HMM profile generation

As the first step towards constructing the pMHC database, we obtained the entire PDB archive as a single FASTA file with each PDB entry split by chain, https://www.rcsb.org/pdb/download/download.do. This PDB database file contained 126580 entries.

To identify all MHC class I entries from the PDB database file, we used a HMM profile from Pfam. This HMM profile is called MHC_I.hmm (accession number: PF00129), and it includes the α1 and α2 domains of the α chain from the MHC class I family. We then used hmmsearch from HMMER (version 3.1) to align the PDB database file against the HMM profile from Pfam.

```
hmmsearch -o alignment.out -E 0.00001 --noali MHC_I.hmm <pdb_filename.fsa>
```

To discard false positive "hits" the E-value threshold was set to 0.00001 and the option --noali was selected to omit the alignment from the output, hence speeding up the procedure and lowering the output volume. This threshold yielded 700 PDB entries which matched the MHC class I HMM profile.

All identified entries were then aligned to the MHC class I HMM profile from Pfam using hmmalign. This step produced a multiple sequence alignment (MSA) that helps when inspecting insertions that might cause problems with the template-based modeling.

```
hmmalign -o alignment.out --trim --amino MHC_I.hmm <sequences.fsa>
```

We here included the options --trim which excludes non-homologous residues from the protein terminals and --amino which specifies the type of sequences provided. Doing this we found 16 non-redundant entries included insertions at specific positions and we therefore constructed and in-house HMM profile for the identified entries. This new HMM profile was constructed using hmmbuild, with the MSA as input file.

```
hmmbuild --amino --symfrac 0 MHC_I_complete.hmm <alignment_file>
```

We here included the options --symfrac which considers each position/column as a consensus column and the --amino which specifies the type of sequences provided. The resulting HMM profile included 181 positions and was named MHC_I_complete.hmm.