Epitope Fingerprinting Report

Date: November 1st, 2024

Work Package: Epitope Fingerprinting for Phycoerithrin antibody (mouse IgG1, PE4-14D10 (Miltenyi))

General comments

The PE antibody coated magnetic beads were used for epitope fingerprinting. The epitope was determined from using NGS data of peptide phage of the naïve ENTE-1 library selected on the antibody coupled to MultiSort MicroBeads (Miltenyi, Order No. 130-090-757).

Sequence pools of 284,127 resp. 346,933 sequences from a first and a second selection round were analysed for the statistical enrichment of motifs corresponding to potential epitopes in the protein. Excellent enrichment of sequences corresponding to a structural epitope was found.

This report is based on data analysed earlier. In recent years epitopic has improved many statistical analysis tools. The analysis of the data is now not only faster, but we are also able to produce even more details of the antibodies' binding to the antigen.

As an internal challenge, we managed to run the analysis again, revealing more details in a much shorter time. With the excellent datasets, it took only about eight hours for the entire work presented here.

The results show/explain:

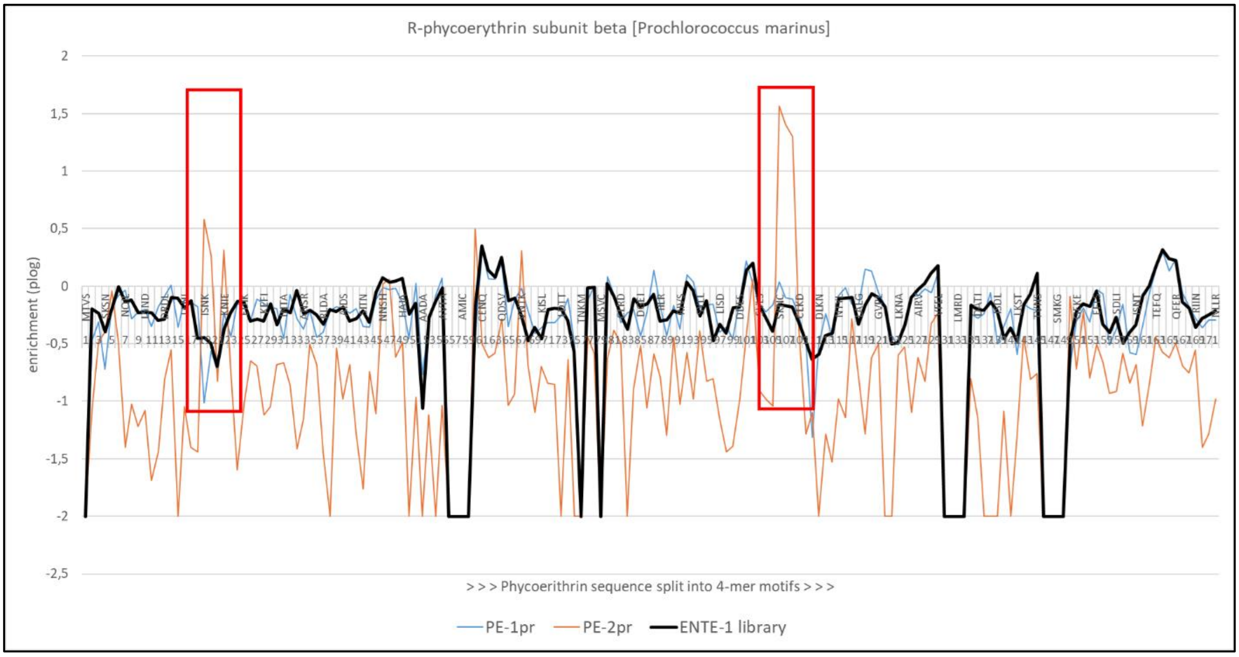
- **(Structural) epitope details by amino acids recognized**
- **Structural confirmation with Alpha-Fold structure**

Statistical Analyses

Standard analysis based on the frequency of 4-mer motifs of the antigen sequence resulted in an overview of potential epitopes.

The graphic below shows the statistical enrichment of the consecutive 4-mer peptides of the R-Phycoerithrin protein sequence from *Prochlorococcus marinus* in the data sets after one and two selection rounds. The Y-axis shows the enrichment of 4-mers over the theoretically expected value defined by the library design. The scale is log10, so several motifs occur 10- to 50-fold more frequently in the NGS datasets than expected. These expected values can be calculated, because the naïve library is constructed from just one codon per amino acid and behaves stable even upon replication. The black line shows the actually measured statistical data from >2 Mio. sequences of the naïve library. Deviations are due to variations in the coupling yield of codon building blocks during the synthesis.

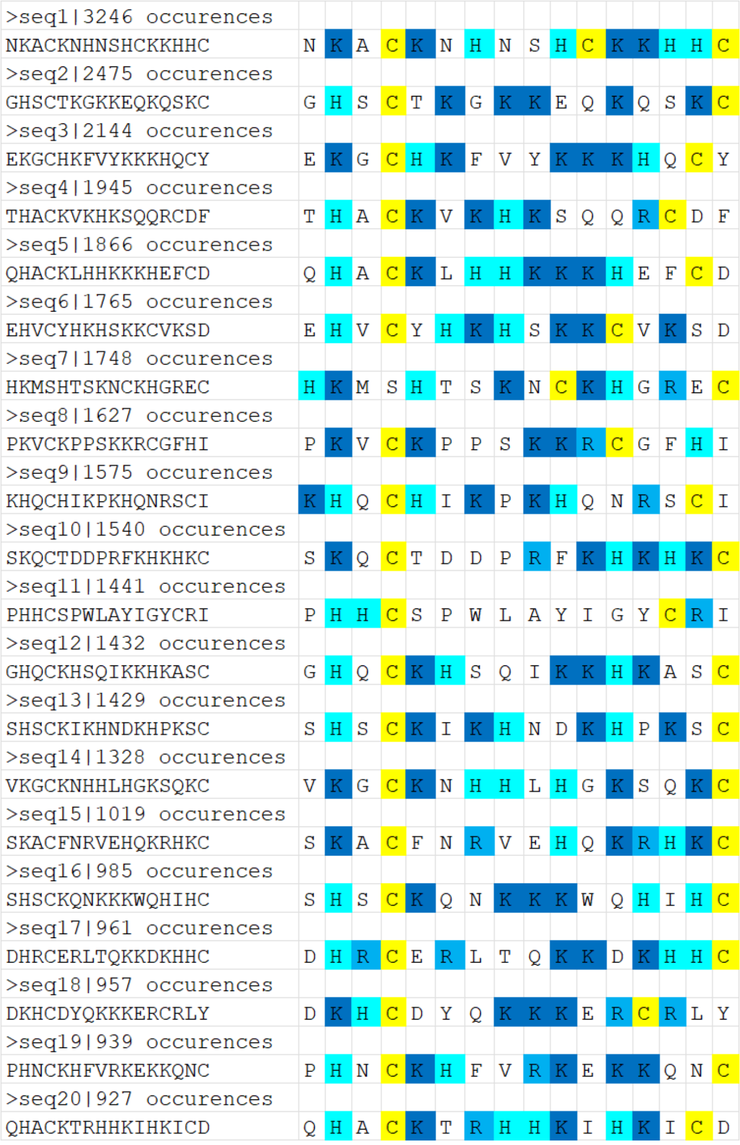
The boxes indicate motifs within the protein enriched more than expected. Such peaks could belong to positions in the protein bound by the antibody, and the sequences containing them were further analysed.



This “enrichment” curve shows that these positions represent the motifs:
19-ISNKNIE-25, 106-SKNCLK-111
It can be seen below that the first motif’s enrichment of about fivefold is likely to be caused by similarity to the amino acids of the second motif (Ser, Asp, Lys) which enriched fiftyfold above the theoretically expected enrichment.

The next step in the analysis is the retrieval of full-length sequences from the selected library pool, which are containing enriched motifs. These are usually hundreds of similar sequences. They are aligned with and without the antigen sequence. In most cases the alignments were at least cured for sequences found in the datasets only once, twice or even more frequently. If a sequence is found at least three to five times, it is less likely to contain typical NGS sequencing errors.

Looking at the 20 most enriched sequences of the second selection round, it is already visible that the PE-antibody is recognizing patterns with multiple Lys but not Arg:



>seq1|3246 occurences																
NKACKNHNSHCKKHHC	N	K	A	C	K	N	H	N	S	H	C	K	K	H	H	C
>seq2|2475 occurences																
GHSCTKGKKEQKQSKC	G	H	S	C	T	K	G	K	K	E	Q	K	Q	S	K	C
>seq3|2144 occurences																
EKGCHKFVYKKKHQCY	E	K	G	C	H	K	F	V	Y	K	K	K	H	Q	C	Y
>seq4|1945 occurences																
THACKVKHKSQQRCDF	T	H	A	C	K	V	K	H	K	S	Q	Q	R	C	D	F
>seq5|1866 occurences																
QHACKLHHKKKHEFCD	Q	H	A	C	K	L	H	H	K	K	K	H	E	F	C	D
>seq6|1765 occurences																
EHVCYHKHSKKCVKSD	E	H	V	C	Y	H	K	H	S	K	K	C	V	K	S	D
>seq7|1748 occurences																
HKMSHTSKNCKHGREC	H	K	M	S	H	T	S	K	N	C	K	H	G	R	E	C
>seq8|1627 occurences																
PKVCKPPSKKRCGFHI	P	K	V	C	K	P	P	S	K	K	R	C	G	F	H	I
>seq9|1575 occurences																
KHQCHIKPKHQNRSCI	K	H	Q	C	H	I	K	P	K	H	Q	N	R	S	C	I
>seq10|1540 occurences																
SKQCTDDPRFKHKHKC	S	K	Q	C	T	D	D	P	R	F	K	H	K	H	K	C
>seq11|1441 occurences																
PHHCSPWLAYIGYCRI	P	H	H	C	S	P	W	L	A	Y	I	G	Y	C	R	I
>seq12|1432 occurences																
GHQCKHSQIKKHKASC	G	H	Q	C	K	H	S	Q	I	K	K	H	K	A	S	C
>seq13|1429 occurences																
SHSCKIKHNDKHPKSC	S	H	S	C	K	I	K	H	N	D	K	H	P	K	S	C
>seq14|1328 occurences																
VKGCKNHHLHGKSQKC	V	K	G	C	K	N	H	H	L	H	G	K	S	Q	K	C
>seq15|1019 occurences																
SKACFNRVEHQKRHKC	S	K	A	C	F	N	R	V	E	H	Q	K	R	H	K	C
>seq16|985 occurences																
SHSCKQNKKKWQHIHC	S	H	S	C	K	Q	N	K	K	K	W	Q	H	I	H	C
>seq17|961 occurences																
DHRCERLTQKKDKHHC	D	H	R	C	E	R	L	T	Q	K	K	D	K	H	H	C
>seq18|957 occurences																
DKHCDYQKKKERCRLY	D	K	H	C	D	Y	Q	K	K	K	E	R	C	R	L	Y
>seq19|939 occurences																
PHNCKHFVRKEKKQNC	P	H	N	C	K	H	F	V	R	K	E	K	K	Q	N	C
>seq20|927 occurences																
QHACKTRHHKIHKICD	Q	H	A	C	K	T	R	H	H	K	I	H	K	I	C	D

Additionally, there are also a few His and Arg and every sequence features two Cys that is most likely forming a loop structure. The selected peptides may require this to mimic a structural epitope with positive charges.

Alignment of sequences with enriched motifs

All sequences containing the amino acid 4-mers of

106-SKNCLK-111

were retrieved from the two selection’s NGS datasets. This identified 273 different in total 5,139 sequences. To keep the number of sequences displayed in the alignment to a minimum, the alignment displays only sequences found at least 3 times.

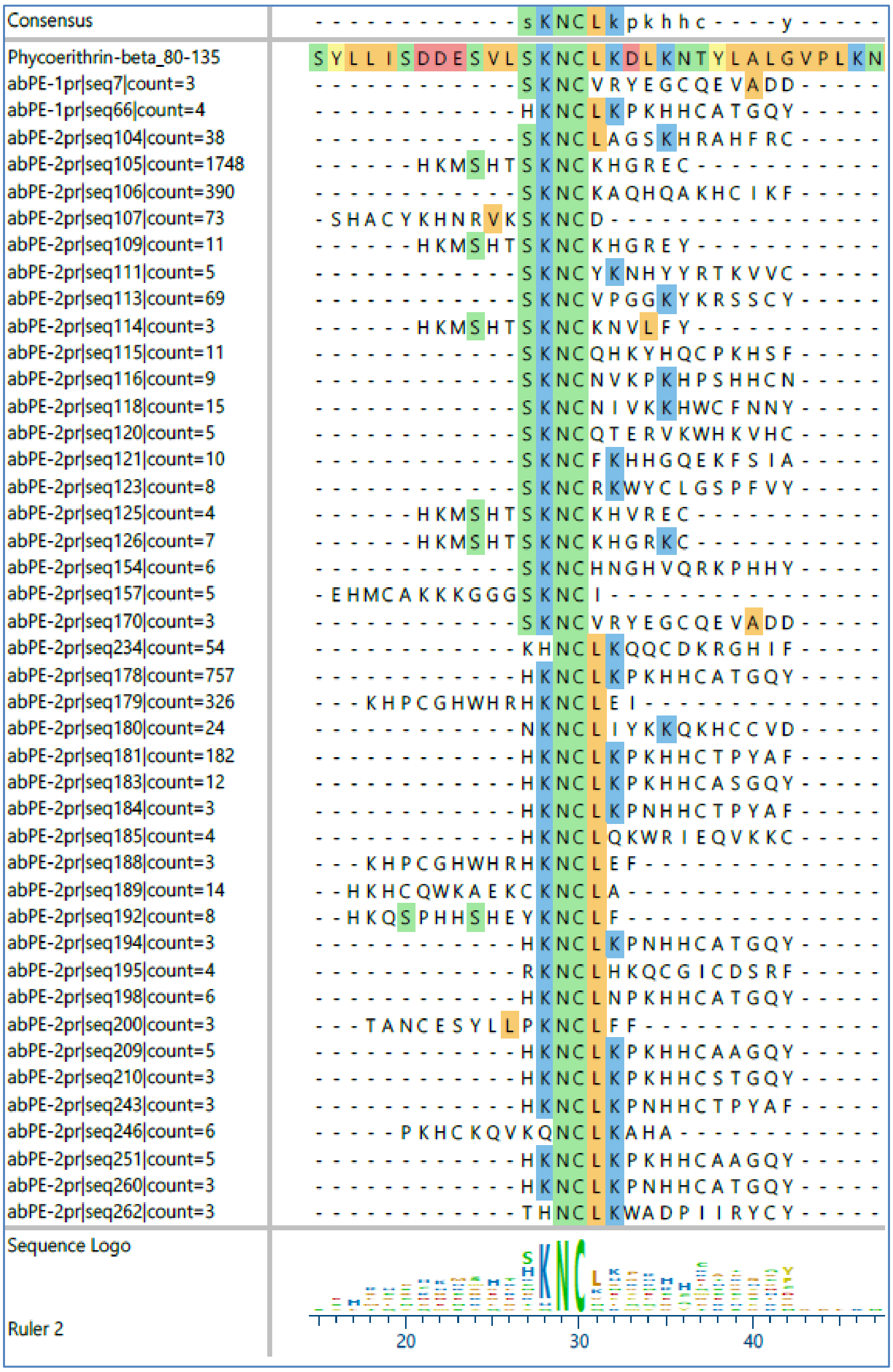
Epitope fingerprinting avoids multiple selection rounds. Therefore, hundreds of similar sequences can be used for the analysis of the spectrum of peptides recognized by the antibody. Due to the library’s stability and a balanced library design the statistical analysis reveals the selected 5,000 sequences, which are only a fraction of the 630,000 sequences in the analysed data sets.

Sequences with 4 - 6 amino acids identity to the antigen sequence of R-PE are strongly enriched and therefore likely not to be present by chance.

Based on the alignment and considering the amino acid residues shared by the enriched sequences the epitope could first be best described as

sKNClKxxK

xx may be 1-3 residues depending on the nature of the residues, e.g. single Pro in several sequences.



```
Consensus                     ------------sKNCLkpkhhc----y-----
Phycoerithrin-beta_80-135     SYLLISDDESVLSKNCLKDLKNTYLALGVPLKN
abPE-1pr|seq7|count=3         ------------SKNCVRYEGCQEVADD-----
abPE-1pr|seq66|count=4        ------------HKNCLKPKHHCATGQY-----
abPE-2pr|seq104|count=38      ------------SKNCLAGSKHRAHFRC-----
abPE-2pr|seq105|count=1748    ------HKMSHTSKNCKHGREC-----------
abPE-2pr|seq106|count=390     ------------SKNCKAQHQAKHCIKF-----
abPE-2pr|seq107|count=73      -SHACYKHNRVKSKNCD----------------
abPE-2pr|seq109|count=11      ------HKMSHTSKNCKHGREY-----------
abPE-2pr|seq111|count=5       ------------SKNCYKNHYYRTKVVC-----
abPE-2pr|seq113|count=69      ------------SKNCVPGGKYKRSSCY-----
abPE-2pr|seq114|count=3       ------HKMSHTSKNCKNVLFY-----------
abPE-2pr|seq115|count=11      ------------SKNCQHKYHQCPKHSF-----
abPE-2pr|seq116|count=9       ------------SKNCNVKPKHPSHHCN-----
abPE-2pr|seq118|count=15      ------------SKNCNIVKKHWCFNNY-----
abPE-2pr|seq120|count=5       ------------SKNCQTERVKWHKVHC-----
abPE-2pr|seq121|count=10      ------------SKNCFKHHGQEKFSIA-----
abPE-2pr|seq123|count=8       ------------SKNCRKWYCLGSPFVY-----
abPE-2pr|seq125|count=4       ------HKMSHTSKNCKHVREC-----------
abPE-2pr|seq126|count=7       ------HKMSHTSKNCKHGRKC-----------
abPE-2pr|seq154|count=6       ------------SKNCHNGHVQRKPHHY-----
abPE-2pr|seq157|count=5       -EHMCAKKKGGGSKNCI----------------
abPE-2pr|seq170|count=3       ------------SKNCVRYEGCQEVADD-----
abPE-2pr|seq234|count=54      ------------KHNCLKQQCDKRGHIF-----
abPE-2pr|seq178|count=757     ------------HKNCLKPKHHCATGQY-----
abPE-2pr|seq179|count=326     ---KHPCGHWHRHKNCLEI--------------
abPE-2pr|seq180|count=24      ------------NKNCLIYKKQKHCCVD-----
abPE-2pr|seq181|count=182     ------------HKNCLKPKHHCTPYAF-----
abPE-2pr|seq183|count=12      ------------HKNCLKPKHHCASGQY-----
abPE-2pr|seq184|count=3       ------------HKNCLKPNHHCTPYAF-----
abPE-2pr|seq185|count=4       ------------HKNCLQKWRIEQVKKC-----
abPE-2pr|seq188|count=3       ---KHPCGHWHRHKNCLEF--------------
abPE-2pr|seq189|count=14      --HKHCQWKAEKCKNCLA---------------
abPE-2pr|seq192|count=8       --HKQSPHHSHEYKNCLF---------------
abPE-2pr|seq194|count=3       ------------HKNCLKPNHHCATGQY-----
abPE-2pr|seq195|count=4       ------------RKNCLHKQCGICDSRF-----
abPE-2pr|seq198|count=6       ------------HKNCLNPKHHCATGQY-----
abPE-2pr|seq200|count=3       ---TANCESYLLPKNCLFF--------------
abPE-2pr|seq209|count=5       ------------HKNCLKPKHHCAAGQY-----
abPE-2pr|seq210|count=3       ------------HKNCLKPKHHCSTGQY-----
abPE-2pr|seq243|count=3       ------------HKNCLKPNHHCTPYAF-----
abPE-2pr|seq246|count=6       -----PKHCKQVKQNCLKAHA------------
abPE-2pr|seq251|count=5       ------------HKNCLKPKHHCAAGQY-----
abPE-2pr|seq260|count=3       ------------HKNCLKPNHHCATGQY-----
abPE-2pr|seq262|count=3       ------------THNCLKWADPIIRYCY-----
Sequence Logo
Ruler 2                                20        30        40
```

Structural mimotope

Since this epitope is definitely not linear, it makes sense to search for enriched discontinuous or structural motifs, too. This alignment results from retrieving sequences with the motif KxxxK (x for any amino acid)

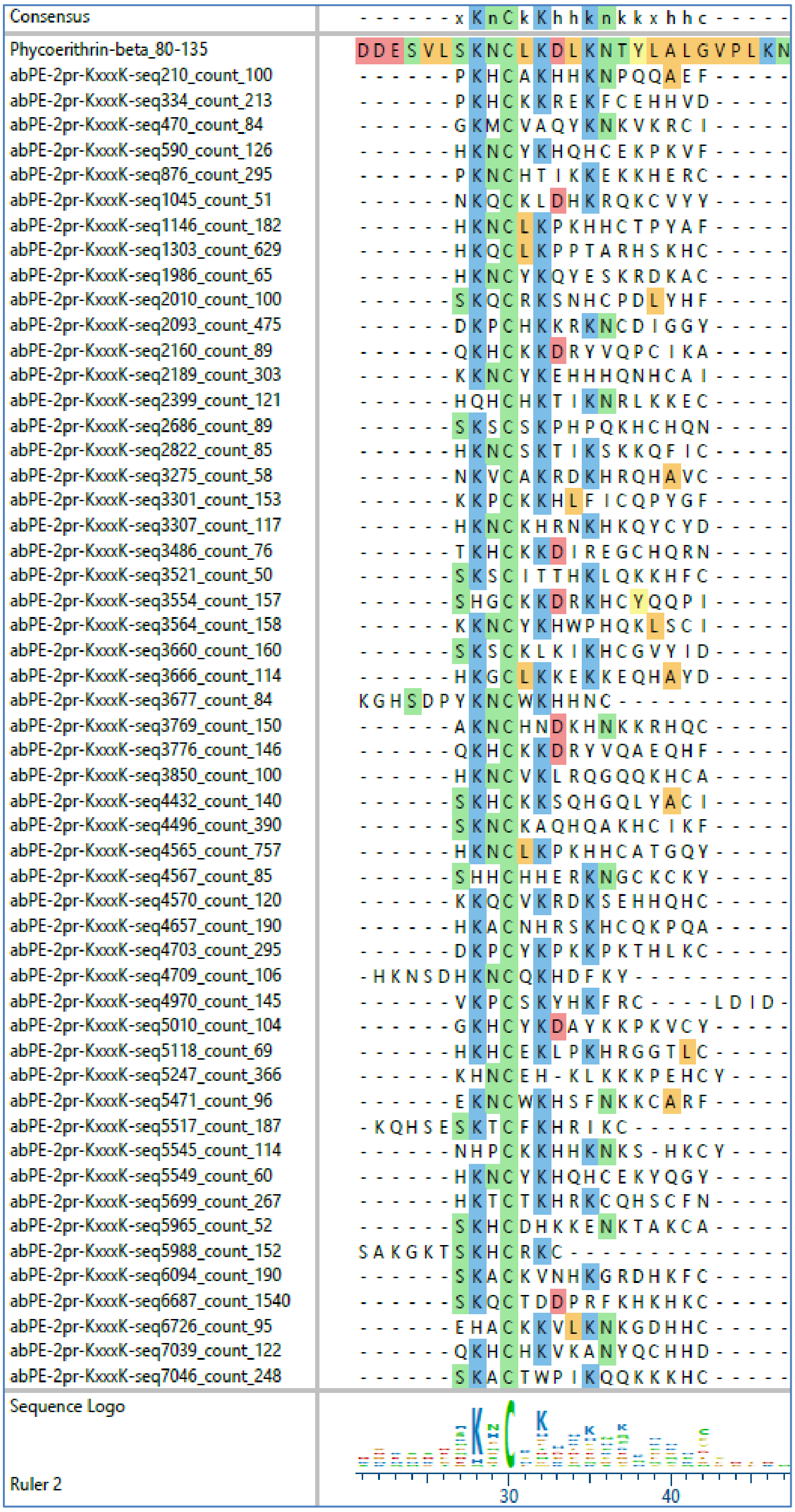
This approach can identify additional sequences by allowing variations of amino acids instead of strict 4-mer motifs. In this case, the search resulted in 7,053 different sequences and 106,870 sequences in total. This amounts to about 30% of the whole dataset in contrast to about 1.2% of the naïve library. The alignment on the right extends the previous one above by only showing the sequences which are found at least 50 times with at least 4 amino acids identity to the antigen sequence. Here it is recognizable that a poly Lys pattern with fixed distances (which matches to the antigen sequence of PE) is probably a better description of the epitope than a fixed motif.

Searches for RxxxK, KxxxR and RxxxR resulted in lower numbers (31,126, 45,354 and 8,907). The peptide epitope structure with three Lys seems to be optimal for binding to the antibody.

Based on this second alignment the epitope would be best described as below.

sKnCxKxxk

The consensus of the shown sequences also reveals that the Cys between the Lys, which is present in ALL sequences, is required for the peptide to mimic the antigen structure for binding by interaction with a C-terminal second Cys.



Consensus	------xKnCkKhhknkkxhhc-----
Phycoerithrin-beta_80-135	DDESVLSKNCLKDLKNTYLALGVPLKN
abPE-2pr-KxxxK-seq210_count_100	------PKHCAKHHKNPQQAEF-----
abPE-2pr-KxxxK-seq334_count_213	------PKHCKKREKFCEHHVD-----
abPE-2pr-KxxxK-seq470_count_84	------GKMCVAQYKNKVKRCI-----
abPE-2pr-KxxxK-seq590_count_126	------HKNCYKHQHCEKPKVF-----
abPE-2pr-KxxxK-seq876_count_295	------PKNCHTIKKEKKHERC-----
abPE-2pr-KxxxK-seq1045_count_51	------NKQCKLDHKRQKCVYY-----
abPE-2pr-KxxxK-seq1146_count_182	------HKNCLKPKHHCTPYAF-----
abPE-2pr-KxxxK-seq1303_count_629	------HKQCLKPPTARHSKHC-----
abPE-2pr-KxxxK-seq1986_count_65	------HKNCYKQYESKRDKAC-----
abPE-2pr-KxxxK-seq2010_count_100	------SKQCRKSNHCPDLYHF-----
abPE-2pr-KxxxK-seq2093_count_475	------DKPCHKKRKNCDIGGY-----
abPE-2pr-KxxxK-seq2160_count_89	------QKHCKKDRYVQPCIKA-----
abPE-2pr-KxxxK-seq2189_count_303	------KKNCYKEHHHQNHCAI-----
abPE-2pr-KxxxK-seq2399_count_121	------HQHCHKTIKNRLKKEC-----
abPE-2pr-KxxxK-seq2686_count_89	------SKSCSKPHPQKHCHQN-----
abPE-2pr-KxxxK-seq2822_count_85	------HKNCSKTIKSKKQFIC-----
abPE-2pr-KxxxK-seq3275_count_58	------NKVCAKRDKHRQHAVC-----
abPE-2pr-KxxxK-seq3301_count_153	------KKPCKKHLFICQPYGF-----
abPE-2pr-KxxxK-seq3307_count_117	------HKNCKHRNKHKQYCYD-----
abPE-2pr-KxxxK-seq3486_count_76	------TKHCKKDIREGCHQRN-----
abPE-2pr-KxxxK-seq3521_count_50	------SKSCITTHKLQKKHFC-----
abPE-2pr-KxxxK-seq3554_count_157	------SHGCKKDRKHCYQQPI-----
abPE-2pr-KxxxK-seq3564_count_158	------KKNCYKHWPHQKLSCI-----
abPE-2pr-KxxxK-seq3660_count_160	------SKSCKLKIKHCGVYID-----
abPE-2pr-KxxxK-seq3666_count_114	------HKGCLKKEKKEQHAYD-----
abPE-2pr-KxxxK-seq3677_count_84	KGHSDPYKNCWKHHNC-----------
abPE-2pr-KxxxK-seq3769_count_150	------AKNCHNDKHNKKRHQC-----
abPE-2pr-KxxxK-seq3776_count_146	------QKHCKKDRYVQAEQHF-----
abPE-2pr-KxxxK-seq3850_count_100	------HKNCVKLRQGQQKHCA-----
abPE-2pr-KxxxK-seq4432_count_140	------SKHCKKSQHGQLYACI-----
abPE-2pr-KxxxK-seq4496_count_390	------SKNCKAQHQAKHCIKF-----
abPE-2pr-KxxxK-seq4565_count_757	------HKNCLKPKHHCATGQY-----
abPE-2pr-KxxxK-seq4567_count_85	------SHHCHHERKNGCKCKY-----
abPE-2pr-KxxxK-seq4570_count_120	------KKQCVKRDKSEHHQHC-----
abPE-2pr-KxxxK-seq4657_count_190	------HKACNHRSKHCQKPQA-----
abPE-2pr-KxxxK-seq4703_count_295	------DKPCYKPKKPKTHLKC-----
abPE-2pr-KxxxK-seq4709_count_106	-HKNSDHKNCQKHDFKY----------
abPE-2pr-KxxxK-seq4970_count_145	------VKPCSKYHKFRC----LDID-
abPE-2pr-KxxxK-seq5010_count_104	------GKHCYKDAYKKPKVCY-----
abPE-2pr-KxxxK-seq5118_count_69	------HKHCEKLPKHRGGTLC-----
abPE-2pr-KxxxK-seq5247_count_366	------KHNCEH-KLKKKPEHCY----
abPE-2pr-KxxxK-seq5471_count_96	------EKNCWKHSFNKKCARF-----
abPE-2pr-KxxxK-seq5517_count_187	-KQHSESKTCFKHRIKC----------
abPE-2pr-KxxxK-seq5545_count_114	------NHPCKKHHKNKS-HKCY----
abPE-2pr-KxxxK-seq5549_count_60	------HKNCYKHQHCEKYQGY-----
abPE-2pr-KxxxK-seq5699_count_267	------HKTCTKHRKCQHSCFN-----
abPE-2pr-KxxxK-seq5965_count_52	------SKHCDHKKENKTAKCA-----
abPE-2pr-KxxxK-seq5988_count_152	SAKGKTSKHCRKC--------------
abPE-2pr-KxxxK-seq6094_count_190	------SKACKVNHKGRDHKFC-----
abPE-2pr-KxxxK-seq6687_count_1540	------SKQCTDDPRFKHKHKC-----
abPE-2pr-KxxxK-seq6726_count_95	------EHACKKVLKNKGDHHC-----
abPE-2pr-KxxxK-seq7039_count_122	------QKHCHKVKANYQCHHD-----
abPE-2pr-KxxxK-seq7046_count_248	------SKACTWPIKQQKKKHC-----
Sequence Logo	
Ruler 2	30 40

PE-antibody epitope location in the protein sequence

This is an overview of the R-PE subunit beta sequence where those regions are marked in yellow, that show enrichment in the statistical analysis of 4-mer motifs in the NGS datasets.

> R-phycoerythrin subunit beta [Prochlorococcus marinus]

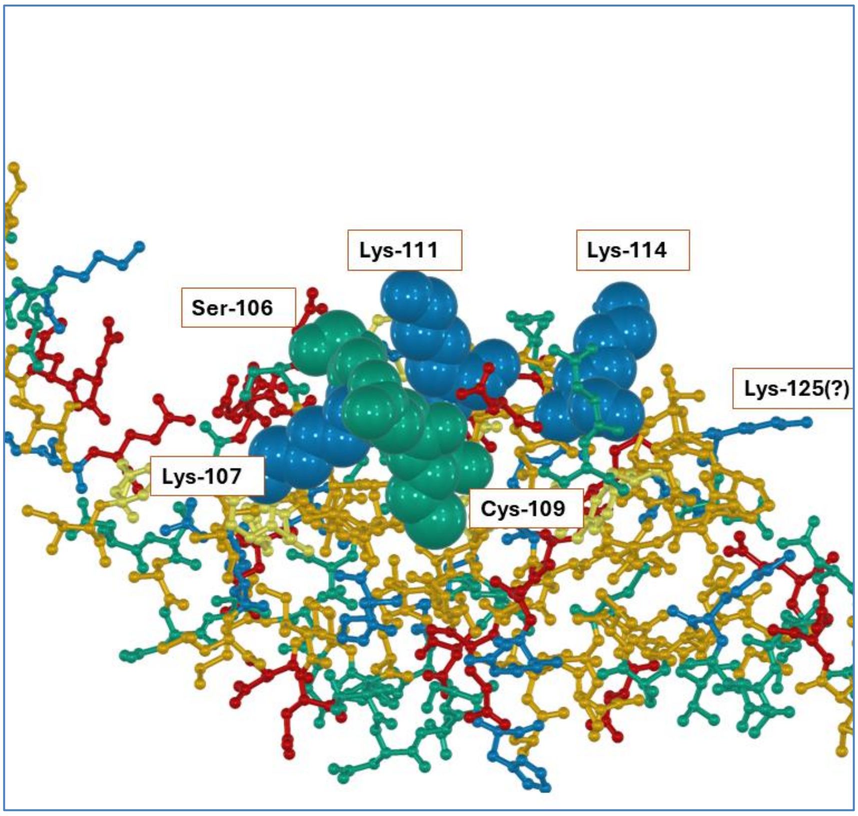
1 MTVSKSNQIL SNDRGLENIS NKNIEDIKEF INTANSRLDA INSITNNSHA IAADAVTAMI
61 CENQDSVNTK ISLDTTNKMS ICLRDGEIIL RIVSYLLISD DESVLS**K**N**C**L **K**DL**K**NTYLAL
121 GVPLKNAIRV FELMRDATIS DLNSTVNSMK GEKEFLPDLI SNTEFQFERI INLLR

Species specificity

Several Phycoerythrins are used in labelling antibodies, but this epitope is specific for the *Prochlorococcus* protein, although many other structures contain neighbouring Lys on their surface.

PE-ab epitope in structure (AlphaFold)

R-phycoerythrin (R-PE) is an intensely bright phycobiliprotein isolated from red algae that exhibits extremely bright red-orange fluorescence and is frequently used, but a crystal structure of R-PE from *Prochlorococcus marinus* has not been reported to the Protein Data Bank. An AlphaFold-structure (A0A1X9PU49-F1) is available. In this structure the epitope's essential residues are shown as spheres. A fourth Lys(125) may also be involved in binding to the antibody, but there is no substantial support from the statistical data.



Contact:

epitopic GmbH – Deutscher Platz 5e - 04103 Leipzig - Germany

E-Mail: info@epitopic.com

Phone +49 341 253 55 160

www.epitopic.com