Electronic Supplementary Data

A novel B- and helper T-cell epitopes-based prophylactic vaccine against *Echinococcus granulosus*

Mohammad M. Pourseif [1,2] , Gholamali Moghaddam [1*] , Hossein Daghighkia [1] , Ahmad Nematollahi [3] , Yadollah Omidi [2,4*]

[1] Department of Animal Sciences, Faculty of Agriculture, University of Tabriz, Tabriz, Iran

[2] Research Center for Pharmaceutical Nanotechnology, Biomedicine Institute, Tabriz University of Medical Sciences, Tabriz, Iran

[3] Department of Pathobiology, Veterinary Collage, University of Tabriz, Tabriz, Iran

[4] Department of Pharmaceutics, Faculty of Pharmacy, Tabriz University of Medical Sciences, Tabriz, Iran

Electronic Supplementary Material: The online version of this article (doi: XXXXXXXXX) contains supplementary material, which is available to authorized users.

* Correspondence authors: Gholamali Moghaddam, Email: ghmoghaddam@tabrizu.ac.ir; Yadollah Omidi, Email: yomidi@tbzmed.ac.ir

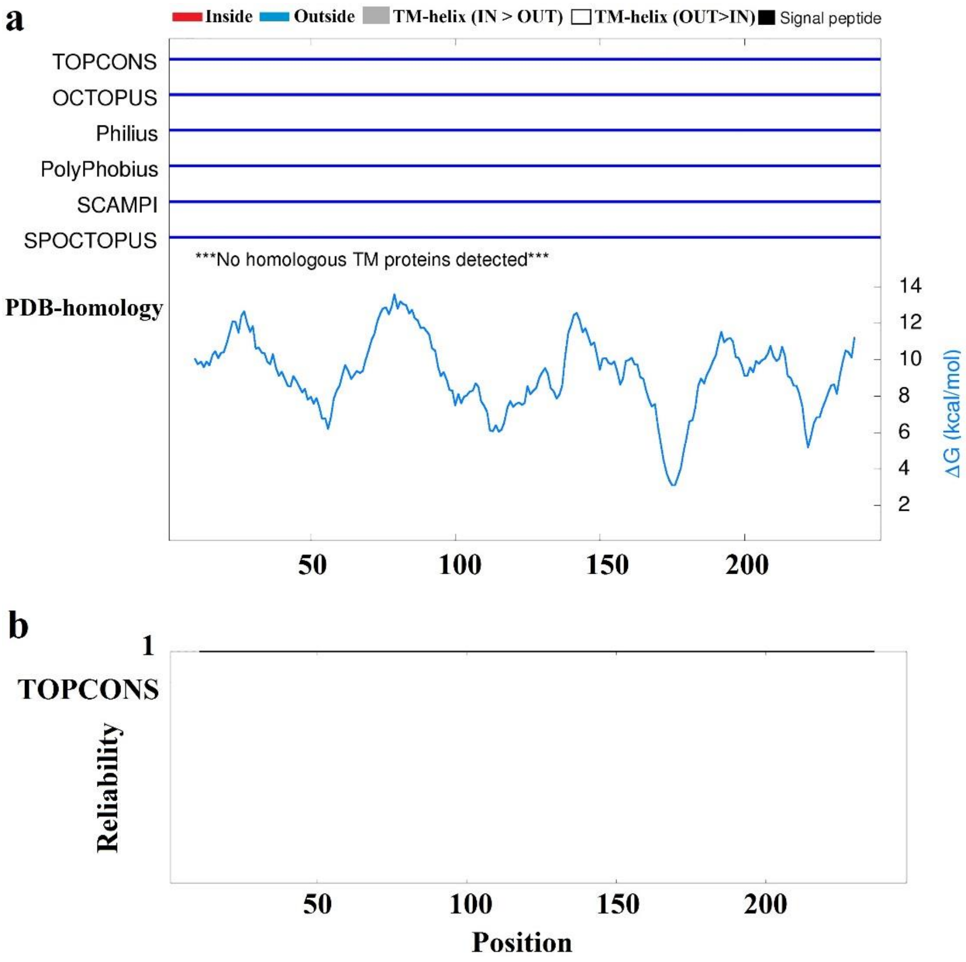


Fig. S1. Prediction of transmembrane topology and ΔG value based on the TOPCONS web-server. **a)** Different prediction algorithms (TOPCONS, OCTOPUS, Philius, PolyPhobius, SCAMPI and SPOCTOPUS). No signal peptide and helical transmembrane were seen in the Eg14-3-3 sequence. **b)** The consensus prediction algorithm (TOPCONS). No transmembrane helix nor signal peptide were predicted.

Fig. S2. Prediction of Eg14-3-3 transmembrane topology using TMpred's ExPASy server. The prediction numerical data were imported into the Excel program. As marked by the yellow arrow, at one position including aa$^{201\text{-}217}$ (inside to outside helix) or aa$^{200\text{-}217}$ (outside to inside helix) was predicted transmembrane helix orientation.

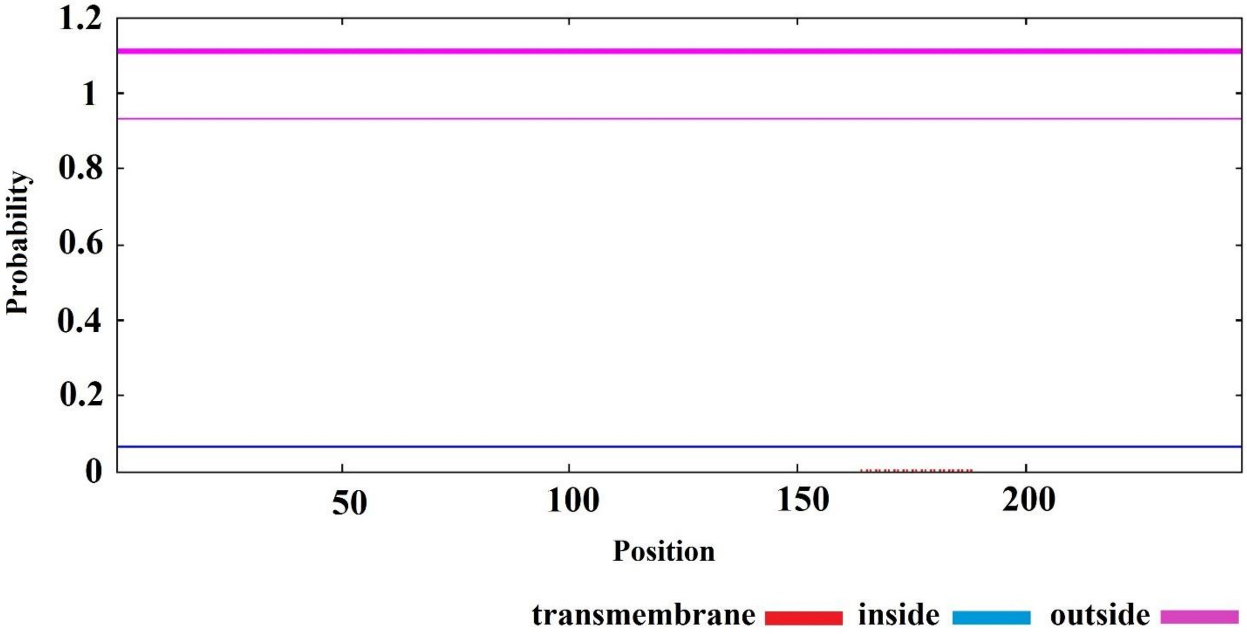


Fig. S3. Transmembrane topology and signal peptide prediction using TMHMM *v*2.0 online server. There is no any signal peptide or transmembrane orientation in the Eg14-3-3 protein sequence.

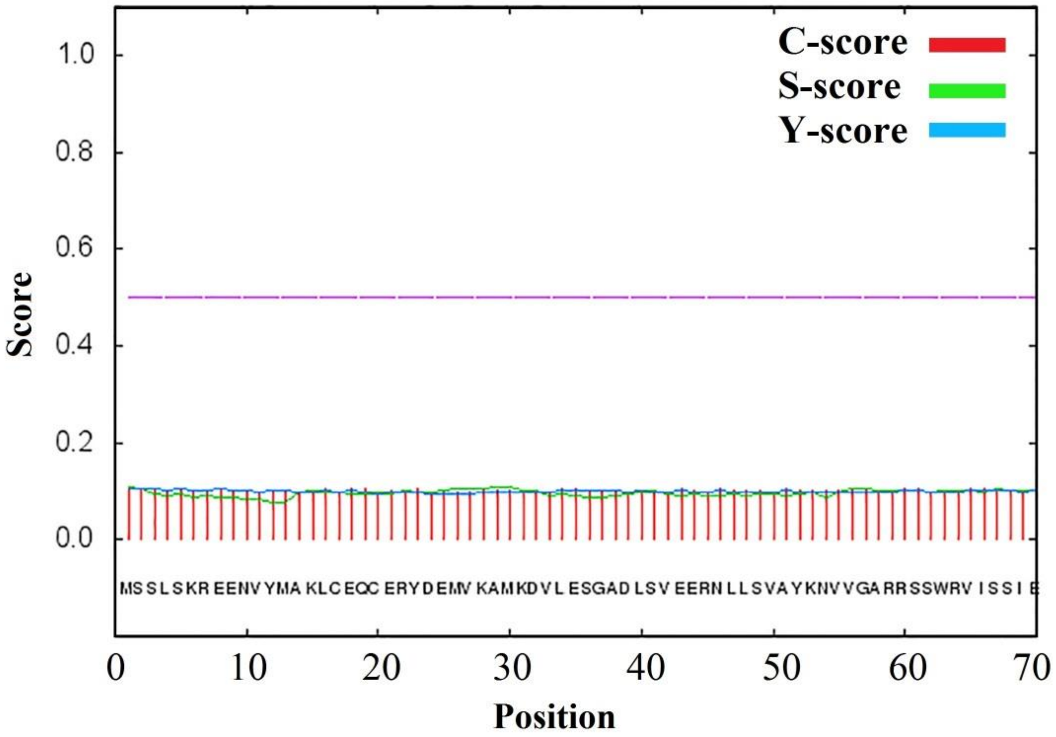


Fig. S4. Signal peptide prediction by means of SignalP web-server. The ‘C-score’ value is related to the cleavage site of the predicted signal peptide. The ‘S-score’ provides a value for the predicted signal peptide and Y-score that is defined as the geometric average between the C- and S-score. None of these values were not statistically more than the threshold.

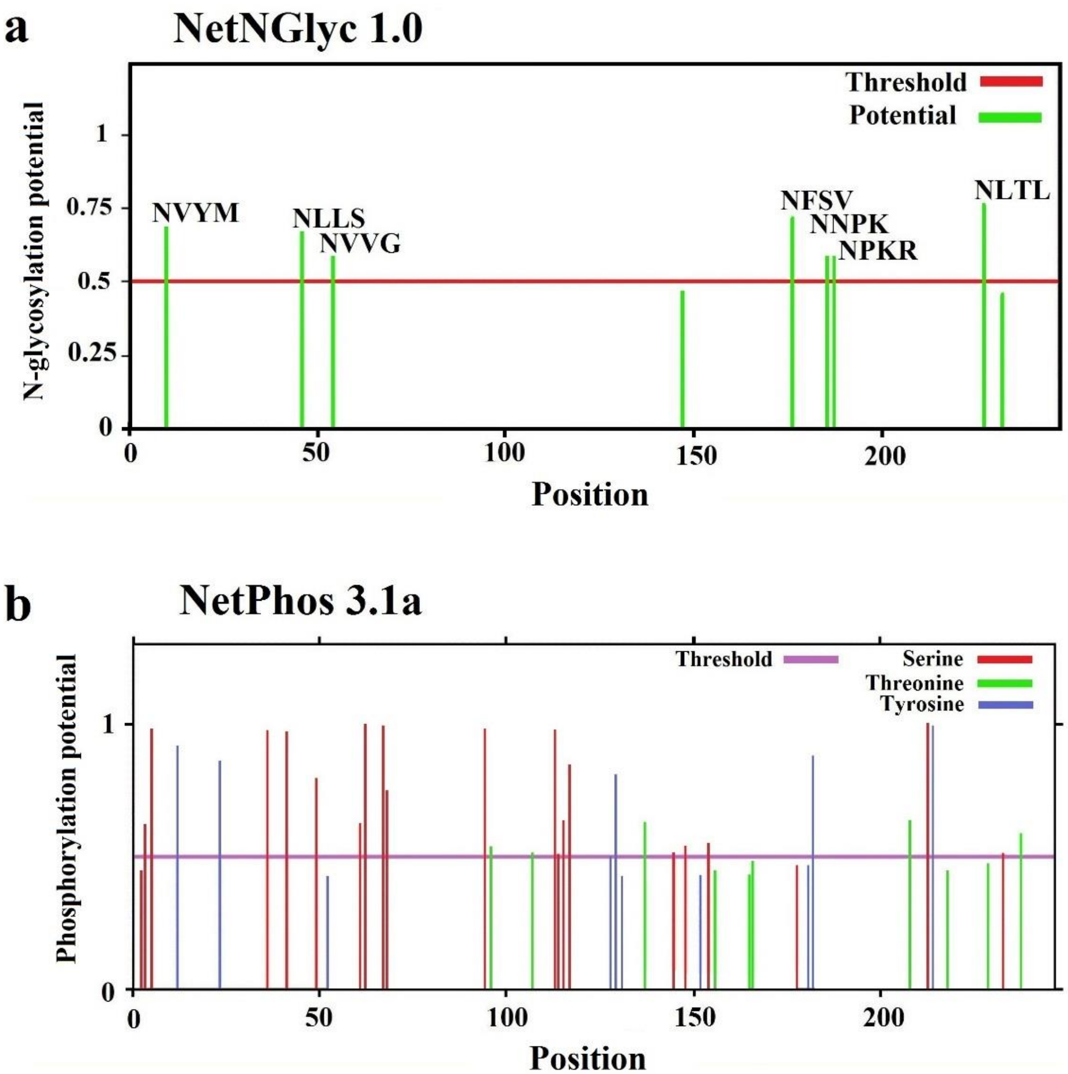
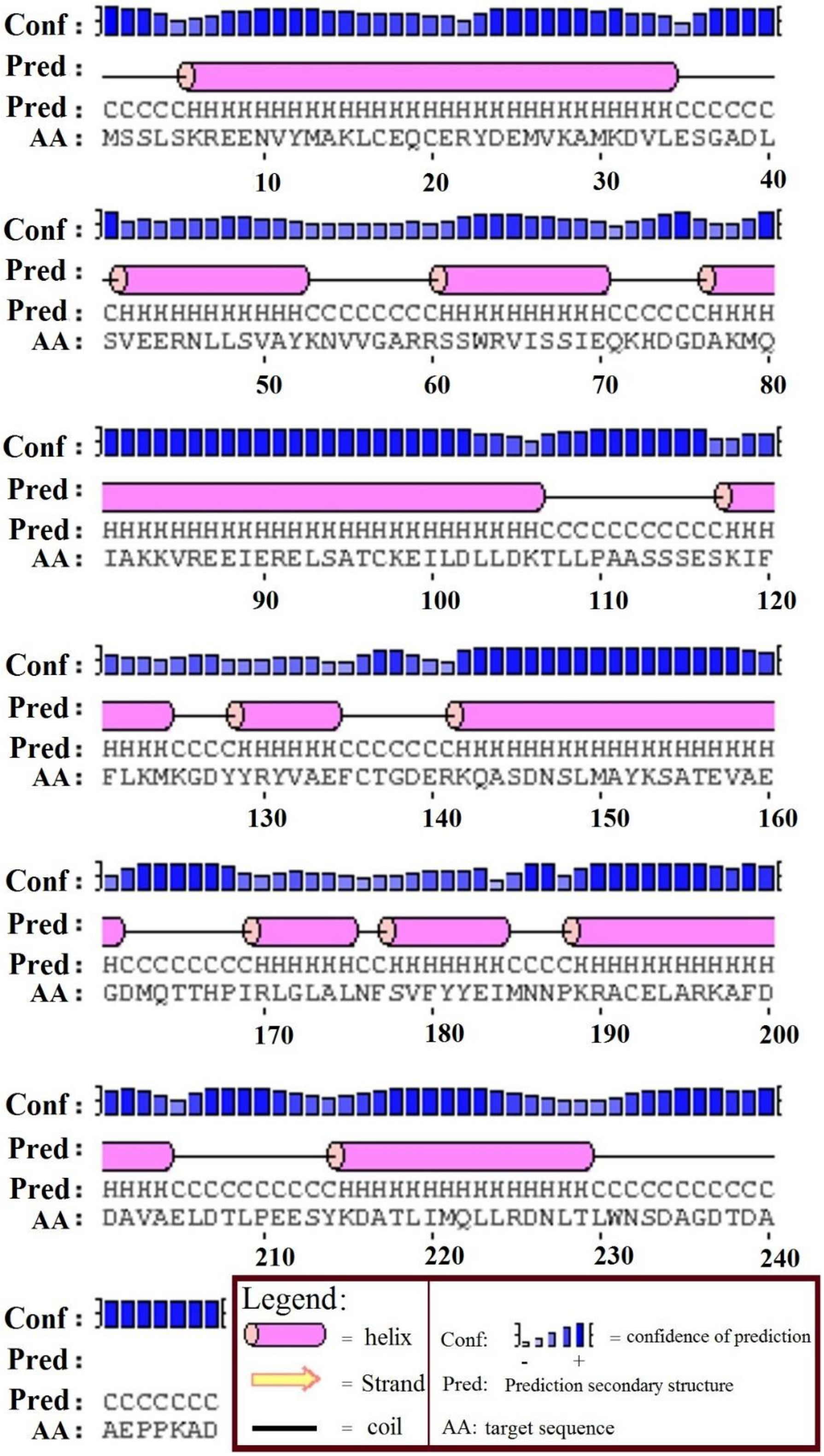


Fig. S5. Prediction of N-linked glycosylation and phosphorylation using NetNGlyc *v*1.0 and NetPhos *v*3.1a web-servers, respectively. **a)** The regions of the sequence that were predicted as possible N-glycosylation are written above the vertical green lines. **b)** The possible serine, threonine and tyrosine phosphorylation sites are indicated as the vertical red, blue and green lines, respectively.



```
Conf :
Pred :
Pred : CCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCC
  AA : MSSLSKREENVYMAKLCEQCERYDEMVKAMKDVLESGADL
                10        20        30        40

Conf :
Pred :
Pred : CHHHHHHHHHHCCCCCCCHHHHHHHHHHCCCCCCHHHH
  AA : SVEERNLLSVAYKNVVGARRSSWRVISSIEQKHDGDAKMQ
                50        60        70        80

Conf :
Pred :
Pred : HHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCHHH
  AA : IAKKVREEIERELSATCKEILDLLDKTLLPAASSSESKIF
                90       100       110       120

Conf :
Pred :
Pred : HHHHCCCCHHHHHHCCCCCCCHHHHHHHHHHHHHHHHHHH
  AA : FLKMKGDYYRYVAEFCTGDERKQASDNSLMAYKSATEVAE
               130       140       150       160

Conf :
Pred :
Pred : HCCCCCCCCHHHHHHCCHHHHHHHCCCCHHHHHHHHHHHH
  AA : GDMQTTHPIRLGLALNFSVFYYEIMNNPKRACELARKAFD
               170       180       190       200

Conf :
Pred :
Pred : HHHHCCCCCCCCCCHHHHHHHHHHHHHHHCCCCCCCCCCC
  AA : DAVAELDTLPEESYKDATLIMQLLRDNLTLWNSDAGDTDA
               210       220       230       240

Conf :
Pred :
Pred : CCCCCCC
  AA : AEPPKAD
```

Fig. S6. The Eg14-3-3 secondary structure. Confidence of prediction for each residue is shown.

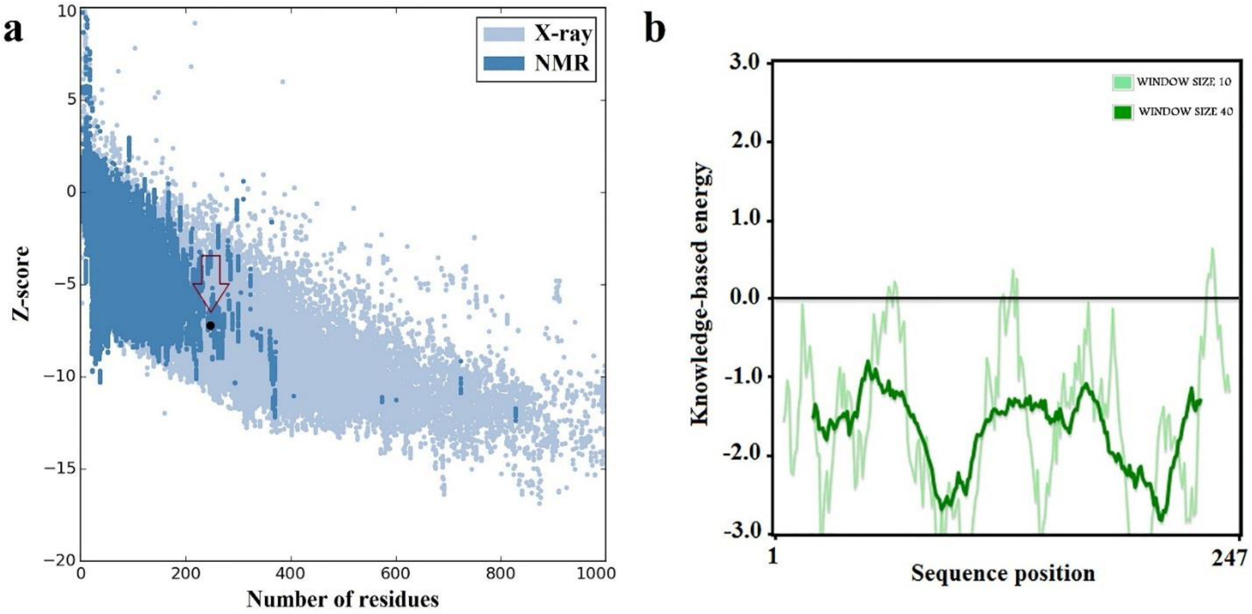


Fig. S7. ProSA-based z-score (**a**) and energy (**b**) plots for the modeled Eg14-3-3. The black dot in the X-ray region indicate the z-score (-7.21) of the model. All parts of the energy plot (in window size 10) exhibit the highly negative energy values. The negative values were related to the stable residues and this confirms the reliability of the modeling.

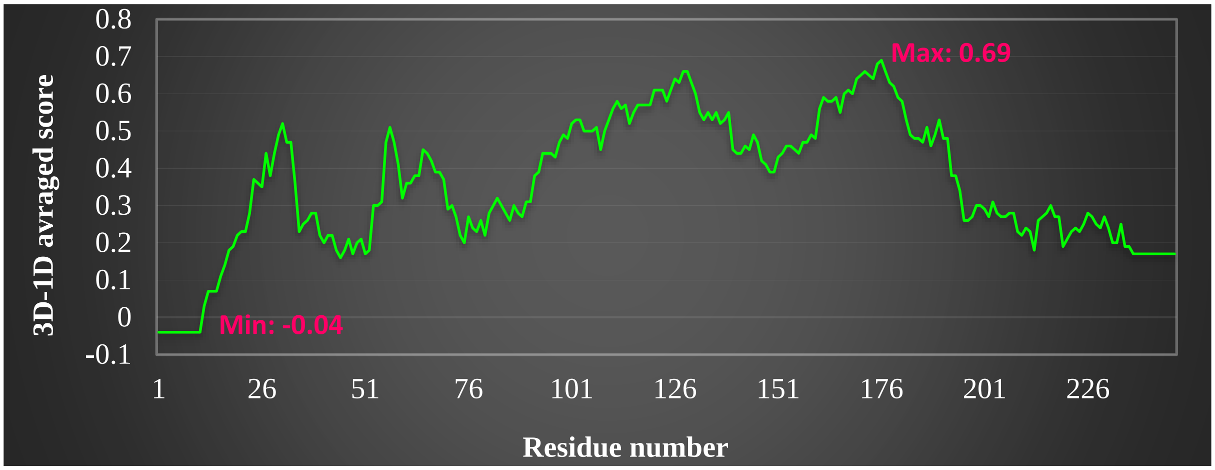


Fig. S8. Validation of quality of the modeled Eg14-3-3 protein. The output of verify3D was plotted based on the 3D-1D average scores for all residues. The minimum value of 3D-1D profile was -0.04 for residues 1-11 (MSSLSKREENV), and the max value of 0.69 relevant to asparagine 176.

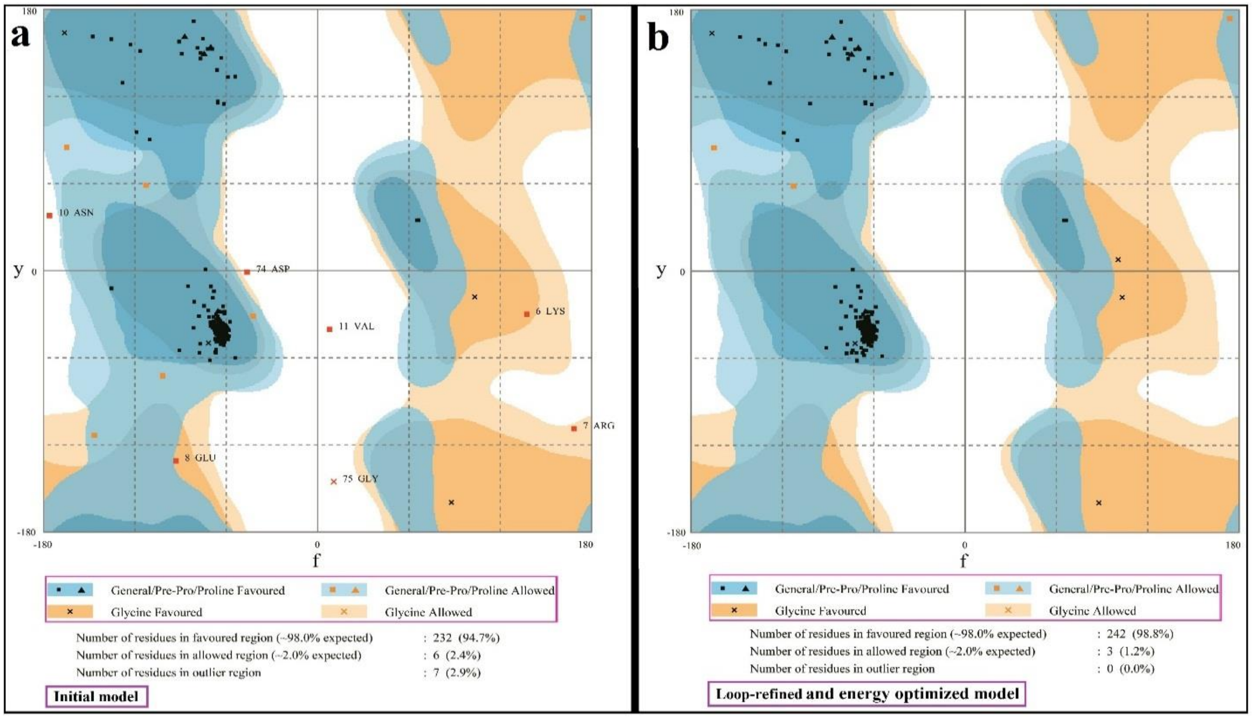


Fig. S9. Ramachandran plot from the computed φ-ψ angels of the protein Eg14-3-3. **a)** Ramachandran plot of the initial predicted model. **b)** Ramachandran plot for the loop-refined and energy minimized model.

Fig. S10. The entropy (Hx) plot of the deduced amino acid sequences of six *E. granulosus* 14-3-3 strains. Amino acid residues with entropy value less than the threshold (>1.0) were known as conserved.

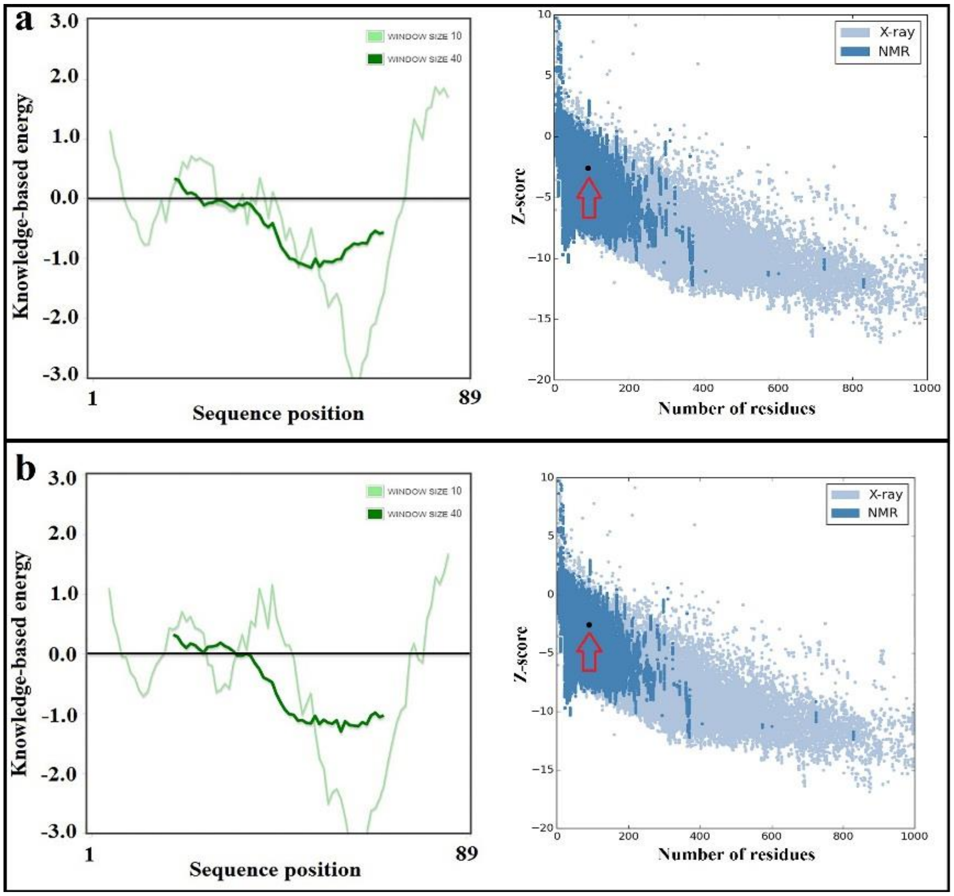
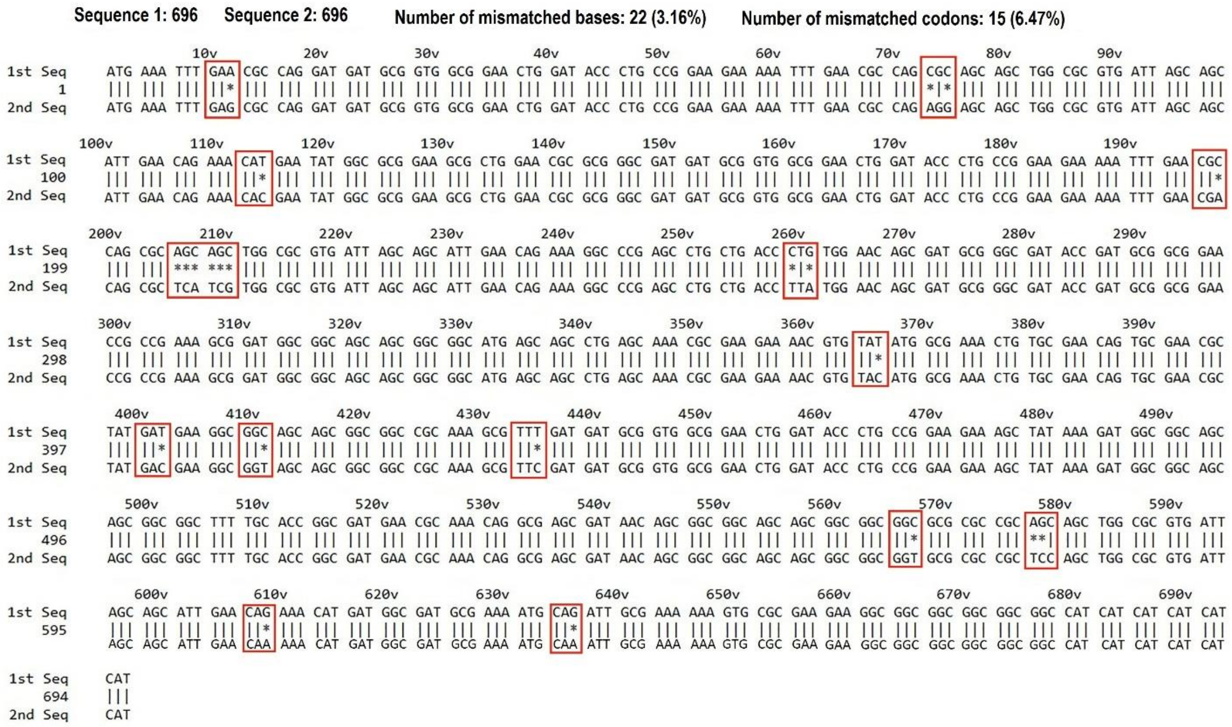
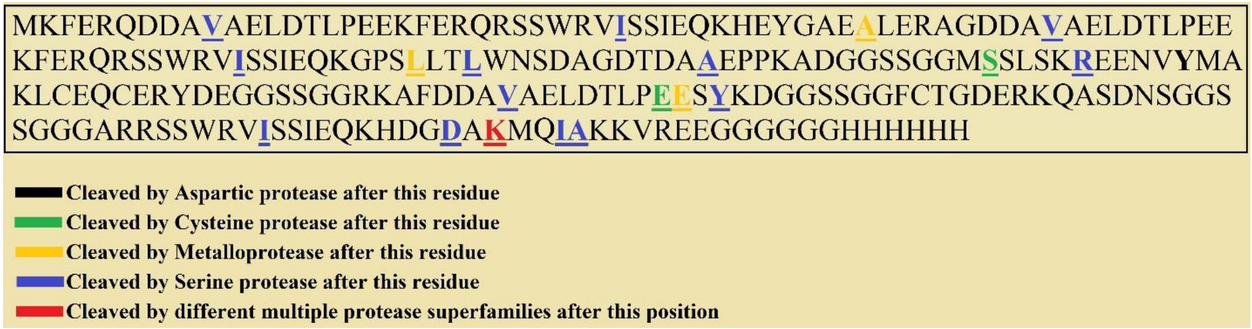


Fig. S11. ProSA-web z-score and energy plot for modeled 3D structures of DLA-DRB1*01501 and 01101 alleles. **a)** The energy plot (left) and ProSA-based z-score (right) of the modeled DLA-DRB1*01501. **b)** The energy plot (left) and ProSA-based z-score (right) of the modeled DLA-DRB1*01101. The negative energy values were correlate to the stable residues. The black dots into the NMR spectroscopy region indicate z-score of the models.



```
Sequence 1: 696    Sequence 2: 696    Number of mismatched bases: 22 (3.16%)    Number of mismatched codons: 15 (6.47%)

                      10v          20v          30v           40v          50v          60v           70v          80v          90v
1st Seq     ATG AAA TTT GAA CGC CAG GAT GAT GCG GTG GCG GAA CTG GAT ACC CTG CCG GAA GAA AAA TTT GAA CGC CAG CGC AGC AGC TGG CGC GTG ATT AGC AGC
      1     ||| ||| ||| ||* ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| *|* ||| ||| ||| ||| ||| ||| ||| |||
2nd Seq     ATG AAA TTT GAG CGC CAG GAT GAT GCG GTG GCG GAA CTG GAT ACC CTG CCG GAA GAA AAA TTT GAA CGC CAG AGG AGC AGC TGG CGC GTG ATT AGC AGC

         100v         110v         120v          130v         140v         150v          160v         170v         180v          190v
1st Seq     ATT GAA CAG AAA CAT GAA TAT GGC GCG GAA GCG CTG GAA CGC GCG GGC GAT GAT GCG GTG GCG GAA CTG GAT ACC CTG CCG GAA GAA AAA TTT GAA CGC
    100     ||| ||| ||| ||| ||* ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||*
2nd Seq     ATT GAA CAG AAA CAC GAA TAT GGC GCG GAA GCG CTG GAA CGC GCG GGC GAT GAT GCG GTG GCG GAA CTG GAT ACC CTG CCG GAA GAA AAA TTT GAA CGA

          200v         210v          220v         230v         240v          250v         260v         270v          280v         290v
1st Seq     CAG CGC AGC AGC TGG CGC GTG ATT AGC AGC ATT GAA CAG AAA GGC CCG AGC CTG CTG ACC CTG TGG AAC AGC GAT GCG GGC GAT ACC GAT GCG GCG GAA
    199     ||| ||| *** *** ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| *|* ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
2nd Seq     CAG CGC TCA TCG TGG CGC GTG ATT AGC AGC ATT GAA CAG AAA GGC CCG AGC CTG CTG ACC TTA TGG AAC AGC GAT GCG GGC GAT ACC GAT GCG GCG GAA

           300v          310v         320v         330v          340v         350v         360v          370v         380v         390v
1st Seq     CCG CCG AAA GCG GAT GGC GGC AGC AGC GGC GGC ATG AGC AGC CTG AGC AAA CGC GAA GAA AAC GTG TAT ATG GCG AAA CTG TGC GAA CAG TGC GAA CGC
    298     ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||* ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
2nd Seq     CCG CCG AAA GCG GAT GGC GGC AGC AGC GGC GGC ATG AGC AGC CTG AGC AAA CGC GAA GAA AAC GTG TAC ATG GCG AAA CTG TGC GAA CAG TGC GAA CGC

             400v         410v         420v          430v         440v         450v          460v         470v         480v          490v
1st Seq     TAT GAT GAA GGC GGC AGC AGC GGC GGC CGC AAA GCG TTT GAT GAT GCG GTG GCG GAA CTG GAT ACC CTG CCG GAA GAA AGC TAT AAA GAT GGC GGC AGC
    397     ||| ||* ||| ||| ||* ||| ||| ||| ||| ||| ||| ||| ||* ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
2nd Seq     TAT GAC GAA GGC GGT AGC AGC GGC GGC CGC AAA GCG TTC GAT GAT GCG GTG GCG GAA CTG GAT ACC CTG CCG GAA GAA AGC TAT AAA GAT GGC GGC AGC

              500v         510v          520v         530v         540v          550v         560v         570v          580v         590v
1st Seq     AGC GGC GGC TTT TGC ACC GGC GAT GAA CGC AAA CAG GCG AGC GAT AAC AGC GGC GGC AGC AGC GGC GGC GGC GCG CGC CGC AGC AGC TGG CGC GTG ATT
    496     ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||* ||| ||| ||| **| ||| ||| ||| ||| |||
2nd Seq     AGC GGC GGC TTT TGC ACC GGC GAT GAA CGC AAA CAG GCG AGC GAT AAC AGC GGC GGC AGC AGC GGC GGC GGT GCG CGC CGC TCC AGC TGG CGC GTG ATT

               600v          610v         620v         630v          640v         650v         660v          670v         680v         690v
1st Seq     AGC AGC ATT GAA CAG AAA CAT GAT GGC GAT GCG AAA ATG CAG ATT GCG AAA AAA GTG CGC GAA GAA GGC GGC GGC GGC GGC GGC CAT CAT CAT CAT CAT
    595     ||| ||| ||| ||| ||* ||| ||| ||| ||| ||| ||| ||| ||| ||* ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
            AGC AGC ATT GAA CAA AAA CAT GAT GGC GAT GCG AAA ATG CAA ATT GCG AAA AAA GTG CGC GAA GAA GGC GGC GGC GGC GGC GGC CAT CAT CAT CAT CAT

1st Seq     CAT
    694     |||
2nd Seq     CAT
```

Fig. S12. Comparison of the vaccine construct mRNA sequence before and after codon optimization. In 15 codons (red boxes), the unfavorable codons were replaced with favorable codons.



MKFERQDDAVAELDTLPEEKFERQRSSWRVISSIEQKHEYGAEALERAGDDAVAELDTLPEE
KFERQRSSWRVISSIEQKGPSLLTLWNSDAGDTDAAEPPKADGGSSGGMSSLSKREENVYMA
KLCEQCERYDEGGSSGGRKAFDDAVAELDTLPEESYKDGGSSGGFCTGDERKQASDNSGGS
SGGGARRSSWRVISSIEQKHDGDAKMQIAKKVREEGGGGGGHHHHHH

▬ Cleaved by Aspartic protease after this residue
▬ Cleaved by Cysteine protease after this residue
▬ Cleaved by Metalloprotease after this residue
▬ Cleaved by Serine protease after this residue
▬ Cleaved by different multiple protease superfamilies after this position

Fig. S13. The predicted cleavage sites of multiple protease families and based on PROSPER web-server. The predicted cleavage sites are colored and described in the legend.

Table S1. The modeling validation values for the MHC 3D-models

Class II MHC alleles	**C-score***	**TM-score¶**	**RMSD**	**ProSA z-score**
DLA-DRB1*01501	1.53	0.93±0.06	1.0±1.0Å	-2.59
DLA-DRB1*01101	1.50	0.92±0.06	1.0±1.0Å	-2.59

* C-score is typically ranged between five and two, where the higher values imply for more reliable model. ¶ Normal range of TM-score is 0-1. TM-score value more than 0.5 is relate to the model with the correct topology.

Table S2. The parameters of codon usage bias before and after vaccine construct optimization.

Parameter	Before optimization	After optimization
CAI	0.95	1.0
Nc	13.52	15.75
tAI	0.3639	0.3661
Overall G/C content (%)	59.05	58.33
G/C content at 1st place (%)	65.94	65.51
G/C content at 2nd place (%)	47.84	47.84
G/C content at 3rd place (%)	63.36	61.63

CAI: Codon adaptation index; **Nc**: Effective number of codons; **tAI**: tRNA adaptation index

Table S3. The predicted Eg14-3-3 cleavage sites of individual proteases and the sequences

Enzyme superfamily	Protease type	Cleavage sequence	position	Score
Aspartic protease	HIV-1 retropepsin	ENVY↓MAKL	122	1.14
Cysteine protease	Cathepsin K	TLPE↓ESYK	157	1.17
	Cathepsin K	GGMS↓SLSK	112	1.14
Metalloprotease	Matrix metallopeptidase-9	GDAK↓MQIA	210	1.19
	Matrix metallopeptidase-9	GPSL↓LTLW	84	1.16
	Matrix metallopeptidase-9	GAEA↓LERA	44	1.11
	Matrix metallopeptidase-9	LPEE↓SYKD	158	0.95
Serine protease	Elastase-2	WRVI↓SSIE	74	1.07
	Elastase-2	WRVI↓SSIE	198	1.07
	Elastase-2	WRVI↓SSIE	31	1.07
	Elastase-2	KMQI↓AKKV	213	1.03
	Elastase-2	MQIA↓KKVR	214	0.98

↓ represents the proteases cleavage sites