frontiers in Genetics

ORIGINAL RESEARCH
published: 15 March 2022
doi: 10.3389/fgene.2022.854571

GLTM: A Global-Local Attention LSTM Model to Locate Dimer Motif of Single-Pass Membrane Proteins

Quanchao Ma[1], *Kai Zou*[1], *Zhihai Zhang*[1] *and Fan Yang*[1,2]*

[1]School of Communications and Electronics, Jiangxi Science and Technology Normal University, Nanchang, China, [2]Artificial Intelligence and Bioinformation Cognition Laboratory, Jiangxi Science and Technology Normal University, Nanchang, China

Single-pass membrane proteins, which constitute up to 50% of all transmembrane proteins, are typically active in significant conformational changes, such as a dimer or other oligomers, which is essential for understanding the function of transmembrane proteins. Finding the key motifs of oligomers through experimental observation is a routine method used in the field to infer the potential conformations of other members of the transmembrane protein family. However, approaches based on experimental observation need to consume a lot of time and manpower costs; moreover, they are hard to reveal the potential motifs. A proposed approach is to build an accurate and efficient transmembrane protein oligomer prediction model to screen the key motifs. In this paper, an attention-based Global-Local structure LSTM model named GLTM is proposed to predict dimers and screen potential dimer motifs. Different from traditional motifs screening based on highly conserved sequence search frame, a self-attention mechanism has been employed in GLTM to locate the highest dimerization score of subsequence fragments and has been proven to locate most known dimer motifs well. The proposed GLTM can reach 97.5% accuracy on the benchmark dataset collected from Membranome2.0. The three characteristics of GLTM can be summarized as follows: First, the original sequence fragment was converted to a set of subsequences which having the similar length of known motifs, and this additional step can greatly enhance the capability of capturing motif pattern; Second, to solve the problem of sample imbalance, a novel data enhancement approach combining improved one-hot encoding with random subsequence windows has been proposed to improve the generalization capability of GLTM; Third, position penalization has been taken into account, which makes a self-attention mechanism focused on special TM fragments. The experimental results in this paper fully demonstrated that the proposed GLTM has a broad application perspective on the location of potential oligomer motifs, and is helpful for preliminary and rapid research on the conformational change of mutants.

Keywords: single-pass membrane protein, dimer motif, Bi-LSTM network, self-attention mechanism, motif localization model

OPEN ACCESS

Edited by:
Lei Wang,
Changsha University, China

Reviewed by:
Xiaoyong Pan,
Shanghai Jiao Tong University, China
Yongxian Fan,
Guilin University of Electronic Technology, China

***Correspondence:**
Fan Yang
kooyang@aliyun.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *14 January 2022*
Accepted: *14 February 2022*
Published: *15 March 2022*

Citation:
Ma Q, Zou K, Zhang Z and Yang F (2022) GLTM: A Global-Local Attention LSTM Model to Locate Dimer Motif of Single-Pass Membrane Proteins. Front. Genet. 13:854571. doi: 10.3389/fgene.2022.854571

INTRODUCTION

Single-pass membrane proteins are one of the most widely classified membrane proteins, composed of a single transmembrane™ helix and several water-soluble domains, and play an important role in cell signaling, motility, and material transport (Rawlings 2016). Compared with the active state of the multi-pass membrane protein is located within the TM helical bundle, the single TM helix of single-pass membrane protein was initially considered as a merely hydrophobic anchor (Zviling et al., 2007). However, the TM helix of single-pass membrane protein has been verified in making crucial contributions to the protein-protein interaction in recent years.

The intramembrane helix-helix interaction of single-pass membrane protein was firstly confirmed in the dimerization process of human glycophorin A (GpA). In the 3D model for the homo-dimer of human GpA, researchers observed the most helix contact points occurred in the GxxxG motif of TMD (Russ and Engelman 2000). Moreover, the statistical result indicated that the GxxxG motif was one of the significant expression residue pairs in the TM domain (Senes et al., 2000), and these single-pass membrane proteins have a high homo-dimerization tendency when their TM domain contains GxxxG motif (Brosig and Langosch 1998). Except for the GxxxG motif, the polar residue and the leucine zipper also confirmed their irreplaceability in the assembly of oligomeric complexes (Li et al., 2012). The interhelical hydrogen bond of the polar residue directly influences their dimerization degree (LaPointe et al., 2013). The leucine zipper is a $(abcdefg)_n$ heptad repeat motif with leucine at every fourth position and hydrophobic residues at every first position. This "knobs-into-holes" type of side-chain packing facilitates self-associates of the TM domain (Oates et al., 2010). Significantly, the conformational change of single-pass membrane protein as typically receptor activation basis selectively regulated cellular signaling (Hubert et al., 2010). Many diseases are directly related to the dysfunction of transmembrane receptor proteins, research of oligomers offers the opportunity to design drug targets and develop new pharmaceuticals (Cymer and Schneider 2010).

The amino acid residues frequency of the TM domain was used to distinguish different homo-oligomer forms in the earliest oligomer prediction model (Song and Tang 2005); their prediction results confirmed the importance of residue composition for protein quaternary structures. To avoid losing important sequence context information of protein sequence, the pseudo-amino acid composition (PseAAC) was proposed to replace the simple amino acid composition (Zhang et al., 2006). Discrete wavelet transformation was used to decompose digit signals of protein primary structure into different coefficients, and screen out effective global context features (Qiu et al., 2011). This global feature description method combined with a decision-tree algorithm obtained outstanding prediction accuracy (Sun et al., 2012). Moreover, the functional domain was discovered to be involved in molecular evolution in recent years. The functional domain information has been confirmed to improve the prediction performance, but the application of these oligomer prediction models was limited in the poor interpretability. For single-pass membrane proteins, an interpretability motif discovery approach was employed to locate their potential oligomer motifs by corresponding oligomer prediction results.

In previous functional motif detection studies, researchers mainly adopted rigorous statistical formulation to search for overexpression subsequence patterns (Liang et al., 2012). TMSTAT directly calculated the frequency of all pairs and triplets of residues to screen out overexpression subsequence patterns in the TM domain (Senes et al., 2000). A regular expressions algorithm was used to more precisely specify special residues position and interval size in SLiMFinder (Edwards et al., 2007). As researchers realized the complexity of nearby residues dependence, Markovian models were gradually used to discover potential motif patterns, such as NestedMICA (Dogruel et al., 2008), weighted hmm (Song and Gu 2015), and HH-MOTiF (Prytuliak et al., 2017). Note that these oligomer motifs as biologically defined anchors or landmarks are limited in a sequence interval. The discriminative motif discovery models DEME (Redhead and Bailey 2007) and DlocalMotif (Mehdi et al., 2013) introduced spatial confinement scores of each subsequence pattern to distinguish unrelated subsequence patterns and local functional motifs. DiMotiF proposed peptide-pair encoding (PPE) to probabilistic segmentation variable-length subsequence patterns and screened out positively related subsequences as potential motifs after annotating possible secondary structures of these subsequences (Asgari et al., 2019). Although these above search algorithms have strong statistical analysis ability to detect subtle subsequence pattern signals from large datasets, these motif discovery approaches cannot define their corresponding biological function for discovered subsequence patterns.

In this paper, we propose a motif localization model called GLTM to locate potential dimer motifs in the dimer prediction process. The Global-Local Bi-LSTM structure was the fundamental component of our motif localization model, and this idea of bilayer structure referred to the influence of highly conserved subsequence patterns and TM domain context information on oligomerization. Combined with the advantage of a Global-Local structure and the character of one-hot encoding, GLTM achieved a new data enhancement on the data preprocessing module. Additionally, new positional penalization was proposed to encourage a self-attention mechanism focused on known subsequence patterns. In the benchmark dataset, GLTM reached 97.5% accuracy and successfully located most key residue with self-focus and position penalization. Moreover, we discuss the existing deficiencies and application prospects of the motif localization model in the dimerization study of residue mutations.

MATERIALS AND METHODS

Dataset

The Membranome database was the first comprehensive resource on single-pass membrane proteins and is widely used to assist

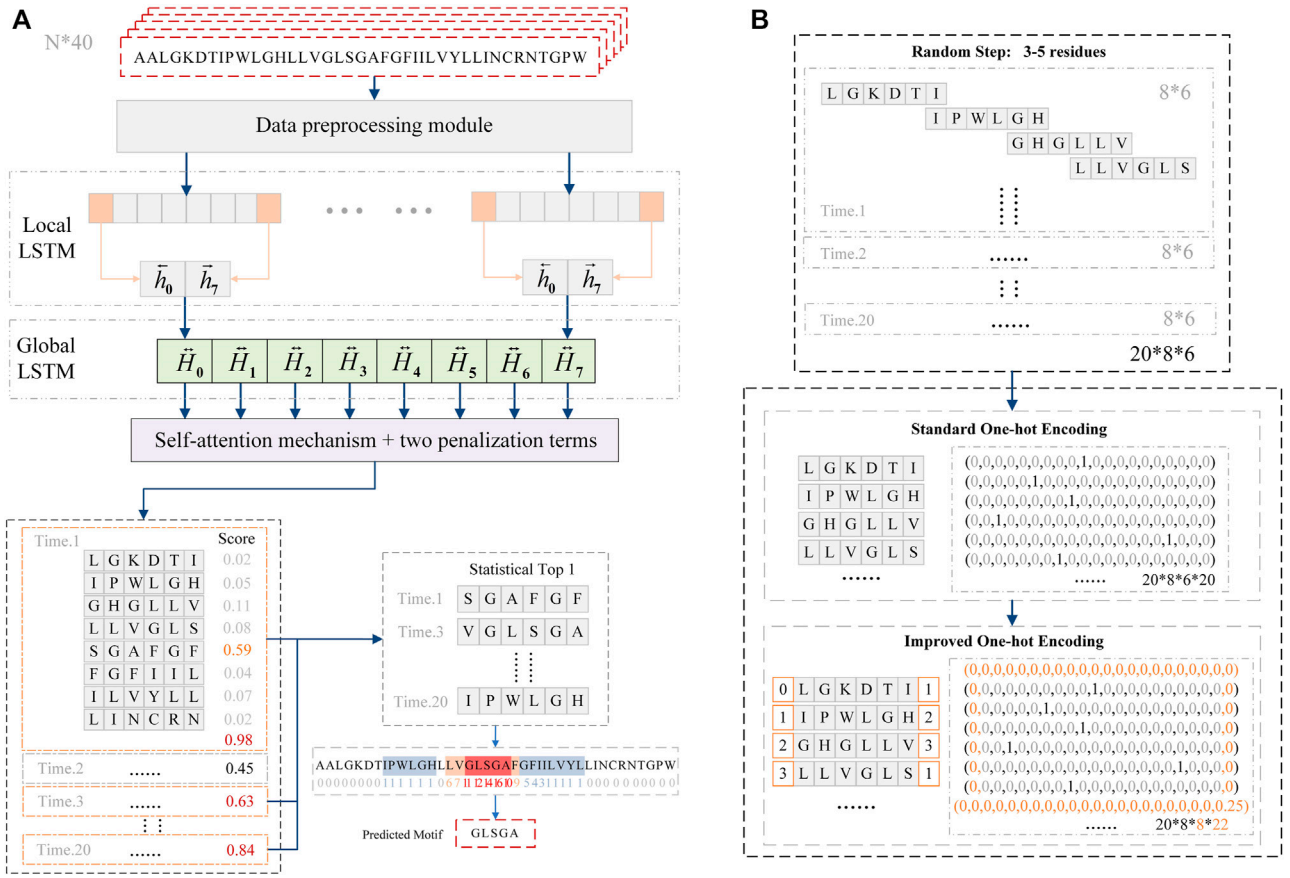


FIGURE 1 | (A) GLTM model composed of data preprocessing module, local Bi-LSTM layer, global Bi-LSTM layer, and self-attention layer. After continuous twenty times prediction, the model chose the subsequence fragment that was predicted more than 10 times as potential motif. **(B)** Data preprocessing module used no-fixed step and improved one-hot encoding to encode original sequence fragment.

analysis and computational modeling of single-pass membrane protein and their complexes (Lomize et al., 2017). The Membranome database collects and compiles diverse data of single-pass membrane proteins, including amino acid sequence, domain architecture, protein topology, and oligomeric states. More importantly, Membranome contains known key residues involved in the homo-dimerization interface according to both mutagenesis studies and computational models.

A new benchmark dataset was established and used for training and testing our motif localization model. Firstly, 334 homo-dimers, which were verified by nuclear magnetic resonance (NMR), mutagenesis experiments, crystal structures of dimers, or other validation methods of TM helix association, were collected from Membranome. Secondly, the orthologs of these 334 homo-dimers with similar oligomerization tendencies were collected from UniProt. Thirdly, chosen dimer motifs were spatially confined in the TM domain, and the C-terminal region of the TM domain participated in helix-helix interactions. Forty residues length of dimer fragment and no-dimer fragment were intercepted from each collected single-pass membrane protein sequence. Finally, the R_{1937} benchmark dataset collected 524 dimer fragments, 1,413 no-dimer fragments, and 24 known motif positions based on 70% maximal identity.

Construction of GLTM Model

In bioinformatics areas, machine learning models widely used k-mers as the protein sequences representation method. Fixed-length subsequences were segmented from the original sequence and regarded as units of biological sequences to encoding in the k-mers treatment method. However, the directly one-hot encoding for subsequence units ignores these strong coupling effects between different positions in the oligomer research of TM protein (Liang et al., 2012). This means that the representation method of short sequence fragments needs to intensify the context information of the TM domain for the oligomer prediction task. Hence, an improved k-mers treatment method was proposed to intensify the independence of every residue based on Global-Local Bi-LSTM bilayer structure.

GLTM consists of the data preprocessing module, local Bi-LSTM layer, global Bi-LSTM layer, and self-attention layer (**Figure 1A**). The first data preprocessing module used the random step selection approach to segment the original sequence and used improved one-hot encoding to represent

these repeated expression residues. Standard one-hot encoding used independent binary vector dimensions to respectively represent twenty standard amino acids (Jing et al., 2020). The K length of local subsequences was converted to a k*20 binary vector by standard one-hot encoding. Our improvement strategy takes advantage of the LSTM network, memory cell of LSTM accepts previous output and cell states as input, and transmits current output and cell states to the next memory cell, this Bi-LSTM structure effectively utilizes sequence context information. Referred to the idea of one-hot encoding, two new binary vector dimensions were proposed to represent repeatedly residues information between contiguous windows, and two window states were appended in every local window to represent repetitious residues numbers. Therefore, the bidirectional feature extraction process preferentially accepted repetition residues information on the local Bi-LSTM layer, and original local subsequences were encoded to k+2*22 binary vectors (**Figure 1B**).

After the data preprocessing module finished subsequences encoding, the encoded vectors directly input into their corresponding local window in the local Bi-LSTM layer. The next global Bi-LSTM layer only accepted the final state output of every local window to extract oligomerization features. Significantly, the weight redistribution process of the self-attention mechanism was the most critical function to locate motif. In order to redress these false weight redistribution processes, new penalization terms were proposed and applied in the last self-attention layer.

Two Penalization Terms in Attentional Mechanism

The self-attention mechanism was widely applied in deep learning, and the redistributive weight of subsequence represented its importance degree for prediction results. Hence, in our motif localization model GLTM, the highest weight of local subsequence was regarded as the potential oligomer motif. When well-trained, GLTM had high prediction accuracy in recognizing dimer fragments. However, accurately locating motifs was always difficult in our previous experiments. This underlying problem, named shortcut learning, is a common deep learning symptom. Shortcut learning typically shows that the deep learning model usually chooses unintended features in prediction results without restricted conditions. Position penalization and self-focus penalization terms were proposed to reduce these fault localization of unintended subsequence patterns.

$$A(x) = softmax\left(W_{s2}tanh\left(W_{s1}H(x)^T\right)\right) \tag{1}$$

GLTM randomly chooses n local window numbers from each sequence fragment, and the feature number of a local window is set as u in each unidirectional. Global Bi-LSTM hidden state $H(x)$ is a weight matrix with a shape of n-by-$2u$. The calculation of annotation vector $A(x)$ needs to set an arbitrary hyperparameter d_a. The weight matrix W_{s1} is sized d_a-by-$2u$, and the matrix W_{s2} has the shape 1-by- d_a. The $softmax(*)$ ensures all elements of annotation vector $A(x)$ sum up to 1.

$$s_i = \frac{e^{c-\left|cen(x,i)-l(x)\right|}}{\sum_n e^{c-\left|cen\left(x,j\right)-l(x)\right|}}, \quad if\ l(x) \neq \varnothing \tag{2}$$

$$S(x) = (s_1, s_2, \cdots s_n) \tag{3}$$

The window position score vector $S(x)$ of these known dimer motifs was calculated in the data preprocessing module. Symbol c is an arbitrarily constant parameter, $cen(x)$ represents the window center-positive of corresponding local subsequence, and $l(x)$ is the center of these known oligomer motifs.

$$P(x) = \begin{cases} \left\|A(x)A(x)^T - I\right\|_2^2, if\ \ s = \varnothing \\ \left\|S(x) - A(x)\right\|_2^2, if\ \ s \neq \varnothing \end{cases} \tag{4}$$

$$L(\theta) = \arg\min_{\theta}\Big(\sum_{i=1}^{m}\Big(\left\|y_i - sigmoid\left(A(x_i,\theta)H(x_i,\theta)\right)\right\|_2^2 + \alpha P(x_i,\theta)\Big)\Big) \tag{5}$$

Self-focus penalization term enhances single-window weight by minimizing the disparity between $A(x)A(x)^T$ and an identity matrix. Position penalization is used to learn known motif distribution by minimizing the disparity between annotation vector $A(x)$ and window position score vector $S(x)$ for these known dimer motifs.

RESULTS AND DISCUSSION

Visualization Result of 26 Known Dimers

In order to verify our model performance, we visualized prediction results and localization results for these containing key residues sequences in **Figure 2**. Note that the same sequence fragment has hundreds of digital matrix representations in the encoding stage. GLTM chose the highest weight local subsequence as a predicted dimer motif when this sequence representation was predicted to dimers and repeated this process twenty times to obtain the more robust localization result. Three color regions were used to mark different localization degrees for the dimer motif, the blue region represents that a subsequence has been predicted to be a dimer motif, the orange region represents more than five predictions as a dimer motif. The subsequences with the most robust prediction result, predicted more than 10 times, comprise the red region. These key residues involved in known dimerization are signalled by a black underline.

We show the prediction performance of GLTM with the different window size and number parameters in **Table 1**, and three evaluation indices were both more than 90% in all experiments. Most known key residues were steadily located in visualization results, in particular for the GxxxG motif of glycophorin A and YxxxxT motif of ζζ which belong to these overexpression subsequence patterns. Only mere unconventionality motifs were successfully located. It may cause by the scarcity of special dimer samples, and this guess was repeatedly verified in the following experiments.

Effect of Two Penalization Terms

In previous experiments, we discovered these successfully located motifs lower than a quarter of the known key residues. In order to

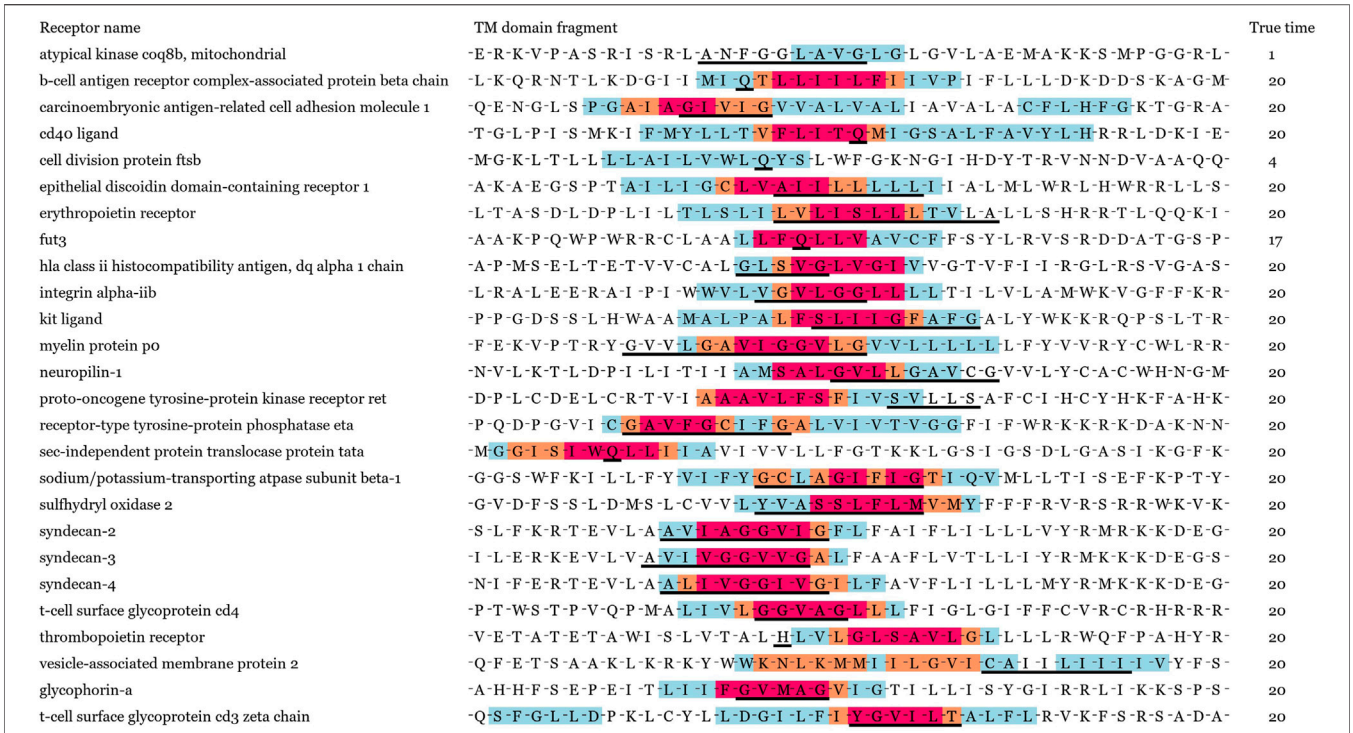


FIGURE 2 | Visualization result of 26 known dimers. True time represents the total true prediction number in twenty prediction results. Black solid line shows key residues for known dimerization process. Different levels of potential motifs have been labeled in red, orange, and blue, respectively. The red denotes to the most important residues in dimerization.

TABLE 1 | Accuracy performance of the model with different window size and window number.

Window number	5 residue lengths			6 residue lengths			7 residue lengths		
	Accuracy	Precious	Recall	Accuracy	Precious	Recall	Accuracy	Precious	Recall
9	0.966	0.934	0.942	0.971	0.936	0.956	0.969	0.932	0.954
10	0.96	0.921	0.932	0.974	0.956	0.947	0.971	0.938	0.954
11	0.961	0.926	0.928	0.965	0.934	0.936	0.974	0.942	0.96
12	0.975	0.94	0.968	0.973	0.929	0.969	0.974	0.945	0.959

enhance the localization accuracy, we proposed two penalization terms to reduce mislocated subsequences, one was self-focus penalization, and the other was position penalization. The self-focus penalization was proposed to distinguish the critical local subsequence in the weight redistribution process. However, diversified oligomer motif localization only relied on self-focus penalization was insufficient. Position penalization was used to encourage the local window weight distribution to approximate the corresponding motif position distribution for these known dimer motifs.

In order to compare the localization performance with different penalization combinations, we showed the localization results of part known dimer sequences in **Figure 3**. Moreover, we drew the located subsequences position distribution of these dimer fragments and no-dimer fragments in **Figure 4**. Obviously, without self-focus penalization and position penalization, the located subsequence distribution for dimer fragment and no-dimer fragment had the same crest position. This means that the weight redistribution process focused on the specific position information rather than subsequence patterns. This tendency deviated from our oligomer motif localization principle. Two penalizations were both successfully reduced the unintended feature extraction for specific position information. However, part end-terminal subsequences were mislocated as potential motifs only with self-focus penalization. With self-focus and position penalization, GLTM reaches outstanding localization accuracy and stability in motif localization tasks.

Dimer Motif Localization of TNF Receptor Superfamily

The tumor necrosis factor receptors superfamily (TNFRSF) is one of the most important single-pass membrane protein families.

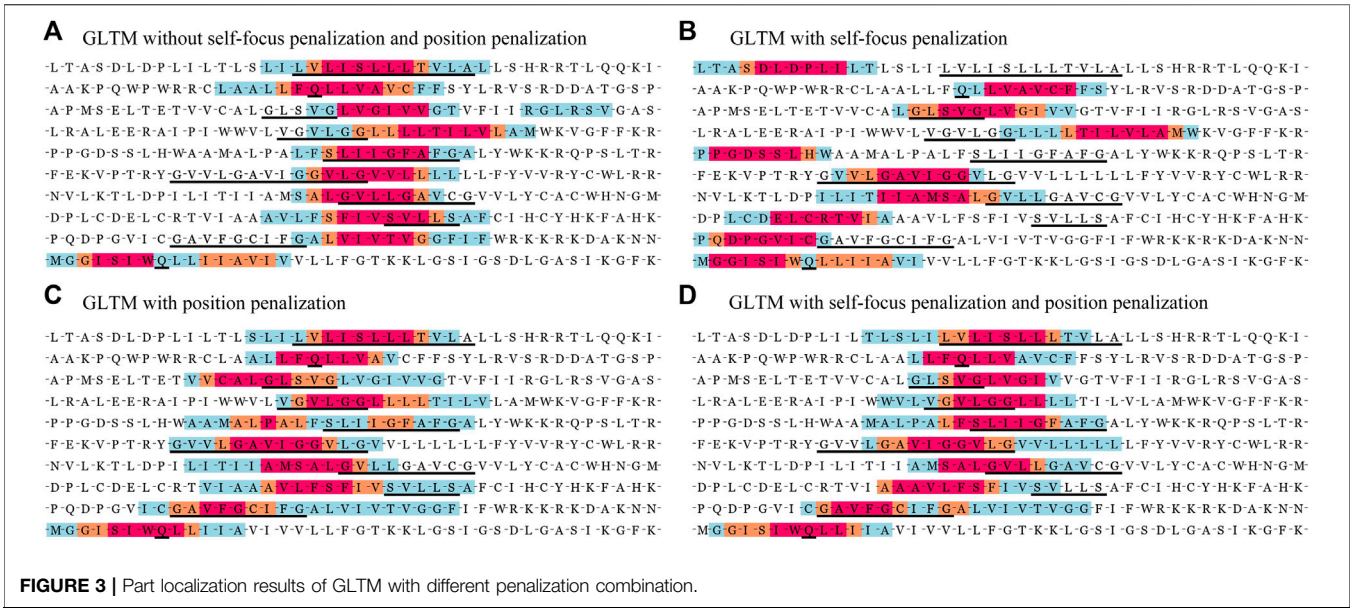


FIGURE 3 | Part localization results of GLTM with different penalization combination.

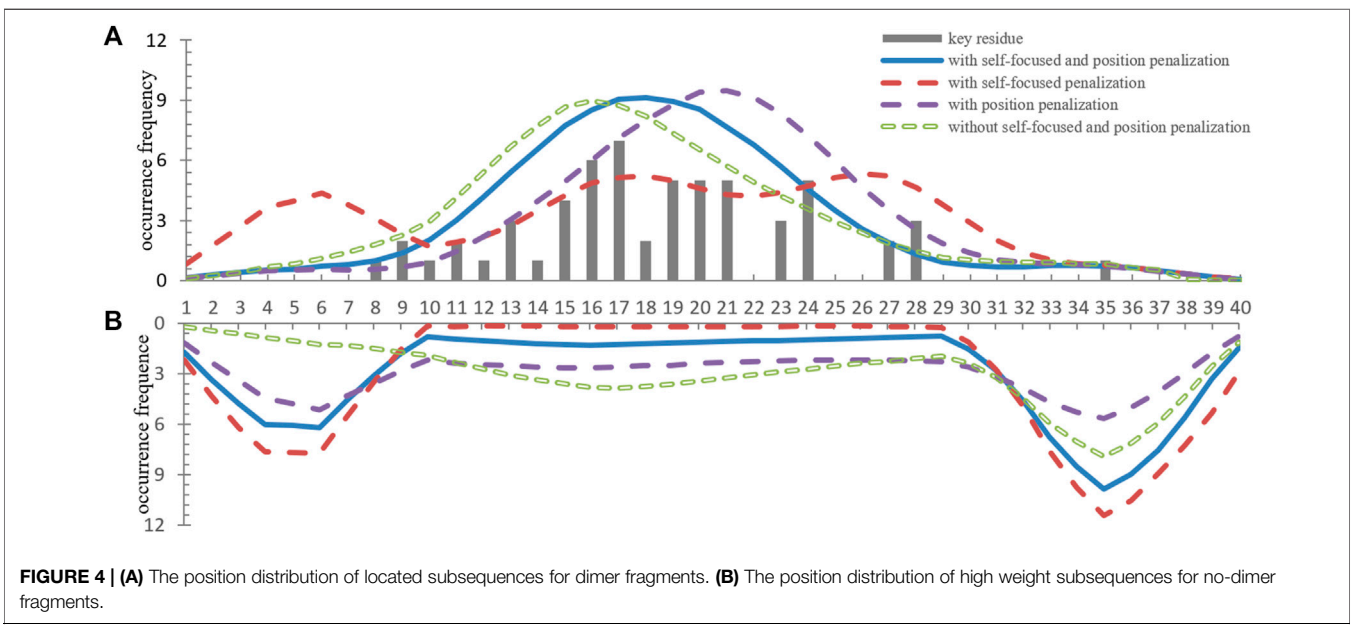
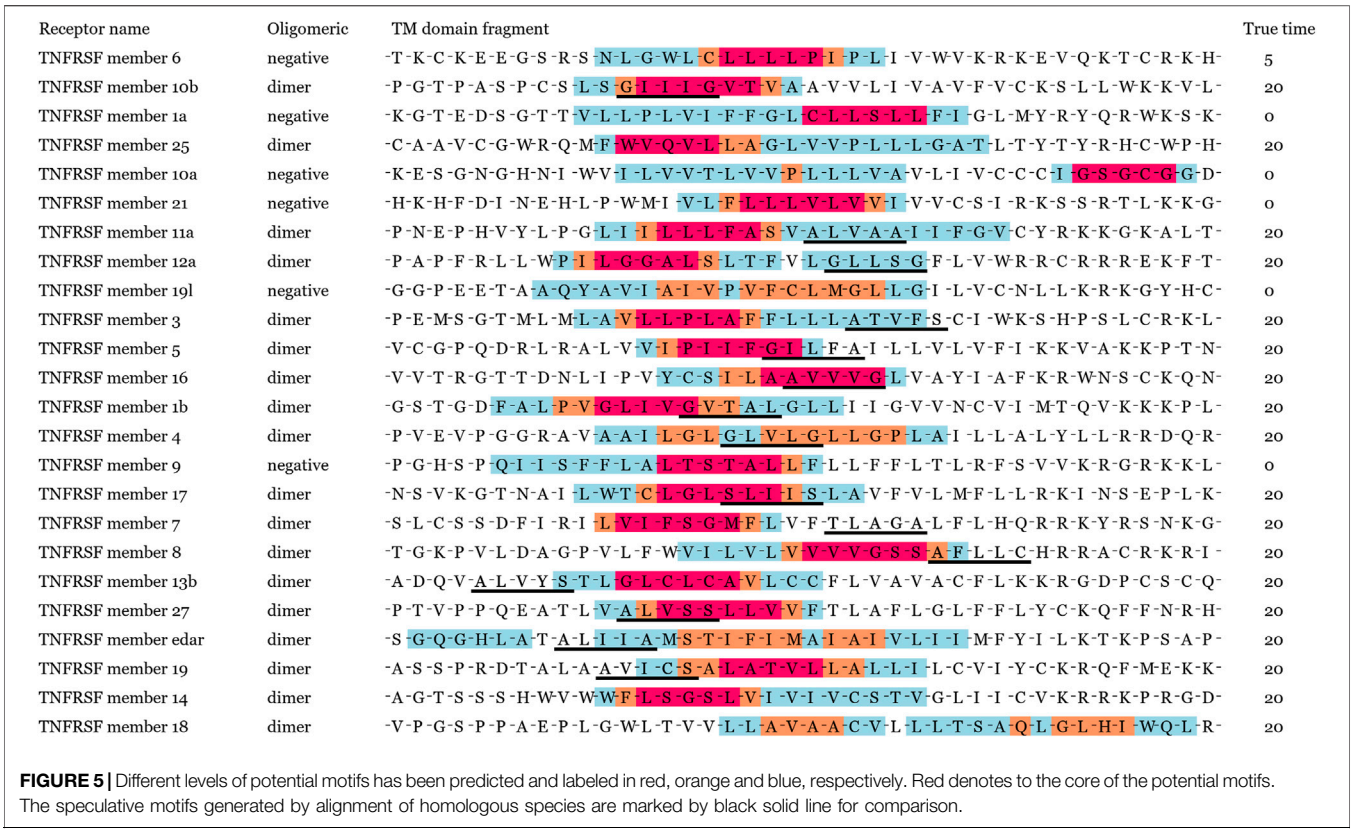


FIGURE 4 | (A) The position distribution of located subsequences for dimer fragments. **(B)** The position distribution of high weight subsequences for no-dimer fragments.

Most TNF receptors are candidates for antibody-based immunotherapy. A recently growing number of studies showed some tumor necrosis factor receptors play an active role in receptor signaling. In driving signaling, dimerization is an essential process which participates in the assembly of higher-order structures (Pan et al., 2019). In recent dimerization research, part potential dimer motifs of TNFRSF were speculated by alignment of TNFRSF sequences from various organisms (Zhao et al., 2020). These speculated dimer motifs referenced to prior biological knowledge had high credibility.

In order to verify our motif localization performance in the TNFRSF dataset, these TM sequences of TNFRSF were collected from UniProt version 2020_10. In the prediction results, partial TM sequences were falsely predicted to dimerize, and these subsequences of high weight were also marked in **Figure 5**. False prediction results were caused by the whole hydrophobicity discrepancy between training samples. Moreover, we noticed the most speculated dimer motifs was the GxxxG motif for TNFRSF, the known subsequence patterns information of the polar residue and the leucine zippers influenced specific GxxxG motif localization in position penalization.

We designed contrast experiments to verify the localization effect of position penalization. We set three new training datasets



Receptor name	Oligomeric	TM domain fragment	True time
TNFRSF member 6	negative	-T-K-C-K-E-E-G-S-R-S-N-L-G-W-L-C-L-L-L-L-P-I-P-L-I-V-W-V-K-R-K-E-V-Q-K-T-C-R-K-H-	5
TNFRSF member 10b	dimer	-P-G-T-P-A-S-P-C-S-L-S-G-I-I-I-G-V-T-V-A-A-V-V-L-I-V-A-V-F-V-C-K-S-L-L-W-K-V-L-	20
TNFRSF member 1a	negative	-K-G-T-E-D-S-G-T-T-V-L-L-P-L-V-I-F-F-G-L-C-L-L-S-L-L-F-I-G-L-M-Y-R-Y-Q-R-W-K-S-K-	0
TNFRSF member 25	dimer	-C-A-A-V-C-G-W-R-Q-M-F-W-V-Q-V-L-L-A-G-L-V-V-P-L-L-L-G-A-T-L-T-Y-T-Y-R-H-C-W-P-H-	20
TNFRSF member 10a	negative	-K-E-S-G-N-G-H-N-I-W-V-I-L-V-V-T-L-V-V-P-L-L-L-V-A-V-L-I-V-C-C-C-I-G-S-G-C-G-G-D-	0
TNFRSF member 21	negative	-H-K-H-F-D-I-N-E-H-L-P-W-M-I-V-L-F-L-L-L-V-L-V-I-V-V-C-S-I-R-K-S-S-R-T-L-K-K-G-	0
TNFRSF member 11a	dimer	-P-N-E-P-H-V-Y-L-P-G-L-I-I-L-L-L-F-A-S-V-A-L-V-A-A-I-I-F-G-V-C-Y-R-K-K-G-K-A-L-T-	20
TNFRSF member 12a	dimer	-P-A-P-F-R-L-L-W-P-I-L-G-G-A-L-S-L-T-F-V-L-G-L-L-S-G-F-L-V-W-R-R-C-R-R-R-E-K-F-T-	20
TNFRSF member 19l	negative	-G-G-P-E-E-T-A-A-Q-Y-A-V-I-A-I-V-P-V-F-C-L-M-G-L-L-G-I-L-V-C-N-L-L-K-R-K-G-Y-H-C-	0
TNFRSF member 3	dimer	-P-E-M-S-G-T-M-L-M-L-A-V-L-L-P-L-A-F-F-L-L-L-A-T-V-F-S-C-I-W-K-S-H-P-S-L-C-R-K-L-	20
TNFRSF member 5	dimer	-V-C-G-P-Q-D-R-L-R-A-L-V-V-I-P-I-I-F-G-I-L-F-A-I-L-L-V-L-V-F-I-K-K-V-A-K-K-P-T-N-	20
TNFRSF member 16	dimer	-V-V-T-R-G-T-T-D-N-L-I-P-V-Y-C-S-I-L-A-A-V-V-V-G-L-V-A-Y-I-A-F-K-R-W-N-S-C-K-Q-N-	20
TNFRSF member 1b	dimer	-G-S-T-G-D-F-A-L-P-V-G-L-I-V-G-V-T-A-L-G-L-L-I-I-G-V-V-N-C-V-I-M-T-Q-V-K-K-K-P-L-	20
TNFRSF member 4	dimer	-P-V-E-V-P-G-G-R-A-V-A-A-I-L-G-L-G-L-V-L-G-L-L-G-P-L-A-I-L-L-A-L-Y-L-L-R-R-D-Q-R-	20
TNFRSF member 9	negative	-P-G-H-S-P-Q-I-I-S-F-F-L-A-L-T-S-T-A-L-L-F-L-L-F-F-L-T-L-R-F-S-V-V-K-R-G-R-K-K-L-	0
TNFRSF member 17	dimer	-N-S-V-K-G-T-N-A-I-L-W-T-C-L-G-L-S-L-I-I-S-L-A-V-F-V-L-M-F-L-L-R-K-I-N-S-E-P-L-K-	20
TNFRSF member 7	dimer	-S-L-C-S-S-D-F-I-R-I-L-V-I-F-S-G-M-F-L-V-F-T-L-A-G-A-L-F-L-H-Q-R-R-K-Y-R-S-N-K-G-	20
TNFRSF member 8	dimer	-T-G-K-P-V-L-D-A-G-P-V-L-F-W-V-I-L-V-L-V-V-V-V-G-S-S-A-F-L-L-C-H-R-R-A-C-R-K-R-I-	20
TNFRSF member 13b	dimer	-A-D-Q-V-A-L-V-Y-S-T-L-G-L-C-L-C-A-V-L-C-C-F-L-V-A-V-A-C-F-L-K-K-R-G-D-P-C-S-C-Q-	20
TNFRSF member 27	dimer	-P-T-V-P-P-Q-E-A-T-L-V-A-L-V-S-S-L-L-V-V-F-T-L-A-F-L-G-L-F-F-L-Y-C-K-Q-F-F-N-R-H-	20
TNFRSF member edar	dimer	-S-G-Q-G-H-L-A-T-A-L-I-I-A-M-S-T-I-F-I-M-A-I-A-I-V-L-I-I-M-F-Y-I-L-K-T-K-P-S-A-P-	20
TNFRSF member 19	dimer	-A-S-S-P-R-D-T-A-L-A-A-V-I-C-S-A-L-A-T-V-L-L-A-L-L-I-L-C-V-I-Y-C-K-R-Q-F-M-E-K-K-	20
TNFRSF member 14	dimer	-A-G-T-S-S-S-H-W-V-W-W-F-L-S-G-S-L-V-I-V-I-V-C-S-T-V-G-L-I-I-C-V-K-R-R-K-P-R-G-D-	20
TNFRSF member 18	dimer	-V-P-G-S-P-P-A-E-P-L-G-W-L-T-V-V-L-L-A-V-A-A-C-V-L-L-L-T-S-A-Q-L-G-L-H-I-W-Q-L-R-	20

FIGURE 5 | Different levels of potential motifs has been predicted and labeled in red, orange and blue, respectively. Red denotes to the core of the potential motifs. The speculative motifs generated by alignment of homologous species are marked by black solid line for comparison.

that include the different known motifs' information. The RA dataset included the information of the known GxxxG motif, the polar residue, and the leucine zippers. The RB dataset only utilized the information of the known GxxxG motif, and the RC dataset had the information of polar residue and leucine zippers. High position score subsequences were collected from the training set, and their residue occurrence frequency was calculated as the reference subsequences in **Figure 6**. The located subsequences represented the residue occurrence frequency for these located subsequences. Besides these originally richly "blue" residues, the position penalization enhanced the specific motif localization performance according to supplied motif information.

The Influence of Sequence Context for Its Dimerization

Oligomer motifs were usually simplified as a helix-helix interactions paradigm, but more and more studies have certified that these subsequence frames cannot simply be regarded as a surrogate tool for oligomer state determination (Li et al., 2012). Other residues also influence helix-helix interactions besides oligomer motifs. For instance, the TM domain context highly determines the thermodynamic stability of TM helix-helix interactions than local GxxxG motif in glycophorin A (Bano-Polo et al., 2012). The SDS-PAGE analysis of glycophorin A mutants demonstrated that the C-terminal region residues were also important for their helix packing (Bano-Polo et al., 2012). Partial residues deletion and replacement will damage oligomerization to different degrees (Orzaez et al., 2000). Moreover, researchers guessed the distance between the dimerization motif and the flanking charged residues play a key role in the stability of TM helix-helix interactions. We chose 17 sequence fragments to research oligomerization based on previous residue mutation experiments of glycophorin A and ζζ. The first fifteen sequence fragments had confirmed their dimerization degree in previous biological experiments, and the dimerization interface of the last seven mutants was destroyed by residue replacement.

Most mutants of single hydrophobic residue replacement were predicted to dimerize in **Figure 7**. Although the prediction results of single residue mutants differ widely from the actual dimerization degree, other mutants were successfully predicted to not dimerize when the hydrophobic residues had been massively replaced. Significantly, the GxxxG motif and YxxxxT motif were stable when located in most mutants. This visualization results demonstrated that GLTM captured these overexpression subsequence patterns and considered sequence context information in oligomer prediction. Current experiments were limited in the lack of oligomer data. The motif localization model has broad application prospects in mutant oligomerization research with the rapid growth of sequencing data.

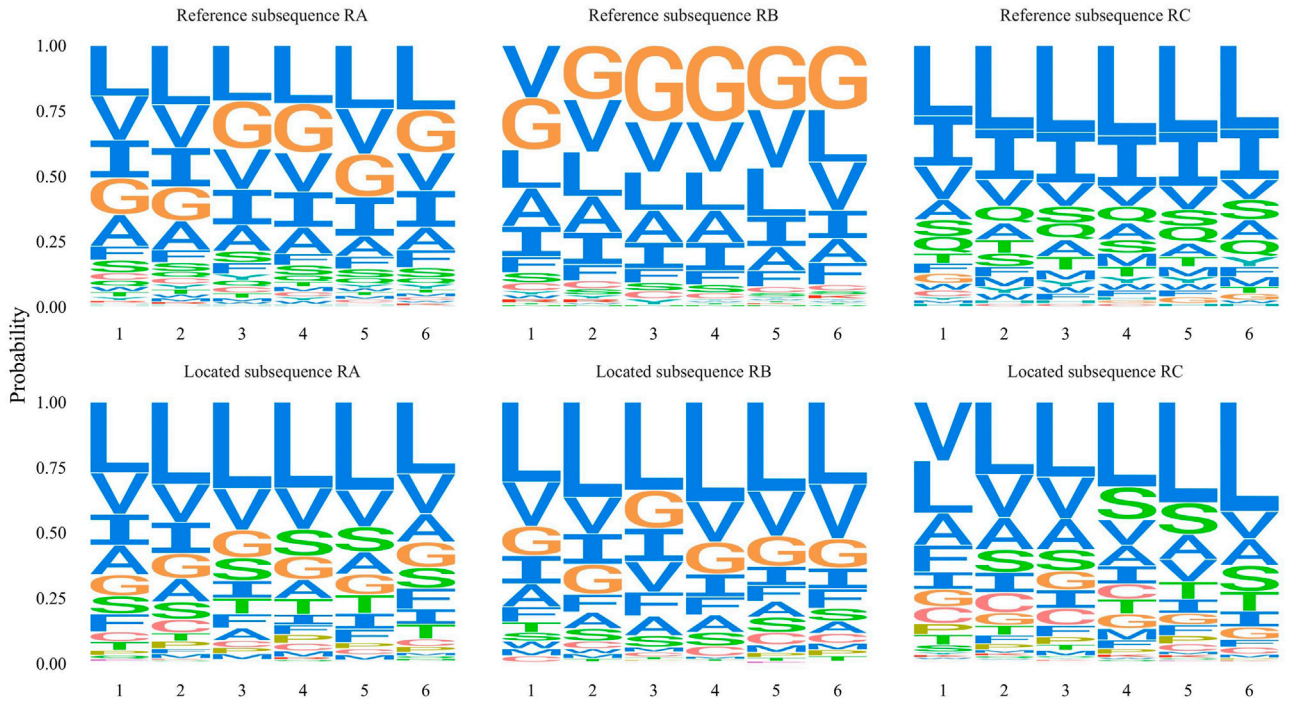


FIGURE 6 | Three contrast experiments respectively used RA, RB, and RC training datasets which include different known motifs information. Corresponding reference subsequence represents the residues frequency of high window score subsequences. The located subsequence represents the frequency of 20 amino acid residues for all located subsequences.

Receptor name	Oligomerize	TM domain fragment	True time
glycophorin-a	dimer	-A-H-H-F-S-E-P-E-I-T-L-I-I-F-G-V-M-A-G-V-I-G-T-I-L-L-I-S-Y-G-I-R-R-L-I-K-K-S-P-S-	20
glycophorin-a I21K	dimer	-A-H-H-F-S-E-P-E-I-T-L-I-I-F-G-V-M-A-G-V-K-G-T-I-L-L-I-S-Y-G-I-R-R-L-I-K-K-S-P-S-	20
glycophorin-a T23D	no-dimer	-A-H-H-F-S-E-P-E-I-T-L-I-I-F-G-V-M-A-G-V-I-G-D-I-L-L-I-S-Y-G-I-R-R-L-I-K-K-S-P-S-	0
glycophorin-a T23K	no-dimer	-A-H-H-F-S-E-P-E-I-T-L-I-I-F-G-V-M-A-G-V-I-G-K-I-L-L-I-S-Y-G-I-R-R-L-I-K-K-S-P-S-	0
glycophorin-a L25D	no-dimer	-A-H-H-F-S-E-P-E-I-T-L-I-I-F-G-V-M-A-G-V-I-G-T-I-D-L-I-S-Y-G-I-R-R-L-I-K-K-S-P-S-	0
glycophorin-a I21K/L25D	no-dimer	-A-H-H-F-S-E-P-E-I-T-L-I-I-F-G-V-M-A-G-V-K-G-T-I-D-L-I-S-Y-G-I-R-R-L-I-K-K-S-P-S-	2
t-cell surface glycoprotein cd3 zeta chain	dimer	-Q-S-F-G-L-L-D-P-K-L-C-Y-L-L-D-G-I-L-F-I-Y-G-V-I-L-T-A-L-F-L-R-V-K-F-S-R-S-A-D-A-	20
t-cell surface glycoprotein cd3 zeta chain L18A	no-dimer	-Q-S-F-G-L-L-D-P-K-L-C-Y-L-L-D-G-I-A-F-I-Y-G-V-I-L-T-A-L-F-L-R-V-K-F-S-R-S-A-D-A-	0
t-cell surface glycoprotein cd3 zeta chain D15E	no-dimer	-Q-S-F-G-L-L-D-P-K-L-C-Y-L-L-E-G-I-L-F-I-Y-G-V-I-L-T-A-L-F-L-R-V-K-F-S-R-S-A-D-A-	0
t-cell surface glycoprotein cd3 zeta chain Y21A	no-dimer	-Q-S-F-G-L-L-D-P-K-L-C-Y-L-L-D-G-I-L-F-I-A-G-V-I-L-T-A-L-F-L-R-V-K-F-S-R-S-A-D-A-	0
t-cell surface glycoprotein cd3 zeta chain T26A	no-dimer	-Q-S-F-G-L-L-D-P-K-L-C-Y-L-L-D-G-I-L-F-I-Y-G-V-I-L-A-A-L-F-L-R-V-K-F-S-R-S-A-D-A-	0
glycophorin-a I21K/T23D/L25D	no-dimer	-A-H-H-F-S-E-P-E-I-T-L-I-I-F-G-V-M-A-G-V-K-G-D-I-D-L-I-S-Y-G-I-R-R-L-I-K-K-S-P-S-	1
glycophorin-a I12D/I21K/T23D/L25D	no-dimer	-A-H-H-F-S-E-P-E-I-T-L-D-I-F-G-V-M-A-G-V-K-G-D-I-D-L-I-S-Y-G-I-R-R-L-I-K-K-S-P-S-	9
glycophorin-a I9D/I12D/I21K/T23D/L25D	no-dimer	-A-H-H-F-S-E-P-E-D-T-L-D-I-F-G-V-M-A-G-V-K-G-D-I-D-L-I-S-Y-G-I-R-R-L-I-K-K-S-P-S-	15
t-cell surface glycoprotein cd3 zeta chain D15E/L18A	no-dimer	-Q-S-F-G-L-L-D-P-K-L-C-Y-L-L-E-G-I-A-F-I-Y-G-V-I-L-T-A-L-F-L-R-V-K-F-S-R-S-A-D-A-	0
t-cell surface glycoprotein cd3 zeta chain D15E/L18A/I20E	no-dimer	-Q-S-F-G-L-L-D-P-K-L-C-Y-K-L-E-G-I-A-F-E-Y-G-V-I-L-T-A-L-F-L-R-V-K-F-S-R-S-A-D-A-	11
t-cell surface glycoprotein cd3 zeta chain D15E/L18A/I20E/A27E	no-dimer	-Q-S-F-G-L-L-D-P-K-L-C-Y-K-L-E-G-I-A-F-E-Y-G-V-I-L-T-E-L-F-L-R-V-K-F-S-R-S-A-D-A-	20

FIGURE 7 | Visualization results of 17 mutants. The labels of first 11 mutants were confirmed in biological experiments, and the labels of last six mutants were speculated to be by their destroyed dimerization interface.

CONCLUSION

In this paper, we propose an attention-based Global-Local structure Bi-LSTM model named GLTM to locate potential dimer motif. The three main components of GLTM can be summarized as follows: The first component was data preprocessing module, this module improved one-hot encoding to achieve a new data enhancement approach of subsequence segmentation; The secondary global-local Bi-LSTM structure was proposed to respectively extract local subsequence patterns and global context features; Proposed position and self-focus penalization reduce these irrelevant subsequences localization in tertiary attention mechanism layer. GLTM successfully located the most known key residues in the established benchmark dataset. In comparative experiments, the visualization results demonstrated the effectiveness of our proposed position and self-focus penalization. Different from the oligomer

motif discovery method, our motif localization model achieved end-end motif localization function without multiple homologous sequences alignment. More importantly, our motif localization model has broad application prospects in the research of mutant oligomerization.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

QM and FY contributed to conception and design of this study. KZ and ZZ performed and implemented the analysis. QM and FY wrote and edited the manuscript. All the authors helped with the draft and reviewed the manuscript before approving for publication.

FUNDING

The National Natural Science Foundation of China (62163017), the Natural Science Foundation of Jiangxi Province of China (20202BAB202009), the Key Science Foundation of Educational Commission of Jiangxi Province of China (GJJ160768), the Scholastic Youth Talent Program of Jiangxi Science and Technology Normal University (2016QNBJRC004), and the Graduate Innovation Fund Project of Education Department of Jiangxi province of China (YC2021-S756).

REFERENCES

Asgari, E., McHardy, A. C., and Mofrad, M. R. K. (2019). Probabilistic Variable-Length Segmentation of Protein Sequences for Discriminative Motif Discovery (DiMotif) and Sequence Embedding (ProtVecX). *Sci. Rep.* 9, 3577. doi:10.1038/s41598-019-38746-w

Bañó-Polo, M., Baeza-Delgado, C., Orzáez, M., Marti-Renom, M. A., Abad, C., and Mingarro, I. (2012). Polar/Ionizable Residues in Transmembrane Segments: Effects on helix-helix Packing. *PLoS One* 7, e44263. doi:10.1371/journal.pone.0044263

Brosig, B., and Langosch, D. (1998). The Dimerization Motif of the Glycophorin A Transmembrane Segment in Membranes: Importance of glycine Residues. *Protein Sci.* 7, 1052–1056. doi:10.1002/pro.5560070423

Cymer, F., and Schneider, D. (2010). Transmembrane helix-helix Interactions Involved in ErbB Receptor Signaling. *Cell Adhes. Migration* 4, 299–312. doi:10.4161/cam.4.2.11191

Doğruel, M., Down, T. A., and Hubbard, T. J. (2008). NestedMICA as an Ab Initio Protein Motif Discovery Tool. *BMC Bioinformatics* 9, 19. doi:10.1186/1471-2105-9-19

Edwards, R. J., Davey, N. E., and Shields, D. C. (2007). SLiMFinder: a Probabilistic Method for Identifying Over-represented, Convergently Evolved, Short Linear Motifs in Proteins. *PLoS One* 2, e967. doi:10.1371/journal.pone.0000967

Hubert, P., Sawma, P., Duneau, J.-P., Khao, J., Hénin, J., Bagnard, D., et al. (2010). Single-spanning Transmembrane Domains in Cell Growth and Cell-Cell Interactions. *Cel Adhes. Migration* 4, 313–324. doi:10.4161/cam.4.2.12430

Jing, X., Dong, Q., Hong, D., and Lu, R. (2020). Amino Acid Encoding Methods for Protein Sequences: A Comprehensive Review and Assessment. *Ieee/acm Trans. Comput. Biol. Bioinf.* 17, 1918–1931. doi:10.1109/TCBB.2019.2911677

LaPointe, L. M., Taylor, K. C., Subramaniam, S., Khadria, A., Rayment, I., and Senes, A. (2013). Structural Organization of FtsB, a Transmembrane Protein of the Bacterial Divisome. *Biochemistry* 52, 2574–2585. doi:10.1021/bi400222r

Li, E., Wimley, W. C., and Hristova, K. (2012). Transmembrane helix Dimerization: Beyond the Search for Sequence Motifs. *Biochim. Biophys. Acta (Bba) - Biomembranes* 1818, 183–193. doi:10.1016/j.bbamem.2011.08.031

Liang, J., Naveed, H., Jimenez-Morales, D., Adamian, L., and Lin, M. (2012). Computational Studies of Membrane Proteins: Models and Predictions for Biological Understanding. *Biochim. Biophys. Acta (Bba) - Biomembranes* 1818, 927–941. doi:10.1016/j.bbamem.2011.09.026

Lomize, A. L., Lomize, M. A., Krolicki, S. R., and Pogozheva, I. D. (2017). Membranome: a Database for Proteome-wide Analysis of Single-Pass Membrane Proteins. *Nucleic Acids Res.* 45, D250–D255. doi:10.1093/nar/gkw712

Mehdi, A. M., Sehgal, M. S. B., Kobe, B., Bailey, T. L., and Bodén, M. (2013). DLocalMotif: a Discriminative Approach for Discovering Local Motifs in Protein Sequences. *Bioinformatics* 29, 39–46. doi:10.1093/bioinformatics/bts654

Oates, J., King, G., and Dixon, A. M. (2010). Strong Oligomerization Behavior of PDGFβ Receptor Transmembrane Domain and its Regulation by the Juxtamembrane Regions. *Biochim. Biophys. Acta (Bba) - Biomembranes* 1798, 605–615. doi:10.1016/j.bbamem.2009.12.016

Orzáez, M., Pérez-Payá, E., and Mingarro, I. (2000). Influence of the C-Terminus of the Glycophorin A Transmembrane Fragment on the Dimerization Process. *Protein Sci.* 9, 1246–1253. doi:10.1110/ps.9.6.1246

Pan, L., Fu, T., Zhao, W., Zhao, L., Chen, W., Qiu, W., et al. (2019). Higher-Order Clustering of the Transmembrane Anchor of DR5 Drives Signaling. *Cell* 6, 1477. doi:10.1016/j.cell.2019.02.001

Prytuliak, R., Volkmer, M., Meier, M., and Habermann, B. H. (2017). HH-MOTiF: De Novo Detection of Short Linear Motifs in Proteins by Hidden Markov Model Comparisons. *Nucleic Acids Res.* 45, W470–W477. doi:10.1093/nar/gkx341

Qiu, J., Sun, X., Suo, S., Shi, S., Huang, R., Liang, R., et al. (2011). Predicting Homo-Oligomers and Hetero-Oligomers by Pseudo-Amino acid Composition: An Approach From Discrete Wavelet Transformation. *Biochimie* 7, 1132–1138. doi:10.1016/j.biochi.2011.03.010

Rawlings, A. E. (2016). Membrane Proteins: Always an Insoluble Problem. *Biochem. Soc. Trans.* 44, 790–795. doi:10.1042/BST20160025

Redhead, E., and Bailey, T. L. (2007). Discriminative Motif Discovery in DNA and Protein Sequences Using the DEME Algorithm. *BMC Bioinformatics* 8, 385. doi:10.1186/1471-2105-8-385

Russ, W. P., and Engelman, D. M. (2000). The GxxxG Motif: A Framework for Transmembrane helix-helix Association. *J. Mol. Biol.* 296, 911–919. doi:10.1006/jmbi.1999.3489

Senes, A., Gerstein, M., and Engelman, D. M. (2000). Statistical Analysis of Amino Acid Patterns in Transmembrane Helices: the GxxxG Motif Occurs Frequently and in Association with β-branched Residues at Neighboring Positions. *J. Mol. Biol.* 296, 921–936. doi:10.1006/jmbi.1999.3488

Song, T., and Gu, H. (2015). Discovering Short Linear Protein Motif Based on Selective Training of Profile Hidden Markov Models. *J. Theor. Biol.* 377, 75–84. doi:10.1016/j.jtbi.2015.03.010

Song, J., and Tang, H. W. (2005). Support Vector Machines for Classification of Homo-Oligomeric Proteins by Incorporating Subsequence Distributions. *J. Mol. Struct. Theochem.* 1-3, 97–101. doi:10.1016/j.theochem.2005.02.002

Sun, X., Shi, S., Qiu, J., Suo, S., Huang, S., and Liang, R. (2012). Identifying Protein Quaternary Structural Attributes by Incorporating Physicochemical Properties into the General Form of Chou's PseAAC via Discrete Wavelet Transform. *Mol. Biosyst.* 12, 3178–3184. doi:10.1039/c2mb25280e

Zhang, S. W., Pan, Q., Zhang, H. C., Shao, Z. C., and Shi, J. Y. (2006). Prediction of Protein Homo-Oligomer Types by Pseudo Amino acid Composition:

Approached with an Improved Feature Extraction and Naive Bayes Feature Fusion. *Amino Acids* 4, 461–468. doi:10.1007/s00726-006-0263-8

Zhao, L. Q., Fu, L., Pan, A., Piai, A., and Chou, J. J. (2020). The Diversity and Similarity of Transmembrane Trimerization of TNF Receptors. *Front.Cell Develop. Biol.* 8, 569684. doi:10.3389/fcell.2020.569684

Zviling, M., Kochva, U., and Arkin, I. T. (2007). How Important Are Transmembrane Helices of Bitopic Membrane Proteins? *Biochim. Biophys. Acta (Bba) - Biomembranes* 1768, 387–392. doi:10.1016/j.bbamem.2006.11.019

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.