AIMS Biophysics

AIMS Biophysics, 8(2): 165–181.
DOI: 10.3934/biophy.2021013
Received: 25 February 2021
Accepted: 13 April 2021
Published: 20 April 2021

http://www.aimspress.com/journal/biophysics

Research article

Sequence–function correlation of the transmembrane domains in NS4B of HCV using a computational approach

Ta-Chou Huang and Wolfgang B. Fischer*

Institute of Biophotonics, School of Biomedical Science and Engineering, National Yang Ming Chiao Tung University, Taipei, Taiwan

* **Correspondence:** Email: wfischer@ym.edu.tw; Tel: +886228267394; Fax: +886228235460.

Abstract: An algorithm is applied to propose a sequence–function correlation of the transmembrane domains (TMDs) of the non-structural protein 4B (NS4B) of hepatitis C virus (HCV). The putative sequence of the TMDs is obtained using 20 available secondary structure prediction programs (SSPPs) with different lengths of the overall amino acid sequence of the protein as input. The results support the notion of four helical TMDs. Whilst the region of the first TMDs leaves room for speculation about an additional TMD, the other three TMDs are consistently predicted. Structural features and the role of each of the TMDs is proposed by applying pairwise sequence alignment using BLAST on the level (i) protein sequence alignment and consequent (ii) function-related alignment. Sequence identity with those TMDs of proteins involved in Ca-homeostasis and generation of replication vesicles, such as Nsp3 of corona viruses, murine coronavirus especially mouse hepatitis virus (MHV), middle east respiratory syndrome coronavirus (MERS), severe acute respiratory syndrome coronavirus (SARS-CoV) and SARS-CoV-2, are suggested. Focusing the search on those proteins in particular and their TMDs playing an active role in their mechanism of function, such as transporters, pumps, viral channel forming protein Vpu of human immunodeficiency virus type 1 (HIV-1) and mediators, suggests TMDs 2 and 4 to have functional roles in NS4B, as well as additionally TMD1 and 3 in case of vesicle formation.

Keywords: secondary structure prediction; sequence alignment; structure–function correlation; bioinformatics; viral membrane protein; NS4B of HCV

1. Introduction

Computational methods can play an important role in bridging the sequence–function–structure gap [1–3]. With a series of structure prediction programs helical structures within globular proteins [4] or structures such as helical transmembrane domains (TMDs) [5,6] can be predicted. Far fewer programs are available predicting β-sheet as the membrane spanning motif. Beyond this level of predicting the secondary structure, obtaining the tertiary or even quaternary structure is within reach [7]. In terms of obtaining functional information, sequence alignment is a sophisticated tool to compare unknown sequences with those, for which for example more information upon its function and structure is available [8,9]. The alignment can be done solely on the sequence as such, or focused on particular sequence motifs, such as charged amino acids in DNA binding proteins, to obtain spatial information about the protein [10].

The combined search, for secondary structure as well as functional information, is applied to the viral non-structural protein 4 (NS4B) of hepatitis C virus (HCV). NS4B is a 261 amino acid membrane protein. It is expressed in the infected cell as part of the approximately 3000 amino acid long polyprotein [11]. Consequent cleavage of the polyprotein by viral and cellular proteases release the protein together with 9 others. The role of the polytopic membrane protein is to support the generation of viral replication organelles [12]. NS4B is proposed to harbor two amphipathic helices on either side of its core region. The core region harbors at least four TMDs based on secondary structure prediction programs (SSPPs) applied to this region [13–15]. TMDs 2 and 3 are suggested to be held together by a nucleoside-triphosphatase motif [16,17]. Eventually a fifth TMD is also proposed from experimental results [18]. The oligomeric state is experimentally proposed to be that of at least a dimer or trimer with higher oligomeric states [19].

A sequence of steps is proposed to make a prediction of the putative structural and functional features of NS4B by applying bioinformatics tools. Secondary structure prediction programs (SSPPs) are used to identify the putative structural motif of the TMDs and their length. The proposed sequences of NS4B is aligned with proteins available in the protein structure and sequence data base (Protein Structure Data Bank, UniProt) to identify reasonable sequence overlap and to propose functional features. In addition, with the outcome of the alignment the search is narrowed into the search for functionally related proteins. This also includes membrane protein Nsp3 of corona viruses. Together with the functional and structural analysis the functional role of individual TMDs is suggested. The data are compared with those obtained when using the reported sequence of TMDs (hither forth denoted as r-TMDs) in the literature [17]

2. Materials and methods

2.1. Secondary structure prediction

The sequence of NS4B is based on that of HCV strain JFH-1(2a), GenBank accession number AB047639 [20] and used to predict consensus TMDs, hither forth called p-TMDs, using SSPPs such as CCTOP[21], DAS (cutoff@1.7) [22] and DAS-TMfilter [23], HMMTop [24], MemBrain [25], Memsat [26] and MEMSAT3 [27], OCTOPUS [28–30] and SPOCTOPUS [28,29], Philius [31], Phobius [28] and PolyPhobius [32], Pro and Prodiv [33], Scampi [34] and ScampiMsa [34], SPLIT 4.0 [35], TMHMM 2.0 [36], TMMOD [37], TMpred [38], TOPOCONS [39], PSIPRED 4.0 [40].

The reported sequences of the TMDs reported in the literature (r-TMDs) were as follows [17]: r-TMD1 70–90: LPGNPAVASMMAFSAALTSPL; r-TMD2 112–131: PAGATGFVVSGLVGAAVGSI; r-TMD3 137–154: LVDILAGYGAGISGALVA; r-TMD4 172–190: LPGILSPGALVVGVICAAI.

Targeted TMDs used for alignment were taken from membrane proteins involved in Ca-homeostasis and grouped as (i) Ca-ion conducting channels: the calcium release-activated calcium channel (Orai) from *Drosophila melanogaster* (PDB ID: 4HKR), L-type, high voltage-activated Ca_v channel ($Ca_v1.1$) from rabbit (PDB ID: 3JBR), human voltage-dependent L-type calcium channel ($Ca_v1.2$) (Uniprot Q13936), human ryanodine receptor 1 (RyR1) (Uniprot P21817), human sarcoplasmic/endoplasmic reticulum Ca^{2+}-permeable cation channel inositol-1,4,5-triphosphate receptor type 1 (IP_3R1) (Uniprot Q14643); (ii) Ca-transporters: mitochondrial calcium uniporter (MCU) from *C. elegans* (PDB ID: 5ID3), rabbit skeletal-muscle sarcoplasmic/endoplasmic reticulum Ca^{2+}-ATPase (SERCA) (PDB ID: 3TLM); (iii) Ca-mediators: SERCA inhibiting rabbit phospholamban (PLN) (PDB ID: 2KYV), human stromal interaction molecule 1 (STIM1) mediating store-operated Ca^{2+} entry (Uniprot Q13586); and (iv) Nsp3 of coronaviruses being involved in the formation of replication vesicles: mouse hepatitis virus (MHV) (Uniprot: Q66WN6, a.a. 2062-2841), SARS-CoV (Uniprot: P0C6X7, a.a. 1997-2745), SARS-CoV-2 (NCBI Reference Sequence: YP_009725299.1, a.a. 1202-1945), MERS-CoV (Uniprot: K0BWD0, a.a. 1953-2750).

2.2. Sequence alignment and structure prediction

Protein-BLAST (https://blast.ncbi.nlm.nih.gov) was used to find similar sequence segments from existing structures in its default mode. The searching was based on the default database excluding data bases like Models (XM/XP), Non-Redundant RefSeq Proteins (WP) and Uncultured/Environmental Sample Sequences. The maximum target sequences were set to 20,000. Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) was used in its default mode to search for proteins with similar function by sequence alignment and to create the phylogram. Sequence Identity and similarity of each aligned sequence pair was calculated using Software Sequence Manipulation Suite (SMS) (http://www.bioinformatics.org/sms2/index.html). All programs were used in their default mode.

3. Results

An algorithm is proposed to predict structure – function relations of the transmembrane domains (TMDs) of the viral membrane protein NS4B (Figure 1). Taking the primary structure of the protein as input from outside sources, e.g. experiments, proposals of the stretch of putative secondary structural motifs is made. With the stretch of amino acids including the secondary structural motif a two-level sequence alignment is applied. On a first level, proposal of the function of the investigated protein is made by comparison with proteins of known function which, at best, also includes structural information. On a second level, with the knowledge from the first level, the search space is narrowed to focus only on those proteins with specific functions and in this study on their TMDs. Other information from outside experimental sources can also enter at this level. The algorithm is not limited to the protein investigated in this study.

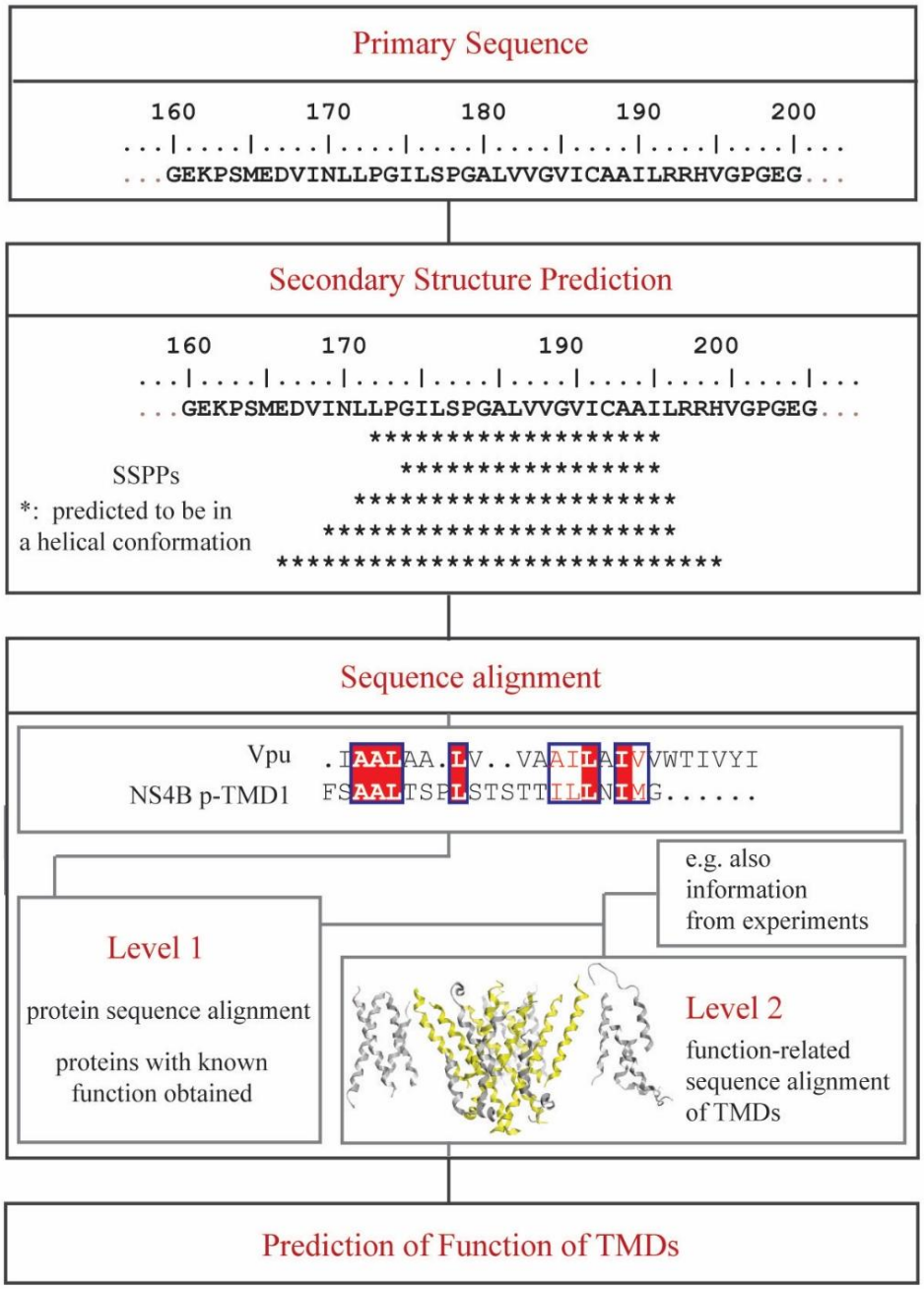

Figure 1. Generalized algorithm for predicting structural and functional of the TMDs of NS4B of HCV. Information of the primary structure is used for secondary structure prediction followed by a two-level sequence alignment protocol. Level 1 performs a sequence alignment on the protein level, while level 2 performs sequence alignment based on either the output information of level 1 or based on information from experimental studies about putative function of the protein and consequent confinement of the alignment to the TMDs.

3.1. Secondary structure prediction

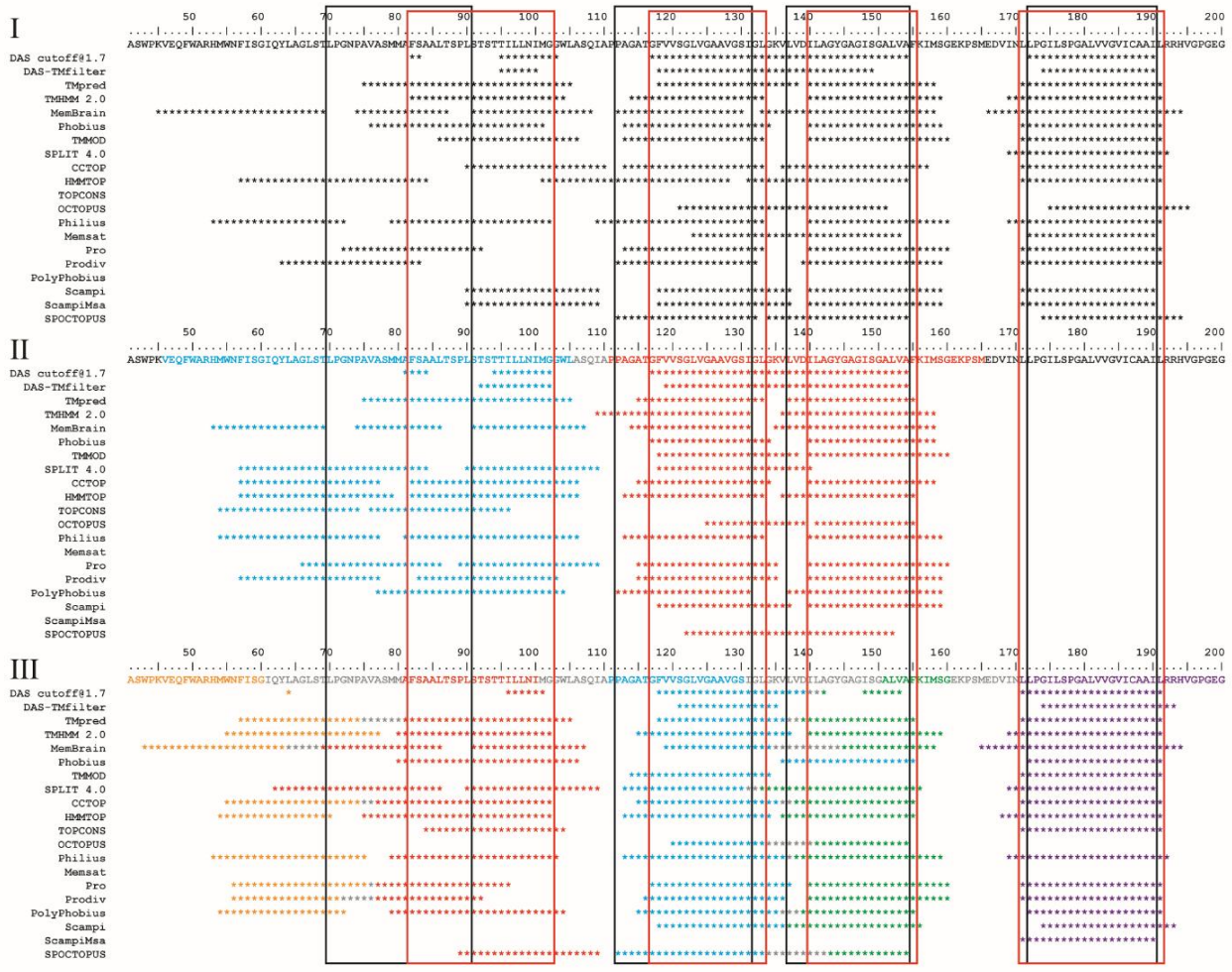

Figure 2. Identification of the TMDs using secondary structure prediction programs (SSPPs) with (I) the entire length of the NS4B sequence, (II) two regions (blue and red) each containing the anticipated TMDs, and (III) the individual proposed stretches which are supposed to contain the respective TMDs (yellow, orange, blue, green, pink). Red boxes mark the sequence selected as TMDs, named p-TMD. Black boxes mark the reported sequence for the TMDs [17], referred to in this study as r-TMDs. Colored residues, and consequent results shown as stars by the SSPPs, represent the sequence used in the SSPPs. Overlapping residues in the sequences used in the SSPs are marked in grey.

A series of 20 SSPPs is used to search for a consensus region of the putative TMDs within the sequence of NS4B (Figure 2). A consensus is reached for a particular amino acid to be part of the TMD if more than 19 of the SSPPs propose this amino acid to be in a helical conformation (see also Figure 3 for details). When using the full-length sequence of NS4B consisting of 261 amino acids, the best consensus is reached for a stretch around I169 to H194, here denoted as p-TMD4 (e.g. red box in Figure 2). A stretch from A110 to G160 is also predicted by all of the programs to containing a helical motif. By visible inspection of the chart (Figure 1, I), it is intriguing to define a short loop for

G132 to D139. Thus, two TMDs are predicted for this region, denoted as p-TMD2 and 3 (e.g. red boxes in Figure 2). This observation is supported by the fact that the full stretch with its 50 amino acids is too long to be a TMD, which is usually in the range of 17–25 amino acids [41,42]. Another helical motif is proposed between residues W-50 to Q-108 by 15 of the SSPPs (Figure 2,I). From those, two of them even propose two helical domains. The stretch overlaps with the stretch proposed to harbor AH2 (S42–G66 [17]). By visible inspection, a consensus sequence M80 to Q108 is chosen as TMD1 which includes the prediction of most of the programs as well as the putative second helix (e.g. red box in Figure 2).

Using the individual sequence which potentially harbors p-TMDs 2 and 3, A106 to M165, supports the existence of two TMDs as mentioned above (Figure 2, II). Using a separate stretch V46 to A110 enhances the impression of the existence of two possible TMDs.

Focusing on the individual stretches harboring potentially the TMDs reveals that the prediction for p-TMDs 2 (blue), 3 (green), and 4 (purple) are suggested for most of the programs within a narrow region of the amino acid sequence (Figure 2, III). The stretch for p-TMD1 is also predicted when including the amino acid sequence from I61 to A110 (Figure 2, III, red). Using a separate stretch A41 to M80, nine programs propose the aforementioned second helical stretch (Figure 2, III, orange).

A similar study using nAChR, GLIC and LMP1 of Epstein-Barr virus shows that independent of the length of the amino acid sequence used, the prediction of the TMDs is similar (Suppl. Figure 1 and 2). In most cases the predictions coincide with those proposals found based on experimental evidence in the literature (black boxes in Supplemental Figures 1 A,B,C).

Counting all the hits for the amino acids being in a helical conformation from the 20 programs defines p-TMD4 of NS4B from L171 to L191 (Figure 3). This definition and those for the other TMDs are shown as red boxes in Figure 1. Also, the region containing p-TMDs 2 and 3, P112 to G160, show distinct patterns for two maxima as well as the N terminal region (W50 to I109). Allowing for a threshold that at least 11 programs suggest a helical stretch the following TMD stretches are proposed:

p-TMD1 from F82 to G102 (FSAALTSPLS TSTTILLNIM G),
p-TMD2 from G117 to L133 (GFVVSGLVGA AVGSIGL), and
p-TMD3 from I140 to F155 (ILAGYGAGIS GALVAF), and
p-TMD4 from L171 to L191 (LLPGILSPGA LVVGVICAAI L).

Residues G134 to D139 (GKVLVD) are seen as a short loop between p-TMD2 and 3 based on their potency to form salt bridges. When using the sequence of the three loops between the p-TMDs (G103–T116, G134–D139, K156–N170) neither sequence alignment (BLAST) nor motif search servers (InterPro [43], PROSITE [44]) identify any similarity with membrane proteins.

The percentage of overlaps of the p-TMDs with the r-TMDs reported in the literature [17] (see black boxes in Figure 2) decreases in the order TMD4 (90.5 %) > TMD3 (79 %) > TMD2 (68.2 %) > TMD1 (27.4 %).

As a result, while prediction of TMD2, 3, and 4 are fairly similar, there is some ambiguity for predicting TMD1.

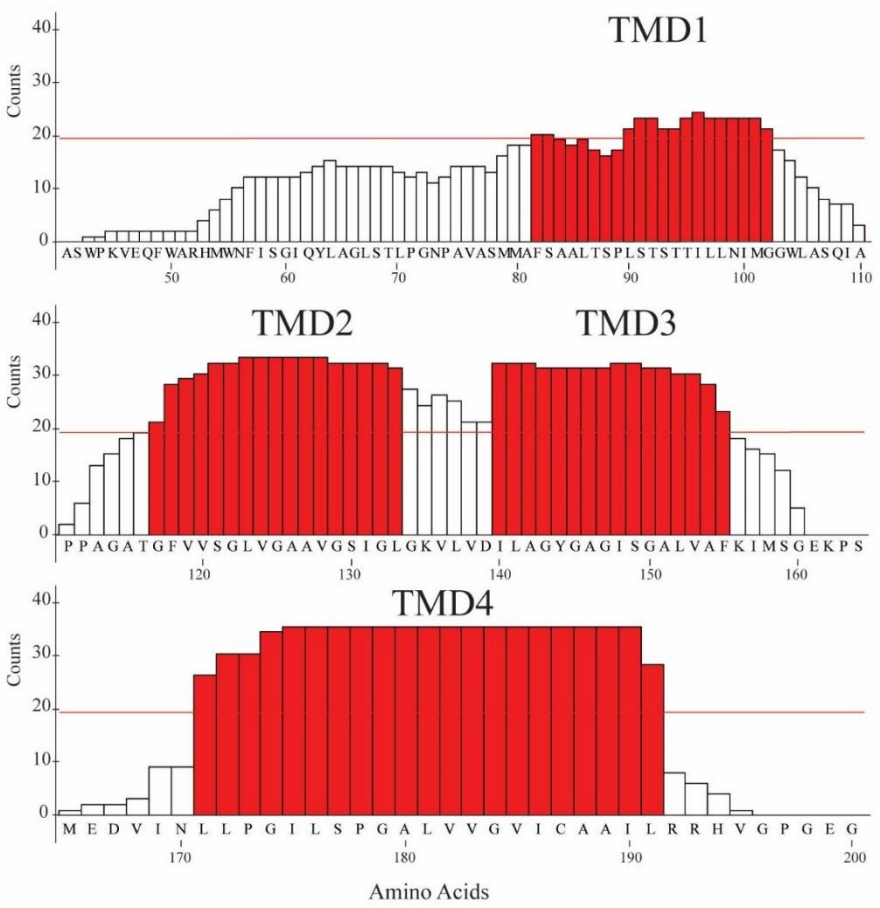

Figure 3. Frequency of the amino acids of NS4B identified to be in a helical conformation by the secondary structure prediction programs and used in this study. The 'counts' are derived from the data presented in Figure 1 I and III by counting the number of SSPPs which identify the respective amino acids to be in a helical conformation. The red columns depict the results when using a threshold of equal and more than 19 programs predict the helical motif for the amino acid defining the p-TMDs.

3.2. Sequence alignment and structural Bioinformatics with experimentally derived sequence data

The software BLAST is used to identify proteins which show sequence identity with especially the TMDs of NS4B and for which structural information is available. With the latter information a structural functional proposal of the p-TMDs of NS4B is given. Using full length sequence, NS4B [20] aligns with 80 % identity with itself when embedded in a polyprotein (P27958) (Table 1,i).

Limiting the subject sequence to the p-TMDs containing segments, F82–L191 of NS4B, results in no alignment (Table 1,ii). For the stretch including the r-TMDs, L70 to I190, an enzyme, succinate-CoA ligase, is identified (Suppl. Table 1).

Table 1. Best scored target proteins derived from multiple sequence alignment with NS4B using BLAST. The length of the NS4B sequence used varies from (i) its full length, to (ii) a stretch which includes the p-TMDs (F82 to L191), to (iii) the individual sequences of the four p-TMDs. Information is provided about their function, the percentage identity (I) and similarity (S), the availability of a crystal structure or the uniport accession number (PDB ID / Uniprot), the oligomeric state (Oligo. state) and whether the alignment is with the TMDs of the target protein (TMDs at target protein). cov (covered): the scoring is reported as the percentage number of identical residues identified in the alignment derived from the percentage of the TMD stretch of NS4B used in BLAST.

	Segment	Target protein	Function	I (cov) [%]	PDB ID / Uniprot[a] (I,S)	Oligo. state	TMDs at target protein
i	Full-length	NS4B	Unknown	80 (11)	2JXF[b]/P27958	-	-
ii	p-TMDs + loops	-	-	-	-	-	-
iii	p-TMD1	NADH dehydrogenase	Transferase	71 (80)	-/M9N917	2[c]	√
		GM10282	Transporter	67 (85)	-/B4ICG1	2[d]	√
	p-TMD2	chromate efflux transporter	Transporter	100 (94)	-/A0A536PSC5	2[e]	√
		ABC transporter permease	Transporter	83 (94)	-/A0A1Q7W0R8	4[f]	√
		Na^+-dependent transporter	Transporter	73 (88)	-/A0A1R4K8V1	4[g] (2+2)	√
		Ca^{2+}/Na^+ antiporter	Transporter	71 (82)	-/A0A3M2BD73 (I: 65/S: 80)[h]	2[i]	√
	p-TMD3	ABC transporter permease	Transporter	89 (100)	-/A0A522S9V5	4	√
		Na^+/H^+ antiporter	Transporter	63 (100)	-/J9GPP0 (I: 48/S: 66)	2[j]	√
		chromate transporter	Transporter	89 (81)	-/A0A413L363	2	√

Continued on next page

p-TMD4	Na^+-translocating pyrophosphatase	Transporter	82 (90)	-/A0A2D5BVY3	2[k]	√
	Na^+ transporter	Transporter	78 (80)	-/A0A520M9G4	2	√
	Na^+/H^+ antiporter	Transporter	72 (80)	-/A0A0X1TJ61 (I: 63/S: 72)	2	√
	ABC transporter permease	Transporter	71 (80)	-/U2NSW6 (I=S: 96)	4	√
	Vpu protein	Ion channel	71 (80)	-/A0A076V5C1 (I: 80/S: 86)	4-5[l]	√
	ion channel protein	Ion channel	67 (100)	-/W9FQ61 (I: 93/S: 97)	2[m]	-

[a]: The information either directly displayed in the BLAST result pages, or the sequence information of target protein found in the Uniprot sequence databases; [b]: solution NMR structure of NS4B(40-69) DOI: 10.1128/JVI.02663-08; [c]: DOI: 10.1016/j.cell.2017.07.050; [d]: DOI: 10.1016/j.bbamem.2010.08.013; [e]: DOI: 10.1007/7171_2006_087; [f]: DOI: 10.1074/jbc.M310785200; [g]: DOI: 10.1124/mi.4.1.38; [h]: Using other sequence information with the same function and organism which found in the Uniprot database due to no fully identical sequence exists; [i]: DOI: 10.1126/science.1215759; [j]: DOI: 10.7554/eLife.03579; [k]: DOI: 10.1093/pcp/pcy054; [l]: DOI: 10.1039/c5mb00798d; [m]: DOI: 10.7554/eLife.36629.

Aligning individual ranges which include the individual sequences of the p-TMDs, identifies sequence identities with membrane proteins in the range of 67–100 % identity (Table 1,iii). The top scored protein is a chromate efflux transporter with an identity of 100 % for p-TMD2, followed by an ABC transporter permease (89 %) for TMD3. All other membrane proteins are mainly transporters for the other p-TMDs. p-TMDs 1 and 4 also show sequence identity with a transferase (71 %) and ion channels (Vpu of HIV-1 (71 %) and an unknown ion channel (67 %)), respectively. Using the individual r-TMDs, transport proteins are proposed for r-TMD2 and 4 as well (Suppl. Figure 1). Selecting r-TMDs2 and 4, for the latter also a Ca/Na antiporter is identified.

As a result, when focusing on the TMD sequences, an increased overlap with the TMDs of other membrane proteins such as transporters, transferase and ion channels is achieved. One of the overlapping proteins, is a Na^+/Ca^{2+} exchanger, involved in Ca-homeostasis. This supports evidence from experiments, showing that NS4B could be involved in Ca-homeostasis of the endoplasmic reticulum [45].

3.3. Sequence alignment with function-related structure-based sequences

Three classes of membrane proteins, (i) to (iii) (see Materials and Methods) knowingly involved in Ca-homeostasis are chosen, to probe functional correlation of their reported functional TMDs (Figure 4) with the p-TMDs of NS4B. Values such as identity (I) and similarity (S) are obtained for the p-TMDs and the selected proteins. The values for each of the p-TMDs for all the 11 proteins are plotted as single values or as an averaged if the protein consists of a few subunits to identify the differences between the TMDs. The values for I are in the range of 5.1 (Orai) to 20.2 ± 5.9 (IP3R), those of S in the range 14.0 (Orai) to 52.2 (PLN) (Table 2).

Counting the matches of the p-TMDs of NS4B which have overlap with either pore lining TMDs or are identified to be involved in ion translocation at the protein site, focusing on classes (i)

and (ii) proteins identifies the sequence as follows. There are 3 matches for p-TMD1, 4 matches for p-TMD2, 3 matches for p-TMD3 and 8 matches for p-TMD4 when looking at I and S values (Table 2). p-TMD2 and 4 overlap with TMDs of PLN as well as STIM1, respectively. A similar distribution is also obtained for the r-TMDs (1 match (r-TMD1), 6 matches (r-TMD2), 1 match (r-TMD3), 8 matches (r-TMD4) (Suppl. Table 2). For classes (iii) r-TMD2 and 3 have overlap with domain II of PLN, and r-TMD4 with STIM1.

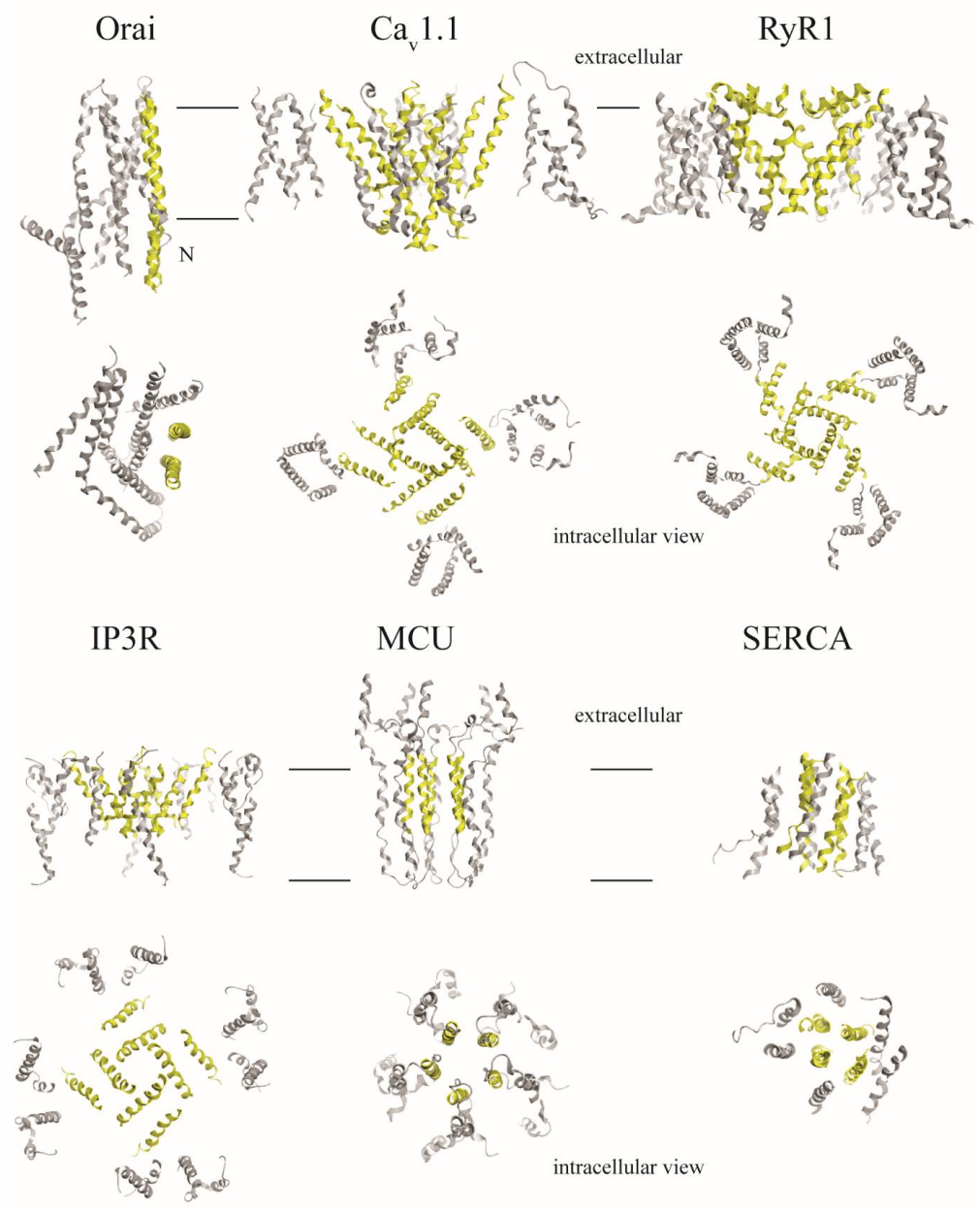

Figure 4. Structures of the TMDs of the proteins involved in Ca-homeostasis and used for sequence alignment. The TMDs of Orai (PDB ID: 4HKR), $Ca_v1.1$ (PDB ID: 3JBR), RyR1 (PDB ID: 5J8V), IP3R (PDB ID: 3JAV), MUC (PDB ID: 5ID3) and SERCA (PDB ID: 3TLM) are shown with the TMDs involved in the function of the proteins and used for sequence alignment with the p-TMDs of NS4B are highlighted in yellow. Black lines mark the approximate boundaries of the lipid membrane.

Table 2. Sequence alignment of p-TMDs. Calculation of p-TMD identity (I) and similarity (S) using Sequence Manipulation Suite of Clustal Omega. Each of the p-TMDs of NS4B is compared with either pore lining or Ca-binding domains of selected target membrane proteins: Orai, $Ca_v1.1$, $Ca_v1.2$, RyR1, IP_3R, MCU, SERCA, PLN, STIM1. Standard deviation is given if the target protein consists of multiple domains or more than one TMD which is identified as pore lining (PL) or identified as Ca-binding site (CaBS). Notation: e.g. 1(4) x 1 (TMD1) = the protein is a homo tetramer and only one of the subunits is consequently used (1(4)), and has only on TMD (x 1), and the TMD is TMD1 (1). For Ca_v channels the PL domains TMD5 and 6 (5/6) are used, for RyR1 also TMD5 and 6 as well as the pore region (marked as (2 + 1) because of two TMDs plus one pore helix). SERCA consists of three domains, from which one of them having 10 TMDs with 4 of them involved in Ca-binding (1x4). For PLN the name of the TMD is given as domain II (DOI: 10.1073/pnas.1016535108). Dark shaded boxes indicate p-TMDs of NS4B which show highest values of I/S with PL domains of target protein, light shaded boxes indicate the corresponding data for CaBS of target protein. mv = maximum value. For SERCA two p-TMDs of NS4B are shaded for I due to their almost similar values.

	TMDs of target protein	I/S	p-TMD1	p-TMD2	p-TMD3	p-TMD4
Orai	1(4) x 1 (1)	I	15.8	7.0	5.1	13.5
		S	26.3	14.0	23.1	29.7
$Ca_v1.1$	4 x 2 (5/6)	I	11.9 ± 4.6	10.2 ± 4.9	11.4 ± 3.3	12.6 ± 4.9
		S	19.5 ± 5.6	29.0 ± 10.8	23.0 ± 6.2	29.6 ± 8.3
$Ca_v1.2$	4 x 2 (5/6)	I	13.9 ± 3.2	12.3 ± 5.8	11.5 ± 2.6	13.9 ± 7.0
		S	21.7 ± 3.8	30.9 ± 13.6	22.1 ± 7.6	37.0 ± 6.5
RyR1	1(4) x (2+1) (5/6)	I	12.5 ± 2.0	8.1 ± 3.5	17.7 ± 4.8	12.3 ± 7.3
		S	18.6 ± 9.5	18.4 ± 4.0	44.8 ± 19.2	34.5 ± 11.8
IP_3R	1(4) x 2 (5/6)	I	20.2 ± 5.9	7.2 ± 0.5	11.9 ± 4.8	14.8 ± 5.2
		S	29.9 ± 4.9	25.6 ± 9.0	33.0 ± 4.5	32.5 ± 3.5
MCU	1(5) x 1 (2)	I	11.1	15.0	10.5	9.1
		S	14.8	25.0	21.1	27.3
SERCA	1 x 4	I	10.4 ± 2.8	15.0 ± 6.2	10.2 ± 2.5	14.7 ± 6.9
		S	28.8 ± 2.7	40.0 ± 7.8	21.5 ± 8.0	35.6 ± 10.9
PL/mv			3	2	3	7
Ca/mv				2		1
Total			3	4	3	8
PLN	domain II	I	12.0	9.1	9.1	13.0
		S	28.0	45.5	40.9	52.2
STIM1	1(2) x 1	I	6.9	16.7	14.3	20.0
		S	20.7	37.5	47.6	40.0

Vpu of human immunodeficiency virus type 1 (HIV-1) is identified as a member of the class of channel forming proteins, also called viroporins, with low ion selectivity [46,47]. Analysis of the sequence alignment on the basis of the individual p-TMDs with specifically the TMD of Vpu (Figure 5)

reveals that p-TMD4 has the largest I value of 25.9 % followed by p-TMD1 with 22.2 % (Suppl. Table 3). On the basis of S values, p-TMD2 with 56.5 % stands out compared to p-TMD4 of 51.9 %.

TMDs 2 and 4 are proposed to have an active role in the function of NS4B. Small changes in the sequence, as for p-TMDs versus r-TMDs, do not affect the results very much. It is like a second experiment under similar conditions.

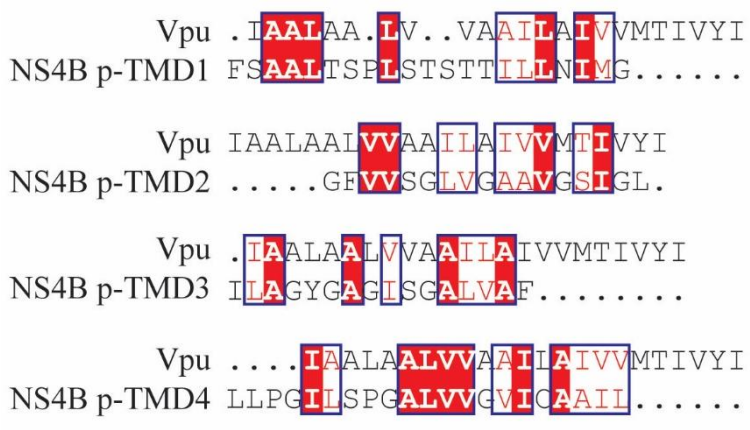

Figure 5. Sequence alignment on the basis of the individual p-TMDs with specifically the TMD of Vpu of HIV-1 (UniProt A0A076V5C1). Red boxes indicate identical residues, red letters similar residues. The blue frame indicates global similarity > 0.7.

3.4. Sequence alignment with Nsp3 of coronaviruses

The sequences of the two TMDs in each of the four Nsp3 proteins are all predicted within narrow stretches of the amino acid sequence by the secondary structure prediction programs (Suppl. Figure 4). Since the prediction of TMD2 includes parts of the proposed ECTO domain [48], two type of stretches are chosen for pairwise alignment, one which refers to the stretch proposed by the secondary structure prediction programs which then overlaps with the stretch found to be part of the ECTO domain (TMD2*), and one stretch in which this part is not included (TMD2).

The sequences for both TMDs hypothesizes a sister group for SARS and SARS2 with differences in sequence similarity of 12–19 % (Suppl. Figure 5). MERS is shown as an outgroups sequence difference of 30–39 % to the sister group especially for TMD2* and TMD1, respectively (Suppl. Figure 5A,B). From this clade, MHV is hypothesized to have 36–40 % differences. For TMD2-(-E) the SARS sequences and MHV for a clade (33 %) from which MERS shows a difference of 40 % in the sequence (Suppl. Figure 5C).

Individual alignment of the four TMDs of NS4B with the two TMDs of Nsp3 of coronaviruses (Clustal Omega) (for sequence see Suppl. Tab. 4) reveals that p-TMDs 2 (7 times) and 4 (5 times) of NS4B show the largest number of highest I / S values amongst its four TMDs (Suppl. Table 6, and Suppl. Table 5). Using the stretch which includes the ECTO domain, p-TMD1 (4 times) shows the largest increase in number of hits (+3) compared to the value without the ECTO domain. It also shows highest numbers together with p-TMD 2 (4 times) and TMD4 (5 times). Choosing the r-TMDs for alignment (no ECTO domain) TMD1 (7 times) and r-TMD2 (6 times) are most often the highest values for I and S (Suppl. Tables 7 and 8). Including the ECTO domain reveals that r-TMDs 1 and 3 (6

and 5 times, respectively) show the highest number of hits with the latter showing the largest increase (+3) amongst the TMDs of NS4B.

p-TMDs 1 and potentially 3 show similarity to the ECTO domain of Nsp3. p-TMDs 4 and 2 are likely to be indifferent to the changes in the alignment protocol.

4. Discussion

With the advance of sequencing the genome rapidly an enormous amount of protein sequences is at hand. Increasingly slower is the pace of converting this information into functional, then structural and even dynamical information of the protein leading to the sequence–structure–function gap [2]. This gap is even more pronounced for membrane proteins, despite their importance in current therapy [49]. Computational tools can be seen as a complementary tool to the experimental techniques to enhance the speed of delivery of essential information [1–3]. Besides generating software able to predict particular features it is also essential to develop algorithms which can be used to achieve meaningful information. In this respect, a flow-chart is presented using bioinformatic tools to achieve potential structure–function correlation. This chart predicts the structural features of the target protein on a first level, here e.g. transmembrane helical motifs, and correlates this motif with information available in sequence and structure data banks on the basis of protein sequence alignment (level 1) and confined alignment (level 2) on functional information obtained from level1 and also potentially from experimental sources.

4.1. Reliability of the data

Multiple resources for obtaining a consensus stretch of the length of the TMDs are used [50]. Sequence alignment programs such as BLAST are widely used to align identified proteins as well as nucleotides with each other to obtain functional information about the query sequence and with it the protein. BLAST shows good performance by combining speed of processing with reliability of the data [51]. The use of one sequence alignment software is compensated by the use of different lengths of amino acid sequences to probe the stability of the data.

4.2. Sequence – structure relation

Predicting the stretch of the TMDs, a large number of programs are used. These programs vary in the algorithm applied to propose the stretches and thus provide proposals which might differs considerably from each other. A threshold of more than 50 % of the programs propose the TMD stretch, is chosen. In addition, a criterion is to allow for a stretch to be long enough to span the bilayer. In some cases, the alignment overlap of all the programs is so that the p-TMDs are almost similar to the r-TMDs predicted by others. In one case the overlap is rather poor, a finding which is also observed for the TMDs of other membrane proteins [50]. It has been argued that this suggest anomalous behavior in terms of the folding of the stretch and consequently on the mechanics and function of the protein. In this respect, TMD1 of NS4B is to some respect close to a region where it is reported that the protein could flip through the membrane [15,17,52].

4.3. Sequence–function–relation

The relation is based on the correlation between amino acid sequence and the biological function followed by structural conservation, or phrased differently, that structural evolution is slower than the amino acid evolution [53]. The same functional and structural feature can be anticipated, if the amino acid sequence most similar. Still, non-homologous proteins in respect to their amino acid sequence can adopt the same structure as long as a particular structural motif is needed to perform a specific task [54]. This is particularly the case for the myoglobin / hemoglobin relation where both proteins share very little sequence identity but a lot of structural similarity [55]. It is anticipated that this is in order to take care of the potential discrepancy between amino acids and structural features by using the protocol of using sequence alignment to identify common features amongst proteins for which structural or functional data are available. At this end, the bioinformatic approach identifies a correlation of the NS4B sequence with sequences of proteins of which their functions are found to be involved in regulation of Ca-homeostasis. Thus, the presented data support experimental findings [20].

In addition, similarity of the TMDs 1 and 3 of NS4B with the TMDs of Nsp3 proteins proposes a similar function for NS4B of being involved in the formation of the replication vesicles. Combining the findings, involvement in Ca-homeostasis and vesical formation, Ca-regulation seems to be important for the latter.

5. Conclusions

Based on these matching results, the proposal of TMDs 2 and 4 playing an active part in homeostasis either as being part of a channeling structural feature or being part of a transporting mechanism, seems plausible. In addition, parts of the stretch containing TMD1 and TMD3 are potentially involved in the formation of replication vesicles.

Experimentally the involvement of NS4B into Ca-homeostasis is proposed, and it would also have been proposed by using the bioinformatics approach. Structural prediction shows the ambiguity for the first TMD, and this sequence stretch in particular, as the programs vary quite substantially in predicting a TMD stretch. Overall, a bioinformatics approach gives thorough evidence of the function and structure of a membrane concerning its transmembrane region.

Based on secondary structure prediction and a two-level sequence alignment approach, the functional role of the TMDs of NS4B is proposed. This algorithm is readily transferable to other proteins and when automatized a valuable tool for predicting structure – function correlations.

Acknowledgments

WBF thanks the Ministry of Science and Technology Taiwan (MOST-104-2112-M-010-001-MY3) for financial support. We are grateful to the National Center for High-performance Computing Taiwan for computer time and facilities.

Conflict of interest

The authors declare no conflict of interests.

References

1. KcDB (2017) Recent advances in sequence-based protein structure prediction. *Brief Bioinform* 18: 1021–1032.
2. Rost B, Sander C (1996) Bridging the protein sequence-structure gap by structure predictions. *Annu Rev Biophys Biomol Struct* 25: 113–136.
3. Schwede T (2013) Protein modeling: What happened to the "protein structure gap"? *Structure* 21: 1531–1540.
4. Torrisi M, Pollastri G, Le Q (2020) Deep learning methods in protein structure prediction. *Comput Struct Biotechnol J* 18: 1301–1310.
5. Almeida JG, Preto AJ, Koukos PI, et al. (2017) Membrane proteins structures: A review on computational modeling tools. *BBA Biomembr* 1859: 2021–2039.
6. Korkmaz S, Duarte JM, Prlić A, et al. (2018) Investigation of protein quaternary structure via stoichiometry and symmetry information. *PLoS One* 13: e0197176.
7. Nealon JO, Philomina LS, McGuffin LJ (2017) Predictive and experimental approaches for elucidating protein–protein interactions and quaternary structures *Int J Mol Sci* 18: 2623.
8. Chowdhury B, Garai G (2017) A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics* 109: 419–431.
9. Saw AK, Tripathy BC, Nandi S (2019) Alignment-free similarity analysis for protein sequences based on fuzzy integral. *Sci Rep* 9: 2775.
10. Cherstvy AG (2009) Positively charged residues in DNA-binding domains of structural proteins follow sequence-specific positions of DNA phosphate groups. *J Phys Chem B* 113: 4242–4247.
11. Moradpour D, Penin F (2013) Hepatitis C virus proteins: from structure to function, *Hepatitis C Virus: from Molecular Virology to Antiviral Therapy*, Berlin: Springer, 113–142.
12. Paul D, Madan V, Ramirez O, et al. (2018) Glycine zipper motifs in hepatitis C virus nonstructural protein 4B are required for the establishment of viral replication organelles. *J Virol* 92: e01890–01817.
13. Lundin M, Monné M, Widell A, et al. (2003) Topology of the membrane-associated hepatitis C virus protein NS4B. *J Virol* 77: 5428–5438.
14. Lundin M, Lindström H, Grönwall C, et al. (2006) Dual topology of the processed hepatitis C virus protein NS4B is influenced by the NS5A protein. *J Gen Virol* 87: 3263–3272.
15. Palomares-Jerez F, Nemesio H, Villalaín J (2012) The membrane spanning domains of protein NS4B from hepatitis C virus. *BBA Biomembr* 1818: 2958–2966.
16. Einav S, Elazar M, Danieli T, et al. (2004) A nucleotide binding motif in hepatitis C virus (HCV) NS4B mediates HCV RNA replication. *J Virol* 78: 11288–11295.
17. Gouttenoire J, Penin F, Moradpour D (2010) Hepatitis C virus nonstructural protein 4B: a journey into unexplored territory. *Rev Med Virol* 20: 117–129.
18. Gouttenoire J, Montserret R, Paul D, et al. (2014) Aminoterminal amphipathic α-Helix AH1 of hepatitis C virus nonstructural protein 4B possesses a dual role in RNA replication and virus production. *PLoS Pathog* 10: e1004501.

19. Yu GY, Lee KJ, Gao L, et al. (2006) Palmitoylation and polymerization of hepatitis C virus NS4B protein. *J Virol* 80: 6013–6023.
20. Fogeron ML, Jirasko V, Penzel S, et al. (2016) Cell-free expression, purification, and membrane reconstitution for NMR studies of the nonstructural protein 4B from hepatitis C virus. *J Biomol NMR* 65: 87–98.
21. Dobson L, Reményi I, Tusnády GE (2015) CCTOP: a consensus constrained TOPology prediction web server. *Nucleic Acids Res* 43: W408–W412.
22. Cserzö M, Wallin E, Simon I, et al. (1997) Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng* 10: 673–676.
23. Cserzo M, Eisenhaber F, Eisenhaber B, et al. (2004) TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. *Bioinformatics* 20: 136–137.
24. Tusnády GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics* 17: 849–850.
25. Shen H, Chou JJ (2008) MemBrain: improving the accuracy of predicting transmembrane helices. *PLOS One* 3: e2399.
26. Jones DT, Taylor WR, Thornton JM (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* 33: 3038–3049.
27. Jones DT (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* 23: 538–544.
28. Käll L, Krogh A, Sonnhammer ELL (2004) A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 338: 1027–1036.
29. Daley DO, Rapp M, Granseth E, et al. (2005) Global topology analysis of the Escherichia coli inner membrane proteome. *Science* 308: 1321–1323.
30. Viklund H, Elofsson A (2008) OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* 24: 1662–1668.
31. Reynolds SM, Käll L, Riffle ME, et al. (2008) Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput Biol* 4: e1000213.
32. Käll L, Krogh A, Sonnhammer ELL (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* 21: i251–i257.
33. Viklund H, Elofsson A (2004) Best α-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci* 13: 1908–1917.
34. Peters C, Tsirigos KD, Shu N, et al. (2016) Improved topology prediction using the terminal hydrophobic helices rule. *Bioinformatics* 32: 1158–1162.
35. Juretić D, Zoranić L, Zucić D (2002) Basic charge clusters and prediction of membrane protein topology. *J Chem Inf Comput Sci* 42: 620–632.
36. Krogh A, Larsson B, Von Heijne G, et al. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305: 567–580.
37. Kahsay RY, Gao G, Liao L (2005) An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. *Bioinformatics* 21: 1853–1858.
38. Hofmann K, (1993) TMbase-A database of membrane spanning protein segments. *Biol Chem* 374: 166.

39. Tsirigos KD, Peters C, Shu N, et al. (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res* 43: W401–W407.
40. Buchan DWA, Jones DT (2019) The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Res* 47: W402–W407.
41. Hildebrand PW, Preissner R, Frömmel C (2004) Structural features of transmembrane helices. *FEBS Lett* 559: 145–151.
42. Saidijam M, Azizpour S, Patching SG (2018) Comprehensive analysis of the numbers, lengths and amino acid compositions of transmembrane helices in prokaryotic, eukaryotic and viral integral membrane proteins of high-resolution structure. *J Biomol Struct Dyn* 36: 443–464.
43. Blum M, Chang HY, Chuguransky S, et al. (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* 49: D344–D354.
44. Sigrist CJA, Cerutti L, Hulo N, et al. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* 3: 265–274.
45. Li S, Ye L, Yu X, et al. (2009) Hepatitis C virus NS4B induces unfolded protein response and endoplasmic reticulum overload response-dependent NF-kappaB activation. *Virology* 391: 257–264.
46. Carrasco L (1978) Membrane leakiness after viral infection and a new approach to the development of antiviral agents. *Nature* 272: 694–699.
47. Mehnert T, Routh A, Judge PJ, et al. (2008) Biophysical characterisation of Vpu from HIV-1 suggests a channel-pore dualism. *Proteins* 70: 1488–1497.
48. Lei J, Kusov Y, Hilgenfeld R (2018) Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. *Antiviral Res* 149: 58–74.
49. Flock T, Venkatakrishnan AJ, Vinothkumar KR, et al. (2012) Deciphering membrane protein structures from protein sequences. *Genome Biol* 13: 160.
50. Cuthbertson JM, Doyle DA, Sansom MSP (2005) Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Eng Des Sel* 18: 295–308.
51. Yan R, Xu D, Yang J, et al. (2013) A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Sci Rep* 3: 2619.
52. Palomares-Jerez MF, Nemesio H, Franquelim HG, et al. (2013) N-terminal AH2 segment of protein NS4B from hepatitis C virus. Binding to and interaction with model biomembranes. *BBA-Biomembr* 1828: 1938–1952.
53. Illergård K, Ardell DH, Elofsson A (2009) Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. *Proteins* 77: 499–508.
54. Sousounis K, Haney CE, Cao J, et al. (2012) Conservation of the three-dimensional structure in non-homologous or unrelated proteins. *Hum Genomics* 6: 1–10.
55. Doolittle RF (1981) Similar amino acid sequences: chance or common ancestry? *Science* 214: 149–159.

AIMS Press