EHG 51

ENCYCLOPEDIA OF THE HUMAN GENOME

2000

Transmembrane Domains

Transmembrane domains#membrane anchored#hydropathy#exon distribution#
XTX/XCX-enriched exons

Ramasarma, T.
T. Ramasarma
Department of Biochemistry
Indian Institute of Science, Bangalore 560 012, India

Joshi, N. V.
N. V. Joshi
Centre for Ecological Sciences
Indian Institute of Science, Bangalore 560 012, India

One or more passes of polypeptide chains of the lipid bilayer, each consisting of about 25 hydrophobic residues with occasional presence of a polar residue, anchor integral proteins in cell membrane. These are known as transmembrane domains that participate in functions of these proteins in some unspecified way. Encoded by random exon makeup, sequences of these short stretches are highly variable with hardly any repetition.

Distinctive architecture of membrane proteins

The human genome must be distinctive in some way. The "evolution of novel extracellular and transmembrane architecture" is the greatest innovation in the human lineage, according to the analysis of the International Human Genome Sequence Consortium (2001). Of the 32000 identified genes in the human genome, transmembrane proteins account for about 20%, a relatively high proportion compared to other species.

Acquiring new potentials depends on arrangement of the polypeptide chain around the membrane and its relation to the rest of the protein, implied by architecture. And this placing the protein in the membrane is programmed in the gene by nucleotide sequences encoding hydrophobic domains. Occupying high or low proportion of the total protein, these domains are not mere anchors but participate decisively in actions of some proteins. And it is all the more interesting that they possess uncomplicated helical structures of about 30Å in length with little or no help from the side-chains. Sequences, exon make-up and contributions in membrane activities of some examples of these little, versatile structures are briefly described here.

Membrane spanning proteins

Transmembrane (TM) proteins, also known as integral membrane proteins, are embedded in the lipid bilayer with the polypeptide chain crossing the membrane. Fig. 1 is a stylized diagram showing possible membrane passes. An extended polypeptide chain (Fig 1.1) even with hydrophobic residues normally can not pass the hydrophobic environment of the membrane because of its polar peptide units. But its segment can do so after forming hydrogen bonds between them, attaining thus a stable structure (α-helix or β-sheet; Fig. 1.2, 1.8). Also a pass of the membrane seems always completed leaving no loose end within the lipid-bilayer. Examples of β-sheet with multipass barrel-type arrangement are known in some channel proteins with about 10-12 residues for each hydrophobic stretch. A short sheet-turn-sheet loop structure half way into the lipid-bilayer (Fig. 1.7) along with other helices is considered typical of the pore region of many channels. Proteins with single, four and seven TM passes are common. Other examples of upto 14 and 17 spans also occur. It is appropriate to refer these as spanins appending the number of times the chain crosses the membrane.

The hydropathic analysis

A stretch of about 25 hydrophobic residues in a protein ideally fits a single α-helical pass of membrane lipid bilayer. The hydropathic analysis of Kyte and Doolittle (1982) is a convenient method of determining such stretches showing as positive peaks when the values of hydropathy of residues are plotted against the sequence number of the polypeptide (see Fig. 2 for an example). This method is widely used for predicting existence, location and number of TM domains. The following authors developed other algorithms: Klein, Kanehisa and De Lisi (1985), von Heijne (1992), Peron and Argos (1994), and Casadio and Sander (1995). Differences in number of domains are encountered using these methods. Words such as 'putative', 'predicted', 'potential' or 'purported' are used to qualify sequences thus identified.

Presence of a couple of hydrophilic residues is not uncommon in several sequences claimed to be TM domains. It is difficult sometimes to understand why such are included while in other cases with long enough stretches of hydrophobic residues are ignored. Presence of glycation sites in the vicinity and release of expected peptides on protease action from the connecting loops provide confirmatory evidence. Some helical sequences as short as 15 and some others as long as 40 residues are known and are referred as negative and positive 'hydrophobic mismatch', respectively (Monne and von Heijne,

2001). Such domains are likely to be accommodated in the membrane by thinning or thickening of its lipid layers and by tilting and bending of the helices (Fig. 1.5, 1.6). Indeed more will be learnt on membrane placement of these domains by studies on the lipids in the vicinity. Attachment to the membrane of these proteins is beyond doubt. Helical nature of folding is also probably correct as supported by the data of crystal structure, circular dichroism and two-dimensional NMR spectroscopy of some membrane proteins. Placing variable amounts of protein exposed in the two aqueous phases thus must serve some purpose.

Monospanins

One span across will suffice to firmly place a protein in the membrane. This simple design is used by families of proteins such as enzyme-linked receptors for signal transduction, transport proteins for moving compounds and ions in and out of the cell, and cell-surface proteins employed as recognition, linker and adhesion molecules. Fig.3 shows distribution of some examples of the polypeptide chain across the membrane.

Signals received by extracellular domains from a variety of ligands such as growth factors and peptide hormones are transferred into the cell where the protein kinase activity of intracellular domains becomes active (Fig. 3 a,b). How does the bit of membrane-locked α-helix with hardly any help expected from its hydrophobic residues transfer the signal into the cell? Such stimulation of ligand-sensitive kinase activity was retained in the absence of the lipid in a purified preparation of the receptor. The only link between the two bulk portions of the polypeptide on the two sides is the tiny TM domain. Signal transfer function must reside in its helical polypeptide backbone. A possible mode of action is mobilization and transfer of electrons in α-helix across its intrinsic supramolecular structure of helical sequences of alternating peptide group and hydrogen bond (HN-C=O..HN-C=O..) (Ramasarma, 2000).

Only short intracellular tails, sufficient to recognize the coated pits, are employed in transport proteins, both for internalization (LDL receptor) and externalization (IgA receptor) (Fig 3 c,d). The strategy of non-covalent dimerization of the membrane-bound polypeptides is used in the examples of receptors of insulin and transferrin (Fig. 3 b,e).

A large number of monospanin proteins that recognize and bind to cell surface structures (e.g., CD4, integrins and selectins) and form functional supramolecular complexes are known. These polypeptide chains (N-terminal) are extended outside the cell where action occurs by way of recognition of other molecules, substrates and signals. Extracellular placement of C-terminal of the polypeptide is also found in the examples of transferrin receptor, MRFP, corin (Fig.3.e,f,g).

Dispanins

The two-membrane spans occur in a few membrane-bound proteins. They have no common type of action. Some are proteins channel ions: a chloride channel (e.g., CFTCR), inward rectifier potassium channel, calcium homoeostasis endoplasmic reticulum protein and sodium channel (DEG-ENaC) (Fig. 4 a). Others have unrelated

enzyme activities such as acylCoA:cholesterol acyl transferase (ACAT), tyrosine phosphatase and an ecto-ATPase, uridine diphosphatase (Fig. 4 b). Both their N- and C-terminals are in the cytoplasm with the connecting loop as the extracellular domain.

Used in the small subunit c in F_0 part of mitochondrial ATP synthase this architecture plays a pivotal role in the process of energy transduction. This enzyme complex is considered a tiny molecular motor with a concentric ring of ten molecules of subunit c (Fig. 4c) surrounding the F_1-γ-subunit (Fig. 4d) that constitutes the rotor. And the subunit c acts as the link between electron transport and ATP synthase, both membrane-based. Thus this dispanin complex undergoes an extraordinary mechanical rotation as a part of the process of transferring energy. Nothing strikes in the sequences as different from other TM domains and the two of these are coded by two separate exons.

Trispanins

Occurrence of three spans in a membrane protein is infrequent. The example of leukotriene-C-4 synthase, involved in pathogenesis of asthma, has the middle span rather long and shown as tilted (Fig. 5). It is encoded by exon 3 up to the two polar residues in its middle and then changes to exon 4. The TM domains are near the N-terminal in a long chain in another example (muscle popeye gene product). In what way the three membrane domains contribute in making these proteins active besides anchoring is not known.

Tetraspanins

Proteins that span the membrane four times forming two extracellular loops, the second one usually large (e.g., CD9 antigen, Fig 5 b), are referred as transmembrane 4 superfamily (TM4SF), simplified as tetraspanins. Known for their action as molecular facilitators in signaling, adhesion, differentiation and proliferation, they bring together large molecular complexes by interacting with proteins such as integrins and other receptors. They are unfairly labeled as 'promiscuous' because their liaison with other proteins is widespread, albeit not entirely random. The tetraspanins include a large number of CD proteins (9, 37, 53, 63, 81 and 82), receptors of GABA and glycine and also some ion channels. In the voltage-dependent potassium channel the center of the associated four helices forms a gate.

Pentaspanins

Proteins known to span the membrane five times are few. The examples of CD 47, AC133, hcLcA2, M83 and Cig30 have extended N-terminal extracellular domains (Fig 5 c). They bind to integrins, slectins and other adhesion molecules. Their functions appear to be similar to tetraspanins. The need for five spans, however, remains ambiguous.

Hexaspanins

Proteins with six spans have a variety of actions as enzymes (phosphatidate phosphatase, type III adenyl cyclase), as channel proteins (HCN2), as transporters (ZnT-3, PI transfer

protein) and as growth factor activators (LMP1). The last two spans of LMP1 are necessary for anchoring the protein but not sufficient for its action on nFκB. This strategy of assignment of protein functions to different spans seems to be used in this and other multispanin proteins.

A growing superfamily of ATP-dependent proteins that translocate amphiphilic and lipophilic substrates belong to this category. Adenylyl cyclase (type III), multi drug resistance-ATPase and cystic fibrosis TM regulator belong to this group. A half size transporter containing a single nucleotide-binding domain, ABCG1, has six TM helices coded by five exons 11-15 with two exons contributing partly to the sequences of domains II and IV (Langmann et al., 2000).

The 7-TM proteins, heptaspanins

By far the best-known TM proteins are the G protein-coupled receptors on the cell surface, characterized by seven membrane spans with N-terminal outside and C-terminal inside the cell (Fig. 6 a). An arrangement of clockwise connectivity of the helices believed to be oriented perpendicular to the membrane was proposed. This provides a membrane-embedded surface of the receptor protein (Fig. 6b). There are several receptor families for the ligands such as noradrenaline, acetylcholine, serotonin, peptides, glycoprotein hormones, adenosine, prostaglandin E2 and thromboxane A2. A good proportion of the polypeptide in these is conserved, in contrast to the monospanins. It is also utilized to build the seven TM domains and the loop between V and VI domains for the β-adrenergic receptor (Emorine etal, 1989), and all these are coded in one unusually long exon (Table 1.7). Binding a ligand on the surface of helices outside the cell leads to dissociating a subunit of G protein acting as the second messenger system inside the cell. Here lies the enigma. How is the information carried through the simple architecture provided by TM helices? In the case of adrenergic receptors, it was found that the essential amide group of noradrenaline binds to helix III. Taking advantage of rotation of its C1-C7 bond (see Fig. 6 b) noradrenaline can bind its other essential group, meta-OH either to helix V or VII. It was proposed that such a choice of helix-pairs may be used by different ligands of the multiple subtypes of these receptors as in adrenergic system.

Multispanins

Multiple spans beyond seven are known in some proteins. Generally these are channel proteins. Reports on 8- and 9-spans are rare. Several 10-span proteins are known as transporters of amino acids and also as cotransporters of chloride and bicarbonate. With even number spans both N- and C-terminals have to face the same side of the membrane, and they are more commonly inside the cell. The surface of the helices of the TM domains is connected by short loops and these few residues are therefore important in the action of the receptor. For example, glucose-6-phosphate transporter utilizes the polypeptide to build it's10 spans (I-X) of residues ranging 18 - 30 with short connecting loops (Pan et al., 1999). These are encoded by 8 exons (1-8) thus: I -1; II -2; III & IV - 3, V -4; VI - 5, VII - 6, VIII - 7; IX - 7 & 8 (G/TG); X - 8. Larger number of spans of 11, 12, 13, 14 and 17, are known to exist in some transporter/channel proteins. Any

arrangement of such a large number of spans is expected to provide multiple sites needed for action of these proteins.

Exon analysis of the nucleotide sequences of cDNA corresponding to transmembrane domains

Dominating the small stretches of TM domains are hydrophobic amino acids, I, V, L, F, C, M, A, G and W. Hydroxy amino acids, S, T and Y and also the helix-breaker, P, occur frequently in these helices. With this many residues, these can have innumerable sequences. And they do, conserving only hydrophobicity. No repetition of a sequence or a part of it was found in the examples of TM domains studied (Table 1). These are coded by exons, necessarily differing in sequences, in many ways: one exon coding for one or more domains; one domain coded by two exons with the split occurring between residues and in some cases between the nucleotides of a triplet. No doubt desired sequences are fused thus, but hardly any repeating units are noticed. In the example of CD9 antigen each of the four domains (I-IV) is coded by two exons (1-8) thus: I - 1 & 2; II -2 & 3 (triplet G/GA); III - 4 & 5; IV - 7 & 8 (Boucheix et al., 1991) (Table 1.6). No generalization is possible with choice of the sequences so random in these domains.

The hydrophobic residues, F, L, I, M, and V, are coded by the second letter T, and the second letter C codes for S, P, T, and A in triplets in cDNA. An analysis of % nucleotide present in the second position showed the expected abundance of the pyrimidines, T followed by C, in the domain sequences (Table 1). Indeed values combined for T and C account for about 70%. Thus these stretches in the exons do show a repeating pattern of XTX or XCX. Animals also share the architecture of multiple TM domains and thus the feature that makes human genome distinct is something beyond these domains.

References

Boucheix C, Benoit P, Bachet P, Billard M. Worthington RE, Gagnon J and Uzan G (1991) Molecular cloning of CD9 antigen, a new superfamily of cell surface proteins. *Journal of Biological Chemistry*, **266**, 117-122.

Casadio RB, Fariselli R and Sander C (1995) Transmembrane helices predicted at 95% accuracy. *Protein Science* **4**, 521-533.

Emorine LJ, Marullo S, Briend-Sutrn, MM, Patey, G, Tate K, Delavier-Klutchko C and Strosberg AD (1989) Molecular characterization of the human beta 3-adrenergic receptor. *Science* **245**,1118-1121.

Klein P, Kanehisa M and Di Lisis C (1985) The detection and classification of membrane-spanning proteins. *Biochimica Biophysica Acta* **815**, 468-476.

Langmann T, Porsch-Orzcurumez M, Unkelbach, U, Kulcken J and Schmitz, G (2000) Genomic organization and characterization of the promoter of the human ATP-binding cassette transporter-G1 (ABCG1) gene. *Biochimica Biophysica Acta* **1494**, 175-180.

Monne, M and von Heijne G (2001) Effects of 'hydrophobic mismatch' on the location of transmembrane helices in the ER membrane *FEBS Letters* **496**,96-100.

Kyte J and Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**, 105-132.

Pan CJ, Lin B and Chou JY (1999) Transmembrane topology of human glucose 6-phosphate transporter. *Journal of Biological Chemistry* **274**, 13865-13869.

Peron B and Argos P (1994) Prediction of transmembrane segments in proteins utilizing multiple sequence alignments. *Journal of Molecular Biology* **237**, 182-192.

Ramasarma T (2000) In praise of the hydrogen bond In: Lal M, Lillford PJ, Naik VM and Prakash V (ed) *Supramolecular and Colloidal Structures in Biomaterials and Biosubstrates* pp. 450-462. London, U.K, Imperial College Press and the Royal Society.

von Heijne G (1992) Membrane protein structure prediction: hydrophobic analysis and the positive-inside rule. *Journal Molecular Biology* **225**, 487-494.

Further reading

Baldwin JM (1993) The probable arrangement of the helices in G protein-coupled receptors. *The EMBO Journal* **12**,1693-1703.

Benovic JL, Bovier M, Caron, MG, Lefkowitz RL (1980) Regulation of adenylc cyclase-coupled β-adrenergic receptors. *Annual Reviews of Cell Biology* **4**,405-428.

Maecker HT, Todd SC and Levy S (1997) The tetraspanin superfamily: molecular facilitators. *FASEB Journal* **11**, 428-442.

Ramasarma T (1996) Transmembrane domains participate in functions of integral membrane proteins. *Indian Journal of Biochemistry and Biophysics* **33**, 20-29.

Stock D, Leslie AGW and Walker JE (1999) Molecular architecture of the rotary motor in ATP synthase. *Science* **286**, 1700-1705.

Yardley Y and Ulrich A (1998) Growth factor receptor tyrosine kinases. *Annual Reviews of Biochemistry* **57**,473-478.

Table 1. Sequences of amino acids of TM domains and of nucleotides of the corresponding cDNA segments with their distribution in exons. Chromosomal location of the gene and the percent values of nucleotides T, C, A and G occurring in second position of the code are given in parentheses. T and C represent hydrophobic residues. Notice the exon split occurs between residues and also within the triple (example 6).

---

1. **α-Platelet-derived growth factor receptor** (chromosome: 4q11-q13)

TM domain I (T:C:A:G::79:21: 00:00)

```
 L  T  V  A  A  A  V  L  V  L  L  V  I  V  I  I  S  L  I  V  L  V  V  I  W
CTCACGGTGGCTGCTGCAGTCCTGGTGCTGTTGGTGATTGTGATCATCTCACTTATTGTCCTGGTTGTCATTTGG
|← exon 10                                                                exon 10 →|
```

2. **Insulin receptor β-subunit** (chromosome: 19p13.3-p13.2)

TM domain I (T:C:A:G::74: 04: 04:18)

```
 I  I  I  G  P  L  I  F  V  F  L  F  S  V  V  I  G  S  I  Y  L  F  L
ATTATCATCGGCCCCCTCATCTTTGTCTTTCTCTTCACTGTTGTCATTGGAACTATTTATCTATTCCTG
|←exon 14                                                        exon 14→|
```

3. **LDL receptor** (chromosome 19p13.3)

TM domain I (T:C:A:G:: 77: 09: 00:14)

```
 A  L  S  I  V  L  P  I  V  L  L  V  F  L  L  L  G  V  F  L  L  W
CTCCTGTCCATTGTCCTCCCCATC. GTGCTCCTCGTCTTCCTTTGCCTGGGGTCTTCCTTCTATGG
|←exon 16                →| ←exon 17                          exon 17 →|
```

4. **Transferrin receptor** (chromosome: 3q26.2-qter)

TM domain I (T:C:A:G::46:08:15:31)

```
 S  G  S  I  C  Y  G  T  I  A  V  I  V  F  F  L  I  G  F  M  I  G  Y  L  G  Y
AGTGGAAGTATCTGCTATGGGACTATTGCTGTGATCGTCTTTTTCTTGATTGGATTTATGATTGGC TACTTGGGCTAT
| ←exon 3                                                          exon 3 → |
```

5. **Fo-ATPase subunit c** (P1 form) (chromosome: 2-pter-2qter)

TM domain I (T:C:A:G::34:28:04:34)

```
 F  I  G  A  G  A  A  T  V  G  V  A  G  S  G  A  G  I  G  T  V  F  G  S  L  I  I  G  Y  A
TTTATTGGTGCTGGGGCAGCCACAGTTGGTGTGGCTGGTTCAGGGGCTGGCATTGGAACCGTG TTTGGCAGCTTGATCATTGGCTATGCC
| ←      exon 4                                                                exon 4 →|
```

TM domain II (T:C:A:G::62:24:03:11)

```
 L  F  S  Y  A  I  L  G  F  A  L  S  E  A  M  G  L  F  C  L  M  V  A  F  L  I  L  F  A  M
CTCTTCTCCTATGCCATTCTTGGCTTTGCCCTGTCTGAGGCCATGGGGCTTTTCTGTTTGATGGTCGCCTTCCTCATCCTCTTCGCCATG
|←       exon 5                                                              exon 5 →|
```

6. **CD9 antigen** (chromosome: 12p13)

TM domain I (T:C:A:G::62:14:05:19)

```
 L  L  F  G  F  N  F  I  F  W  L  A  G  I  A  V  L  A  I  G  L
CTGCTGTTCGGATTTAACTTCATCTTCTGG..CTTGCCGGGATTGCTGTCCTTGCCATTGGACTA
|← exon 1                    →|←exon 2                        →|
```

TM domain II (T:C:A:G::57:14:10:19)

F Y T G V Y I L I G A G A L M M L V G F L

TTCTACACAG..GAGTCTATATTCTGATCGGAGCCGGCGCCCTCATGATGCTGGTG GGCTTCCTG

|← exon 2 →|←exon 3 exon 3→|

TM domain III (T:C:A:G::38:21:33:08)

V I F A I E I A A A I W G Y S H K D E V I K E V

GTGATATTCGCCATTGAAATAGCTGCGGCCATCTGGGGATATTCCCACAAGGATGAG..GTGATTAAGGAAGTC

| ←exon 4 exon 4 →| ← exon 5 →|

TM domain IV (T:C:A:G::63:12:00:25)

A V G I G I A V V M I F G M I F S M I L C C A I

GCAGTGGGCATCGGCATTGCCGTGGTCATG..ATATTTGGCATGATCTTCAGTATGATCTTGTGCTGTGCTATC

| ← exon 7 →|← exon 8 exon 8→|

7. **Adrenergic receptor β3** (8p11.1-8p12)

TM domain I (T:C:A:G::48:37:04:11)

A A L A G A L L A L A V L A T V G G N L L V I V A I A

GCGGCCCTAGCCGGGGCCCTGCTGGCGCTGGCGGTGCTGGCCACCGTGGGAGGCAACCTGCTGGTCATCGTGGCCATCGCC

| exon 1 exon 1→|

TM domain II (T:C:A:G::56:26:11:05)

N V F V T S L A A A D L V M G L L V V

AACGTGTTCGTGACTTCGCTGGCCGCAGCCGACCTGGTGATGGACTCCTGGTGGTG

|←exon 1 →|

TM domain III (T:C:A:G::41:32:09:18)

L W T S V D V L C V T A S I E T L C A L A V

CTGTGGACCTCGGTGGACGTGCTGTGTGTGACCGCCAGCATCGAAACCCTGTGCGCCCTGGCCGTG

|←exon 1 →|

TM domain IV (T:C:A:G::43:36:04:17)

T A V V L V W V V S A A V S F A P I M S Q W W

ACAGCTGTGGTCCTGGTGTGGGTCGTGTCGGCCGCGGTGTCGTTTGCGCCCATCATGAGCCAGTGGTGG

|←exon 1 exon 1→|

TM domain V (T:C:A:G::59:27:14:00)

Y V L L S S S V S F Y L P L L V M L F V Y A

TACGTGCTGCTGTCCTCCTCCGTCTCCTTCTACCTTCCTCTTCTCGTGATGCTCTTCGTCTACGCG

|←exon 1 →|

TM domain VI (T:C:A:G::54:23:05:18)

T L G L I M G T F T L C W L P F F L A N V L

ACCTTGGGTCTCATCATGGGCACCTTCACTCTCTGCTGGTTGCCCTTCTTTCTGGCCAACGTGCTG

|←exon 1 →|

TM domain VII (T.C.A.G: 38.24.19.19)

A F L A L N W L G Y A N S A F N P L I Y C

GCTTTCCTTGCCCTGAACTGGCTAGGTTATGCCAATTCTGCCTTCAACCCGCTCATCTACTGC

|←exon 1 →|

Legends for Figures

Figure 1. Arrangement of polypeptide chain in membrane spans. Membrane bilayer is represented as two lines with middle broken line. The possible membrane passes are shown: 1. extended polypeptide chain normally not found; 2. α-helix shown as a box; 3. short helix, negative mismatch; 4. long helix, positive mismatch; 5. tilted helix; 6.bent helix; 7. half occupied sheet characteristic of channel proteins; 8. β-sheet shown as parallel arrows.

Figure 2. Hydropathic plot according to Kyte and Doolittle (1982). The hatched peak, corresponds to the hydrophobic residues given below, the purported membrane span of insulin precursor protein.

Figure 3. Distribution of polypeptide chain in some typical monospanin proteins. The membrane span is shown as a box with the polypeptide chain extended into the extracellular and cytoplasmic sides. The number of residues of each domain is given. a. α-platelet-derived growth factor receptor; b. insulin receptor; c. low density lipoprotein receptor; d. polyIg receptor; e. transferrin receptor; f. membrane-type frizzled-related protein; g. corin; h. CD4 protein

Figure 4. Architecture of multimeric subunit c of F_0-ATPase and some dispanin proteins. a. DEC/ENaC, a sodium channel; b. UDPase, an ecto ATPase; c. P1 form of c-subunit of F_0-ATPase. (Where shown by arrow the signal peptide is clipped. The hatched helix forming the outer ring has the conserved residue E); d. arrangement of the 10 subunits around the two helices of subunit γ of F_0-ATPase which is a part of the rotary unit, as viewed from one side of the membrane.

Figure 5. Distribution of the polypeptides of tri-, tetra- and penta-spanin proteins: a. leukotriene C4 synthase (long middle helix is shown tilted); b. CD9 antigen; c. M83 protein. Hydrophilic residues occurring within the helices are shown.

Figure 6. β-adrenergic receptor; distribution of the polypeptide (a), and the proposed architecture with clockwise connectivity of the helices viewed from the extracellular side (b). Critical for activity are the short loops on the extracellular side and the loop between helices V-VI and the C-terminal chain. Noradrenaline is proposed to bind the helices as shown in (b), and its meta-OH can bind to either helix V or VII by rotation of the molecule indicated by arrow

Glossary

Transmembrane: across the membrane, spanning the membrane lipid bilayer

Hydropathy: degree of hydrophobicity; proportion of non-polar residues

Spanins: short stretches of polypeptide pass in membrane lipid bilayer

Domain: a distinctive portion of the protein identified with a specific role

Extracellular: outside the cell (projecting out of the plasma membrane)

Intracellular: inside the cell (projecting into the cytoplasm)

Transmembrane Domains

T. Ramasarma[1] and N. V. Joshi [2]
Department of Biochemistry[1] and Centre for Ecological Sciences[2], Indian Institute of Science, Bangalore 560 012, India

The human genome must be distinctive in some way. The greatest innovation in the human lineage is the "evolution of novel extracellular and transmembrane architecture" according to the analysis of the International Human Genome Sequence Consortium (2001). Of the 32000 identified genes in the human genome, transmembrane proteins that carry out multiple chores for the cell account for about 20%. This is a relatively high proportion compared to other species. The arrangement of the polypeptide chain around the membrane and its relation to the rest of the protein, as implied by the word, architecture, will have a bearing on acquiring new potentials. And this is programmed in the gene by suitable nucleotide sequences that code for the required hydrophobic domains for placing the protein in the membrane. These domains are not mere anchors but participate decisively in actions of some proteins. Their versatility, sequences and exon make-up are briefly described here.

Membrane spanning proteins

Transmembrane proteins, also known as integral membrane proteins, are embedded in the lipid bilayer with the polypeptide chain crossing the membrane. Fig. 1 is a stylized diagram showing possible membrane passes. An extended polypeptide chain even with hydrophobic residues normally can not pass the hydrophobic environment of the membrane because of its polar peptide units. But its segment can do so after forming hydrogen bonds between them, attaining thus a stable structure (α-helix or β-sheet). Also a pass of the membrane seems always completed leaving no loose ends within the lipid-bilayer.

Helical structures that predominate in membrane domains are invariably hydrophobic.

Examples of β-sheet are known in some channel proteins and about 10-12 residues will suffice for each hydrophobic stretch in the multipass barrel-type arrangement. A short sheet-turn-sheet loop structure half way into the lipid-bilayer along with other helices is considered typical of the pore region of many channels (Soman et al., 1995).

Proteins with single and seven transmembrane passes are common. Amazingly other examples of 2-6, 8-14 and 17 also occur. Referring these as spanins appending the number of times the chain crosses the membrane is appropriate (Maecker et al., 1997). Placing variable amounts of protein exposed in the two aqueous phases thus must serve some purpose.

The hydropathic analysis

A stretch of about 25 hydrophobic residues in a protein ideally fits a single α-helical pass of membrane lipid bilayer. The hydropathic analysis of Kyte and Doolittle (1982) is a convenient method of determining such stretches showing as positive peaks when the values of hydropathy of residues are plotted against the sequence number of the polypeptide (see Fig. 2 for an example). The hydrophobicity values of amino acids in decreasing order are given below: isoleucine (+4.5), valine (+4.2), leucine (+3.8), phenylalanine (+2.8), cysteine (+2.5), methionine (+1.9), alanine (+1.8), proline (+1.6), tyrosine (+1.3), tryptophan (+0.9), glycine (-0.4), threonine (-0.7), serine (-0.8), histidine (-3.2), aspartate (-3.5), glutamate (-3.5). asparagine (-3.5), glutamine (-3.5), lysine (-3.9) and.arginine (-4.5). This method is widely used for predicting existence, location and number of transmembrane domains. The following authors developed other algorithms albeit not much in use: Klein, Kanehisa and De Lisi (1985), von Heijne (1992), Person and Argos (1994), and Casadio and Sander (1995). Differences in number of domains and sequences are encountered using these methods. Words such as 'putative', 'predicted', 'potential' or 'purported' are used to qualify sequences thus identified.

Presence of a couple of possibly functional hydrophilic residues is not uncommon in several sequences claimed to be transmembrane domains by the hydropathic analysis. It is difficult sometimes to understand why such are included while in other cases with long enough stretches are ignored. Presence of glycation sites in the vicinity and release of expected peptides on protease action from the connecting loops provide confirmatory evidence. Referred to as 'hydrophobic mismatch' (Kilian, 1998) between the membrane and the domain, some sequences, assumed to be helical, as short as 15 (negative) and some as long as 40 residues (positive), are known (von Heijne, 2000). Such domains are likely to be accommodated in the membrane by thinning or thickening of its lipid layers and by tilting and bending of the helices (Fig. 1). Indeed more will be learnt of membrane placement of these domains by studies on the lipids in the vicinity. Attachment to the membrane of these proteins is beyond doubt and the conjecture of helical nature of folding is probably correct as found by the data of crystal structure, circular dichroism and two-dimensional NMR spectroscopy of some membrane proteins.

Functions of transmembrane proteins

A membrane is a place for vigorous cellular activity and not a passive divider between two aqueous phases. Proteins studded on the membrane vary in their density from low in the plasma membrane to high in the inner membrane of mitochondria. Selectively anchored in these sites these proteins carry out a wide range of functions. These include catalysis (more commonly hydrolases) that aid in material transfer, transport across the membrane by binding and internalizing ligands and by forming channels, receptor-linked transduction of signals from hormones and growth factors, electron transport through the complexes and energy transfer, recognition and adhesion of cell surface matrix.

Most of these are selective functions for which membrane localization is meaningful. Mitochondrial inner membrane is chosen for capturing energy of electron transport in the form of ATP. Plasma membrane is a veritable 'beehive' of activity transferring signals and materials across. The membrane-based proteins elegantly adapt their basic catalytic

potential, and in this transmembrane domains have a role to play, whether occupying high or insignificant proportion of the total protein. And it is all the more interesting that they possess uncomplicated helical structures of about 30Å in length with little or no help from the side-chains. Contribution of these little structures in membrane activities will be described here with some examples.

Monospanins

One span across will suffice to firmly place a protein in the membrane. This simple design is used by a number of families of proteins such as enzyme-linked receptors for signal transduction, transport proteins for moving compounds and ions in and out of the cell, and cell-surface proteins employed as recognition, linker and adhesion molecules. The distribution of the polypeptide chain across the membrane for a set of representative samples is shown schematically in Fig.3. These polypeptide chains (N-terminal) are extended outside the cell where most action occurs by way of recognition of other molecules, substrates and signals take place. Extracellular placement of C-terminal of the polypeptide is also known (e.g., transferrin receptor, MRFP, corin).

Long enough polypeptides possessing kinase activities, adapted to respond to the signals from a variety of ligands such as growth factors and peptide hormones, are built in the intracellular domains (C-terminal) where signal transfer into the cell occurs. Stimulation of ligand-sensitive kinase activity was retained in purified preparation of the receptor in the virtual absence of the lipid. The only link between the two bulk portions of the polypeptide on the two sides is the tiny transmembrane domain. Hydrophobic residues being of little help signal transfer function must reside in its α-helical polypeptide backbone. How does the bit of membrane-locked α-helix transfer the signal into the cell? Signal-linked electron mobilization and transfer across the intrinsic supramolecular structure of helical sequences of alternating peptide group and hydrogen bond (..HN-C=O..HN-C=O..) in α-helix, had been surmised to form the basis of actions of these transmembrane domains (Ramasarma, 1985).

Only short intracellular tails sufficient to recognize the coated pits are employed in transport proteins, both for internalization (LDL receptor) and externalization (IgA receptor). Non-covalent dimers of the membrane-bound polypeptides are also used (e.g., insulin receptor, transferrin receptor and MHC class II). A large number of monospanin proteins exist that recognize and bind to cell surface structures such as integrins and selectins, and form functional supramolecular complexes. The cells utilize the extraordinary variation of the extracellular domains of these proteins, obtained by the constituent mosaic of exons

Dispanins

The two-membrane spans occur in a few membrane bound proteins with no identifiable common type of action. Some are proteins channel ions: a chloride channel (eg., CFTCR), inward rectifier potassium channel, calcium homoeostasis endoplasmic reticulum protein and sodium channel (DEG-ENaC). Others have enzyme activities bering no relationship to each other such as acylCoA:cholesterol acyl transferase

(ACAT), tyrosine phosphatase and an ecto-ATPase, uridine diphosphatase. Both their N- and C-terminals are in the cytoplasm with the connecting loop as the extracellular domain (Fig 4).

Used in the small subunit c in F_0 part of mitochondrial ATP synthase this architecture plays a pivotal role in the process of energy transduction. This enzyme complex is considered a tiny molecular motor with a concentric ring of ten molecules of subunit c surrounding the F_1-γ-subunit (Fig. 4d) that constitutes the rotor. And the subunit c acts as the link between the membrane-based electron transport and the ATP synthase. Thus this dispanin complex does an extraordinary rotation in transferring energy.

Trispanins

Occurrence of three spans in a membrane protein is infrequent. Two examples are leukotriene-C-4 synthase involved in pathogenesis of asthma and muscle popeye gene product. The example shown in Fig. 5 has the middle span rather long and thereby possibly tilted with two polar residues in its middle. The transmembrane domains are near the N-terminal in a long chain in another example. The question whether the membrane helices contribute in making these proteins active besides anchoring remains to be answered.

Tetraspanins

Proteins that span the membrane four times forming two extracellular loops (Fig 5b) are referred to as transmembrane 4 superfamily (TM4SF) simplified as tetraspanins (Mecker et al, 1997). Known for their action as molecular facilitators in signaling, adhesion, differentiation and proliferation, they bring together large molecular complexes by interacting with proteins such as integrins and other receptors. They are unfairly labeled as 'promiscuous' because their liaison with other proteins, is widespread, albeit not entirely random. The tetraspanins include a large number of CD proteins (9,37, 53, 63, 81 and 82), receptors of GABA and glycine and also some ion channels. The center of the associated four helices forms a gate in the voltage-dependent potassium channel (Zhou et al, 2001).

Pentaspanins

The few examples of proteins, CD 47, AC133, hcLcA2 and Cig30, known to span the membranes five times have extended extracellular domains (Fig 5c). They also bind to integrins, slectins and other adhesion molecules. Their functions appear to be similar to tetraspanins. The need for five spans, however, remains ambiguous.

Hexaspanins

Proteins with six spans have a variety of actions -enzyme activity (phosphtidate phosphatase, type III adenyl cyclase), a channel (HCN2), a transporter (ZnT-3, PI transfer protein) and growth factor activation (LMP1). The last two spans of LMP1 are necessary for anchoring the protein but not sufficient for its action on nFκB. The growing

superfamily of ATP-dependent translocation of amphiphilic and lipophilic substrates belong to this category that include multi drug resistance-ATPase and cystic fibrosis transmembrane regulator. A half size transporter with single nucleotide-binding domain, ABCG1, with the characteristic Walker motif, coded by single exon 3, has six transmembrane helices coded by five exons.

The 7-TM proteins, heptaspanins

By far the best-known transmembrane proteins are the G protein-coupled receptors on the cell surface, characterized by seven membrane spans with N-terminal on the outside and C-terminal inside the cell (Fig. 6). There are several receptor families for the ligands such as noradrenaline, acetylcholine, serotonin, peptides, glycoprotein hormones, adenosine, prostaglandin E2 and thromboxane A2. Binding a ligand to the membrane-embedded surface of the receptor protein leads to dissociating a subunit of G protein on the other side of the membrane with no obvious interaction or chemical reaction. The G protein acts as the second messenger system (Baldwin, 1993). Here is the enigma. How is the information carried inside the cell through the architecture provided by transmembrane helices, believed to be oriented perpendicular to the membrane? In an arrangement of clockwise connectivity of the helices (Baldwin, 1993) helix III, binds to the essential amide group of noradrenaline (Benovic et al., 1980). Either helix V or VII may bind to the other essential group, meta-OH, giving trans-α or trans-β forms of the noradrenaline, respectively, taking advantage of rotation of its C1-C7 bond (see insert, Fig. 6b). It was surmised that choice of helix-pairs may thus be put to use by the multiple subtypes of these receptors (Ramasarma, 1995)

Multispanins

Multiple spans beyond seven are known in some proteins. Generally these are used for channel proteins. Reports are rare on 8- and 9-spans. Several 10-span proteins are known as transporters of amino acids and glucose-6-phosphate and also as cotransporters of chloride and bicarbonate. With even number spans both N- and C-terminals have to face the same side of the membrane and they are more commonly inside the cell. The surface of the helices of the transmembrane domains are connected by short loops and these few amino acids residues are therefore important in the action of the receptor for transmitting the signal or providing a channel. Larger number of spans of 11, 12, 13, 14 and 17 are known to exist in some transporter/channel proteins, decided mostly by the hydropathic analysis. In the case of sulfonyl urea receptor (SUR1) values of 15, 16 and 17 are obtained with different methods. Analysis showed that even these short stretches are invariably encoded by different exons, with one or two of these coded by two exons with split between residues and in some case within the triplet. The arrangement of such a large number of spans is still to be worked out but it is surmised to provide for multiple sites needed for action of these proteins.

Other method neural network hidden Markov

Also occur in other exons

Exon Analysis

References

Further Reading

XX(T/C)XX

Legends for figures

Table 1

Sequences of transmembrane domain (amino acids and cDNA)	%			
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20	T	C	A	G
1. α-Platelet-derived growth factor receptor (4q11-q13)				
L T V A A A V L V L L V I V I I S L I V				
CTCACGGTGGCTGCTGCAGTCCTGGTGCTGTTGGTGATTGTGATCATCTCACTTATTGTC				
| exon 10 |				
L V V I W				
CTGGTTGTCATTTGG	79	21	0	0
| exon 10 (contd) |				
2. Insulin receptor β-subunit (19)				
G V P F R H L L L V L Q L a L L P A A T				
GGAGTCCCTTTTAGGCACTTGCTTCTGGTGCTGCAACTGGCGCTCCTCCCAGCAGCCACT	50	30	10	10
| exon 1 | exon 2 |				
3. LDL receptor				
A L S I V L P I V L L V F L L L G V F L L W				
CTCCTGTCCATTGTCCTCCCCATCGTGCTCCTCGTCTTCCTTTGCCTGGGGTCTTCCTTCTATGG	77	9	0	14
| exon 16 | exon 17 |				
4. Fo-ATPase subunit c (P1 form, TMD 1 and 2)				
FIGAGAATVGVAGSGAGIGTVFGSLIIGYAL				
LFSYAILGFALSEAMGLFCLMVAFLILFAM				

X X T/C A X

ENCYCLOPEDIA OF THE HUMAN GENOME
2000

Transmembrane Domains

T. Ramasarma[1] and N. V. Joshi [2]
Department of Biochemistry[1] and Centre for Ecological Sciences[2], Indian Institute of Science, Bangalore 560 012, India

The human genome must be distinctive in some way. The "evolution of novel extracellular and transmembrane architecture" is the greatest innovation in the human lineage, according to the analysis of the International Human Genome Sequence Consortium (2001). Of the 32000 identified genes in the human genome, transmembrane proteins account for about 20%. This is a relatively high proportion compared to other species. Acquiring new potentials depends on arrangement of the polypeptide chain around the membrane and its relation to the rest of the protein, implied by architecture. And this placing the protein in the membrane is programmed in the gene by nucleotide sequences encoding hydrophobic domains. Occupying high or low proportion of the total protein, these domains are not mere anchors but participate decisively in actions of some proteins. And it is all the more interesting that they possess uncomplicated helical structures of about 30Å in length with little or no help from the side-chains. Sequences, exon make-up and contributions in membrane activities of some examples of these little versatile, structures are briefly described here.

Membrane spanning proteins

Transmembrane (TM) proteins, also known as integral membrane proteins, are embedded in the lipid bilayer with the polypeptide chain crossing the membrane. Fig. 1 is a stylized diagram showing possible membrane passes. An extended polypeptide chain (Fig 1.1) even with hydrophobic residues normally can not pass the hydrophobic environment of the membrane because of its polar peptide units. But its segment can do so after forming hydrogen bonds between them, attaining thus a stable

structure (α-helix or β-sheet; Fig. 1.2, 1.8). Also a pass of the membrane seems always completed leaving no loose end within the lipid-bilayer. Examples of β-sheet with multipass barrel-type arrangement are known in some channel proteins with about 10-12 residues for each hydrophobic stretch. A short sheet-turn-sheet loop structure half way into the lipid-bilayer (Fig. 1.7) along with other helices is considered typical of the pore region of many channels (Soman et al., 1995). Proteins with single, four and seven TM passes are common. Other examples of upto 14 and 17 spans also occur. It is appropriate to refer these as spanins appending the number of times the chain crosses the membrane (Maecker et al., 1997).

The hydropathic analysis

A stretch of about 25 hydrophobic residues in a protein ideally fits a single α-helical pass of membrane lipid bilayer. The hydropathic analysis of Kyte and Doolittle (1982) is a convenient method of determining such stretches showing as positive peaks when the values of hydropathy of residues are plotted against the sequence number of the polypeptide (see Fig. 2 for an example). This method is widely used for predicting existence, location and number of TM domains. The following authors developed other algorithms: Klein, Kanehisa and De Lisi (1985), von Heijne (1992), Peron and Argos (1994), and Casadio and Sander (1995). Differences in number of domains are encountered using these methods. Words such as 'putative', 'predicted', 'potential' or 'purported' are used to qualify sequences thus identified.

Presence of a couple of hydrophilic residues is not uncommon in several sequences claimed to be TM domains. It is difficult sometimes to understand why such are included while in other cases with long enough stretches of hydrophobic residues are ignored. Presence of glycation sites in the vicinity and release of expected peptides on protease action from the connecting loops provide confirmatory evidence. Some helical sequences as short as 15 and some others as long as 40 residues are known and are referred as negative and positive 'hydrophobic mismatch', respectively (Monne and von Heijne, 2001). Such domains are likely to be accommodated in the membrane by thinning or thickening of its lipid layers and by tilting and bending of the helices (Fig. 1.5, 1.6). Indeed more will be learnt on membrane placement of these domains by studies on the lipids in the vicinity. Attachment to the membrane of these proteins is beyond doubt. Helical nature of folding is also probably correct as supported by the data of crystal structure, circular dichroism and two-dimensional NMR spectroscopy of some membrane proteins. Placing variable amounts of protein exposed in the two aqueous phases thus must serve some purpose.

Monospanins

One span across will suffice to firmly place a protein in the membrane. This simple design is used by families of proteins such as enzyme-linked receptors for signal transduction, transport proteins for moving compounds and ions in and out of the cell, and cell-surface proteins employed as recognition, linker and adhesion molecules. Fig.3 shows distribution of some examples of the polypeptide chain across the membrane.

Signals received by extracellular domains from a variety of ligands such as growth factors and peptide hormones are transferred into the cell where the protein kinase activity of intracellular domains becomes active (Fig. 3 a,b). How does the bit of membrane-locked α-helix with hardly any help expected from its hydrophobic residues transfer the signal into the cell? Such stimulation of ligand-sensitive kinase activity was retained in the absence of the lipid in a purified preparation of the receptor (Yarden and Ulrich, 1988). The only link between the two bulk portions of the polypeptide on the two sides is the tiny TM domain. Signal transfer function must reside in its helical polypeptide backbone. A possible mode of action is mobilization and transfer of electrons in α-helix across its intrinsic supramolecular structure of helical sequences of alternating peptide group and hydrogen bond (HN-C=O..HN-C=O..) (Ramasarma, 1999).

Only short intracellular tails, sufficient to recognize the coated pits, are employed in transport proteins, both for internalization (LDL receptor) and externalization (IgA receptor) (Fig.3c,d). The strategy of non-covalent dimerization of the membrane-bound polypeptides is used in the examples of receptors of insulin and transferrin (Fig. 3b,e).

A large number of monospanin proteins that recognize and bind to cell surface structures (e.g., CD4 integrins and selectins) and form functional supramolecular complexes are known. These polypeptide chains (N-terminal) are extended outside the cell where action occurs by way of recognition of other molecules, substrates and signals. Extracellular placement of C-terminal of the polypeptide is also found in the examples of transferrin receptor, MRFP, corin (Fig.3,e,f,g).

Dispanins

The two-membrane spans occur in a few membrane-bound proteins. They have no common type of action. Some are proteins channel ions: a chloride channel (e.g., CFTCR), inward rectifier potassium channel, calcium homoeostasis endoplasmic reticulum protein and sodium channel (DEG-ENaC) (Fig.4a). Others have unrelated enzyme activities such as acylCoA:cholesterol acyl transferase (ACAT), tyrosine phosphatase and an ecto-ATPase, uridine diphosphatase (Fig. 4b). Both their N- and C-terminals are in the cytoplasm with the connecting loop as the extracellular domain.

Used in the small subunit c in F_0 part of mitochondrial ATP synthase this architecture plays a pivotal role in the process of energy transduction. This enzyme complex is considered a tiny molecular motor with a concentric ring of ten molecules of subunit c (Fig. 4c) surrounding the F_1-γ-subunit (Fig. 4d) that constitutes the rotor (Stock et al., 1999). And the subunit c acts as the link between electron transport and ATP synthase, both membrane-based. Thus this dispanin complex undergoes an extraordinary mechanical rotation as a part of the process of transferring energy. Nothing strikes in the sequences as different from other TM domains and the two of these are coded by two separate exons (Dyer and Walker, 1993).

Trispanins

Occurrence of three spans in a membrane protein is infrequent. The example of leukotriene-C-4 synthase (Welch et al., 1994), involved in pathogenesis of asthma,

has the middle span rather long and shown as tilted (Fig. 5). It is encoded by exon 3 up to the two polar residues in its middle and then changes to exon 4. The TM domains are near the N-terminal in a long chain in another example (muscle popeye gene product). In what way the three membrane domains contribute in making these proteins active besides anchoring is not known.

Tetraspanins

Proteins that span the membrane four times forming two extracellular loops, the second one usually large (e.g., CD9 antigen, Fig 5b), are referred as transmembrane 4 superfamily (TM4SF), simplified as tetraspanins (Maecker et al, 1997). Known for their action as molecular facilitators in signaling, adhesion, differentiation and proliferation, they bring together large molecular complexes by interacting with proteins such as integrins and other receptors. They are unfairly labeled as 'promiscuous' because their liaison with other proteins is widespread, albeit not entirely random. The tetraspanins include a large number of CD proteins (9,37, 53, 63, 81 and 82), receptors of GABA and glycine and also some ion channels. In the voltage-dependent potassium channel the center of the associated four helices forms a gate (Zhou et al, 2001).

Pentaspanins

Proteins known to span the membrane five times are few. The examples of CD 47, AC133, hcLcA2, M83 and Cig30 have extended N-terminal extracellular domains (Fig 5c). They bind to integrins, slectins and other adhesion molecules. Their functions appear to be similar to tetraspanins. The need for five spans, however, remains ambiguous.

Hexaspanins

Proteins with six spans have a variety of actions as enzymes (phosphatidate phosphatase, type III adenyl cyclase), as channel proteins (HCN2), as transporters (ZnT-3, PI transfer protein) and as growth factor activators (LMP1). The last two spans of LMP1 are necessary for anchoring the protein but not sufficient for its action on nFκB (Hatzivassiliou et al., 1998). This strategy of assignment of protein functions to different spans seems to be used in this and other multispanin proteins.

A growing superfamily of ATP-dependent proteins that translocate amphiphilic and lipophilic substrates belong to this category. Adenylyl cyclase (type III), multi drug resistance-ATPase and cystic fibrosis TM regulator belong to this group. A half size transporter containing a single nucleotide-binding domain, ABCG1, has six TM helices coded by five exons 11-15 with two exons contributing partly to the sequences of domains II and IV (Langmann et al., 2000).

The 7-TM proteins, heptaspanins

By far the best-known TM proteins are the G protein-coupled receptors on the cell surface, characterized by seven membrane spans with N-terminal outside and C-terminal inside the cell (Fig. 6a). An arrangement of clockwise connectivity of the helices believed to be oriented perpendicular to the membrane was proposed

(Baldwin, 1993). This provides a membrane-embedded surface of the receptor protein (Fig. 6b). There are several receptor families for the ligands such as noradrenaline, acetylcholine, serotonin, peptides, glycoprotein hormones, adenosine, prostaglandin E2 and thromboxane A2. A good proportion of the polypeptide in these is conserved, in contrast to the monospanins. It is also utilized to build the seven TM domains and the loop between V and VI domains for the β-adrenergic receptor (Emorine etal, 1989), and all these are coded in one unusually long exon (Table 1.7). Binding a ligand on the surface of helices outside the cell leads to dissociating a subunit of G protein acting as the second messenger system inside the cell (Baldwin, 1993). Here lies the enigma. How is the information carried through the simple architecture provided by TM helices? In the case of adrenergic receptors, it was found that the essential amide group of noradrenaline binds to helix III (Benovic et al, 1980). Taking advantage of rotation of its C1-C7 bond (see Fig. 6b) noradrenaline can bind its other essential group, meta-OH either to helix V or VII. It was proposed that such a choice of helix-pairs may be used by different ligands of the multiple subtypes of these receptors as in adrenergic system (Ramasarma, 1995)

Multispanins

Multiple spans beyond seven are known in some proteins. Generally these are channel proteins. Reports on 8- and 9-spans are rare. Several 10-span proteins are known as transporters of amino acids and also as cotransporters of chloride and bicarbonate. With even number spans both N- and C-terminals have to face the same side of the membrane, and they are more commonly inside the cell. The surface of the helices of the TM domains is connected by short loops and these few residues are therefore important in the action of the receptor. For example, glucose-6-phosphate transporter utilizes the polypeptide to build it's10 spans (I-X) of residues ranging 18 - 30 with short connecting loops (Pan and Chou, 1999). These are encoded by 8 exons (1-8) thus: I -1; II -2; III & IV - 3, V -4; VI - 5, VII - 6, VIII - 7; IX - 7 & 8 (G/TG); X - 8. Larger number of spans of 11, 12, 13, 14 and 17, are known to exist in some transporter/channel proteins. Any arrangement of such a large number of spans is expected to provide multiple sites needed for action of these proteins.

Exon analysis of the nucleotide sequences of cDNA corresponding to transmembrane domains

Dominating the small stretches of TM domains are hydrophobic amino acids, I, V, L, F, C, M, A, G and W. Hydroxy amino acids, S, T and Y and also the helix-breaker, P, occur frequently in these helices. With this many residues, these can have innumerable sequences. And they do, conserving only hydrophobicity. No repetition of a sequence or a part of it was found in the examples of TM domains studied (Table 1). These are coded by exons, necessarily differing in sequences, in many ways: one exon coding for one or more domains; one domain coded by two exons with the split occurring between residues and in some cases between the nucleotides of a triplet. No doubt desired sequences are fused thus, but hardly any repeating units are noticed. In the example of CD9 antigen each of the four domains (I-IV) is coded by two exons (1-8) thus: I - 1 & 2; II -2 & 3 (triplet G/GA); III - 4 & 5; IV - 7 & 8 (Boucheix et al., 1991) (Table 1.6). No generalization is possible with choice of the sequences so random in these domains.

The hydrophobic residues, F, L, I, M, and V, are coded by the second letter T, and the second letter C codes for S, P, T, and A in triplets in cDNA. An analysis of % nucleotide present in the second position showed the expected abundance of the primidines, T followed by C, in the domain sequences (Table 1). Indeed values combined for T and C account for about 70%. Thus these stretches in the exons do show a repeating pattern of XXTXX or XXCXX. Animals also share the architecture of multiple TM domains and thus the feature that makes human genome distinct is something beyond.

References (suggested reading *)

Baldwin JM (1993) The probable arrangement of the helices in G protein-coupled receptors. *The EMBO Journal* **12**,1693-1703.

Benovic JL, Bovier M, Caron, MG, Lefkowitz RL (1980) Regulation of adenylc cyclase-coupled β-adrenergic receptors. *Annual Reviews of Cell Biology* **4**,405-428. *

Boucheix C, Benoit P, Bachet P, Billard M. Worthington RE, Gagnon J and Uzan G (1991) Molecular cloning of CD9 antigen, a new superfamily of cell surface proteins. *Journal of Biological Chemistry*, **266**, 117-122.

Casadio RB, Fariselli R and Sander C (1995) Transmembrane helices predicted at 95% accuracy. *Protein Science* **4**, 521-533.

Dyer, MR and Walker JE (1993) Sequences of members of the human gene family for the c subunit of mitochondrial ATP synthase. *Biochemical Journal*, **293**, 51-64.

Emorine LJ, Marullo S, Briend-Sutrn, MM, Patey, G, Tate K, Delavier-Klutchko C and Strosberg AD (1989) Molecular characterization of the human beta 3-adrenergic receptor. *Science* **245**,1118-1121.

Hatzivassiliou E, Miller WE, Raab-Traub N, Keiff E and Mosialos G (1998) *Journal of Immunology* **160**, 1116-1121.

Klein P, Kanehisa M and Di Lisis C (1985) The detection and classification of membrane-spanning proteins. *Biochimica Biophysica Acta* **815**, 468-476.

Langmann T, Porsch-Orzcurumez M, Unkelbach, U, Kulcken J and Schmitz, G (2000) Genomic organization and characterization of the promoter of the human ATP-binding cassette transporter-G1 (ABCG1) gene. *Biochimica Biophysica Acta* **1494**, 175-180.

Maecker HT, Todd SC and Levy S (1997) The tetraspanin superfamily: molecular facilitators. *FASEB Journal* **11**, 428-442. *

Monne, M and von Heijne G (2001) Effects of 'hydrophobic mismatch' on the location of transmembrane helices in the ER membrane *FEBS Letters* **496**,96-100.

Kyte J and Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**, 105-132.

Pan CJ, Lin B and Chou JY (1999) Transmembrane topology of human glucose 6-phosphate transporter. *Journal of Biological Chemistry* **274**, 13865-13869.

Peron B and Argos P (1994) Prediction of transmembrane segments in proteins utilizing multiple sequence alignments. *Journal of Molecular Biology* **237**, 182-192.

Ramasarma T (1996) Transmembrane domains participate in functions of integral membrane proteins. *Indian Journal of Biochemistry and Biophysics* **33**, 20-29. *

Ramasarma T (2000) In praise of the hydrogen bond In: Lal M, Lillford PJ, Naik VM and Prakash V (ed) *Supramolecular and Colloidal Structures in Biomaterials and Biosubstrates* pp. 450-462. London, U.K, Imperial College Press and the Royal Society.

Soman CKV, McCann JA and Brown AM (1995) *Protein Engineering*, 8, 397-401

Stock D, Leslie AGW and Walker JE (1999) Molecular architecture of the rotary motor in ATP synthase. *Science* **286**, 1700-1705.

von Heijne G (1992) Membrane protein structure prediction: hydrophobic analysis and the positive-inside rule. *Journal Molecular Biology* **225**, 487-494.

Welsch DJ, Creely DP, Hauser SD, Mathis KJ, Krivi GG and Isakson PV (1994) Molecular cloning and expression of human leukotriene-C-4 synthase. *Proceedings of the National Academy of Sciences (USA)* **91**, 9745-9749

Yardley Y and Ulrich A (1998) Growth factor receptor tyrosine kinases. *Annual Reviews of Biochemistry* **57**,473-478.

Zhou M, Morais-Cabral JH, Mann S and Mackinnon R (2001) Potassium channel receptor site for the inactivator gate and quarternary amine inhibition. *Nature* **411**, 657-667.

Legends for Figures

Figure 1. Arrangement of polypeptide chain in membrane spans. Membrane bilayer is represented as two lines with middle broken line. The possible membrane passes are shown: 1. extended polypeptide chain normally not found; 2. α-helix shown as a box; 3. short helix, negative mismatch; 4. long helix, positive mismatch; 5. tilted helix; 6.bent helix; 7. half occupied sheet characteristic of channel proteins; 8. β-sheet shown as parallel arrows.

Figure 2. Hydropathic plot according to Kyte and Doolittle (1982). The hatched peak, corresponds to the hydrophobic residues given below, the purported membrane span of insulin precursor protein.

Figure 3. Distribution of polypeptide chain in some typical monospanin proteins. The membrane span is shown as a box with the polypeptide chain extended into the extracellular and cytoplasmic sides. The number of residues of each domain is given. a. α-platelet-derived growth factor receptor; b. insulin receptor; c. low density lipoprotein receptor; d. polyIg receptor; e. transferrin receptor; f. membrane-type frizzled-related protein; g. corin; h. CD4 protein

Figure 4. Architecture of multimeric subunit c of F_0-ATPase and some dispanin proteins.
a. DEC/ENaC, a sodium channel; b. UDPase, an ecto ATPase; c. P1 form of c-subunit of F_0-ATPase. (Where shown by arrow the signal peptide is clipped. The hatched helix forming the outer ring has the conserved residue E); d. arrangement of the 10 subunits around the two helices of subunit γ of F_0-ATPase which is a part of the rotary unit, as viewed from one side of the membrane.

Figure 5. Distribution of the polypeptides of tri-, tetra- and penta-spanin proteins: a. leukotriene C4 synthase (long middle helix is shown tilted); b. CD9 antigen; c. M83 protein. Hydrophilic residues occurring within the helices are shown.

Figure 6. β-adrenergic receptor; distribution of the polypeptide (a), and the proposed architecture with clockwise connectivity of the helices viewed from the extracellular side (b). Critical for activity are the short loops on the extracellular side and the loop between helices V-VI and the C-terminal chain. Noradrenaline is proposed to bind the helices as shown in (b), and its meta-OH can bind to either helix V or VII by rotation of the molecule indicated by arrow

Table 1. Sequences of amino acids of TM domains and of nucleotides of the corresponding cDNA segments with their distribution in exons. Chromosomal location of the gene and the percent values of nucleotides T, C, A and G occurring in second position of the code are given in parentheses. T and C represent hydrophobic residues. Notice the exon split occurs between residues and also within the triplet (example 6).

---

1. α-Platelet-derived growth factor receptor (chromosome:4q11-q13)

```
 L  T  V  A  A  A  V  L  V  L  L  V  I  V  I  I  S  L  I  V  L  V  V  I  W
CTCACGGTGGCTGCTGCAGTCCTGGTGCTGTTGGTGATTGTGATCATCTCACTTATTGTCCTGGTTGTCATTTGG
|← exon 10                                                          exon 10 →|
```

(T:C:A:G::79:21: 00:00; TM domain I)

2. Insulin receptor β-subunit (chromosome: 19p13.3-p13.2)

```
 I  I  I  G  P  L  I  F  V  F  L  F  S  V  V  I  G  S  I  Y  L  F  L
ATTATCATCGGCCCCCTCATCTTTGTCTTTCTCTTCACTGTTGTCATTGGAACTATTTATCTATTCCTG
|←exon 14                                                      exon 14→|
```

(T:C:A:G::74: 04: 04:18; TM domain I)

3. LDL receptor (19p13.3)

```
 A  L  S  I  V  L  P  I  V  L  L  V  F  L  L  L  G  V  F  L  L  W
CTCCTGTCCATTGTCCTCCCCATC. GTGCTCCTCGTCTTCCTTTGCCTGGGGTCTTCCTTCTATGG
|←exon 16          →| ←exon 17                          exon 17 →|
```

(T:C:A:G:: 77: 09: 00:14; TM domain I)

4. Transferrin receptor (chromosome: 3q26.2-qter)

```
 S  G  S  I  C  Y  G  T  I  A  V  I  V  F  F  L  I  G  F  M  I  G  Y  L  G  Y
AGTGGAAGTATCTGCTATGGGACTATTGCTGTGATCGTCTTTTTCTTGATTGGATTTATGATTGGC
TACTTGGGCTAT
| ←exon 3                                                          exon 3 →
|
```

(T:C:A:G:: 46: 08:15:31; TM domain I)

5. Fo-ATPase subunit c (P1 form) (chromosome: 2-pter-2qter)

```
 F  I  G  A  G  A  A  T  V  G  V  A  G  S  G  A  G  I  G  T  V
TTTATTGGTGCTGGGGCAGCCACAGTTGGTGTGGCTGGTTCAGGGGCTGGCATTGGAACCGTG
| ←      exon 4                                                  →

 F  G  S  L  I  I  G  Y  A
TTTGGCAGCTTGATCATTGGCTATGCC
                    exon 4 →|
```

(T:C:A:G::34:28:04:34; TM domain I)

```
 L  F  S  Y  A  I  L  G  F  A  L  S  E  A  M  G  L  F  C  L  M  V
```



```
CTCTTCTCCTATGCCATTCTTGGCTTTGCCCTGTCTGAGGCCATGGGGCTTTTCTGTTTGATGGTC
|←        exon 5                                                    →

 A  F  L  I  L  F  A  M
GCCTTCCTCATCCTCTTCGCCATG
                  exon 5 →|
```

(T:C:A:G:: 62: 24: 03: 11: TM domain II)

6. CD9 antigen (chromosome: 12p13)

```
 L  L  F  G  F  N  F  I  F  W  L  A  G  I  A  V  L  A  I  G  L
CTGCTGTTCGGATTTAACTTCATCTTCTGG..CTTGCCGGGATTGCTGTCCTTGCCATTGGACTA
|← exon 1                        →|←exon 2                    →|
```

(T:C:A:G:: 62:14:05:19; TM domain I)

```
 F  Y  T  G  V  Y  I  L  I  G  A  G  A  L  M  M  L  V  G  F  L
TTCTACACAG..GAGTCTATATTCTGATCGGAGCCGGCGCCCTCATGATGCTGGTG GGCTTCCTG
|← exon 2   →|←exon 3                                    exon 3→|
```

(T:C:A:G::57:14:10:19; TM domain II)

```
 V  I  F  A  I  E  I  A  A  A  I  W  G  Y  S  H  K  D  E  V  I  K  E  V
GTGATATTCGCCATTGAAATAGCTGCGGCCATCTGGGGATATTCCCACAAGGATGAG..GTGATTAAGGAAGTC
| ←exon 4                                          exon 4  →| ← exon 5        →|
```

(T:C:A:G::38:21:33:08; TM domain III)

```
 A  V  G  I  G  I  A  V  V  M  I  F  G  M  I  F  S  M  I  L  C  C  A  I
GCAGTGGGCATCGGCATTGCCGTGGTCATG..ATATTTGGCATGATCTTCAGTATGATCTTGTGCTGTGCTATC
| ← exon 7                    →|← exon 8                          exon 8→
```

(T:C:A:G::63:12:00:25; TM domain IV)

7. Adrenergic receptor β3 (8p11.1-8p12)

```
 A  A  L  A  G  A  L  L  A  L  A  V  L  A  T  V  G  G  N  L  L  V  I  V
GCGGCCCTAGCCGGGGCCCTGCTGGCGCTGGCGGTGCTGGCCACCGTGGGAGGCAACCTGCTGGTCATCGTG
| exon 1                                                          →

 A  I  A
GCCATCGCC
exon 1  →|
```

(T:C:A:G::48:37:04:11; TM domain I)

```
 N  V  F  V  T  S  L  A  A  A  D  L  V  M  G  L  L  V  V
AACGTGTTCGTGACTTCGCTGGCCGCAGCCGACCTGGTGATGGACTCCTGGTGGTG
|←exon 1                                          →|
```

(T:C:A:G:: 56:26:11:05; TM domain II)

```
 L  W  T  S  V  D  V  L  C  V  T  A  S  I  E  T  L  C  A  L  A  V
CTGTGGACCTCGGTGGACGTGCTGTGTGTGACCGCCAGCATCGAAACCCTGTGCGCCCTGGCCGTG
|←exon 1                                                    →|

 T  A  V  V  L  V  W  V  V  S  A  A  V  S  F  A  P  I  M  S  Q  W  W
ACAGCTGTGGTCCTGGTGTGGGTCGTGTCGGCCGCGGTGTCGTTTGCGCCCATCATGAGCCAGTGGTGG
|←exon 1                                                    →|
```

(T:C:A:G::41:32:09:18; TM domain III)

```
 Y  V  L  L  S  S  S  V  S  F  Y  L  P  L  L  V  M  L  F  V  Y  A
TACGTGCTGCTGTCCTCCTCCGTCTCCTTCTACCTTCCTCTTCTCGTGATGCTCTTCGTCTACGCG
|←exon 1                                                    →|
```

(T:C:A:G::59:27:14:00; TM domain IV)

```
 T  L  G  L  I  M  G  T  F  T  L  C  W  L  P  F  F  L  A  N  V  L
ACCTTGGGTCTCATCATGGGCACCTTCACTCTCTGCTGGTTGCCCTTCTTTCTGGCCAACGTGCTG
|←exon 1                                                    →|
```

(T:C:A:G::54:23:05:18; TM domain VI)

```
 A  F  L  A  L  N  W  L  G  Y  A  N  S  A  F  N  P  L  I  Y  C
```

GCTTTCCTTGCCCTGAACTGGCTAGGTTATGCCAATTCTGCCTTCAACCCGCTCATCTACTGC
|←exon 1 →|
(T.C.A.G: 38.24.19.19; TM domain VII)

Prof. T. Ramasarma
/ T.Ramasarma
Department of Biochemistry
Apts.(res) Indian Institute of Science
18th Cross, Malleswaram
Bangalore-560012 (India)
(India)
phone: +91-80-3092538
+91-80-3346134
Email:
indiratangirala@rediffmail.com

T. Indira
408 Shashikiran
Bangalore-560055
phone:
trs@biochem.iisc.ernet.in

July 4,

2001
Mr. Michael Calais
Project Manager
Nature Publishing Group
The Macmillian Building
4 Crinan Street
London, NY 9XW, UK

Dear Mr. Calais:

Ref: EHG-51 Transmembrane Domains

I am attaching herewith the first draft of the manuscript and scan of 6 figures. I will mail the 2 copies of the hard copy, original drawings and the disc after hearing from you.

I have added Dr. NV Joshi as a co-author and he contributed greatly to improve this article.

I have not used the recommended typing style, as I am not good at this and could not get the services of a professional at this time.

Dr. Joshi and I are aware of excessive length. Even the text came to 5-single space- typed pages. Then we have references, a Table and 6 figures. If all the spanin groups have to be touched we needed this length. After taking a look at the matter I request your editors to suggest what to cut. I suppose we can do some trimming but it is better that the full version be seen first. Await your opinion.

I have taken extra time to reach this stage. I had not realized how vast is the available literature expaned since my article in 1996 and how difficult it is to present a comprehensive view on this subject. I am sorry for the delay and many unavoidable things contributed.

Yours sincerely,

(T. Ramasarma)

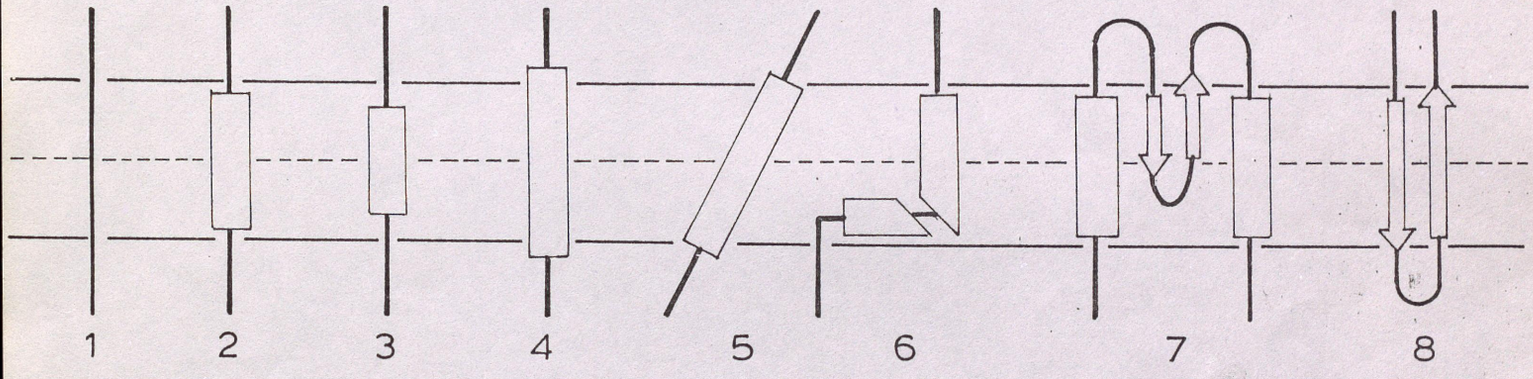

Fig. 1 Ramasarma/Joshi (EHG-51)

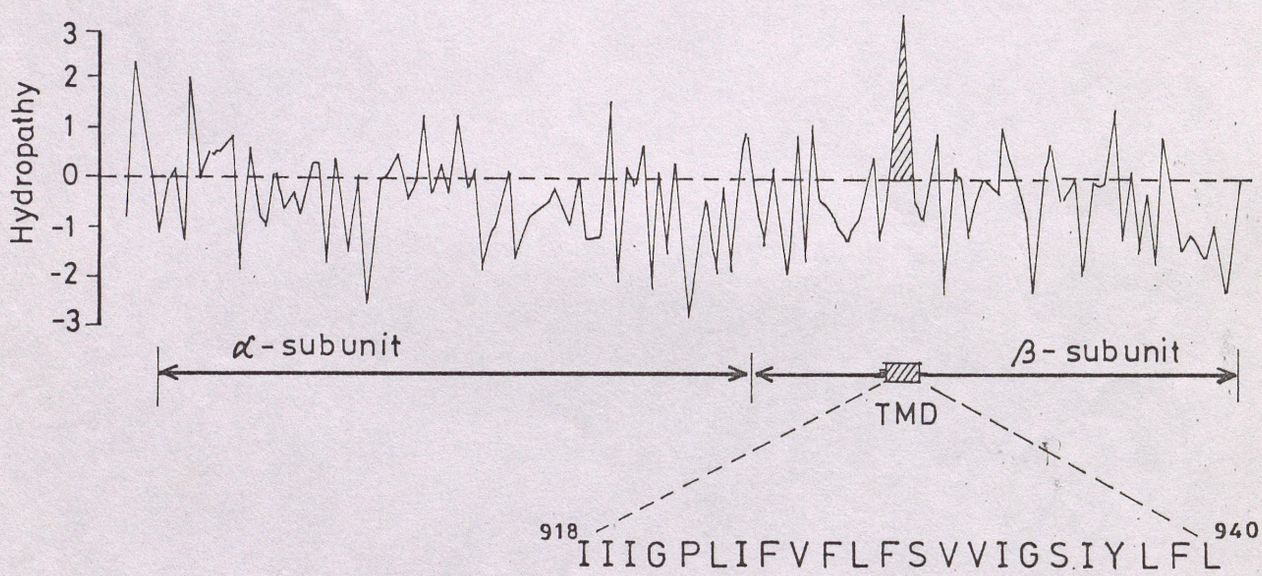

Fig. 2 Ramasarma/Joshi (EHG-51)

#2

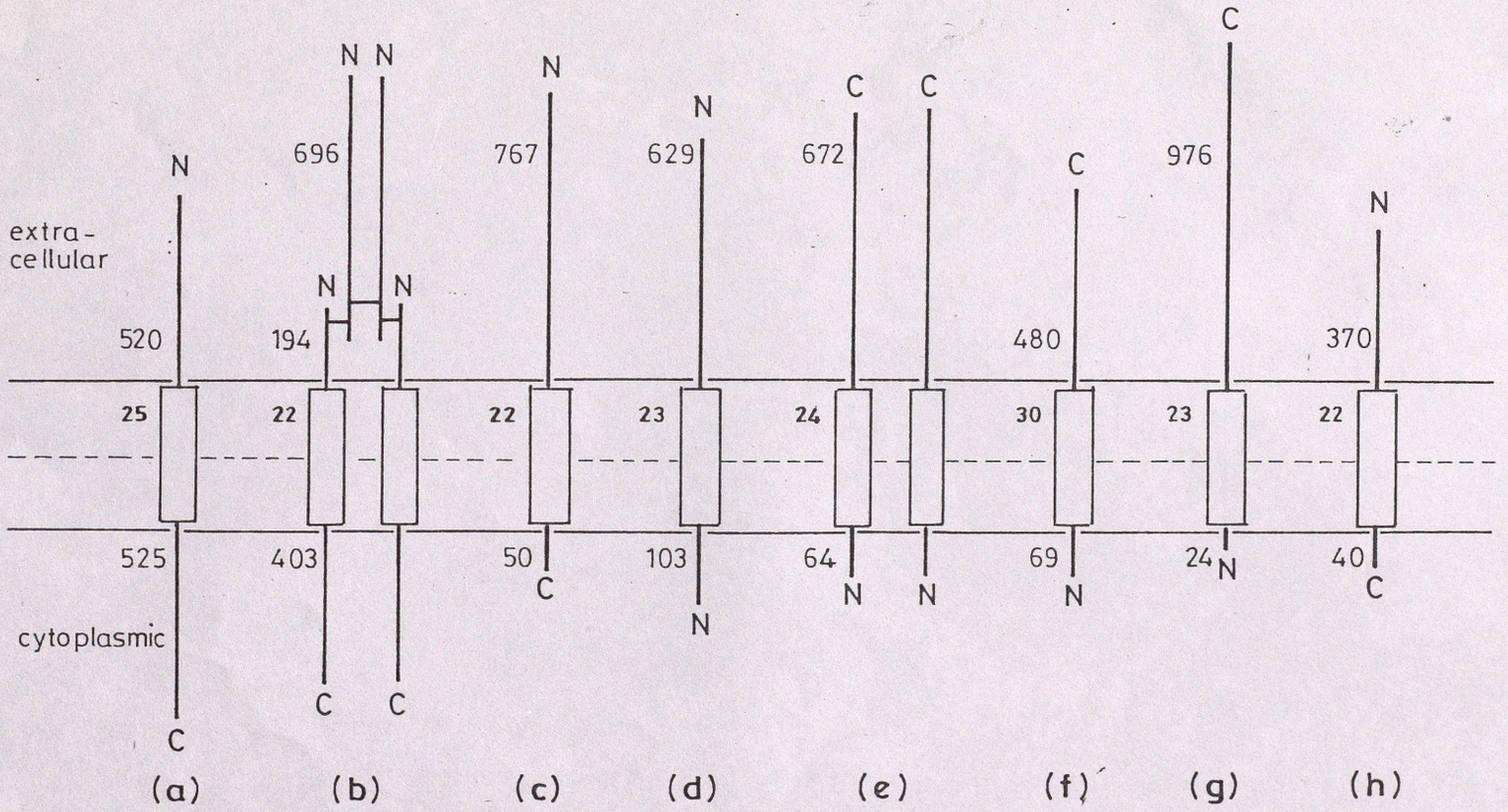

Fig. 3 Ramasarma/Joshi (EHG-51)

#3

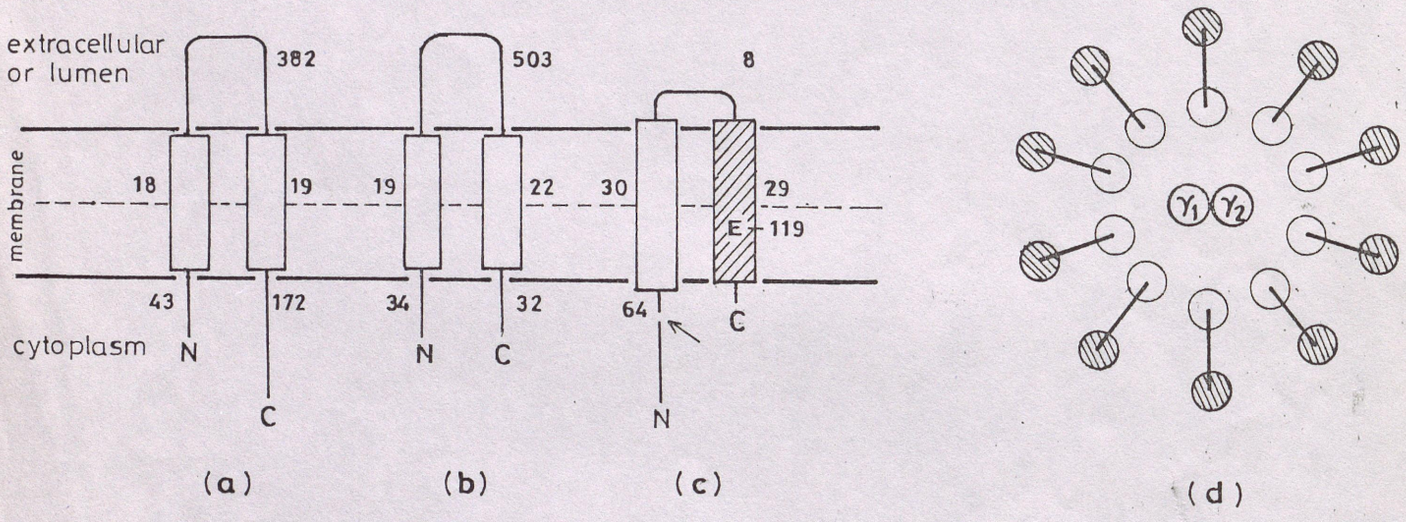

Fig: 4 Ramasarma/Joshi (EHG-51)

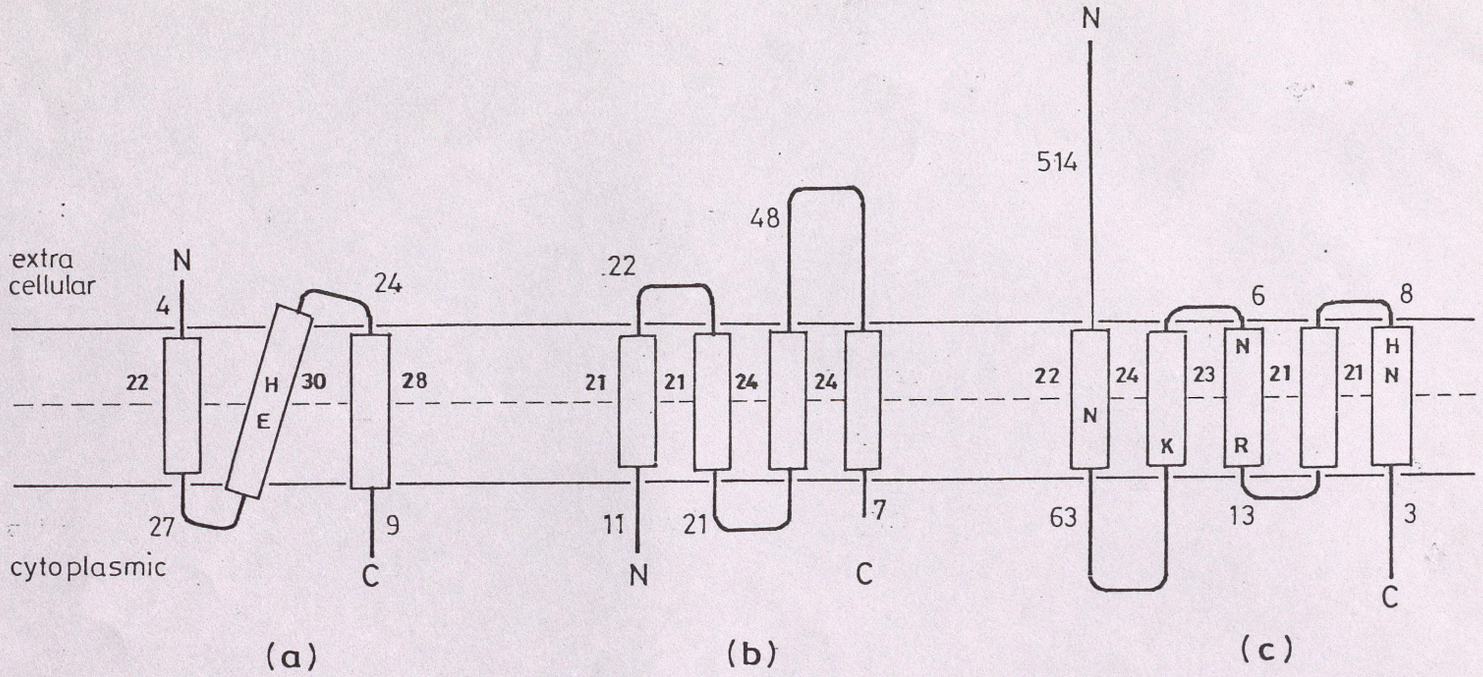

Fig. 5 Ramasarma/Joshi (EHG-51)

#5

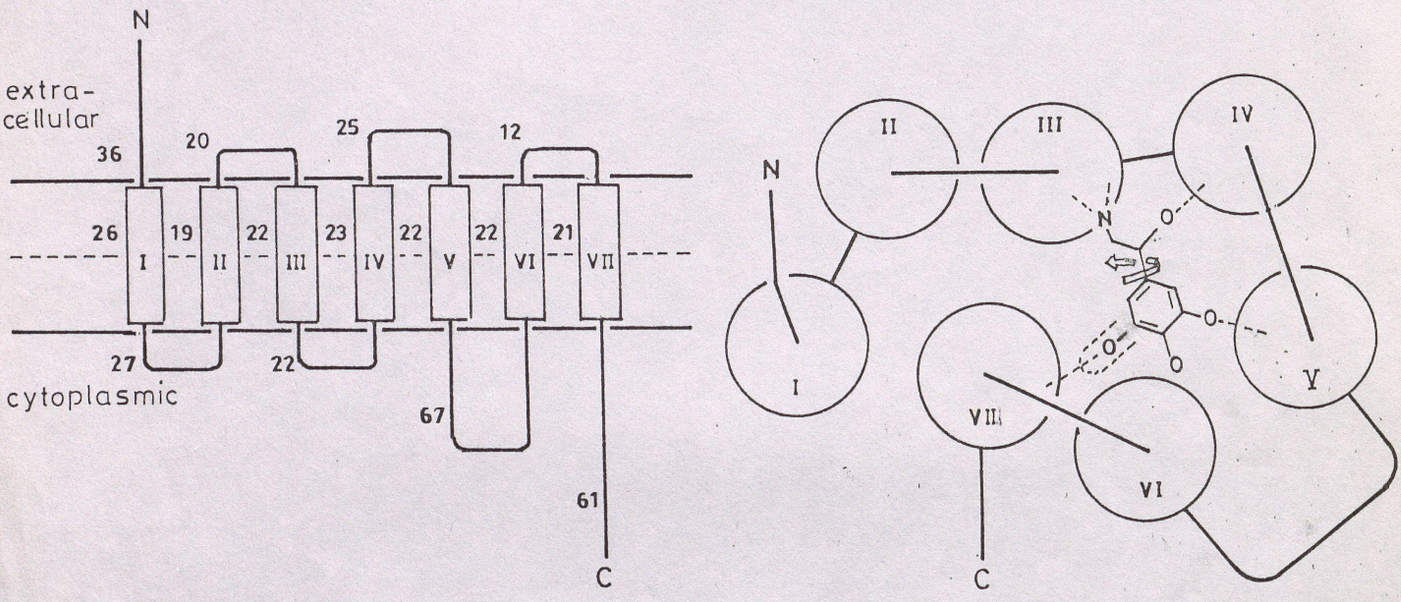

fig. 6 Ramasarma/Joshi (EHG-51)