Identification and Evaluation of transmembrane protein prediction tools for designing novel drug targets

[*,1]Sudhakar Malla; [2]Dr. B. Venkata Raman

[*,1]Research Scholar, Biotechnology, Bharathiar University, Coimbatore; Dept of Biotechnology, Indian Academy Degree College, Bangalore.

[2]Dean, R&D, CBIT, Proddatur Kadapa District, A.P

Abstract:

Conventional medicine is always incomplete without pharmaceuticals and novel biologics and the drug design and discovery always relies on the selective binding of the drug molecule to its target. To design a drug towards a receptor, one needs to know the structure of the receptor. Structurally proteins are very unique in nature and exist as secretory and transmembrane type. Transmembrane proteins are those which are embedded within the lipid bilayers and serve as best drug targets. Inorder, to design a novel drug towards any proteins which alter the biochemical pathway, we need to identify whether it is transmembrane or secretory and if transmembrane is it soluble or not. To identify these prerequisite parameters, we identified some transmembrane prediction tools and evaluated taking an aminoacid sequence as example. We found TMpred was found to be more reliable in terms of data acquisition and result output.

Key words: Transmembrane proteins, TMHMM, Phobius, TMpred.

Introduction: Transmembrane helices prediction among the integral membrane proteins is now considered to be a vital component of bioinformatics. Many methods as of date are more successful in predicting the individual transmembrane helices and also the complete topology of the desired protein (Von Heijne, 1999). These prediction methods are more applicable in genome analysis and are used in extracting the global strategy of protein evolution. Two decades back the prediction methods are based solely on the hydrophobicity analysis only. But now the recent analysis methods depends on the abundance of positively charged residues in the sequences (Almén MS, 2009). These two signals (charge and hyrophobicity) enhances the prediction methods more effectively. Transmembrane proteins are those proteins which are located within the lipid layers of the cells. Structurally they are made up of transmembrane spanning regions which pass by the lipid layers(Michalik, Marcin; 2017). Each protein is assigned with a specific function and there are multiple families of these proteins. These are classified into main forms called channels and carriers. Channels are those which form a constant pore like stricture across the plasma membrane and aids in diffusion of molecules at a faster pase. On the other hand carriers are those which bind to the solutes which passes across the membrane to undergo a conformational change. They are also called as transporters of pumps (Bracey MH, 2002). The lipid molecules which are predominant components of bilayer are hydrophobic in nature and as such the portion of the transmembrane protein which is placed

within the bilayer have residues which are nonpolar (Murzin AG, 1994). And these residues form a coil, or helix, which is hydrophobic in nature and very stable.

Transmembrane proteins are made of three regions or domains namely the domain in the bilayer, the domain outside the cell, and the domain inside the cell. Even though a cell membrane seems to be fluid like the transmembrane proteins orientation does not change as these proteins are very large in size nullifying the orientation rate. These transmembrane (TM) proteins play vital role in the functioning of cells. Some of the functions are communication via signalling, controlling the exchange of materials across the membrane. There are a second class of proteins called secretory proteins which aids in export from the cell. These are sorted by a small peptide sequence which drives them to the correct destination (Carpenter, E. P; 2008). These proteins are transported via vesicles or protein conducting transmembrane channels. Some of the proteins are non secretory in nature and stay inside the cell only due to non availability of compatible transported protein. As such they are involved in cellular metabolic activities like growth and survival of the individual cells.

Most of the drugs which are made towards many diseases target the membrane proteins by altering the cellular signalling (Rabiner,L. 1989). Most of the newly designed technologies in the drug therapeutics always try to target these TM regions and often call them as druggable regions. Finding suitable drug targets especially TM proteins is very crucial at the time, which aids in designing novel therapeutics against many infectious and metabolic disorders. Novel approaches in biomedical research and development are usually very slow due to the lack of knowledge on whether the protein is TM or secretory in nature. That is it is very difficult to differentiate the druggable and undruggable targets. In the recent days defying the homology relationships among the sequences is very much essential for research. By this the orthologs sequences can be analysed and is very useful for computational biology and annotation of genomes. This not only identifies the drug targets but also identifies the phylogenetic evolution of proteins within the infectious organisms (Ka¨ ll,L., 2005).

In the current review, we tried to survey the different methods of TM protein prediction servers and their evaluation in terms of biological understanding and technological capacity which will be useful for the pharmacotherapy. Different TM prediction servers like TMHMM, SignalP, Phobius, TMpred and SOSUI are described and evaluated basing on their output.

Methods:

Many software platforms were designed in the recent years to predict the TM protein structures. In this review some important tools were briefly described along with their output mode of evaluation. Histidine kinase of the competence regulon, ComD [Streptococcus mutans; NCBI Reference Sequence: NP_722221.1] was used in the evaluation process.

>NP_722221.1 histidine kinase of the competence regulon, ComD [Streptococcus mutans UA159]MNEALMILSNGLLTYLTVLFLLFLFSKVSNVTLSKKELTLFSISNFLIMIAVTVNVNLFYPAEPLYFIALSIYLNRQN SLSLNIFYGLLPVASSDLFRRAIIFFILDGTQGIVGSSIITTYMIEFAGIALSYLFLSVFNVDIGRLKDSLTKMKVKKRLIPMNI TMLLYYLLIQVLYVIESYNVIPTLKFRKFVVIVYLILFLILISFLSQYTKQKVQNEIMAQKEAQIRNITQYSQQIESLYKDIRS FRHDYLNILTSLRLGIENKDLASIEKIYHQILEKTGHQLQDTRYNIGHLANIQNDAVKGILSAKILEAQNKKIAVNVEVSSK IQLPEMELLDFITILSILCDNAIEAAFESLNPEIQLAFFKKNGSIVFIIQNSTKEKQIDVSKIFKENYSTKGSNRGIGLAKVNHI LEHYPKTSLQTSNHHHLFKQLLIIK

TMHMM: it is a novel method used for predicting the transmembrane helices which is based on a hidden Markov model. This was developed by Anders Krogh and Erik Sonnhammer. This software predicts all the membrane proteins in a large collection of most of the sequenced genomes and also provides a statistics of the frequency of proteins with their topology variations (A. Krogh, 2001). The major advantage of this server is it is based on the Hidden Markov Model (HMM) as such it aids in modelling the helix length. In contrast many of the methods designed earlier are based on setting both the upper and lower limits for the length of a membrane helix. The query sequence was restricted to about 8000 amino acids only.

From the output it was found that the desired protein was a TM protein. If the whole sequence was labelled as inside or outside, then it signifies it contains no membrane helices. The prediction also shows the most probable location and orientation of transmembrane helices in the sequence. The predicted number of TMHMM were found to be 6. The graphical report can be seen in the diagram below.

```
# NP_722221.1 Length: 441
# NP_722221.1 Number of predicted TMHs:  6
# NP_722221.1 Exp number of AAs in TMHs: 124.71377
# NP_722221.1 Exp number, first 60 AAs:  41.85395
# NP_722221.1 Total prob of N-in:        0.78634
# NP_722221.1 POSSIBLE N-term signal sequence
NP_722221.1     TMHMM2.0        inside      1     6
NP_722221.1     TMHMM2.0        TMhelix     7    24
NP_722221.1     TMHMM2.0        outside    25    38
NP_722221.1     TMHMM2.0        TMhelix    39    61
NP_722221.1     TMHMM2.0        inside     62    80
NP_722221.1     TMHMM2.0        TMhelix    81   103
NP_722221.1     TMHMM2.0        outside   104   117
NP_722221.1     TMHMM2.0        TMhelix   118   140
NP_722221.1     TMHMM2.0        inside    141   159
NP_722221.1     TMHMM2.0        TMhelix   160   179
NP_722221.1     TMHMM2.0        outside   180   193
NP_722221.1     TMHMM2.0        TMhelix   194   213
NP_722221.1     TMHMM2.0        inside    214   441
```

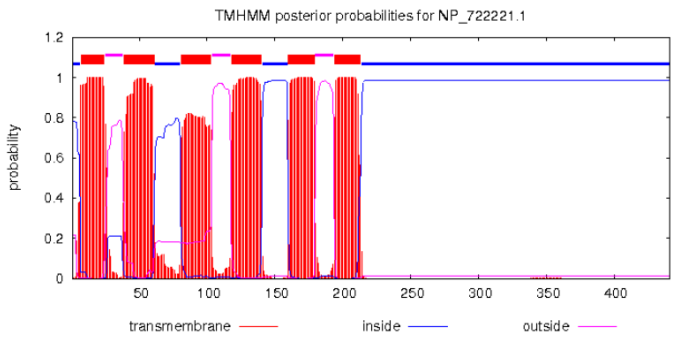

Figure 1: Result output of the TMHMM server. Left: Image showing the number of predicted TMHs. Right: Graphic report showing the TM helices. Red colour depicts the TM status.

SignalP: this server predicts not only the presence but also the location of cleavage sites within the signal peptides. It identifies the cleavage sites from different organisms including prokaryotes and eukaryotes. SignalP is based on neural network–based method which discriminates the signal peptides from the transmembrane regions. It is based on a deep convolutional neural network. In order to predict potential signal peptides of proteins, the D-score from the SignalP output is used for discrimination of signal peptide versus non-signal peptide (Henricson,A., 2005). This score has been shown to be the most accurate [Klee and Ellis, 2005] in an evaluation study of signal peptide predictors. In this it filters the sequences and restricts the user to select between the eukaryota and prokaryota to make the output in an effective manner. It means it allows a customized organism run which is absent in other cases. The precise prediction is given

whether the desired protein is a signal peptide or others. Since the selected protein is a TM in nature, the output in this case is not recognized, as it predicts the signal peptides in highlight. The maximum number of proteins to be queried is about 5000.

Protein type	Signal peptide (Sec/SPI)	TAT signal peptide (Tat/SPI)	Lipoprotein signal peptide (Sec/SPII)	Other
Likelihood	0.0337	0.0007	0.0337	0.9319

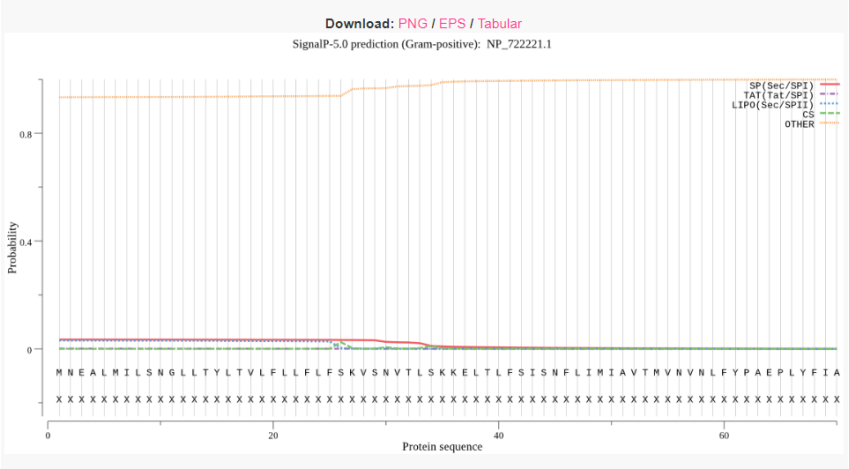

Figure 2: Result output from SignalP server. Image showing the probability or likeliness to be TM or signal peptide. Graphic output predicting that the protein is others (orange line) which means TM in nature.

Phobius: this server combines both the transmembrane topology and signal peptide predictions. It provides easy and accurate method to predict the signal peptides and transmembrane topology (Ka¨ ll,L., 2004). It also aids in making an optimal choice between the transmembrane segments and signal peptides, and also allows for constrained and homology-enriched predictions (Lao,D.M., 2002). This server stands ahead of others by predicting the TM as well as whether they are cytoplasmic or non cytoplamsic. Even signal peptide if any can also be predicted from the output. The desired protein was found to be TM in nature from the graphic representation below. This adds more beneficial than TMHMM. Even position of the aminoacid is also given by this server.

Prediction of NP_722221.1

```
ID   NP_722221.1
FT   TOPO_DOM      1      5       NON CYTOPLASMIC.
FT   TRANSMEM      6     26
FT   TOPO_DOM     27     37       CYTOPLASMIC.
FT   TRANSMEM     38     61
FT   TOPO_DOM     62     66       NON CYTOPLASMIC.
FT   TRANSMEM     67     90
FT   TOPO_DOM     91     96       CYTOPLASMIC.
FT   TRANSMEM     97    114
FT   TOPO_DOM    115    119       NON CYTOPLASMIC.
FT   TRANSMEM    120    140
FT   TOPO_DOM    141    160       CYTOPLASMIC.
FT   TRANSMEM    161    180
FT   TOPO_DOM    181    191       NON CYTOPLASMIC.
FT   TRANSMEM    192    210
FT   TOPO_DOM    211    441       CYTOPLASMIC.
//
```

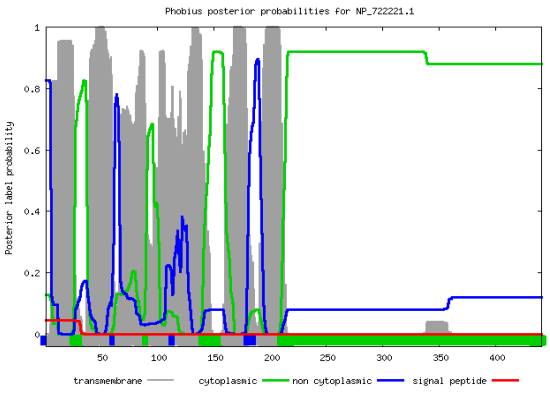

Figure 3: Result output from Phobius server. Image showing the probability or likeliness of a protein to be cytoplasmic or non cytoplasmic. Graphic output predicting that the protein is others (orange line) which means TM in nature.

When using the TMHMM and SignalP platforms, there is a minor overlap between these two predictions. When applied to about five complete proteomes, it was found about 30–65% of all predicted signal peptides and 25–35% of all predicted transmembrane topologies overlap which impairs about 5–10% of predictions (Klee,E.W. 2005). Hence biologists use Phobius which combines transmembrane topology and signal peptide predictions. This method was found to make an optimal choice between transmembrane segments and signal peptides. Even most of the conventional topology predictors predict signal peptides as transmembrane segments but the results vary a lot from the actual. The problem was resolved by removing proteins with transmembrane segments when predicting signal peptides. But however, the number of errors due to cross predictions seems to be same for the two kinds of predictors and moreover the gain was as high as the loss by such methods. But with the aid of Phobius, a better discrimination was made which forced the predictor to chose between the two types of features (Krogh,A., 2001). And false classifications of signal peptides were reduced from 26 to 4% from TMHMM's (4) and false classifications of transmembrane helices were reduced from SignalP 2.0's (5) 19 to 8%.

TMpred: This server makes a prediction of the membrane-spanning regions and also their orientation (Daley,D.O., 2005). It is based on the algorithm of TMbase, a database of naturally occurring transmembrane proteins during where the predictions are made using a combination of several weight-matrices for scoring (Nielsen,H., 1997). The result output is much consolidated predicting the possible transmembrane helices along with scores above 500. The result also suggests the best model with topology and best score. The graphic output also clearly depicts the orientation of the protein from inside outside and inside outside.

```
1.) Possible transmembrane helices
==================================
The sequence positions in brackets denominate the core region.
Only scores above  500 are considered significant.

Inside to outside helices :   7 found
     from        to    score center
  71 (  71)  89 (  89)   1639     79
 102 ( 107) 125 ( 125)   1984    115
 165 ( 165) 189 ( 183)   1411    175
 183 ( 183) 206 ( 201)   1375    193
 224 ( 226) 244 ( 242)   1536    234
 258 ( 258) 275 ( 275)   2860    267
 403 ( 403) 421 ( 421)    193    413

Outside to inside helices :   7 found
     from        to    score center
  71 (  71)  90 (  87)   2077     79
 102 ( 104) 127 ( 122)   1611    114
 130 ( 134) 159 ( 150)    507    142
 186 ( 186) 206 ( 202)   1674    194
 224 ( 226) 244 ( 242)   1074    234
 258 ( 258) 277 ( 277)   2528    268
 403 ( 403) 422 ( 422)    277    411
```

```
3.) Suggested models for transmembrane topology
===============================================
These suggestions are purely speculative and should be used with
EXTREME CAUTION since they are based on the assumption that
all transmembrane helices have been found.
In most cases, the Correspondence Table shown above or the
prediction plot that is also created should be used for the
topology assignment of unknown proteins.

2 possible models considered, only significant TM-segments used

*** the models differ in the number of TM-helices ! ***

-----> STRONGLY prefered model: N-terminus outside
 7 strong transmembrane helices, total score : 11717
 # from   to length score orientation
 1   71   90 (20)    2077 o-i
 2  102  125 (24)    1984 i-o
 3  130  159 (30)     507 o-i
 4  165  189 (25)    1411 i-o
 5  186  206 (21)    1674 o-i
 6  224  244 (21)    1536 i-o
 7  258  277 (20)    2528 o-i

------> alternative model
 6 strong transmembrane helices, total score : 10399
 # from   to length score orientation
 1   71   89 (19)    1639 i-o
 2  102  127 (26)    1611 o-i
 3  165  189 (25)    1411 i-o
 4  186  206 (21)    1674 o-i
 5  224  244 (21)    1536 i-o
 6  258  277 (20)    2528 o-i
```

```
2.) Table of correspondences
============================
Here is shown, which of the inside->outside helices correspond
to which of the outside->inside helices.
  Helices shown in brackets are considered insignificant.
  A "+"  symbol indicates a preference of this orientation.
  A "++" symbol indicates a strong preference of this orientation.

         inside->outside | outside->inside
   71-  89 (19) 1639     |   71-  90 (20) 2077 ++
  102- 125 (24) 1984 ++  |  102- 127 (26) 1611
                         |  130- 159 (30)  507 ++
  165- 189 (25) 1411 ++  |
  183- 206 (24) 1375     |  186- 206 (21) 1674 ++
  224- 244 (21) 1536 ++  |  224- 244 (21) 1074
  258- 275 (18) 2860 ++  |  258- 277 (20) 2528
(  403- 421 (19)  193    ) |( 403- 422 (20)  277  + )
```

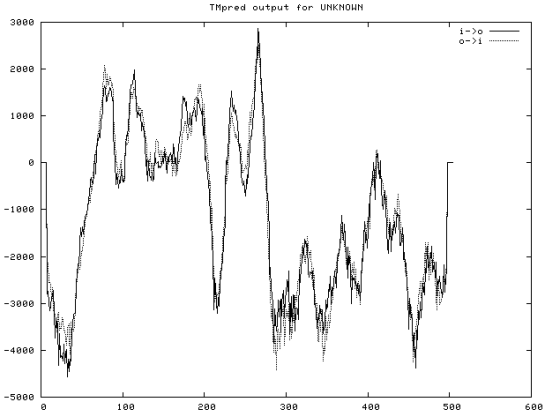

Figure 4: Result output from TMpred server. Image showing the probability or likeliness of a protein to be cytoplasmic or non cytoplasmic. Graphic output predicting that the protein is others (orange line) which means TM in nature.

SOSUI: SOSUI on the other hand constructs a system for predicting the membrane proteins by indexing the amino acids. Amphiphilicity plots of most of the membrane proteins of known 3D structure indicates that this type of plot is very useful in finding the end region of transmembrane helices (Hirokawa T., 1998). The accuracy was found to be 99% and the corresponding value for the transmembrane helix prediction was 97%. The server accepts minimum 20amino acids and maximum to about 5000 amino acids. It predicts a part of the secondary structure of the query proteins and also determines whether the protein of interest is a soluble or a transmembrane protein (Mitaku S., 1999). The system SOSUI for the discrimination of membrane proteins and soluble ones together with the prediction of transmembrane helices was developed, in which the accuracy of the classification of proteins was 99% and the corresponding value for the transmembrane helix prediction was 97% (Mitaku S., 2002). The output signified that the protein of query was a soluble [protein and has no signal peptide. The number of TM proteins were found to be 6.

Conclusion:

Predicting the structure and functions of most of the integral membrane proteins seems to be easy only for water-soluble globular proteins. But to understand extensively about various transmembrane proteins and secretory proteins, it is very much necessary to identify them at structural level. Only then novel drugs can be designed and they can be used as drug targets.

However, our knowledge of the membrane protein structural features are very limited owing to the low resolution information which was available as of now. In the recent years many bioinformatics tools were designed and used for predicting the transmembrane nature of the proteins. Such prediction not only aids in finding the structure of the proteins but also whether they are cytoplasmic or not and soluble or not. Such vast deciphering of the structures aid sin designing novel drugs towards them, which can be used to design drugs. The current study we found out that all the 5 models were well used by the users and among them TMpred was mostly used owing to its output features.

References:

1. A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. Journal of Molecular Biology, 305(3):567-580, January 2001.
2. Almén MS, Nordström KJ, Fredriksson R, Schiöth HB (2009). "Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin". BMC Biol. 7: 50. doi:10.1186/1741-7007-7-50. PMC 2739160. PMID 19678920.
3. Bracey MH, Hanson MA, Masuda KR, Stevens RC, Cravatt BF (November 2002). "Structural adaptations in a membrane enzyme that terminates endocannabinoid signaling". Science. 298 (5599): 1793–6.
4. Carpenter, E. P.; Beis, K.; Cameron, A. D.; Iwata, S. (2008). "Overcoming the challenges of membrane protein crystallography". Current Opinion in Structural Biology. 18 (5): 581–586. doi:10.1016/j.sbi.2008.07.001. PMC 2580798. PMID 18674618.
5. Daley,D.O., Rapp,M., Granseth,E., Melen,K., Drew,D. and vonHeijne,G. (2005) Global topology analysis of the Escherichia coli inner membrane proteome. Science, 308, 1321–1323.
6. Henricson,A., Ka¨ ll,L. and Sonnhammer,E.L.L. (2005) A novel transmembrane topology of presenilin based on reconciling experimental and computational evidence. FEBS J., 272, 2727–2733.
7. Hirokawa T., Boon-Chieng S., and Mitaku S., Bioinformatics, **14** 378-9 (1998) SOSUI: classification and secondary structure prediction system for membrane proteins.
8. Ka¨ ll,L., Krogh,A. and Sonnhammer,E.L.L. (2004) A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol., 338, 1027–1036.
9. Ka¨ ll,L., Krogh,A. and Sonnhammer,E.L.L. (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics, 21(Suppl. 1), 251–257.
10. Klee,E.W. and Ellis,L.B.M. (2005) Evaluating eukaryotic secreted protein prediction. BMC Bioinformatics, 6, 256.
11. Krogh,A., Larsson,B., vonHeijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol., 305, 567–580.
12. Lao,D.M., Arai,M., Ikeda,M. and Shimizu,T. (2002) The presence of signal peptide significantly affects transmembrane topology prediction. Bioinformatics, 18, 1562–1566.
13. Michalik, Marcin; Orwick-Rydmark, Marcella; Habeck, Michael; Alva, Vikram; Arnold, Thomas; Linke, Dirk (2017-08-03). "An evolutionarily conserved glycine-tyrosine motif forms a folding core in outer membrane proteins". PLOS One. 12 (8): e0182016.
14. Mitaku S., Hirokawa T. Protein Eng. **11** (1999) Physicochemical factors for discriminating between soluble and membrane proteins: hydrophobicity of helical segments and protein length.

15. Mitaku S., Hirokawa T., and Tsuji T., Bioinformatics, **18** 608-16 (2002) Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces.

16. Murzin AG, Lesk AM, Chothia C (March 1994). "Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis". J. Mol. Biol. 236 (5): 1369–81. doi:10.1016/0022-2836(94)90064-7. PMID 8126726.

17. Nielsen,H., Engelbrecht,J., Brunak,S. and vonHeijne,G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int. J. Neural Syst., 8, 581–599.

18. Rabiner,L. (1989) A tutorial on hidden Markov models and selected aplications in speech recognition. Proceedings of the IEEE, 77, 257–286.

19. Von Heijne G. A Day in the Life of Dr K. or How I Learned to Stop Worrying and Love Lysozyme: a tragedy in six acts. J Mol Biol. 1999 Oct 22;293(2):367-79.