Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 3, pp. 683~693

Prediction of Transmembrane Protein Topology Using Position-specific Modeling of Context-dependent Structural Regions

Sang-Mun Chi[1)]

Abstract

This paper presents a new transmembrane protein topology prediction method which is an attempt to model the topological rules governing the topogenesis of transmembrane proteins. Context-dependent structural regions of the transmembrane protein are used as basic modeling units in order to effectively represent their topogenic roles during transmembrane protein assembly. These modeling units are modeled by means of a tied-state hidden Markov model, which can express the position-specific effect of amino acids during transmembrane protein assembly. The performance of prediction improves with these modeling approaches. In particular, marked improvement of orientation prediction shows the validity of the proposed modeling. The proposed method is available at http://bioroutine.com/TRAPTOP.

Keywords : hidden Markov model, position-specific effect, topogenesis of transmembrane proteins, topology prediction

1. Introduction

Many researches have been performed on the prediction of transmembrane protein (TMP) topology from amino acid sequence alone. The accuracy of these prediction methods can improve more when the methods are guided by a deeper understanding of TMP assembly mechanisms. Recently, much progress has been made toward understanding the important details of how the machinery of membrane protein assembly works (Higy et al., 2004; White and Heijne, 2004). These papers explain that the topogenesis of transmembrane proteins, i. e.

1) Assistant Professor, Department of Computer Science, College of Multimedia, Kyungsung University, 110-1 Daeyeon-dong, Nam-gu, Pusan, 608-736, Korea
E-mail: smchiks@ks.ac.kr

translocation and insertion of proteins into the membrane, is governed by the charged residues and hydrophobicity of the membrane spanning regions. But, many topogenic rules found for TMP assembly have not been effectively exploited in current prediction methods.

HMM (hidden Markov model)-based approaches currently give the best performance (Möller et al., 2001; Chen et al., 2002; Arai et al., 2004). The increasing number of reliable membrane protein structural data makes the HMM-based procedures more promising approaches because these methods use structural data for training the prediction methods (Jones et al., 1994; Tusnady and Simon 1998, Krogh et al., 2001). But, comparison results showed that only around 50% of membrane proteins was predicted with correct topology for the best methods (Möller et al., 2001; Chen et al., 2002; Arai et al., 2004).

One of the reasons for the low prediction rates of current HMM-based method is that the topogenic determinants of TMP assembly are not well modeled in current HMM-based methods. Current HMM-based methods use no context-dependent modeling units. Thus, they are lack of the modeling for unique topogenic role of each context-dependent structural region such as the first loops, intermediate loops, and the last loops. Also, current HMM-based method use shared probability for modeling structural regions, i.e., every state in a modeling unit is tied. This tying cannot represent changing properties of amino acids inside modeling regions. But, experimental analysis of membrane proteins has shown that the position-specific distribution of amino acids is topogenically crucial, and the amino acid distributions in membrane proteins have different characteristic biases with the position of each structural part.

The present work makes a HMM structure suitable for incorporating topogenic rules such as the positive-inside rule, the hydrophobicity of membrane spanning region (MSR), folding properties of N-terminal segment, and the length of polypeptide. To model these topogenic determinants effectively, structural regions in different context are defined by separate modeling units. Multiple states with state-dependent probabilities are used for modeling this changing amino acids distribution with the position of amino acids in each structural part. Since the increased number of parameters by using multiple states needs more data for parameter estimation, the present work tied states to reconcile the detail modeling with the requirement to have enough data.

2. Position-specific Modeling of Transmembrane Proteins

2.1 Definition of Context-dependent Structural Regions

This work defines context-dependent structural regions and uses multiple states with state-dependent probabilities in order to effectively express the following

topogenic determinants for TMP assembly : (a) charged residues flanking the hydrophobic core of amino acid sequence, (b) hydrophobicity of signal sequence, (c) folding properties of N-terminal segment (Higy et al., 2004; White and Heijne, 2004; Goder and Spiess, 2001).

Charged residues affect the orientation of the first hydrophobic sequence, signal sequence, which plays an important role in protein topogenesis by orienting itself in the translocon and the membrane (Goder and Spiess, 2003; Goder et al., 2004). Internal transmembrane domains also follow the charge rule, although less stringently than the most N-terminal signal. Insertion of positive charge residues into short exoplasmic loops of model proteins caused individual hydrophobic domains not to insert at all, showing that internal charges can be topogenically active (Gafvelin et al., 997). Furthermore, charged residues have position-specific effects on N-tail translocation ; the C-terminal end of N-terminal tail is more critical for translocation than central and N-terminal regions (Whitley et al., 1995) ; the effect on helical hairpin formation of C-terminally flanking Lys and Asp residues decreases with the distance from the hydrophobic core of amino acid sequence (Hermansson et al., 2001). To effectively model these position-specific effects, this work defines new modeling units, context-dependent units. A context-dependent unit is a modeling unit in which the same structural region is defined as separate modeling units if the region has different left and right context. For example, the cytoplasmic loops are classified to three different modeling units depending on the context of the loops : (a) the first N-tail cytoplasmic loop, _i_M, has no structural region in the left but a MSR in the right when the amino acid sequence is read from N-terminus to C-terminus, (b) cytoplasmic loop between transmembrane helices, M_i_M, has a MSR in the left and right, (c) the last cytoplasmic loop, M_i_ has a MSR in the left but no structural region in the right. Similarly, context-dependent outer membrane loops, _o_M, M_o_M, M_o_, are defined. The context-dependent units make it easy to determine the position of amino acids that are in close proximity to the MSR - C-terminal end for the first loop, N-terminal end for the last loop, and both ends for intermembrane loops, while this proximity cannot be determined using a common modeling unit for different context in current prediction methods. In addition to the advantage in modeling position-specific effect, the context-dependent unit can separately represent the unique topogenic role of the first loop, loop between MSRs, and the last loop. A multiple state HMM with state-dependent probabilities is used to model these context-dependent units for the representation of the position-specific effect. This HMM structure will be described in the next section.

Hydrophobicity of the signal sequence is the another topogenic determinant. The natural direction of movement of a nascent chain of TMPs is for N-terminus to move from cytoplasm toward exoplasm. Driven by a local electrical potential, the signal sequence may invert its orientation and translocate the C-terminal sequence.

Increased hydrophobicity slows down inversion by stabilizing the initial bound state (Goder and Spiess, 2003). It has been proposed that a hydrophobicity gradient within the apolar core of the signal affects orientation, rather than the overall hydrophobicity. The more hydrophobic end of the signal sequence has been found to be more efficiently translocated (Harley et al., 1998). Another position-specific feature of amino acid distribution is that there is a distinct center-to-end heterogeneity in the distribution of apolar amino acids, with the aromatic residues F, Y, and W concentrated at the ends and the aliphatic residues L, I, and V more often found near the center (Heijne, 1994). For these reasons, the modeling unit for the signal sequence and MSR should be capable of representing the position-specific heterogeneity in the distribution of amino acids. Thus, this work represents MSRs with a multiple state HMM for the use of the position-specific modeling.

Folding of sequences N-terminal to the internal signal sequence may sterically prevent translocation of the N-terminus irrespective of charge distribution (Goder and Spiess, 2003). A polypeptide chain needs to be unfolded for translocation and that the folding properties of the N-terminal domain influence protein orientation. Thus, the structural regions related with these properties should be treated differently from the other non-related regions in order to incorporate this properties in the model. Since the present work uses context-dependent modeling units, the first loop model is capable of containing this principle implicitly. This is an another need for the context-dependent units. The topologies generated by the same signal sequence seem to depend on the length of C-terminal loop next to the signal sequence, with predominantly N-terminal translocation for very short proteins and reaching a maximum of ~55% C-terminal translocation for polypeptides of ~300 amino acids or more (Goder and Spiess, 2003). As the polypeptide grows longer, the signal sequence inverts its orientation driven by a local electrical potential acting on the flanking charges. Experiments showed that topology actually depended on the time of translation rather than on the length of the polypeptide (Goder and Spiess, 2003). Since the time of translation depends on various biological conditions, it is impossible to know the time of translation from amino acid sequence alone. Thus, the present work makes two independent model for each short and long loop instead of using the time of translation.

2.2 HMM Structures for Position-specific Modeling

A HMM can have suitable structures for position-specific modeling of the context-dependent structural regions. The HMM is a Markov chain where the output observation is a random variable generated according to an output probabilistic function associated with each state (Durbin et al, 1998). The present work uses multiple states with state-dependent transition and output probabilities in order to model the position-specific statistical properties of structural part,

whereas a common output distribution for every state of modeling regions has been used in the previous HMM based methods (Jones et al., 1994; Tusnady and Simon 1998, Krogh et al., 2001).

Since context-dependent structural regions are modeled by state-dependent output probability in this work, there are large number of parameters to be estimated. These large number of parameters need enough data for estimation. For alleviating insufficient data and robust estimation of probabilities, parts of the states are tied. The tied-states are made to share output probabilities. Figure 1 shows example HMM structures of context-dependent units, where, _i_M and M_i_M are the first and intermembrane loops in cytoplasm, M_o_M and M_o_ the intermembrane and last loops in non-cytoplasm, i_M_o the MSR from cytoplasm to non-cytoplasm, and i_M_o the MSR from non-cytoplasm to cytoplasm. The states in the same gray rectangle are tied. For the detail modeling of the states near MSR, smaller number of states are tied near MSR than far from MSR. Since the first loops can be preceded by signal sequence, N-terminal end can be close to MSR. For this reason, N-terminal end of the first loops are tied with small number of states. The shared probabilities are estimated from the amino acid sequences which belong to the same tied-state in estimation stage. The detail configures of the HMM structure for each modeling unit are presented in results and discussion.

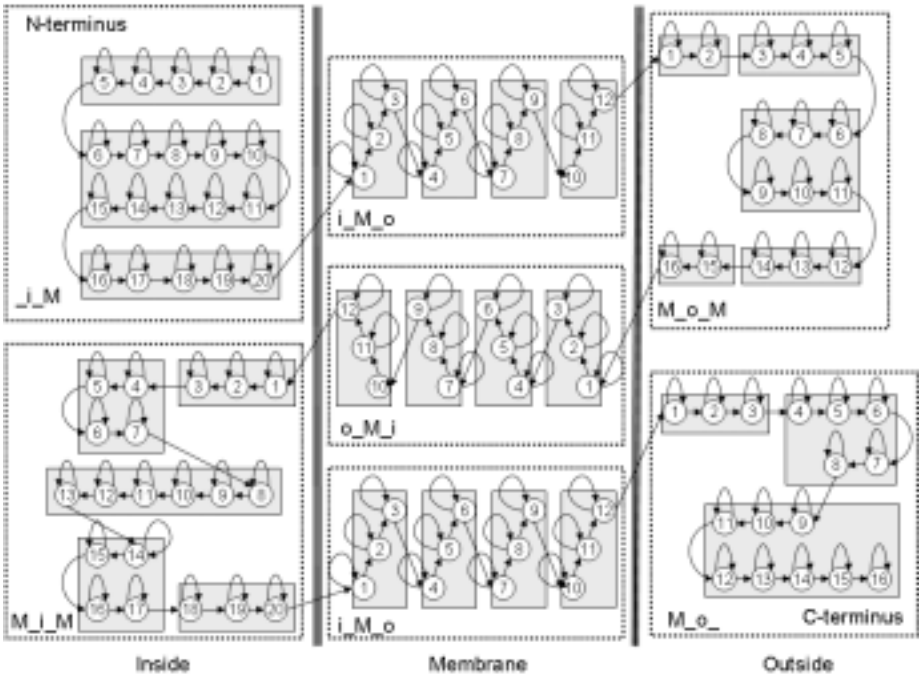

Figure 1. HMM structures for context-dependent units.

3. Prediction Experiments

3.1 Data Sets and Evaluation Criteria

To benchmark the performance of transmembrane protein prediction methods, it is necessary to use a test set of sequences with experimentally confirmed transmembrane regions. A data set (Möller et al., 2000) was downloaded from ftp://ftp.ebi.ac.uk/ databases/testsets/transmembrane. This test set contains 188 proteins with 883 MSRs that have been determined from either their elucidated structures or by fusion experiments. The present work will call this data set MöllerDB. Another data set, TMPDB_alpha_non-redundant(Ikeda et al., 2003), was downloaded from ftp://bioinfo.si .hirosaki-u.ac.jp/TMPDB. The present work will call this data set TMPDB. This data set is composed of 138 prokaryotic and 93 eukaryotic sequences with experimentally characterized topology information. These two data sets have 115 proteins in common. Some data with inconsistent annotation (amino acids, orientation and MSR position) were made to have the annotation in (Möller et al., 2000) before prediction experiments. Although mitochondrial matrix and intermembrane are respectively regarded as an outer and inner membrane in TMPDB, they are regarded reversely in this work following general classification.

To examine the performance of our methods, the rate of correct topology (TOPOLOGY) was mainly used and following accuracies were used additionally - (a) All MSRs (MSR) : the percentage of predicted proteins whose all MSRs are found correctly, (b) Sidedness (SIDE): the percentage of the correct sidedness of the protein's membrane integration when all MSRs were found correctly, (c) Specificity (SP): the percentage of correctly predicted MSRs over the predicted MSRs, (d) Sensitivity (SE): the percentage of correctly predicted MSRs over the true MSRs. For an MSR to be evaluated as correct, two evaluation rules were adopted in the present work. The first rule is that the MSR must share at least nine residues with the reference annotation's MSR (Möller et al., 2001). This rule is used for Table 1 and Table 3 to compare with other methods. Another rule is that the center position of predicted MSR coincided within 11 residues with that of MSR in the actual data (Arai et al., 2004), which is used in Figure 2.

3.2 Prediction of Transmembrane Protein Topology

To show the effectiveness of the proposed modeling units defined in 2.1, the prediction performance of context-independent unit with no position-specific modeling is compared with the proposed modeling units. Table 1 shows the performance of the proposed modeling on MöllerDB consisting 188 proteins with 883 MSRs. Every method in Table 1 adopts length-dependent models for every

modeling unit. Outer loops of lengths up to 15 (half of the total outer loops) are modeled with 3 state HMM, while the outer loops of lengths larger than 16 are modeled with 16 state HMM. Similarly, 3 state HMM is used or inner loops of lengths up to 19 (half of the total inner loops), 20 state HMM for inner loops of lengths larger than 20, 12 state HMM for membrane helices of lengths up to 20 (half of the total membrane helices), 21 state HMM for membrane helices of lengths larger than 21. Since the use of state-dependent output probabilities and length-dependent models increases the number of parameters to be estimated than the other HMM-based methods, MöllerDB and TMPDB are used as train set. CITS (Context-Independent Tied-State) uses context-independent modeling units - inner loop, membrane helix, and outer loop and every state of the modeling units is tied. This method is very similar to the previous HMM-based methods in that it ties every state in a modeling unit and it doesn't use context-dependent units. Thus, CITS can model neither the topogenic effect of context-dependent structural regions, nor position-specific heterogeneity in the distribution of amino acids in structural regions. CDMS (Context-Dependent Multiple States)) uses context-dependent units : (a) the first loop, intermembrane loop, and the last loop for each cytoplasmic and exoplasmic loop, (b) membrane helix from cytoplasm to exoplasm and from exoplasm to cytoplasm. CDMS adopts state-dependent output probabilities in order to model the changing distribution of amino acids with the position of amino acid. The proposed method, TRAPTOP (TRansmembrane Protein TOpology Prediction), uses the same context-dependent modeling units and multiple state HMM that were adopted in CDMS. But, the states are tied for the robust parameter estimation. Details of state-tying of the TRAPTOP are described in Figure 1 and Table 2. Explicit modeling of position-specific effect gives improved performance as can be seen in Table 1. TRAPTOP can predicts topology with 68% accuracy and gives marked improvement of correct sidedness 94%. When the same test set MöllerDB was used in the evaluation test (Möller et al., 2001), prediction accuracies of topology of TMHMM (Krogh et al., 2001), MEMSAT 1.5 (Jones et al., 1994), and HMMTOP (Tusnady and Simon 1998) were 47%, 41% and 36%, respectively. The MSR of TMHMM, MEMSAT 1.5, and HMMTOP were 68%, 53%, and 44%, respectively. Since training data differ from method to method, the comparison cannot be considered strict. But, the tendency of improved orientation prediction of TRAPTOP can be seen.

Table 1. Prediction performance of the proposed method (%).

Methods	TOPOLOGY	MSR	SIDE	SP	SE
CITS	56	64	88	90	95
CDMS	72	73	98	92	96
TRAPTOP	68	72	94	92	95

Table 2. Configuration of state-tying for TRATOP

Context-dependent unit	Tied states
_i_M	[1:5],[6:15],[16:20]
M_i_M	[1:3],[4:7],[8:13],[14:17],[18:20]
M_i_	[1:4],[5:10],[11:20]
_o_M	[1:4],[5:12],[13:16]
M_o_M	[1:2],[3:5],[6:11],[12:14],[15:16]
M_o_	[1:3],[4:8],[9:16]
short i_M_o and o_M_i	[1:3],[4:6],[7:9],[10:12]
long i_M_o and o_M_i	[1:5],[6:10],[11:15],[16:21]

Arabic numbers in brackets mean the state numbers of HMMs shown in Figure 1, and [s:e] represents that all states from state s to state e are tied. Every state is tied in all three-state HMMs, which are the short model of loops.

Table 3. Prediction performance for independent testing data (%).

Methods	TOPOLOGY	MSR	SIDE	SP	SE
CDMS	50	57	88	87	90
TRAPTOP	57	65	88	90	91

Table 3 presents the prediction performance for proteins that were not used for training. CDMS and TRAPTOP used the cross-validation test. Union of MöllerDB and TMPDB (304 proteins with 1477 MSRs) is partitioned into ten subsets possessing roughly equal numbers of proteins, and each group is tested after training on remaining nine subsets. Since the cross-validation result depends on the partition of database, 20 independent tests were performed with random partition of the database. CDMS has much more parameters than other method, which causes the parameter estimation unstable because of the insufficiency of training data. The performance of CDMS degrades rapidly when testing data are not similar to training data as can be seen in Table 3, while the performance degradation of TRATOP is rather mild because it ties states for the robust parameter estimation. The performance of TMHMM 2.0, MEMSAT 1.5 and HMMTOP with independent testing data were lower than 37% in TOPOLOGY and lower than 60% in MSR in the evaluation test (Möller et al., 2001). Since the training and testing data differ from method to method, the results cannot be compared directly. But, the topology prediction performance of TRAPTOP can be considered higher than the other methods since all of the methods used independent test set.

When the test set TMPDB consisting of 231 proteins with 1156 MSRs was used in the test (Arai et al., 2004), the performance of TMHMM 2.0, MEMSAT 1.5 and

HMMTOP were lower than 54.5% in TOPOLOGY and lower than 64.1% in MSR. The performance of methods in the test (Arai et al., 2004) is reverse to the results in the evaluation test (Möller et al., 2001). HMMTOP becomes the best method in the test (Arai et al., 2004), while it was the worst among the three methods in the evaluation test (Möller et al., 2001). This is because the train and test data of each method is not the same. Figure 2 shows the performance of TRAPTOP which has varying number of test data which are contained in training data. TMPDB and MöllerDB were respectively test and train set in Figure 2. Part of TMPDB data were randomly selected and inserted to training data. This procedure was repeated 20 times for each ratio between the number of test data which are contained in training data and the number of test data. Since half of TMPDB is already contained in MöllerDB, the ratio is 0.5 when MöllerDB is training set. As can be seen in Figure 2, the prediction accuracies improve with increasing ratio. TRAPTOP consistently outperforms the other methods in the test (Arai et al., 2004) between the ratio 0.5 and 1. When the independent test data were used by using cross validation in Table 3, the performance of the proposed method slightly better than the other methods in the test (Arai et al., 2004).

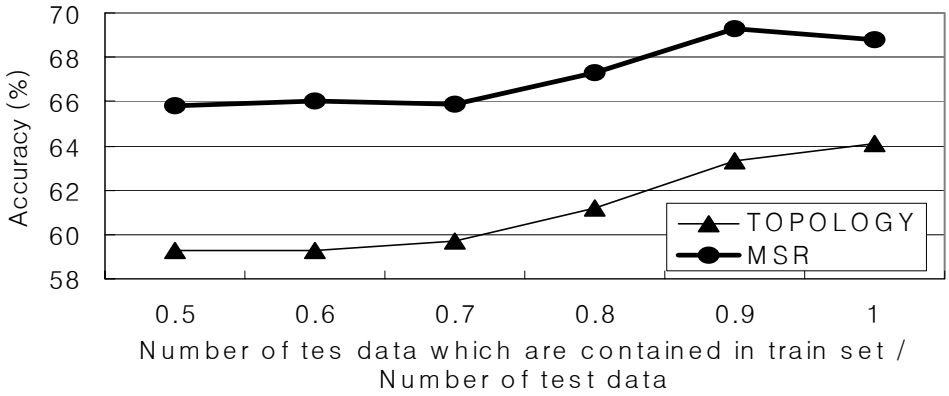

Figure 2. Accuracy of TRAPTOP for different training set.

4. Conclusion

For the effective topology prediction, the modeling unit should be both accurate and trainable; it accurately represents the topogenic role of each structural region, and it has enough data to train the parameters of itself. Although context-dependent units used for

TRAPTOP are accurate and representative, they are less trainable because the position-specific modeling increases the number of parameters to be estimated. To have the advantage of detailed modeling using the large number of parameters,

robust estimation is required. While tied-state HMM alleviates the problem of insufficient data in the present work, more efficient technique for robust estimation should be developed.

Like the time of translation whose information is not contained in amino acid sequence, topology prediction needs another non-sequence information, for example, physical, and biological information related to the biological assembly in natural lipid bilayer milieu. Consequently, it is impossible to completely predict the topology of TMPs solely from amino acid sequence. More study is needed to what other information should be contained together with sequence.

In conclusions, the proposed method performs better than the other HMM based methods for the prediction of TMP topology. In particular, the marked improvement of sidedness determination is obtained by including position-specific modeling of context-dependent structural regions and incorporating grammar into the search stage for probable membrane protein topology. The proposed method has used multiple state HMM possessing their own probabilities in order to model the position-specific heterogeneity in the distribution of amino acids within structural regions while an identical distribution has been used in other HMM based methods. Also, states have been tied for the robust estimation of HMM parameters. The TRAPTOP is available as a prediction server at http://bioroutine.com/TRAPTOP. There is also pointer to the data used in this work.

References

1. Arai, M., Mitsuke, H., Ikeda, M., Xia, J., Kikuchi, T., Satake, M., and Shimizu, T. (2004). ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability, *Nucleic Acids Research,* 32, w390-w393.
2. Chen, C. P., Kernytsky, A., and Rost, B. (2002). Transmembrane helix predictions revisited, *Protein Science*, 11, 2774-2791.
3. Durbin, R. M., Eddy, S. R., Korgh, A. and Mitchison, G. (1998) *Biological Sequence Analysis*, Cambridge University Press, Cambridge, UK.
4. Gafvelin, G., Sakaguchi, M., Andersson, H., and von Heijne, G. (1997). Topological rules for membrane protein assembly in eukaryotic cells, *Journal of Biological Chemistry*, 273, 6119-6127.
5. Goder, V. and Spiess, M. (2001). Topogenesis of membrane proteins: determinants and dynamics. *FEBS Letters*, 504, 87-93.
6. Goder, V. and Spiess, M. (2003). Molecular mechanism of signal sequence orientation in the endoplasmic reticulum, *The EMBO Journal*. 22, 14, 3645-3653.

7. Goder, V., Junne, T., and Spiess, M. (2004). Sec61p contributes to signal sequence orientation according to the positive-inside rule, *Molecular Bbiology of the Cell,* 15, 1470-1478.
8. Harley, C. A., Hot, J. A., Turner, R., and Tipper, D. J. (1998). Transmembrane protein insertion orientation in yeast depends on the charge difference across transmembrane segments, their total hydrophobicity and its distribution. *Journal of Biological Chemistry*, 273, 24963-24971.
9. Hermansson, M., Monne, M., and von Heijne, G. (2001). Formation of Helical Hairpin, *Journal of Molecular Biology,* 313, 1171-1179.
10. Higy, M., Junne, T., and Spiess, M. (2004). Topogenesis of membrane proteins at the endoplasmic reticulum, *Biochemistry*, 43, 12716-12722.
11. Ikeda, M., Arai, M., Okuno, T., and Shimizu, T. (2003). TMPDB: a database of exper imentallly-characterized transmembrane topologies. *Nucleic Acids Research,* 31, 406-409.
12. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, 33, 3038-3049.
13. Krogh, A. Larsson, B. Heijne, G., and Sonnhammer, E. L. L. (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology,* 305, 567-580.
14. Möller, S., Kriventseva, E. V.. and Apweiler, R. (2000). A collection of well characterised integral membrane proteins. *Bioinformatics*, 16, 12, 1159-1160.
15. Möller, S., Croning, M. D. R.. and Apweiler, R. (2001) Evaluation of methods for the prediction of membrane spanning regions, *Bioinformatics*, 17, 7, 646-653.
16. Tusnady, G. E. and Simon, I. (1998). Principals governing amino acid composition of integral membrane proteins: application to topology prediction. *Journal of Molecular Biology,* 283, 489-506.
17. von Heijne, G. (1994). Membrane proteins: from sequence to structure. *Annu. Rev. Biophys. Biomol. Struct.* 23, 167-192.
18. White, S. H. and von Heijne, G. (2004). The machinery of membrane protein assembly, *Current Opinion in structural biology*, 14, 397-404.
19. Whitley, P., Gafvelin, G., and von Heijne, G. (1995). SecA-independent translocation of the periplasmic n-terminal tail of an escherichia coli inner membrane protein, *Journal of Biological Chemistry*, 50, 29831-29835.

[received date : May. 2005, accepted date : Aug. 2005]