A sequence based analysis of the structure of SNARE family N terminal domains
Yuanyuan Hua

A hallmark of eukaryotic cells is the ability to segregate biochemical reactions within membrane-bound organelles. SNARE proteins, the protein family that mediate membrane fusion of vesicles trafficking between organelles, are conserved through phylogeny from yeast to man, as well as throughout the cell from the endoplasmic reticulum to the plasma membrane. SNAREs are integral membrane proteins present on both vesicle and target membranes. Formation of a very stable SNARE complex by 3 α-helices contributed by t-SNAREs located on target membrane, and 1 by v-SNAREs located on the vesicle is proposed to pull the vesicle and target membrane together and may provide the energy to drive fusion of the lipid bilayers. The neuronal SNARE complex, which mediates the fusion of synaptic vesicles with the presynaptic nerve, consists of one helix each from syntaxin 1A and VAMP 2 and two helices from SNAP-25, where VAMP 2 is the v-SNARE. The crystal structure of this complex is a four α-helix bundle consist of 7 conserved leucine-zipper-like layers composed of leucine, isoleucine and valine residues at *a* and *d* positions of septet repeats on each side of a central ionic layer, which is composed of an arginine contributed by VAMP 2, and three glutamate residues contributed by syntaxin 1A and SNAP-25 (Sutton et al. 1998, Fig. 1). Since residues forming both the ionic layer and the hydrophobic layers are highly conservative between species and throughout the cell, this four helix bundle structure is believed to be representative of all SNARE complexes. The four neuronal SNARE complex forming helices have been used as prototypes to define the 4 subclasses of SNAREs: syntaxin, SNAP C, SNAP N and VAMP (Bock et al., 2001).

Syntaxin 1A is a particularly interesting of the three neuronal SNAREs. It consists of a long, α-helix forming N-terminal domain followed by the SNARE complex forming domain called SNARE motif and a transmembrane domain. The N terminal domain is unique to syntaxin 1A among the 3 neuronal SNAREs, and in absence of VAMP 2 and SNAP 25 it is folded into 3 strands of α-helix and back onto the SNARE motif to form a four-helix bundle resembling the SNARE complex. This conformation of syntaxin 1A is referred as the closed conformation of syntaxin 1A, as opposed to the open conformation when SNARE complex form in presence of SNAP 25 and VAMP, and the N terminal domain is extended from the SNARE motif. The three α-helix forming regions of syntaxin 1A N-terminal domain have been termed Ha (aa 28-62), Hb(aa 71-104) and Hc(aa 111-144) respectively. It has been proposed that regulation of exocytosis events makes use of switching of syntaxin 1A between the closed and open conformation, since the interaction of syntaxin 1A with its specific binding partners, VAMP 2 and SNAP-25, can be regulated this way.

Interestingly, all known syntaxin family members and a number of SNARE proteins in the other three SNARE families also have long N terminal sequences preceding the SNARE motif. Although no significant homology was observed between these N terminal sequences, the available structures of all four N-terminal domain containing SNARE proteins (rat syntaxin 1A, mouse syntaxin 6, yeast Sso1p, and yeast Vam3p) present as four-helix bundles. Particularly, while Vam3p and Sso1p belong to the

syntaxin family, syntaxin 6 is classified as a member of the SNAP C family by profiling. This raises the interesting possibility that adoption of the closed conformation by SNARE protein monomers is a common mechanism for regulation of SNARE complex formation and membrane fusion. In this paper the structures of SNARE protein N terminal domains are analyzed based on sequence information available. The results suggest that the closed conformation is a common feature shared by a number of SNARE proteins.

Result

Prediction of the N-terminal structure of syntaxin proteins based on sequence homology

The size of the SNARE family has remained mostly unchanged in yeast, flies and worms, but has increased in mammalian animals such as mice and humans. For the syntaxin family 9 worm unique sequences and 11 mouse ones have been reported so far. In order to focus on the evaluation of available prediction methods, the first part of the paper used only the 11 mouse syntaxin sequences for analysis. The second part of paper analyzes some representative SNAP N, SNAP C and VAMP family proteins.

Since the SNARE complex structure is very characteristic and the known SNARE monomer structures resemble it, it is possible that such structural homology has a sequence basis, and the region of SNARE N-terminals contributing to the four helix bundle are remotely homologous to the SNARE domain. To detecting the potential homology between N terminal domain and SNARE domain of syntaxins Block+, ISREC profilescan and ematrix were used. Since all these programs have SNARE domain signature in the block/profile libraries, scanning the Habc domain of syntaxin 1A (aa 1-144) against these libraries using Block Searcher, ematrix search and profilescan should report the SNARE signature if homology is detected. Since in my search SNARE signature was reported even at the most relaxed penalty level, there does not appear to be homologous relationship between the N terminal and SNARE domains of syntaxins.

Alternatively, since structures of 4 SNARE proteins have been solved as four helix bundles, significant homology between N terminal sequence of a SNARE protein of unknown structure and N terminal sequences of these 4 proteins would suggest formation of four helix bundles. Some routine methods used to detect sequence homology based on multiple alignment are motifs, blocks, profiles and HHM models, listed in increased sensitivity. Since I did not find a web based HHM server (decypher was very slow in response), I chose profilemake in GCG package as the major tool for this analysis. In practice, Vam3p, Sso1p, msyn6 and msyn1a sequences were aligned using ClustalW, the N terminal coiled coil forming sequences were used to build profile using the GCG profilemake program, then the profile was used to search again swissprot containing all the 11 mouse syntaxin sequences used in this analysis. Of the first 400 hits returned for the search, the first 14 sequences were composed syntaxin 1-4 from various organisms and Vam3p and Sso1p, and the remaining sequences had no relationship with SNARE family. Notably, syntaxin 6 was not reported as positive although it is included in profile building, probably because it is distantly related to the other 3 sequences and weighted less in profile building. Profile was also built from aligned msyn1a, Sso1p, Vam3p N terminal sequence, which returned essentially the same result for database search. Finally, a profile was built from aligned sequences of all the 11 mouse syntaxins and used for profilesearch. The first 15 hits returned were syntaxin 1-4 sequences and the remaining 1100 sequences examined were unrelated to SNARE family. Thus the profilemake program was able to detect homology between only syntaxin 1, 2, 3, 4 and Vam3p and Sso1p, but not other syntaxin family members. This implies that syntaxin 2, 3, and 4 are capable of forming four helix bundles, while leaves the structures of other syntaxin proteins unaddressed.

The alignment of syntaxin family members itself was informational, however, in that there are 8 to 9 hydrophobic layers conserved in each of the three regions aligned with syntaxin 1A Ha, Hb and Hc. 70% to 100% of the amino acids in these layers are hydrophobic residues such as Ala, Leu, Ile, Val, and Phe. Formation of these layers could be due simply to the fact that the syntaxin family members are homologous, or they may reflect the coiled coil forming nature of the syntaxin proteins.

Complementary to the profile building approach, blocks and motifs were also built from the Vamp3p, Sso1p and syntaxin 1A N-terminal sequences. Blockmaker focus on finding homology in conserved regions of aligned protein sequences whereas protein profiles are built for the entire length of protein sequence submitted. Since 3 conserved regions in the N terminal of a syntaxin protein are proposed to participate in four helix bundle formation, blockmaker potentially could detect local homology which profile methods fail to detect. However, when N terminal sequences of Vam3p, Sso1p and syntaxin 1a were submitted to blockmaker server no block with proper alignment was generated. MEME, a profile based motif building program from GCG, was used to build motifs from alignment of Vam3p, Sso1p and syntaxin 1A. Of the 3 motifs found, 1 was above 10 aa long and located within a helix forming region, the Hb domain of syntaxin 1A. Database search using this motif returned 15 sequences of syntaxin 1-4 from various organisms, Vamp3p and Sso1p, and other sequences unrelated to SNARE family. Thus block and motif based methods were not able to detect remote homology missed by profile method. Furthermore, block based method appears to required highest level of homology for the sequences submitted as multiple alignment of the three multiple alignment based methods used.

Prediction of the N-terminal structure of syntaxin proteins using sequence based secondary structure prediction methods

Of the numerous sequence based secondary structure prediction algorithms available now, the majority make three state predictions of a residue being in α-helix, β-sheet or random coil states. Thus the region in a syntaxin protein contributing to formation of a four helix bundle would be predicted as in the α-helix state. Result from a few of such programs are shown below, where the syntaxin 1A 1-144 aa sequence was submitted as the gold standard.

nnpredict (Cohen et al., UCSF) (H = helix, E = strand, - = no prediction):
-----HHH----------EEE-------HHHHHHHHHHHH-HHH-HHHHHHHHHHHH--
E----------HHHHHHHHHHHHHH-HHHHH------H--HHHHH---------EEE----
----HHHHHHHHH----------

SOPM (Geourjon and Deléage, IBCP-CNRS) (H = helix, E = strand, c = random coil)
hhhhhhhhhhccccccccceeeeeectthhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhheeeeccccccc
chhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhcccchhhheehcccchhhhhhhhhhhhhhhccchh
Hhhh

SSpro (Pollastri &Baldi, UCI) (H = helix, E = strand, c = random coil)
CHHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
HHHHHHHHHHHHHHHHCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH

Predator (Frischman & Argos, EMBL) (H = helix, E = strand, _ = random coil)
___HHHHHHHHH_____EEEEEE___HHHHHHHHHHHHHHHHHHHHHHH
HHHHHHHHHHH_____HHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
HHHH_____EEEEE_____HHHHHHHHHHHHH_____

Results of similar information content were returned from SSP/NNSSP(Solovyev and Salamov, Baylor College), and PHD (Rost & Sander, EMBL) servers and omitted from the text to save space. Several problems in applying these programs to the SNARE structure prediction problem are exemplified by the results above. First, such programs can only predict helix forming regions which may or may not form coiled coils, so false positive number is high. Second, within the Habc domain some programs predicted fragmented helices and miss true positives. Third, the programs used do not agree with each other, and consensus is hardly possible to make. The single program tested within this category which reported satisfactory result is Jpred (Cuff & Barton, EBI). It predicted helix region in aa 31-61, 71-100 and111-153 that correspond approximately to the Ha, Hb and Hc regions of syntaxin 1a. Jpred takes either aligned sequences or a single sequence, align the sequence(s) to homologous sequences found in database by blast, and analyze the alignment with several algorithms including Jnet prediction, Jnet alignment prediction, Jnet hmm profile prediction, Jnet PSIBLAST pssm profile prediction and Jnet PSIBLAST frequency profile prediction. A consensus is reported together with the results from individual methods. Two features making Jpred potentially powerful are that the program relies on alignment of homologous sequences and that it combines the results of a number of methods. The Jpred documentation states that it also combine results from PHD, PREDATOR, NNSSP, and Zpred in the consensus reported although results from these methods were not shown in the prediction report. Since Jpred prediction takes several hours, probably due to intensive computation required, the mouse syntaxin sequences were not further tested on this server.

Multicoil (Wolf, Kim, & Berger; MIT) and Coils 2.0 Version (Lupas, Dyke & Stock, EMBnet) are programs devoted to predicting potential coiled coil structures. Coils compares a sequence to a database of known parallel two-stranded coiled-coils and report the prediction from 3 different window width, 14, 21, and 28. MultiCoil predicts potential dimeric or trimeric coiled coils in a given amino acid sequence. Thus neither program is tailored to detect coiled coils contributing to a four helix bundle. For the syntaxin 1-144 aa sequence submitted Coils predicted 3 coil regions corresponding Ha, Hb and Hc for 21 aa window width, whereas Multicoil predicted Ha and Hb domains as coiled coil but not Hc domain. For the other 10 mouse syntaxin sequences submitted, most sequences were predicted to contain 3 coiled coil forming regions by both programs. The exceptions are msyn 11, 16, 18 predicted by Multicoil to contain no coil in the N terminal domain, and msyn16, 18 predicted by Coils to contain only 1 coiled coil region. In general the

predicted coiled coil regions correspond to with the Habc region of syntaxin 1A, but the alignments are not exact. Thus based on these two coiled coil prediction programs msyn 1A, 1B, 2, 3, 4, 5, 7, and 13 likely form 4 helix bundles, whereas msyn16, 18 probably do not form 4 helix bundles, msyn11 is more likely to form four helix bundle than not because the result from Coils is positive and Coils appears to performs better in this task. Notably, the 1.0 version of Coils supplied by GCG package predicted msyn 1A structure incorrectly and in general the predictions agree poorly with Coils 2.0 Version.

Finally, since hydrophobic layers are a salient feature of the 4 helix bundle, amphiphilic helix formation can be used as a criteria for potential coiled coil forming regions. For example, Syntaxin 13 is predicted to contain only one region preceding aa 100 with high probability of coiled coil formation by Multicoil, whereas high probability of coiled coil formation almost throughout the N terminal domain was predicted by Coils. To elucidate the confusion the three regions of syntaxin 13 aligned to Ha, Hb and Hc were separately analyzed using the helixwheel program in GCG package, assuming that amino acids in these regions form helices. Since all three regions were predicted to be ambiphilic, I concluded that msyn13 forms four helix bundle in monomer state. Analysis of msyn 11 using such procedure suggests it to be 4 helix bundle forming which agrees with Coils but not Multicoil result.

Prediction of the N-terminal structure of syntaxin proteins based on structures available in PDB database

Two 3D structure prediction methods dependent on PDB database were examined in this work. SAM-T99 (Hughey, Karplus and Krogh, UCSC) perform iterative protein homolog searches in PDB using HMM in order to predict secondary structure, it also performs a SAM-T99 Model Libary Search built from HMM structural alignment. Sequence alignment, PDB database hits and scores, secondary structure prediction and HHM library search results are returned to user. Two syntaxin proteins of known structure, msyn1a and Vam3p, were submitted for standard test. The msyn1a analysis reported 1ez3A (syntaxin 1A, NMR), 1br0A(syntaxin 1A, crystal structure), 1dn1B (nsec1, syn1 complex, crystal structure) and 1fioA(Sso1p, NMR) as PDB hits. The secondary structure prediction is shown below

LLLHHHHHHHHHLLLLLLLLLLLLLLLLLHHHHHHHHHHHHHHHHHHHHHHHHHHHHHLL
LLLLLHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHLLLLLLLLHHHHHHHHHHHHHH
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHLLLLLLHHHHHHHHHLLLLHHHHHHHHHH
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
HHHHHHHHHHHHHHHHLLLEEEEEEHHHHHHHHHHHHEEELLL

For Vam3p search, 1dn1B is the single PDB hit, the secondary structure prediction is

LLLLLLLLLLLLLLLLLLLLLLLLLLLLHHHHHHHHHHHHHHHHHHHHHHHHLLLLLLHHHHHHHH
HHHLLLLLLLLLHHHHHHLEEEELLLLLLHHHHHHHHHHHHHHHHHHLLLLLLLLLLLLLLLLLL
LLLLLLLLHHLLLLLLLLL

Examination of the results reveals that SAM-T99 has limited ability to detect sequence homology. For example, the Vam3p search returned 1dn1B and missed 1ez3A, although both are syntaxin structures. Moreover, whereas the secondary structure of syn1a, of

which the structure is available in PDB database, agrees well with the 3D structure, the prediction of Vam3p secondary structure predicted the second and third coiled coil forming regions of Vam3p incorrectly. Therefore SAM-T99 performs best when structures of closely homologous proteins are available, and it is not satisfactory in the present case.

The second PDB dependent program used, 123D+(Alexandrov, Nussinov & Zimmer, IMMB-NCI), uses substitution matrix, secondary structure prediction, and contact capacity potentials to thread a sequence throughout known protein structures. Therefore it is structural homology rather than sequence homology based. The analysis of msyn3 by 123D+ returned SNARE core complex, stat, Apo E3, ribosome recycling factor, and some other DNA binding proteins as top hits in PDB database. All these structures contain coiled coil domains which correspond to msyn3 in the sequence alignments, suggesting that msyn3 form four helix bundle in monomeric state. The Vam3p full sequence search returned similar result. The Vam3p N terminal sequence search did not return 1br0A, the SNARE core complex structure, yet all hits reported contain coiled coil domains suggesting existence of coiled coil in Vam3p structure. 123D+ also reports secondary structure predictions based on the PDB search result which did not agree exactly with the crystal structure in the Vam3p case. Thus 123D+ captures the general feature of a protein structure well by reporting a number of known structures sharing the same structure, but is not superior to the methods discussed before in predicting structural details such as exact location of the helices contributing to coiled coil.

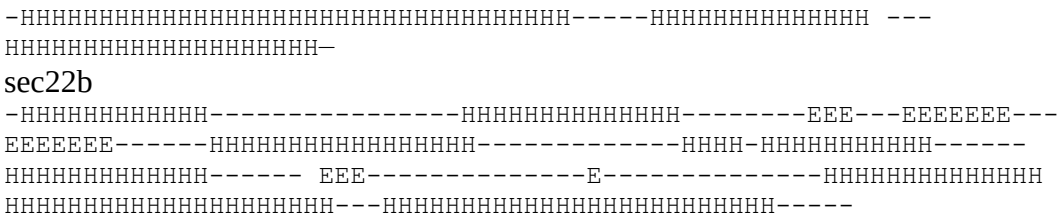
In summary, the mouse syntaxin family were analyzed using a variety of sequence homology detection, secondary structure prediction and 3D structure prediction methods. Msyn 2, 3, 4, 5, 7, 11, 13 likely form four helix bundles similar to the known syntaxin 1A structure, whereas msyn 16 and 18 has only weak tendency to form coiled coils, probably in dimeric or trimeric fashion. Jpred, Coils, helixwheel, and 123D+ returned most informative results of or methods tested, therefore they were used to for further structural predictions of SNARE proteins other than the syntaxin family.

Prediction of the N-terminal structure of SNAP N and VAMP proteins using sequence based secondary structure prediction methods

For the four species examined, mouse, fly, worm and yeast, about 50% of the SNARE proteins in SNAP N, SNAP C and VAMP families contain N-terminal sequences preceding the SNARE motif longer than 150 residues. Orthologs for Sec22, ykt6, vti1, gs28 and membrin exist in all the four species, which account for above 60% of the long N-terminal sequence containing SNAREs outside syntaxin family. The other proteins in this category are more species specific. Due to limit of time and space, I analyzed two representative sequences, mouse membrin from the SNAP N family, and mouse sec22b from the VAMP family, for this part of work instead of attempting to analyze all the 32 potential four helix bundle forming SNARE proteins. The results are shown as follows.

Jpred result
membrin

```
--HHHHHHHHHHHHHHHHHHHH------HHHHHHHHHHHHHHHHHHHHHHHH-------
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH ------------HHHHHHHHHH-HHHH-
```



sec22b

Coils result
Membrin N terminal: 3 coils located at approximately 5-30, 30-55, and 55-120 close to each other.
Sec22b N terminal: no region with significant probability of forming coiled coil was reported.

Helixwheel result
Membrin N terminal: regional amphilicity was observed when the N terminal sequence is analyzed as a single α-helix.
Sec22b N terminal: no amphilicity was observed.

123D+ result
mouse membrin N terminal sequence (first 5 hits)

PDB ID	Z-score	a.a. aligned	% identities	Protein identity
d1pnb.1 106 a.a.	5.90	106	9	Napin Bnib
d1bg1a1 186 a.a.	4.48	140	13	Stat3B Homodimer
d2spca_ 107 a.a.	3.99	99	6	Spectrin (One Repeat Unit)

Msec22b N terminal sequence

PDB ID	Z-score	a.a. aligned	% identities	Protein identity
d1tfe__ 142 a.a.	6.16	136	8	Dimerization Domain Of Ef-Ts
d1bmfa1 131 a.a.	4.43	125	9	Bovine Mitochondrial F1-Atpase
d1opd__ 85 a.a.	4.42	85	7	Histidine-Containing Protein (Hpr)

The results reported by Jpred, Coils and Helixwheel suggest that mouse membrin N terminal sequence contains 3 coiled coil forming α-helices which are probably organized into a four helix bundle together with the SNARE motif, whereas the α-helices in mouse sec22b N terminal sequence are organized differently. The 123D+ program returned coiled coil harboring structures for both proteins, indicating that msec22b may contain dimeric or trimeric coiled coils consist of fewer residues than a four helix bundle. Therefore the four helix bundle structure appears to be shared by proteins in all four SNARE families, although in each family this structure applies to only a subgroup of proteins.

Discussion

Two general approaches of structure prediction are searching for closely homologous proteins whose structure is known, and using specific algorithms designed to predict structure based on sequence information. In the case of SNARE proteins, N terminal sequence homology were detected only between syntaxin 1, 2, 3, 4, and more remotely Vam3p and Sso1p. Therefore structural prediction for the majority of SNARE proteins relies on prediction algorithms. As demonstrated in this work, the current structural prediction programs are far from mature. First, results reported by different programs often contradict with each other. Second, interpretations of the results are often not straightforward. For example, a fragment of secondary structure prediction taken from the results section

LLLLLLLLHHHHHHLEEEELLLLLLHHHHHHHHHHHHHHHH

is not likely an exact description of the structure, and it is hard to extract information from this type of prediction without using complicated, computation intensive algorithms. Third, most of the prediction methods examined in this work do not incorporate data from PDB database, which is an increasingly important source of structural information.

Comparison of the prediction methods examined in this work reveals the following aspects promising in improving the current |methods:

(1) Programs which blast submitted sequence against sequence database, make alignment, then predict structure of the alignment tend to perform better than single sequence prediction.
(2) Programs extracting information from PDB database, for example by homology search or threading, tend to perform better than the other programs.
(3) Programs incorporating results from several algorithms into a consensus tend to perform better than single algorithms.

These observations are probably self-evident considering that the coding of structural information by biological systems is highly redundant.

Reference:

1. Bock JB et al. (2001) Nature. 409:839-41.
2. Sutton RB et al. (1998) Nature. 395:347-53.

Figure 1

Topology and organization of the synaptic fusion complex. **a**, Backbone ribbon drawing of the synaptic fusion complex: blue, VAMP 2; red, syntaxin-1A; green, SNAP-25 (Sn1 and Sn2). **b**, Conformational variability assessed by overlay of the three non-crystallographically related complexes **c**, Organization of the synaptic fusion complex. C α traces (grey), local helical axes, the superhelical axis (black), and layers (0, red; -1, +1 and +2, blue; all others black) are shown for one of the three complexes in the asymmetric unit. **d**, Radii of the three synaptic fusion complexes in the asymmetric unit. (copied from Sutton et al. 1998)