MAPAS: a tool for predicting membrane-contacting protein surfaces

Yuriy Sharikov, Ross C Walker, Jerry Greenberg, Valentina Kouznetsova,

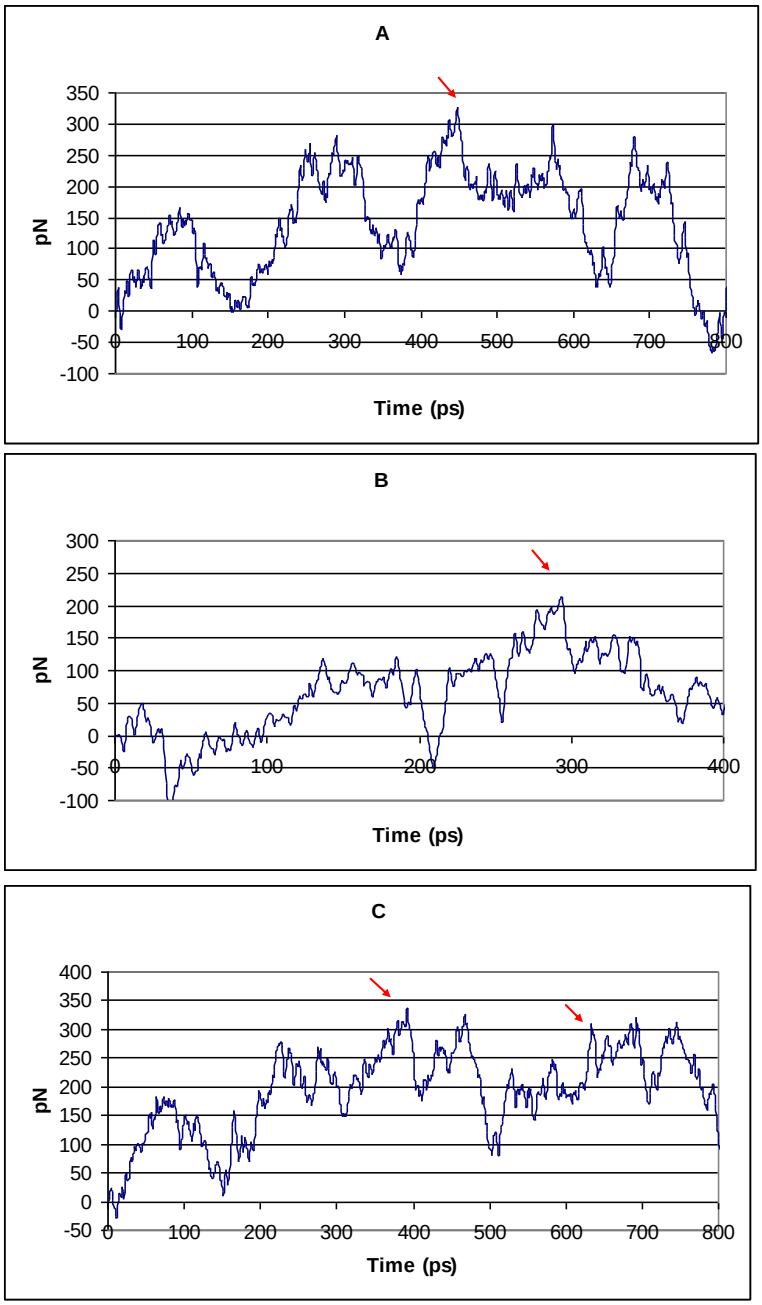
Sanjay Nigam, Mark A Miller, Eliezer Masliah & Igor F Tsigelny

Supplementary figures and text:

Supplementary Figure 1.

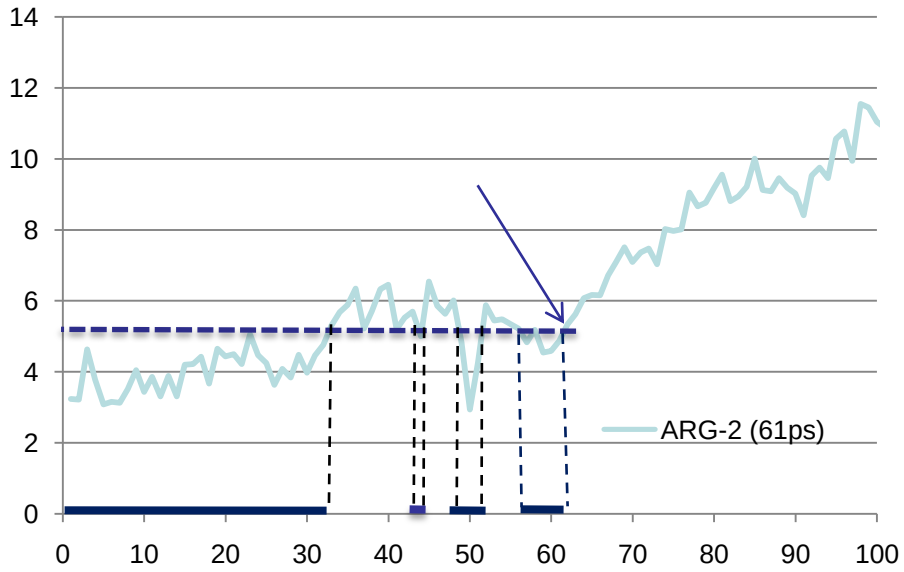
The input module of the web-based program.

Supplementary Figure 2a.



Calculation of Scoring Function: a). Examples of forces applied to the Gly-X-Gly-Gly-Gly-X-Gly peptide during SMD disengagement experiments. Red arrows show the moments of disconnection of the X residues side chains: (A) One residue disengagement for X-Asp (SMD started after 25ps equilibration), (B) One residue disengagement for X – Leu (SMD started after 10 ps equilibration), (C) Two residues disengagement for X – Lys (SMD started after 10 ps equilibration).

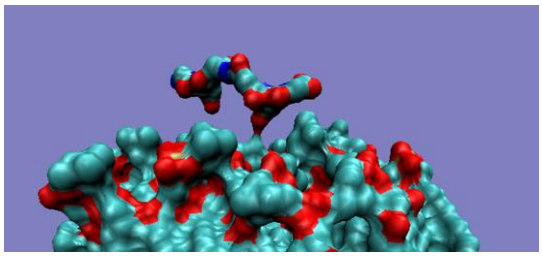
Supplementary Figure 2b.



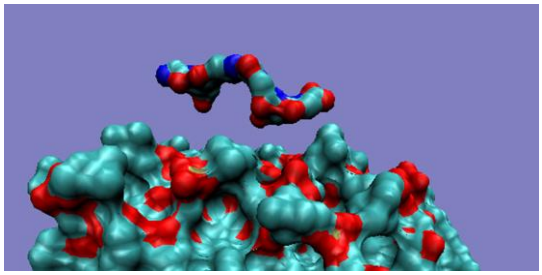
Calculation of Scoring Function: b). Shortest distance between the selected amino acid (arginine in this example) of the probe peptide during SMD disengagement experiments (SMD started after 10 ps unrestrained MD). Arrow shows the time of irreversible disengagement. Highlighted near the X axis are the periods when this amino acid is in contact with the membrane. The residue has some number of disengagement and engagement events before the final time of irreversible disengagement. 5.2 Å boundary is used as a criteria of membrane-peptide contact.

Supplementary Fig. 3.

A

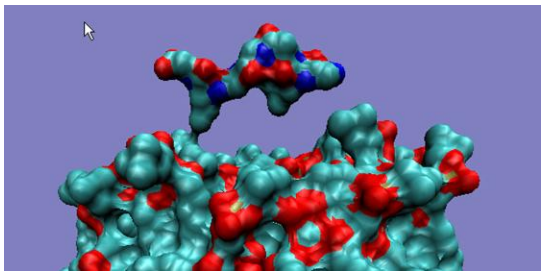


ASP (25ps equilibration) 44ps SMD

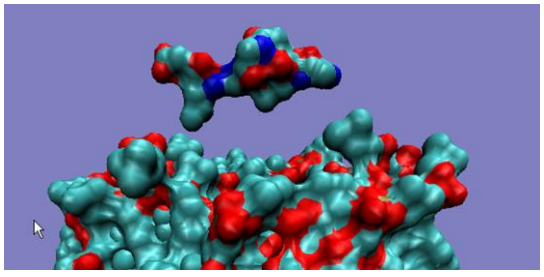


ASP (25ps equilibration) 45ps SMD

B

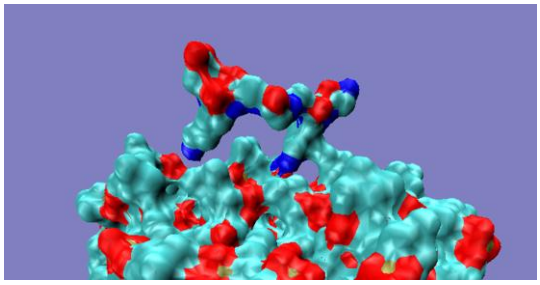


LEU (10ps equilibration) 29ps SMD

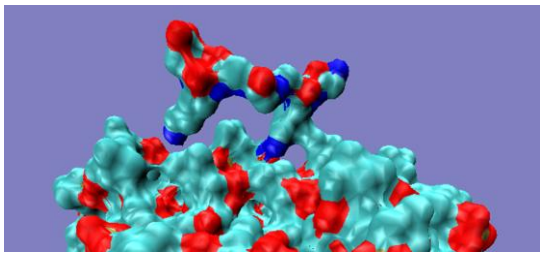


LEU (10ps equilibration) 30ps SMD

C

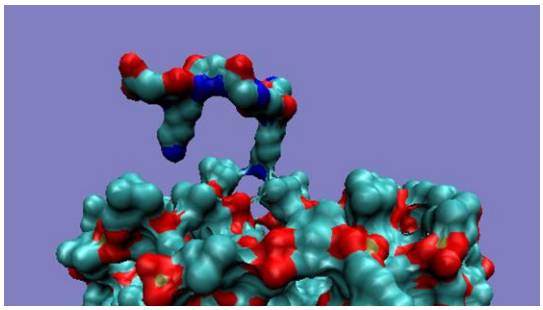


LYS (10ps equilibration) 39ps SMD

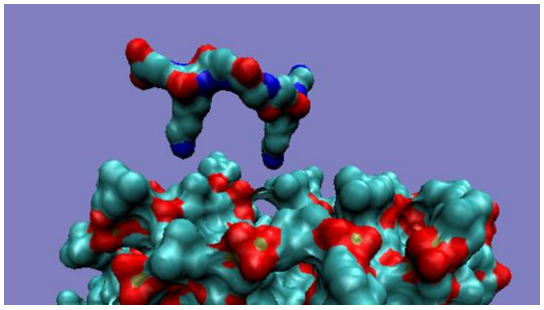


LYS (10ps equilibration) 40ps SMD

D



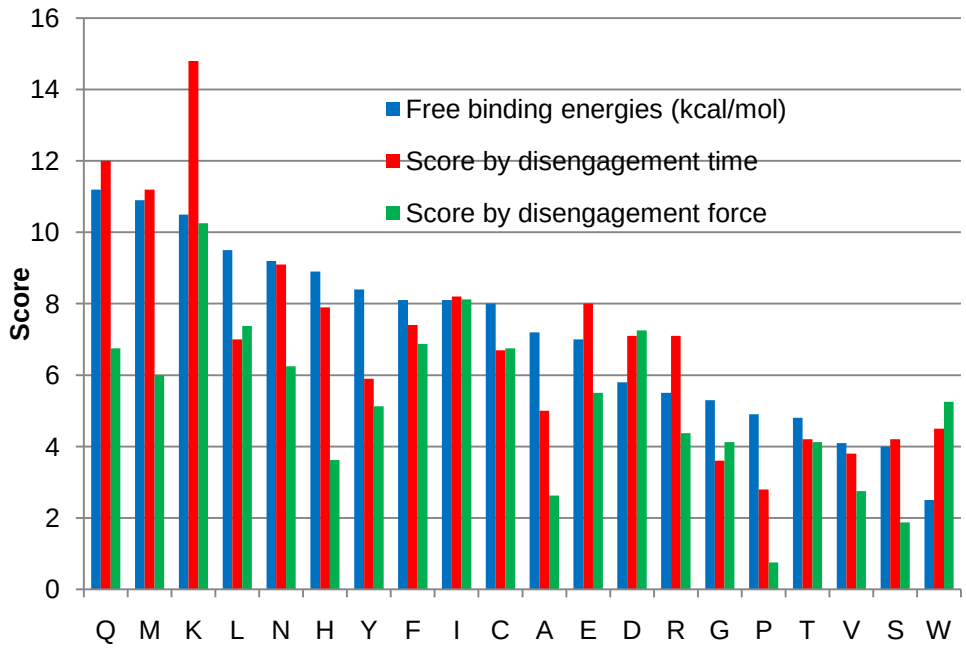
LYS (10ps equilibration) 62ps SMD



LYS (10ps equilibration) 63ps SMD

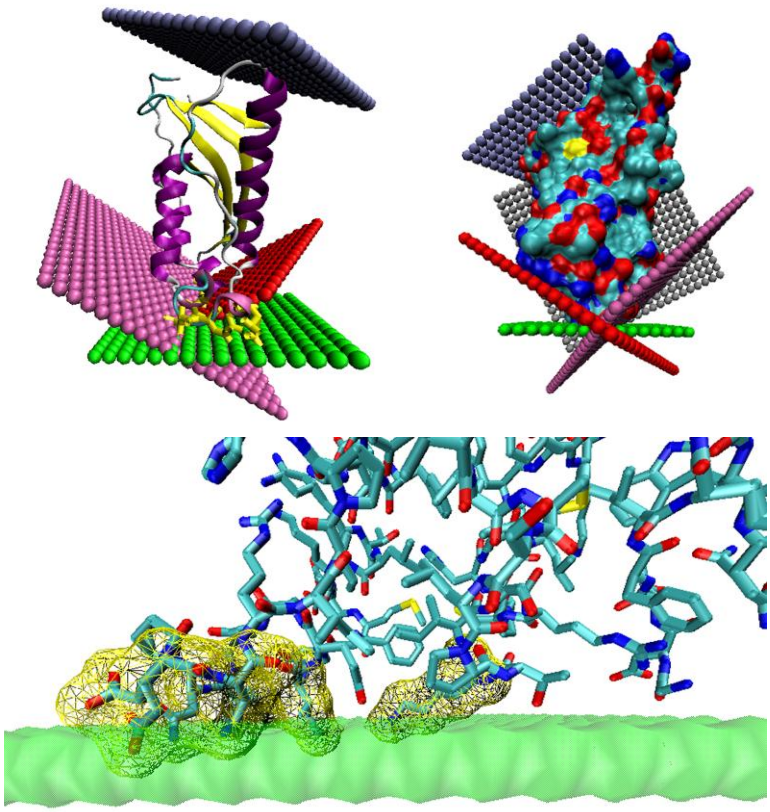
Examples of residues disengagement during SMD:
Before (left column) and after (right column) disengagement.

Supplementary Figure 4.



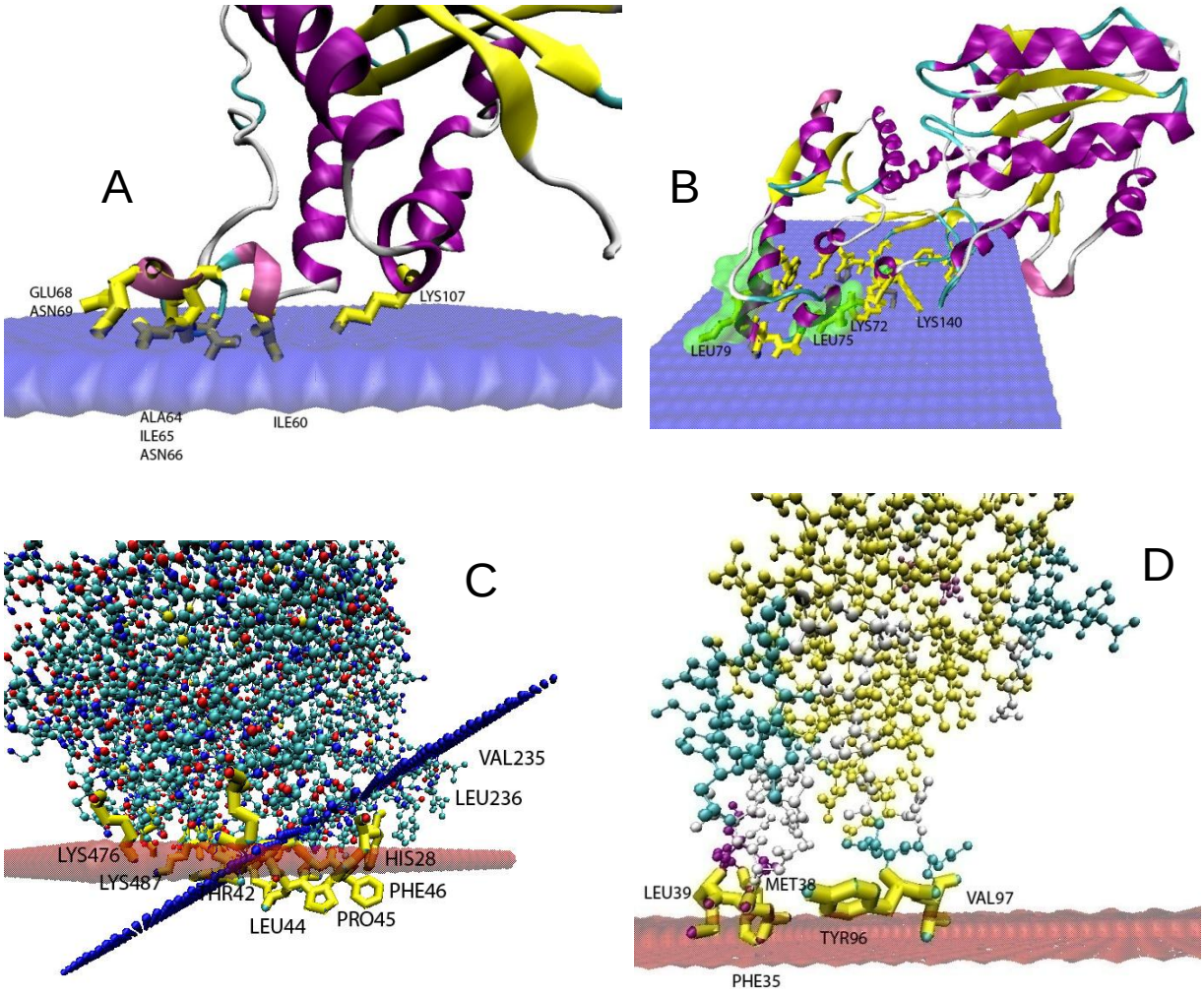
Parameters used for calculation of scoring function: Free energies of interaction of each amino acid with the membrane (dark green) and SMD defined parameters of the residue “membranephilicity” (light green – calculated with disengagement time; violet-calculated with disengagement force).

Supplementary Fig.5.



Selection of the best membrane-contacting flat surface encompassing a protein (1kq6). If the distance of any atom of a residue is within +- 0.5 A from the membrane (green layer) these residues (yellow surfaces) are considered contacting the membrane.

Supplementary Figure 6.



Predicted by MAPAS and proposed on the basis of experiments residues contacting membranes: (A) Protein 1kq6, yellow residues – predicted by MAPAS 100% correspond to the residues proposed on the basis of experiment; (B) Protein 1f0k, yellow residues – predicted by MAPAS, green surfaces – residues proposed on the basis of experiment[6] , concordance 71%; (C) Protein 1tqn (blue plane proposed in experimental paper[7], yellow residues – predicted by MAPAS; (D) Protein 1rlw, yellow residues predicted by MAPAS (5 from 9 proposed on the basis of experiment[8] .

Supplementary Figure 7.

Image constructed by MAPAS program. Model of transmembrane protein OAT1 (human anionic transporter) . MAPAS selected one of the membrane contacting surfaces (red spheres) with high scores: MRS 3.27, MAS 65.49, K_{mpha} 2.01

Supplementary Table 1.

Energies of interaction of amino acids with the membrane and their membrane disengagement scores

Amino Acid	Free energy of amino acid interaction with the membrane (kcal/mol)	Membrane disengagement Score (W) (calculated with 'force' of disengagement)	Membrane disengagement Score (W) (calculated with 'time' of disengagement)
Q	11.2	5.4	12.0
M	10.9	4.8	11.2
K	10.5	8.2	14.8
L	9.5	5.9	7.0
N	9.2	5.0	9.1
H	8.9	2.9	7.9
Y	8.4	4.1	5.9
F	8.1	5.5	7.4
I	8.1	6.5	8.2
C	8.0	5.4	6.7
A	7.2	2.1	5.0
E	7.0	4.4	8.0
D	5.8	5.8	7.1
R	5.5	3.5	7.1
G	5.3	3.3	3.6
P	4.9	0.6	2.8
T	4.8	3.3	4.2
V	4.1	2.2	3.8
S	4.0	1.5	4.2
W	2.5	4.2	4.5

Supplementary Table 2

Comparison of predicted membrane contacts of proteins with experimentally defined contacts

	1pocA	**1kq6A**	**1f0kA**	**1tqnA**	**1rlw_**
Residues defined in publication.	2,14,23, 24,51,53 ,78, 82, 85,92	64-66	72,75,79,80, 82,86,140	28-32,42-54, 218-238,256	34-39, 96-98
Predicted membrane-contacting residues	2,14,23, 24, 78,82,85	58-60, 64-66	73,75,76-83 117,140,141	28-31,42,43 48-54, 216-225, 227-231,233-239	33-40 96-98
Overlap %	70	100	71.4	81.6	100
MRS Score	4.57	4.34	4.29	4.03	4.33
MAS Score	46.84	62.03	56.9	38.0	30.5
K_{mpha}	1.33	1.76	1.88	0.93	1.17
GO-classification (membrane contacting)			0016020 0030259	0016020 0016021 0005624	0016021 0016042
Localization		Endosome membrane	Bacterial gram-negative inner membrane.	Endoplasmic reticulum membrane	Nuclear outer membrane
Reference	4	5	6	7	8

Supplementary Table 3.

Membrane-contact scores for known membrane contacting and random proteins

Random proteins	Membranephilic residues score **(MRS)**	Membranephilic area score **(MAS)**	**Membrane-associated proteins**	Membranephilic residues score (**MRS)**	Membranephilic area score **(MAS)**
1c43_A	2.75	6.10	**1d3h_A**	4.42	46.22
1h10_A	2.13	17.42	**1h3q_B**	3.26	15.56
1eko_A	2.37	26.35	**2a01**	6.00	71.21
1h11_A	1.72	1.37	**2qgc_A**	3.54	29.53
1m7y_A	2.18	14.28	**1tqn_A**	4.03	38.00
1ime_A	1.88	23.32	**1w0f_A**	3.49	52.10
5gch_	1.26	5.02	**1w6k_A**	2.69	34.20
1g9s_A	2.49	34.50	**1tqn_A**	3.98	66.80
1ru2_A	2.19	9.15	**1faq_A**	5.55	82.73
1dad	2.73	14.60	**1uum_A**	4.41	36.12
1ksz	2.56	25.98	**1pub_A**	3.54	55.00
1euu_	2.33	10.30	**1vmo_A**	4.74	54.50
1g2b_A	2.23	24.84	**1xte_A**	4.48	82.61
1oc5_A	1.71	16.87	**1yrk_A**	3.56	59.11
1exg	2.17	5.20	**2ddt_A**	4.14	36.48
1oyo_A	2.18	17.96	**2sn3_A**	4.95	78.02
1kdv_A	1.93	5.08	**1dyn**	4.99	39.21
1sgt	1.84	5.19	**1pmy_A**	4.16	25.03
1c2h_A	2.00	10.66	**2spc_A**	5.58	58.82
1faj_	2.36	29.23	**1g2x_A**	4.13	19.97

SUPPLEMENTARY METHODS.

Steered Molecular Dynamics

For modeling and simulations we created a set of peptides with the common sequence GLY-X-GLY-GLY-GLY-X-GLY where X = one of the 20 naturally occurring L-amino acids. This conformation mimics protein regions containing these residues. We conducted molecular dynamics (MD) simulations of the behavior of each peptide molecule in contact with the POPC membrane in water under normal conditions. The NAMD molecular dynamics program[1] version 2.5 was used with CHARMM27 force-field parameters[2] using periodic boundary conditions at constant pressure (1 atm) and temperature (300 K). The temperature was maintained at 300 K by means of Langevin dynamics using a collision frequency of 1/ps. A fully flexible cell at constant pressure (1 atm) was employed by means of the Nosé-Hoover Langevin Piston algorithm[3,4] as implemented in the NAMD software package. The van der Waals interactions were switched smoothly to zero over the region 10 Å and electrostatic interactions were included via the smooth particle-mesh Ewald summation[5]. The impulse-based Verlet-I/r-RESPA method[6,7] was used to perform multiple time-stepping: 4 fs for the long-range electrostatic forces, 2 fs for short-range non-bonded forces, and 1 fs for bonded forces. The peptide was situated on the surface of the membrane using the Visual Molecular Dynamics (VMD) software[8] with the non-GLY residues contacting lipids of the membrane. The system including the peptide, membrane and water box underwent the following steps of simulations: (1) 10,000 iterations minimization, (2) heating to 310 K at a rate 0.1 degree per ps, (3) equilibration for 10,000 ps. Then the system underwent 100 ps of non-restrained MD to insure that the initial conditions of the membrane-peptide system would not affect the location of the peptide on the membrane. The coordinates of the system were saved at intervals of 10, 25, 50, 75, and 100 ps during the unrestrained MD.

For each of these four snapshots we proceeded to conduct Steered Molecular Dynamics (SMD) with a force applied to the center of masses of the peptide directed perpendicular to the membrane, forcing the peptide away from the membrane. SMD continued from 70 to 150 ps depending on the time of disengagement of the peptide from the membrane. Mechanical forces and minimum distances of each residue to the membrane were measured at 0.1 ps intervals. **Supplementary Fig. 2a** online shows the values of the force (f_d) and **Supplementary Fig. 3b** online shows the time at which the peptide side chain disengaged from the membrane (t_d). The molecular configuration of the system around these minimums was independently examined using various graphical tools (**Supplementary Fig. 3** online). A 'force-defined' (W_1) and 'time-defined' (W_2) membrane disengagement scores were calculated correspondingly as f_d/L_1 and t_d/L_2 (where t is the average time until complete disengagement, and L_1 and L_2 are the scaling coefficients).

Free binding energy calculations

We also calculated the binding free energy between the selected residues of the probe peptide and the membrane lipids using an MM-PBSA approach[9-11]. Here the free energy change upon peptide binding in solution is calculated as the difference between the free energy of binding in vacuum and the free energy of solvation of the peptide, lipid membrane and complex respectively.

$$\Delta G_{Bind}^{Sol} = \Delta G_{Bind}^{Vac} + \Delta G_{Solv}\ Complex\ -\ \Delta G_{Solv}\ Peptide\ + \Delta G_{Solv}\ Lipid \qquad (0.1)$$

where ΔG_{Bind}^{Vac} is the difference between the gas phase energy of the complex, peptide, and lipid membrane adjusted by the difference in entropy

$$\Delta G_{Bind}^{Vac} = E_{MM}\ Complex\ -\ E_{MM}\ Peptide\ + E_{MM}\ Lipid\ \ - T\Delta S \qquad (0.2)$$

and ΔG_{Solv} is calculated by solving the linearized Poisson Boltzman (PB) equation[12,13] for each of the three states, giving the electrostatic contribution to the solvation free energy, and adding an empirical term for hydrophobic contributions.

$$\Delta G_{Solv} = G_{elec}^{\varepsilon=80} - G_{elec}^{\varepsilon=1} + \Delta G_{Hydrophobic} \tag{0.3}$$

where

$$\Delta G_{Hydrophobic} = \gamma SA \tag{0.4}$$

and γ is an empirical atomic solvation parameter, often referred to as the surface tension, and SA is the solvent accessible surface area calculated with a solvent probe radius of 1.4 Å.

To calculate the peptide to lipid membrane binding free energies using the MM-PBSA method we took the snapshots at 10, 25, 50, 75 and 100 ps intervals from the explicit solvent MD trajectories described above. The solvent was removed and then the gas phase molecular mechanics energy, PB solvation energy and surface area was evaluated using CHARMM c33b1[14]. The all atom CHARMM27 force field[2], was used for both the peptide and lipid. All non-bonded interactions were evaluated without truncation although the van der Waals energy difference was scaled by 0.171 based on previous work by Roux *et al*[15]. The PB solvation energy was calculated using CHARMM's PBEQ module. The dielectric constants of the solvent and solute were set to 80.0 and 1.0 respectively. A focused refinement of the PB grid was performed with a grid spacing of 0.45 Å. The atomic radii used were those optimized by Roux and Nina[16]. The solvent accessible surface area was calculated using a probe of radius of 1.4 Å and the empirical solvation parameter 'γ' was set to 0.033 KCal/Mol/Å^2. Since our interest is in the relative differences in binding energy between the various peptides and the entropy difference between the various peptides was expected to be small we did not calculate the final term in equation (0.2).

Since the snapshots used for the PBSA calculations did not uniformly have just one of the residues of interest bound to the surface, or in some cases had a mixture of the residue of interest and one or more of the GLY residues also bound so we had to account for this in our results in order to obtain an effective binding energy for just one residue. This was achieved by first calculating the average binding energies for a poly-glycine with just one glycine bound to the membrane. Visual inspection showed that the 10, 25 and 50ps snapshots all had just one glycine bound and so these were used to obtain the control value for glycine. Then for each of the five snapshots for all the 20 amino acid sequences we performed visual inspection to determine how many of the residue of interest and how many of the remaining glycines were bound to the membrane surface. We then calculated the effective binding energy of a single residue by subtracting off the effect of any glycines that were bound based on the GLY control value and then dividing by the number of the residues of interest that were bound to the membrane. For example the 10ps snapshot of the alanine case was determined to have two ALA residues and two GLy residues in contact with the membrane. Thus the binding energy for a single ALA residue in this snapshot was estimated to be:

$$E_{bind}^{1A} = \frac{E_{calc}^{2A+2G} - 2E_{calc}^{G}}{2} \qquad (0.5)$$

Once the effective binding energy for a single residue in each snapshot was determined the effective binding energy over all five snapshots was then averaged to give a binding energy for each amino acid type. These results are shown in **Supplementary Table 1** online.

Supplementary Fig. 4 online shows the sorted free energies of interaction of amino acids with the membrane and SMD defined parameters of amino acids disengagement from the membrane. One can see that these are correlating reasonably (coefficient of correlation 0.82 between 'time' scores and free energies set).

Encompassing Planes Construction

Through each three C-alpha atoms of the protein we construct a plane (equation *Av+Bx+Cz+D=0*). Then a program calculates if any of the C-alpha atoms of the protein is located on the distance more than 1A from both sides of the plane. If so – such a plane is discarded. This produces a population of protein regions where all C-alpha atoms are located on one side of a plane. This procedure defines all possible planes that encompass the protein 3D structure. Then the program calculates a number of residues that belong to each plane by calculating a distance of each atom of the protein to each encompassing plane (**Supplementary Fig. 5** online). Residues that have any atom within 0.5 A of the surrounding plane are considered to be included to this plane. MAPAS can be tuned to use greater or smaller sets of membranephilic residues for its scoring functions, or thickness of the layer from which the algorithm identifies the residues that belong to a 'best' membrane-contacting plane.

Scoring Parameters

The solvent accessible area of a residue included in a plane was scored as the entire solvent-accessible area of the residue. Each protein is encompassed by a significant number of putative planes (for globular proteins the number can be in the thousands). Scoring functions were developed to select the plane(s) most probable to make a stable contact with a membrane, The first of these, membranephilic residues score, (**MRS**) is calculated using the membrane disengagement score (W) from SMD (previous section): $MRS = (\sum W(Nu_i * Su_i/Su_{imax}) + \sum((W*Nc_i)/ (Nu_i + Nc_i)$. Where Nu_i = the number of uncharged residues in the plane, Nc_i = number of charged residues in the plane, Su_i– solvent accessible surface of the uncharged residue that is included to the plane, Su_{imax}– maximum solvent accessible surface of the uncharged residue that is included to the plane, Nc_i– number of charged residues in the plane. MRC > 4 is a strong support for defining of the protein as membrane contacting.

The second scoring function is membranephilic area score **(MAS)** is based on the assumption that in a stable membrane:protein interface, the preponderance of solvent exposed surface area in the planar protein interface will be contributed by highly membranephilic residues. The algorithm for this calculation is MAS= ($S_{top(plane)}/S_{all(plane)}$); where $S_{top(plane)}$ is the solvent-accessible surface of a selected number (for example 5) of the top membranephilic residues defined by SMD (**Supplementary Table 1** online) in the best putative membranephilic plane; and $S_{all(plane)}$ is the solvent-accessible surface of all residues in this plane. Values of MAS more than 40% usually support the selection of a plane as membrane-contacting.

The third scoring function is based on the assumption that the overall surface properties of a particular protein will also influence the tendency of the most membranephilic region of the protein to form a stable membrane association. Thus, the MAS score is used in conjunction with this assumption to calculate the third scoring function, the coefficient of "membranephilic asymmetry" ($\mathbf{K_{mpha}}$). $K_{mpha}=(MAS)/(S_{top(protein)}/S_{all(protein)})$, where $S_{top(protein)}$ is the solvent-accessible surface of a selected number (for example 5) of the top membranephilic residues of the entire protein, $S_{all(protein)}$ is the solvent-accessible surface of all residues of the protein. $K_{mpha} > 2$ is usually a strong indication that the protein can be membrane-contacting.

We have to note that these scoring functions are experimental, and therefore are in a constant process of improvement and have a number of ways to be improved. Nevertheless using of all of them usually brings reliable prediction of membrane-contacting planes and prediction of membrane contacting proteins.

The program has been tested with the following browsers: Microsoft Internet Explorer, Netscape, Mozilla Firefox, and Opera.

References

1. Kalé L, Skeel R, Bhandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K, Schulten K. (1999) *J Comp Physics* 151:282-312.
2. Feller S, MacKerell A (2000) *J Phys Chem B* 104:7510-7515.
3. Tu K, Tobias DJ, Klein ML. (1995) *Biophys J* 69:2558-2562.
4. Feller S, Zhang Y, Pastor RW, Brooks BR. (1995*J Chem Phys* 103:4613-4621.
5. Essmann U, Perera L & Berkowitz ML. (1995) *J Chem Phys* 103:8577-8593.
6. Tuckerman M, Berne BJ, Martyna GJ. (1992) *J Chem Phys* 97:1990-2001.
7. Grubmüller H, Heller H, Windemuth A, Schulten K. (1991) *Mol Simul* 6:121-142.
8. Humphrey W, Dalke A, Schulten KJ*. Molec. Graphics*, 1996, 14: 33-38.
9. Massova I, Kollman PA. (1999) *J Am Chem Soc* 121:8133-8143
10 Kollman PA, Massova I, Reyes C, Kuhn B, et al. (2000) *Acc Chem Res* 33: 889-897
11. Habtemariam B, Anisimov VM, MacKerell ADj. (2005) *Nuc Acid Res* 33:4212-4222
12. Sharp KA, Honig B. (1990) *Annu. Rev. Biophys Biophys Chem* 19:301-332
13. Davis ME, McCammon JA. (1990) *Chem Rev* 90: 509-521
14. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. (1983) *J Comp Chem* 4: 187-217.
15. Roux B. *Personal Communication*.
16. Nina M, B. Roux B. (1997) *J Phys Chem B* 101: 5239-5248

SUPPLEMENTARY DISCUSSION

Prediction of membrane-contacting surfaces of proteins.

To find out whether MAPAS can reliably define the membrane-contacting regions of the protein surfaces we selected a group of five proteins that were experimentally shown to be membrane-contacting and for which the membrane contacting residues were proposed in corresponding publications[1-5]. **Supplementary Fig. 6** online and **Supplementary Table 2** online show a comparison of membrane-contacting surfaces and sets of residues contacting the membrane for a set of membrane-contacting proteins predicted by the MAPAS software and as proposed on the basis of experiments[1-5]. One can see that in all presented cases the MRS scores are greater than 3. In all cases, the combination of two scoring parameters could reliably predict the best membrane-contacting plane. In all five cases, sets of membrane contacting residues defined by MAPAS and proposed on the basis of experiments[1-5] significantly intersect.

Prediction of membrane-associated proteins.

We next examined the ability of the MAPAS program to identify membrane associated proteins. We selected two sets of proteins from the PDB: one set consisted of random non-membrane-associated proteins while the other included only membrane associated proteins. Random proteins were selected using the following criteria. Proteins excluded from the random set were: (1) proteins that are listed in GO classification with 'membrane-related' codes, for example 0016021 – integral to membrane, 0016020 – membrane, 0019897 – extrinsic to plasma membrane, etc. http://www.geneontology.org/index.shtml Membrane-associated proteins were selected using proteins defined in GO classification as membrane contacting (see **Table 1 below**). **Fig. 1b** and **Supplementary Table 3** online show values for the scoring parameters for random and membrane-

Table 1.

GO-terms used to select membrane-contacting proteins.

Type of membrane	GO cellular component	GO-term
Endoplasmic reticulum membrane	0005783 0005790 0016020 0016021	Endoplasmic reticulum Smooth endoplasmic reticulum Membrane Integral to membrane
Endosome membrane	0005769 0019897	Early endosome Extrinsic to plasma membrane
Eukaryotic plasma membrane	0005886 0005887	Plasma membrane Integral to plasma membrane
Lysosome membrane	0005764 0016021	Lysosome Integral to membrane

associated proteins. One can see that in each case there is a clear distinction between the set of random proteins and the membrane-related proteins. Using proposed scoring methods one can reliably predict that a given protein is membrane contacting and find its putative surface of contact.

Prediction of surface-contacting parts of transmembrane proteins.

Checking the program after all current improvements we found that MAPAS can be used in many cases also for prediction of membrane outer surface contacting parts of the transmembrane proteins. **Supplementary Fig. 7** online shows such a surface (red spheres) predicted by MAPAS with high scores for our model of human organic anion transporter OAT1. We understand that use of MAPAS for transmembrane proteins is in experimental stage and we are working to improve this option. Nevertheless we described this application because even in its current state it can by already useful for researchers.

We would like to note that sometimes false positive scores in studied proteins may arise from conserved functional roles that are unrelated to membrane-contacting functionalities. For example, surfaces between the domains of a

multimeric protein or surfaces designed to bind other specific proteins can each have a high level of membranephilicity.

Although our scores can be used in a complementary manner to rank membrane-contacting proteins and their planes of contact, each of these scores has intrinsic bias. For example more SMD experiments with different types of membranes could be done to improve the reliability of the residue disconnection scoring function. Additionally identification and exclusion of interdomain regions, that can sometimes lead to false positive predictions, could be used in the future to further improve the program.

REFERENCES

1. Lin Y, Nielsen R, Murray D, Hubbell WL, Mailer C, Robinson BH, Gelb MH (1998), Science, 279:1925-1929
2. Cho W, Stahelin RV. 2005. Annu Rev Biophys Biomol Struct. 34:119-151.
3. Sha HA, Walker D, Shi Y, Walker S. (2000) Protein Science, 9:1045-1052
4. Williams PA, Cosme J, Sridhar V, Johnson EF, McRee DE (2000) Mol. Cell 5:121-132
5. Evans J, Murray D, Leslie C, Falke J. (2006) Molecular Biology of the Cell 17:56-66