Internship project 2015

Computational Protein Design of a PDZ Domain using Proteus

In the Biochemistry laboratory of the Ecole Polytechnique

From January 12 to June 30, 2015

By Xingyu CHEN
Master2 student (2014/2015) of In Silico Drug Design, University Paris Diderot – Paris7
Add.: 166 Avenue de La Division Leclerc, 92160 Antony, France
Phone: +33 (0)6 03 81 15 88
Email: xingyu809@gmail.com

Supervised by Professor Thomas SIMONSON
Add.: Department of Biology, Laboratoire de Biochimie (UMR 7654 du CNRS), Ecole Polytechnique, 91128 Palaiseau, France
Phone: +33 (0)1 69 33 48 60 (fax: -- 49 09)
Email: thomas.simonson@polytechnique.fr

ACKNOWLEDGMENTS

I thanks to all members in BIOC and Bioinformatics team members, especially my supervisor Thomas Simonson, who helps me patiently and suggests on my work. I gratefully acknowledge David Mignon for his help in solving technical problems. I gratefully acknowledge Nicolas Panel and Karen Druart for their discussions in my project. I gratefully acknowledge Anke Steinmetz, who always encourages me and gives me her expert suggests in my project.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Context

The Rho (Ras homologous) GTPase (Guanosine Triphosphatase) family is well known for its regulation of actin cytoskeleton dynamics and is crucial for cell migration. Rho GTPase misregulation contributes to cancer cell invasion, tumorigenesis and metastasis[1]. The T-cell lymphoma invasion and metastasis gene 1 (Tiam1), a multiple domain protein, is identified as a guanine nucleotide exchange factor (GEF) that activates the Rho-family GTPase Rac1 (Ras-related C3 Botulinum Toxin Substrate 1)[2]. Tiam1 has important roles in cell biological function, including cell polarity, integrity of adherent junctions[3], tight junctions[4-6], and cell-matrix interactions[7, 8], whose deregulation has been documented in many cancers and Tiam1/Rac1 signalling has been implicated in the oncogenic transformation of cells. Tiam1 can be inhibited by small molecules that bind to its PDZ domain, and block or down regulate the interaction with its target proteins. We focus on the PDZ domain of the Tiam1 protein for searching peptide inhibitors.

1.2 PDZ Structure and Classification

The common structure of PDZ domains comprises six β strands and two α helixes. A groove between the β strand and the α helix can accept a peptide ligand at the C-terminal binding as a beta sheet extend *(figure1)*. The PDZ domains are classified into 3 classes according to the sequences of their binding peptide ligand. All 3 types of PDZ domain prefer a ligand with a hydrophobic residue at 0 position from C-terminal. At the -2 position of the ligand, the class II PDZ domain prefers a hydrophobic residue, whereas the class I PDZ domain prefers the residue S and T, and the class III PDZ domain prefers the residue D and E. The PDZ domain of Tiam1 is the class II PDZ domain. Some precedent studies show that, ligands of Tiam1 PDZ domain has a consensus amino acid sequence motif of $[Y/F/G]_{-2}$-$[Y/F/H/W]_{-1}$-$[W/F/A]_{0}$-COOH[9, 10]. In this project, we use our computational protein design (CPD) method to generate mutants of ligands at the positions -4 to 0, in Tiam1 PDZ complexes.

1.3 Experimental sequences of PDZ peptide ligand

There are not many experimental structures of class II PDZ. We collected two experimental sequence sets of PDZ peptide ligand. We searched the crystallography structures of PDZ domain in Protein Data Bank, and found 143 PDZ structures that are formed in complex. Among these 143

structures, there are only 30 structures of class II PDZ (the PDB codes and the ligand sequences are presented in *table1*). The first set collect the 26 sequences of ligand peptide from class II PDZ complexes. The second set is from a Tiam1-binding peptide library, which contains 69 peptide sequences in total[11]. The sequences in library are "YAAX$_{-4}$X$_{-3}$X$_{-2}$X$_{-1}$X$_0$", in which X represents a given residue type. The sequence logos of the ligand at position -4 to 0 are presented in *figure2*. The size of residue logo represents its population at each position.

1.4 Internship subject

The global goal of the PDZ project is design the peptide inhibitors of Tiam1 PDZ domain for anticancer applications. In my internship project, we want to evaluate the performance of the CPD method, which is developed in our group, by mutagenesis of ligand in Tiam1 PDZ complexes. To this end, we developed different protocols to optimize the parameters of CPD method, the reference energies. We applied these optimized parameters and compared the calculated sequences with experimental sequences.

2. MATERIALS & METHODS

2.1 General CPD method

To perform CPD, our laboratory developed the Proteus software package, which is implemented by xplor, shell, and perl scripts. The method is already published[12]. There are three main steps: system construction, energy matrix calculation, and sequence exploration.

2.1.1 Structural model

The folded protein or peptide is modeled using a fixed backbone and a discrete rotamer library for the side chains; solvent is modelled implicitly. For the unfolded state, we use a extended peptide model, for examples, ALA-X-ALA. We consider that each residue X in the protein interacts only with the nearby backbone and solvent, not with other side chains.

Each residue in the protein is defined as one of three types: residues with fixed side chain are called as "frozen"; the residues with flexible side chains that cannot mutate are called "inactive", and the residues that can be mutated are called as "active".

2.1.2 Energy function

The energy function is calculated using an MM/GBSA model. We approximate the energy further by decomposing it into residue terms. The total energy equals the sum of self-interaction energy of

residue i (E_{ii}) plus the sum of pairwise interaction energy between residue i and residue j (E_{ij}).

$$E = \Sigma E_{ii} + \Sigma E_{ij} \qquad (1)$$

For the unfolded state, we introduce "reference energies" for each residue type X (E_X), which represent the contribution of amino acid X in the unfolded state. Each value includes a contribution E_X^{pept} from an extended peptide model plus an empirical contribution e_x that can be adjusted. The unfolded energy of protein ($E_{unfolded}$) is the sum of the reference energies of each position i (equation (2)).

$$E_{unfolded} = \Sigma E_X(i) \quad (2)$$

2.1.3 Energy matrix

The interaction energies are stored in an energy matrix, used as a look-up table. The self-interaction energy term (E_{ii}) and the reference energy (E_X) are saved in the diagonal of the matrix; the pairwise interaction energy terms (E_{ij}) are saved in off diagonal terms.

2.1.4 Mutation space exploration – MC

The mutation space exploration is achieved by proteus program. In this project, we used a Monte Carlo (MC) method. MC generates sequences and conformations according to a Boltzmann distribution that depends on the folding free energy:

$$E_{folding} = E_{folded} - E_{unfolded} \qquad (3)$$

For selected positions of the protein (or peptide) sequence, we can do single point mutations or multiple mutations and search the rotamer conformations, or only search the rotamer conformations. The results will be analysed for testing and improving the reference energies E_X, or for selecting the interesting mutants.

2.2 Application systems

2.2.1 The syndecan peptide

For preparing the system, we used the Tiam1:peptide Xray complex (PDB: 4GVD_A_D) and minimized slightly with the procedure in below:

1. 200 steps minimization of Steepest Descent (SD) method with 50 kcal protein constraints.
2. 200 steps minimization of SD method with 10 kcal protein constraints.
3. 200 steps minimization of SD method and of Adopted Basis Newton-Raphson) method respective, and with 2 kcal protein constraints.
4. 200 steps minimization of sd method and abnr method respective, and with released protein.

The histidines in complex are treated by the software "PROPKA" and "Reduce". The two histidines in protein are protonated at delta position with the pka value equals to 6.18 for histidine at position 844 and 6.34 for the one at position 847.

We firstly worked with the extended Syndecan1 peptide. The peptide structure was extracted from the minimized complex structure. The sequence is "TKQE$_{-4}$E$_{-3}$F$_{-2}$Y$_{-1}$A$_{0}$-COOH". For the energy matrix calculation, we used solute a dielectric constant of either 4 or 8 for the GB energy term. The atomic surface coefficients were (in kcal/mol/Å^2): alkane atoms = -0.005; hydrophobic atoms = 0; polar atoms = -0.008; aromatic atoms = -0.012; ionic atoms = -0.009. Each diagonal (E_{ii}) and off-diagonal (E_{ij}) term in the energy matrix was computed after 15 steps of Powell minimization (per pair). We ran MC simulations with 5 million steps, to explore single point mutations with all 20 amino acid types allowed at the last five peptide positions. The MC simulation uses a "mono walker" with the temperature equals to 0.6K. The movement probabilities of rotamer (*Rot_Proba* and *Rot_Rot_Proba*) are equal to 0.5. The movement probabilities of mutation (*Mut_Proba*, *Mut_Mut_Proba* and *Mut_Rot_Proba*) are equal to 0.1.

2.2.2 Tiam1 PDZ protein

The second system is the Tiam1 PDZ protein. The structure was extracted from Tiam1:peptide Xray complex (PDB: 4GVD_A_D) after the minimizations. For the energy matrix, we used the same parameters as for Syndecan1 above. We ran MC simulations with 10 million steps, to explore multiple mutations over all the positions of protein (in the absence of the peptide ligand). For the MC simulation, we used "8 walkers" with the temperatures (unit: K) equal to 3, 2, 1.333, 0.888, 0.592, 0.395, 0.263, and 0.175. The movement probabilities of rotamer and mutation are the same as in above. We only selected the sequences generated with the temperature 0.592.

2.2.3 Cask PDZ protein

Our third system is built with the PDZ protein of hCask from the Xray structure (PDB code: 1KWA_A). The missing residues in the file pdb are added by "scwrl". The histidines are treated as protonated at both delta and epsilon positions. We used the same procedure and parameters as for Tiam1 PDZ protein.

2.2.4 Sequence alignment and amino acid composition of PDZ proteins

To calculate the experimental amino acid composition of our proteins, we collected two sets of natural sequences from UNIPROT: one with Tiam1 homologues and the other with hCask homologues. The homologues with sequence identities compared to the target protein of between

60% and 85% were selected. There were 50 homologues of Tiam1 and 126 of Cask.

2.3 Reference energies, method 1: syndecan peptide as the unfolded model

A first method to calculate the reference energies relies on the extended Syndecan1 peptide as the unfolded state backbone model. From our MC peptide simulations (above), we used the reference energies, which are obtained from the precedent study with a peptide fragment in complex with MHC class II protein ($\{E_X^i\}_{MHC}$)[14]. As MC method uses folding free energy to generate mutants (equation (2) and (3) above), and as our peptide is extended, the folded energy (E_{folded}) in our case represents the extended peptide energy of mutants (E_{mut}). The energies at unfolded state are represented by the sum of initial reference energies ($\Sigma_i E_{Xini}^i$). Thus, the MC energy can be represented in equation (4). We calculated the energies of all point mutant respect to alanine (A) with the equation (5) (E_{mut}^{Xi} represent the extended energy of the mutant with given residue type X at position i, E_{mut}^{Ai} represent the extended energy of the mutant with residue type A at position i). The equation (5) can be transformed in equation (6).

$$E_{MC} = E_{mut} - (\Sigma_{j\neq i}\{E_X^i\}_{MHC} + \{E_X^i\}_{MHC}) \qquad (4)$$

$$\{\delta E_{mut}^{pept}\}_X^i = (E_{mut}^{Xi} - \Sigma_{j\neq}\{E_X^i\}_{MHC} - \{E_X^i\}_{MHC}) - (E_{mut}^{Ai} - \Sigma_{j\neq i}\{E_A^i\}_{MHC} - \{E_A^i\}_{MHC}) \qquad (5)$$

$$\{\delta E_{mut}^{pept}\}_X^i = E_{mut}^{Xi} - E_{mut}^{Ai} - \{E_X^i\}_{MHC} + \{E_A^i\}_{MHC} \qquad (6)$$

We obtained energies for all possible single point variants at the last five peptide positions. For each sidechain type X, the energies of the point mutant difference relative to Ala and their averaged value (at the same peptide position) was then taken to represent E_X. For Ala, we can arbitrarily set $E_X = 0$.

2.4 Reference energies, method 2: empirical model

2.4.1. General method

Here, we use a simple extended peptide model to descript the unfolded state. But this simple model needs to be supplemented by an empirical correction. Thus, we add an empirical energy correction e_X to each E_X, to avoid the too abundant occurrence of a given residue type X:

$$E_X = E_X^{ini} + e_X \qquad (7)$$

e_X depends only on the amino acid type. We iteratively modify e_X using the Boltzmann-like relation[12, 14]:

$$e_X^{new} = e_X^{old} + 0.5 \ln f_X^{exp} / f_X^{cal} \qquad (8)$$

where f_X^{exp} represents the experimental composition of amino acid type X (obtained by counting the

occurrences of X in a set of homologs, see above). f_X^{cal} represents the theoretical composition of amino acid type X, obtained from one or more MC simulations. We apply (8) repeatedly; the iterations will be stopped when f_X^{cal} of all amino acid types have converged close to f_X^{exp}. A flowchart of the procedure is presented in *figure3*. In general case, we have 20 acid amino types and e_X values to adjust. But to decrease the number of residue types, we can also group similar residues, so that they share a single e_X value (but different E_X^0 values). To refine the model, we can assign different e_X values to positions that are buried or exposed (in the folded state).

2.4.2 Implementation details for Tiam1, version 1

The experimental amino acid compositions are computed with the sequences of the 50 Tiam1 homologues. The calculated amino acid compositions are generated by a 500 million step MC where all protein positions mutate freely. The MC simulation uses a "multi-walker", Replica exchange protocol, with 8 walkers (the 8 temperatures as in section 2.2.2). For the initial E_X values, we used the lowest self-interaction energies (E_{ii}) of each residue type, averaged over all the protein positions. We set the initial e_X values to zero. We iteratively updated the reference energies until convergence (10 iterations for eps8, and 15 iterations for eps4 but not yet converged).

2.4.3 Implementation details, version 2

The second implementation is with Tiam1 and hCask protein together. We classified residues into 7 groups according to the matrix BLOSUM62[13] *(table2)*.

Each group shares a single e_X value (zero, initially). We also separated the buried and exposed positions using a surface burial fraction threshold equals to 0.2, allowing them to have different e_X values. There are 33 buried and 51 exposed positions in Tiam1, and 34 buried and 40 exposed positions in Cask. Thus, each residue type has two reference energy values: buried reference energy (E_{Xb}) and exposed reference energy (E_{Xe}). The iteration formulas are:

$$\text{Buried residues: } E_{Xb}^{new} = E_{Xb}^{old} + 0.5\ln(f_{gb}^{exp}/f_{gb}^{cal})$$

$$\text{Exposed residues: } E_{Xe}^{new} = E_{Xe}^{old} + 0.5\ln(f_{ge}^{exp}/f_{ge}^{cal})$$

For the experimental compositions of amino acid groups (f_g), we calculated separately for the 50 homologues of Tiam1 PDZ and the 126 homologues of Cask PDZ, and averaged the two. The calculated compositions of amino acid groups are computed from 10 separated MC simulations: 5 for Tiam1 and 5 for Cask, at each iteration. For the 5 MC simulations with each protein, we assigned in order 1 per 5 positions as active to each simulation for the multiple mutation explorations, and the rest 4/5 inactive positions with the side chain flexible. The MC simulations

used the "8 walkers" as in section 2.4.2 above. The initial E_X are used the delta mutant energies, which averaged over the 5 active positions of peptide from the section 2.3. The shell script for reference energy optimization is in *Appendix*.

2.5 Simulations of the Tiam1 PDZ:peptide complex

To test the quality of the reference energies, which are obtained from the different methods above, we applied these reference energies to explore ligand mutants with Tiam1 PDZ complexes. For the simulations, we used 3 different starting structures:

1. The Xray structure: Tiam1 PDZ complex binding with Syndecan1 (PDB: 4GVD_A_D) and we minimized slightly with the procedure describing in the section 2.2.1.
2. The Tiam1 PDZ complex with the peptide ligand, which was selected from the Tiam1-binding library[11] with the best binding affinity score. The sequence of peptide is "$YAAE_{-4}K_{-3}Y_{-2}W_{-1}A_0$". The structure is obtained by 20ns molecular dynamics simulation.
3. Tiam1 PDZ complex binding the syndecan1 mutant with Ala mute to Phe at position 0. We used a 100ns molecular dynamics simulation. The structure used is the structure most similar than the average structure from the molecular dynamics simulation.

For each structure, we built two systems both with the protein flexible (side chain flexible) and frozen (side chain fixed), and calculated the energy matrixes. In each system, ligand peptides are always modeled with the first 3 positions inactive (side chain flexible), and the last 5 positions active (will be mutated). The dielectric constant was either 4 or 8 (the other parameters are as the parameters used in the simulations above).

For sequence exploration, we used the MC method for both multiple mutations and single point mutations at positions -4 to 0 of peptide ligand in Tiam1 PDZ complex. The multiple mutation explorations are achieved by a 10-million-step MC over all active positions. The single point mutations are achieved by 5-million-step MC at each active position. The MC simulations are used the "8 walkers" as described in above.

The interested structures were reconstructed, and minimized by 200 steps using xplor.

3. RESULTS

3.1 Sequence alignments and experimental amino acid compositions

The two sets of homologues were aligned separately by the *Clustal Omega* program in *Uniprot* and viewed with *Jalview*. The alignment with 50 Tiam1 homologues is presented *in figure5*. The alignment with 126 Cask homologues is presented *in figure6*. The identify percentage between Tiam1 and Cask is 23.68%. We eliminated the redundant sequences, and the aligned sequences are ordered by their identify percentages. The amino acid compositions of these two homologue sets are represented *in table4*. Results are shown separately for the buried and exposed protein positions and for groups of similar amino acids.

3.2 Simulations of the syndecan peptide

We simulated the syndecan peptide in its extended conformations with the last 5 positions allowed to mutate, either individually or together. For the single point mutations, we obtained all amino acid types at each active position of peptide. The results are presented by sequence logos (*figure4*). The size of the letters corresponds to the residue type population. The types at each mutated position are sorted by their populations. The letter "h" represents the histidine protonated at delta position (Hid) and "j" represents the histidine protonated at epsilon position (Hie). The energies of all point mutants, relative to Ala, and their averaged values are shown in blue bloc of *table3*. Theses energies are similar at all peptide positions, consistent with its extended structure. They will be applied to Tiam1:peptide complex as the reference energies for the peptide mutation space exploration.

3.3 Simulations of the Tiam1: peptide complex – version1

Next, we considered the peptide in complex with Tiam1. If we assume the extended syndecan peptide is a good model of an unfolded protein, we can use the mutation energies above as reference energies E_X. We explored mutations in the same five peptide positions, individually or together. We used energy matrices corresponding to three slightly different 3D structures. The results are presented by sequence logos. For the results of multiple mutation explorations, the sequences that have the occurrence upper than 0.2% were aligned. The results are viewed in *Jalview* and show in *figure11-13*.

The first structure we used is the Tiam1:syndecan1 Xray complex (slightly minimized). The logos are presented in *figure7*. The peptide positions -1, -3 and -4 vary extensively, especially

position -1. At position 0 (the C-terminus), we only obtained small residues (Ala, Cys, Ser), not aromatic residues Phe and Trp. At position -2, with a frozen protein, we got many Phe, which is the native residue, and its similar residue His (protonated or not). The results obtained using frozen proteins are more native-like than using flexible proteins, as expected. The MC energies of the single point mutants (using eps8 and frozen protein) at position -2 and 0 are presented in *table6*. The 3D structures of top50-population sequences from multiple mutations with frozen protein and eps8 were reconstructed and minimized. A cross-eye stereo figure of *Pymol* is presented in *figure15* with the structures of top3-population sequences and of Xray.

The second structure is an MD model of Tiam1 bound to a peptide from the Tiam1-binding library[11] with the best binding affinity score. The sequence of peptide is "$YAAE_{-4}K_{-3}Y_{-2}W_{-1}A_0$". The results are presented in *figure8*. The obtained residues at position -1 and -3 are again very diverse, and that of the other 3 active positions are conserved. At position 0, we obtained almost only Ala, except a small population of Ser obtained by multiple mutations with a flexible protein. Many protonated His are found at position -2. At position -4, we obtained a large population of Ala and a little Cys.

Comparing the results with the experimental Tiam1-binding library[11] *(figure2b)*, we found the results are not very satisfying. The experimental studies have already obtained consensus sequences with Phe at position 0 of Tiam1 PDZ ligands[11]. But, with both 3D structures, we only obtained small residues (Ala, Cys, and Ser) at position 0. Thus, we also used a third MD structure, Tiam1 binding the syndecan1 mutant with Ala0 changed to Phe (A0F). The resulting sequence logos are presented in *figure9*. At position 0, we found a lot of His, which has similar structure with Phe, by the simulations with frozen protein, and with flexible protein, the residue Cys has the largest population. We did not regain Phe at position 0. In additional simulations that imposed Phe at position 0, we found the Phe residue type has energy about 2-3 kcal/mol too high to be sampled.

3.4 Optimization of empirical reference energies

To improve the peptide design results, we wish to introduce an empirical correction e_X to the reference energies used above. We explored two different methods to optimize the e_X; we only present the more sophisticated one, which gives the best results. With this method, we assign the same e_X value to a group of similar amino acid types, to reduce the number of parameters. At the same time, to refine the model, we assign different values to positions that are buried or exposed in

the folded structure. We adjust the e_X iteratively, so that the computed amino acid compositions match the experimental ones. The computed compositions at each iteration are obtained from ten MC simulations, five each of Tiam1 and Cask, with five different sets of 15 active positions.

The optimized reference energies are presented in green bloc of *table3*. The reference energies of buried residues are very similar to those of exposed residues. They are also close to the reference energies obtained by the simulations of the Syndecan peptide. The calculated amino acids compositions, after convergence (15 iterations), are presented in *table5*. The group compositions agree well with the experimental ones, but within each group, some amino acid compositions agree less well with the experimental values.

3.5 Simulations of the Tiam1: peptide complex – version2

We next used the improved reference energies to explore mutations of the syndecan peptide bound to Tiam1 (Xray structure). The MC energies of the 20 single point mutants (using eps8 and frozen/flexible protein) at position -2 and 0 are presented in *table6 and table7*. The sequence log results are presented in *figure10*. We obtained a lot of hydrophobic residues at each position with all simulations. At position 0, we obtained always only small residues. We have got a large population of aromatic groups at positions -2 and -1, and a large population of hydrophobic residues at position -3. These results are consistent with the experimental results.

By the simulation with frozen protein, large populations of aromatic residues are obtained at position -4 too. At position -3, a lot of acidic and polar residues (Glu and Asp) are found, which are consistent with the native residue Glu. The results of single point mutations with both flexible and frozen protein, and the results of multiple mutations with frozen protein, are similar. We can conclude that these results are more robust than those obtained in the previous simulations (no e_X corrections). For the results of multiple mutation explorations, the sequences that have the occurrence upper than 0.2% were aligned. The results are viewed in *Jalview* and show in *figure14*.

The 3D structures of top50-population sequences from multiple mutations with frozen protein and eps8 were reconstructed and minimized. A cross-eye stereo figure of *Pymol* is presented in *figure15* with the structures of top3-population sequences and of Xray.

We also reconstructed 3D structures of the best 200 conformations of each point mutant (at position 0 and -2) sequence with flexible protein and eps8. The averaged energies over the best 200 conformations and the affinities are presented in *table8*.

DISCUSSION

We studied Tiam1:peptide complexes using two different models for the unfolded state and two sets of reference energies E_X. The first model uses the unfolded peptide backbone to represent the unfolded state (of Tiam1). The mutation energies at each peptide position are similar; the position-average values are used for the E_X. The second method adds an empirical correction. We have two implementations with this method. The first, simpler implementation used Tiam1 (not Cask) and gave results that seem to be of moderate quality (not shown). The second implementation is more sophisticated and the e_X values are more constrained (use of groups, fewer active positions, use of two proteins). It also distinguishes buried and exposed positions (even though E_X is supposed to represent the unfolded state, where everything is exposed). The buried E_X and the exposed E_X of a given residue type are very similar, and similar to the unfolded peptide model (the e_X are small) which is encouraging. We can conclude that these reference energies are fairly general, not depending much on the positions or protein.

In the simulations of the Tiam1:peptide complexes, we observe that with a slight difference of the reference energies, the obtained sequence are significantly different. This reflects the exponential dependence of the populations on the E_X values. With the latest values, the MC results show a good agreement with the experimental peptide sequences and are more robust than the precedent results. But the multiple mutation results with a flexible protein in all the systems are less good, possibly because we did not include the Xray orientation of the sidechains among the possible rotamers. In the next work, we will add these “native” rotamer in our method for the mutation exploration.

REFERENCES

[1] Li, H., Peyrollier, K., Kilic, G. & Brakebusch, C. Rho GTPases and cancer. Biofactors (2013).

[2] Habets GG, Scholtes EH, Zuydgeest D, van der Kammen RA, Stam JC, Berns A, Collard JG. Identification of an invasion-inducing gene, Tiam-1, that encodes a protein with homology to GDP-GTP exchangers for Rho- like proteins. Cell, 1994, 77(4): 537–549

[3] Malliri, A., van Es, S., Huveneers, S. & Collard, J.G. The Rac exchange factor Tiam1 is required for the establishment and maintenance of cadherin-based adhesions. J Biol Chem 279, 30092-8 (2004).

[4] Chen, X. & Macara, I.G. Par-3 controls tight junction assembly through the Rac exchange factor Tiam1. Nat Cell Biol 7, 262-9 (2005).

[5] Mertens, A.E., Rygiel, T.P., Olivo, C., van der Kammen, R. & Collard, J.G. The Rac activator Tiam1 controls tight junction biogenesis in keratinocytes through binding to and activation of the Par polarity complex. J Cell Biol 170, 1029-37 (2005).

[6] Nishimura, T. et al. PAR-6-PAR-3 mediates Cdc42-induced Rac activation through the Rac GEFs STEF/Tiam1. Nat Cell Biol 7, 270-7 (2005).

[7] Shepherd, T.R. et al. The Tiam1 PDZ domain couples to Syndecan1 and promotes cell-matrix adhesion. J Mol Biol 398, 730-46 (2010).

[8] Sander, E.E., van Delft, S., ten Klooster, J. P., Reid, T., van der Kammen, R. A., Michiels, F. & Collard, J. G.. Matrix-dependent Tiam1/Rac signaling in epithelial cells promotes either cell-cell adhesion or cell migration and is regulated by phosphatidylinosi- tol 3-kinase. . J. Cell Biol. 143, 1385–1398. (1998).

[9] Songyang, Z., Fanning, A. S., Fu, C., Xu, J., Marfatia, S. M., Chishti, A. H., Crompton, A., Chan, A. C., Anderson, J. M., and Cantley, L. C. (1997) Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. Science 275, 73–77.

[10] Tonikian, R., Zhang, Y., Sazinsky, S. L., Currell, B., Yeh, J. H., Reva, B., Held, H. A., Appleton, B. A., Evangelista, M., Wu, Y., Xin, X., Chan, A. C., Seshagiri, S., Lasky, L. A., Sander, C., Boone, C., Bader, G. D., and Sidhu, S. S. (2008) A specificity map for the PDZ domain family. PLoS Biol. 6, e239.

[11] Tyson R.Shepherd, al. Distinct Ligand Specificity of the Tiam1 and Tiam2 PDZ Domains. Biochemistry 2011, 50, 1296-1308. DOI: 10.1021/bi1013613.

[12] T., Simonson, T., Gaillard, D., Mignon, M., Schmidt am Busch, A., Lopes, N., Amara, S., Polydorides, A., Sedano, K., Druart, G., Archontis. Computational Protein Design: The Proteus Software and Selected Applications. Journal of Computational Chemistry 2013, 34, 2472–2484

[13] Dardel, F., Képès, F.. Bioinformatique Génomique et post-génomique. ISBN: 978-2-7302-0927-4

[14] Computational design of MHC class II binding epitopes

PDB code	Ligand
1IHJ	TEFCA
1KWA	SYREF
1MFG	LDVPV
1MFL	LDVPV
1N7F	RTYSC
1OBY	NEFYA
1V1T	NEYKV
1W9E	NEFYF
1W9O	NEYYV
1W9Q	NEFAF
1YBO	NEFYA
2EJY	KEYCI
2JOA	RIWWV
2PKU	ESVKI
2PZD	TMFWV
3GCN	--YQF
3GCO	NVYQF
3GDS	NVYYF
3GDU	--YRF
3GDV	--YQF
3HPK	ESVKI
3HPM	ESVKI
3KZE	KEYYA
3OTP	GILQI
3R0H	GQYWV
4GVC	EEFYA
4GVD	EEFYA
4Q2N	WFLDI
4RQZ	LVYQF
4UU5	MERLI

table1. The ligand of class II PDZ (30 in total) collecting from PDB. The first column represents the PDB code and the second column represents the ligand sequences at the last 5 positions.

Group1	Ala, Cys, Ser, Thr	Small
Group2	Glu, Asp, Asn, Gln	Acidic polar
Group3	Hip, Hid, Hie	Basic aromatic
Group4	Ile, Val, Leu, Met	Hydrophobic
Group5	Gly, Pro	No mutable in our system
Group6	Arg, Lys	Basic
Group7	Trp, Phe, Tyr	Aromatic

table2. Amino acid groups and their properties. We classified the residues into 7 groups according to the matrix BLOSUM62.

(kcal/mol)	isolated Syndecan1 peptide								Tiam1 protein		Tiam1+Cask	
eps4	$\{E_X^i\}_{MHC}$	E_X^{-4}	E_X^{-3}	E_X^{-2}	E_X^{-1}	E_X^{0}	E_X^{mean}	E_{ii}^{pept}	E_{ii}^{prot}	E_X^{opt}	E_X^{buried}	E_X^{exp}
ALA	0,0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ARG	-46,9	-47,32	-48,11	-48,67	-48,59	-49,44	-48,42	-49,24	-62,11	-30,49	-48,82	-51,57
ASN	-15,8	-16,06	-16,94	-16,17	-16,27	-16,27	-16,34	-17,00	-23,89	3,64	-16,40	-17,52
ASP	-18,2	-18,55	-19,14	-17,47	-17,81	-15,68	-17,73	-18,98	-26,18	-3,45	-17,79	-18,91
CYS	-0,4	0,10	-0,74	-0,38	-0,54	-1,43	-0,60	-0,74	-0,74	-5,36	-0,60	-0,60
GLN	-12,6	-12,46	-14,09	-12,70	-12,79	-12,94	-13,00	-13,83	-21,46	3,67	-13,06	-14,17
GLU	-15,7	-15,15	-16,32	-14,35	-14,69	-14,95	-15,09	-16,40	-25,35	0,40	-15,15	-16,27
HID	12,7	12,04	11,96	12,27	12,12	12,30	12,14	11,49	6,83	11,33	12,85	10,45
HIE	11,6	10,91	10,91	11,28	11,30	12,72	11,42	10,72	6,28	10,78	12,14	9,74
HIP	0.0	15,23	15,45	14,48	14,08	15,98	15,04	14,86	9,32	13,83	15,76	13,36
ILE	2,9	1,84	1,29	2,24	2,20	0,80	1,67	1,87	4,13	4,79	3,70	2,64
LEU	-3,9	-4,16	-4,78	-4,72	-4,67	-4,67	-4,60	-4,59	-3,13	4,30	-2,57	-3,63
LYS	-5,5	-6,00	-6,71	-7,30	-7,07	-8,32	-7,08	-6,83	-14,01	7,33	-7,48	-10,23
MET	-1,9	-2,22	-2,87	-2,41	-2,42	-2,82	-2,55	-2,72	-2,16	-10,47	-0,52	-1,58
PHE	-1,1	-1,35	-2,17	-0,87	-1,24	-0,45	-1,22	-1,97	-1,99	1,09	0,34	-0,53
SER	-4,4	-4,41	-4,66	-4,28	-4,48	-7,40	-5,05	-5,70	-8,69	10,88	-5,05	-5,05
THR	-7,1	-6,82	-7,90	-7,13	-7,42	-11,83	-8,22	-8,28	-9,97	7,47	-8,22	-8,22
TRP	-0,7	-2,18	-1,95	-1,21	-1,59	-1,60	-1,71	-2,15	-3,06	-18,74	-0,15	-1,02
TYR	-6,8	-7,13	-8,40	-6,72	-7,24	-7,37	-7,37	-8,48	-7,83	-20,19	-5,82	-6,69
VAL	-3,1	-3,79	-4,49	-3,54	-3,94	-6,87	-4,52	-4,44	-2,55	-0,59	-2,50	-3,56
eps8	$\{E_X^i\}_{MHC}$	E_X^{-4}	E_X^{-3}	E_X^{-2}	E_X^{-1}	E_X^{0}	E_X^{mean}	E_{ii}^{pept}	E_{ii}^{prot}	E_X^{opt}	E_X^{buried}	E_X^{exp}
ALA	0,0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ARG	-23,3	-24,24	-25,58	-25,53	-24,98	-26,18	-25,30	-25,31	-37,22	-34,31	-26,71	-28,41
ASN	-5,5	-5,83	-6,34	-5,77	-5,67	-4,58	-5,64	-6,23	-12,28	-9,59	-6,60	-6,56
ASP	-9,4	-9,95	-10,60	-8,59	-8,85	-7,83	-9,17	-9,78	-16,78	-14,80	-10,13	-10,09
CYS	-0,7	-0,52	-1,42	-0,91	-0,90	-1,46	-1,04	-0,93	-1,15	-2,16	-1,04	-1,04
GLN	-3,6	-3,86	-5,79	-4,09	-4,07	-4,28	-4,42	-4,40	-11,58	-8,85	-5,38	-5,34
GLU	-7,7	-7,88	-9,57	-7,13	-7,06	-7,76	-7,88	-8,29	-16,97	-14,38	-8,84	-8,81
HID	14,0	12,55	12,91	13,07	13,13	14,12	13,16	12,87	8,02	7,33	12,70	11,41
HIE	13,3	11,71	12,28	12,64	12,56	13,90	12,62	12,39	7,71	7,02	12,16	10,88
HIP	0.0	15,47	16,22	15,13	15,17	16,65	15,72	15,92	10,08	9,39	15,27	13,98
ILE	3,2	1,73	0,94	2,07	2,16	1,24	1,63	2,16	3,76	3,44	3,12	2,65
LEU	-1,0	-1,52	-2,55	-1,88	-1,86	-1,79	-1,92	-1,80	-0,52	0,15	-0,43	-0,89
LYS	-2,2	-3,22	-4,36	-4,27	-3,89	-5,29	-4,21	-3,60	-11,23	-8,85	-5,61	-7,31
MET	-1,5	-2,17	-2,99	-2,30	-2,12	-2,61	-2,44	-2,29	-2,26	-3,57	-0,94	-1,41
PHE	-1,0	-1,38	-2,40	-1,37	-1,41	-0,55	-1,42	-1,80	-2,23	-3,05	-0,56	-0,77
SER	-2,5	-2,73	-2,87	-2,37	-2,31	-4,00	-2,85	-3,17	-6,26	-4,20	-2,85	-2,85
THR	-3,0	-3,07	-3,89	-3,19	-3,26	-5,72	-3,82	-3,55	-5,59	-4,27	-3,82	-3,82
TRP	-1,8	-3,52	-2,99	-2,21	-3,43	-1,16	-2,66	-2,70	-4,23	-8,33	-1,80	-2,01
TYR	-3,8	-4,19	-5,79	-4,24	-4,74	-3,87	-4,56	-5,28	-4,85	-8,47	-3,71	-3,91
VAL	-1,0	-1,95	-2,64	-1,65	-1,78	-3,25	-2,25	-1,94	-0,62	-0,74	-0,76	-1,23

table3. Reference energy table. In bleu bloc (Isolated peptide model): $\{E_X^i\}_{MHC}$ are the reference energies obtained from a peptide of MHC[14]; E_X^{-4} - E_X^{0}, E_X^{mean} are the point mutant energies at last 5 positions of peptide, respective, and their averaged values; E_{ii}^{pept}, the lowest self-interaction energy, averaged over the peptide. In red bloc (empirical model - method1): E_{ii}^{prot}, the lowest self-interaction energy averaged over the protein; E_X^{opt}, the optimized reference energies. In green bloc (empirical model - method2): E_X^{buried}, the optimized reference energies of buried residues; E_X^{exp}, the optimized reference energies of exposed residues.

fexp	Buried residues						Exposed residues						All residues					
	Tiam1		Cask		Mean		Tiam1		Cask		Mean		Tiam1		Cask		Mean	
ALA	0,071		0,047		0,059		0,073		0,019		0,046		0,065		0,032		0,048	
CYS	0,000	0,151	0,030	0,166	0,015	0,158	0,023	0,320	0,001	0,152	0,012	0,236	0,016	0,232	0,013	0,146	0,015	0,189
SER	0,049		0,044		0,047		0,152		0,053		0,102		0,100		0,045		0,073	
THR	0,031		0,044		0,038		0,073		0,079		0,076		0,050		0,056		0,053	
GLU	0,060		0,063		0,061		0,114		0,097		0,105		0,093		0,073		0,083	
ASP	0,031	0,121	0,040	0,124	0,035	0,123	0,084	0,309	0,040	0,346	0,062	0,328	0,056	0,224	0,036	0,218	0,046	0,222
ASN	0,030		0,007		0,019		0,060		0,087		0,074		0,043		0,045		0,044	
GLN	0,001		0,014		0,008		0,051		0,123		0,087		0,032		0,065		0,048	
HIS	0,002	0,002	0,013	0,013	0,007	0,007	0,014	0,014	0,080	0,080	0,047	0,047	0,008	0,008	0,044	0,044	0,026	0,026
ILE	0,117		0,197		0,157		0,048		0,035		0,041		0,067		0,098		0,082	
VAL	0,132	0,520	0,139	0,572	0,135	0,546	0,038	0,144	0,072	0,175	0,055	0,159	0,067	0,261	0,092	0,319	0,079	0,290
LEU	0,256		0,151		0,204		0,056		0,040		0,048		0,121		0,081		0,101	
MET	0,016		0,084		0,050		0,001		0,028		0,014		0,006		0,048		0,027	
ARG	0,029	0,088	0,006	0,078	0,018	0,083	0,060	0,170	0,130	0,222	0,095	0,196	0,043	0,123	0,065	0,139	0,054	0,131
LYS	0,059		0,072		0,065		0,111		0,092		0,101		0,081		0,074		0,077	
TRP	0,000		0,000		0,000		0,000		0,000		0,000		0,000		0,000		0,000	
PHE	0,060	0,118	0,040	0,040	0,050	0,079	0,002	0,018	0,006	0,008	0,004	0,013	0,022	0,052	0,019	0,020	0,021	0,036
TYR	0,058		0,001		0,029		0,015		0,002		0,009		0,029		0,001		0,015	
GLY	0,000	0,001	0,001	0,007	0,000	0,004	0,019	0,025	0,015	0,018	0,017	0,021	0,070	0,099	0,088	0,114	0,079	0,107
PRO	0,001		0,006		0,003		0,006		0,003		0,004		0,029		0,027		0,028	

table4. Experimental amino acid composition table. The experimental sequences contain 50 Tiam1 homology sequences and 126 hCask homology sequences with identify percentage between 60% and 85%. We separated the buried and exposed residue positions using a buried factor threshold 0.2. In Tiam1, there are 33 buried positions and 51 exposed positions. In Cask, there are 34 buried positions and 40 exposed positions.

	eps8				eps4			
	buried		exposed		buried		exposed	
ALA	0,037		0,057		0,05		0,06	
CYS	0,028	0,164	0	0,2	0,027	0,171	0	0,205
SER	0,056		0,082		0,054		0,088	
THR	0,044		0,06		0,04		0,057	
GLU	0,066		0,165		0,066		0,167	
ASP	0,041	0,137	0,063	0,364	0,041	0,139	0,071	0,372
ASN	0,015		0,078		0,017		0,076	
GLN	0,015		0,058		0,015		0,057	
HIP	0,004		0,043		0,003		0,036	
HID	0,002	0,008	0,008	0,058	0,002	0,007	0,009	0,052
HIE	0,001		0,007		0,002		0,008	
ILE	0,159		0,028		0,156		0,027	
VAL	0,099	0,515	0,038	0,126	0,095	0,511	0,034	0,121
LEU	0,155		0,041		0,16		0,042	
MET	0,102		0,02		0,1		0,017	
ARG	0,054	0,104	0,186	0,241	0,048	0,097	0,177	0,237
LYS	0,049		0,055		0,049		0,06	
TRP	0,001		0		0		0	
PHE	0,05	0,072	0	0,011	0,048	0,074	0	0,013
TYR	0,022		0,011		0,026		0,013	
GLY	0	0	0	0	0	0	0	0
PRO	0		0		0		0	

table5. Calculated amino acid composition table. The calculated sequences are from the last iteration for optimization of empirical reference energies.

	MC energies (simulation1)						MC energies (simulation2)					
	Position -2			Position 0			Position -2			Position 0		
	Averaged	Min	Max	Averaged	Min	Max	Averaged	Min	Max	Averaged	Min	Max
A	-26,482	-32,996	-24,157	-25,669	-35,508	-23,547	-34,646	-35,270	-32,862	-28,925	-55,374	-27,142
C	-24,832	-30,825	-22,544	-	-	-	-31,507	-34,788	-30,877	-	-	-
D	-27,387	-32,548	-24,503	-	-	-	-	-	-	-	-	-
E	-23,999	-29,704	-21,768	-	-	-	-31,762	-32,042	-31,678	-	-	-
F	-21,895	-32,214	-19,752	-	-	-	-28,920	-34,675	-27,142	-	-	-
H	-22,924	-31,404	-20,673	-	-	-	-30,352	-55,270	-28,736	-	-	-
K	-23,821	-43,049	-21,992	-	-	-	-31,722	-32,065	-31,532	-	-	-
L	-27,853	-28,920	-27,300	-	-	-	-31,769	-32,137	-31,448	-	-	-
M	-23,247	-32,900	-21,126	-	-	-	-29,633	-33,569	-27,972	-	-	-
N	-24,488	-32,839	-22,028	-	-	-	-31,784	-35,709	-31,151	-	-	-
Q	-24,625	-30,879	-22,344	-	-	-	-31,999	-32,124	-31,833	-	-	-
R	-21,700	-32,283	-19,592	-	-	-	-30,386	-33,421	-28,968	-	-	-
S	-25,666	-31,248	-23,017	-28,473	-38,976	-25,898	-32,149	-37,082	-31,705	-30,244	-32,260	-28,348
T	-25,409	-29,890	-23,509	-58,745	-61,124	-56,354	-34,706	-34,706	-34,706	-	-	-
h	-22,067	-32,717	-19,805	-	-	-	-30,081	-33,883	-28,379	-	-	-
j	-22,313	-32,275	-19,989	-	-	-	-30,300	-35,833	-28,670	-	-	-

table6. proteus energies (folding free energies) from MC by single point mutant explorations of peptide sequences (at position -2 and 0) with Tiam1:Syndecan1 complex. In proteus, the sequences are generated by the highest energies. The system is built using the minimized structure from Xray (4GVD_A_D). The complete sequence of peptide is "TKQEEFYA". The MC simulations used frozen protein and eps8. In simulation1, we used the reference energies obtained from extended peptide model. In simulation2, we used the reference energies obtained from empirical model with multiple proteins (Tiam1+Cask).

	MC energies (simulation1)						MC energies (simulation2)					
	Position 0			Position -2			Position 0			Position -2		
	Averaged	Min	Max	Averaged	Min	Max	Averaged	Min	Max	Averaged	Min	Max
A	-407,4	-408,5	-403,1	-362,7	-363,7	-358,8	-434,3	-435,4	-429,9	-436,3	-437,3	-432,4
C	-406,8	-407,6	-403,7	-360,7	-361,5	-356,7	-433,2	-434,1	-430,1	-434,5	-435,3	-430,4
D	-416,9	-417,7	-413,6	-368,1	-369,9	-363,6	-446,0	-446,9	-442,7	-443,2	-445,0	-438,8
E	-419,6	-420,6	-415,9	-362,0	-362,9	-357,6	-447,6	-448,6	-443,8	-437,3	-438,2	-432,9
F	-452,0	-452,9	-448,1	-361,5	-362,6	-357,1	-478,9	-479,8	-474,9	-434,3	-435,4	-429,9
H	-416,7	-417,6	-413,6	-364,2	-365,1	-360,5	-445,0	-445,8	-441,9	-437,6	-438,5	-433,9
h	-413,6	-414,5	-410,2	-363,2	-364,1	-359,1	-441,9	-442,7	-438,5	-437,2	-438,0	-433,1
j	-419,7	-420,5	-416,9	-364,0	-364,8	-361,1	-448,3	-449,1	-445,4	-438,1	-438,9	-435,2
I	-415,8	-416,7	-412,5	-368,8	-369,7	-365,3	-440,8	-441,6	-437,5	-441,4	-442,3	-437,9
K	-423,1	-424,0	-420,0	-361,8	-362,6	-358,6	-450,3	-451,2	-447,1	-436,7	-437,6	-433,6
L	-423,6	-424,5	-420,5	-364,7	-365,6	-361,6	-449,1	-450,0	-446,0	-436,9	-437,7	-433,8
M	-416,4	-417,1	-412,5	-358,8	-359,8	-354,5	-441,6	-442,3	-437,7	-431,1	-432,1	-426,7
N	-414,8	-417,2	-409,8	-363,1	-364,0	-359,3	-443,7	-446,1	-438,6	-437,5	-438,4	-433,7
Q	-419,0	-419,9	-415,3	-362,7	-363,6	-359,0	-446,9	-447,8	-443,3	-437,6	-438,5	-433,8
R	-426,8	-427,7	-422,5	-360,7	-361,7	-356,9	-454,2	-455,1	-449,9	-435,5	-436,5	-431,6
S	-409,4	-410,3	-406,0	-364,0	-364,8	-360,3	-435,1	-436,0	-431,7	-438,1	-438,9	-434,4
T	-413,1	-414,0	-409,3	-363,3	-364,2	-359,4	-438,1	-439,0	-434,3	-437,5	-438,4	-433,6
V	-415,6	-416,4	-412,9	-367,8	-368,6	-363,9	-439,9	-440,8	-437,2	-440,5	-441,3	-436,6
W	-458,9	-459,8	-455,1	-361,9	-362,7	-359,0	-486,4	-487,3	-482,6	-435,1	-435,8	-432,2
Y	-486,2	-487,0	-483,3	-361,1	-362,0	-357,1	-512,9	-513,7	-510,0	-434,1	-435,1	-430,1

table7. proteus energies (folding free energies) from MC by single point mutant explorations of peptide sequences (at position 0 and -2) with Tiam1:Syndecan1 complex. In proteus, the sequences are generated by the highest energies. The system is built using the minimized structure from Xray (4GVD_A_D). The complete sequence of peptide is "TKQEEFYA". The MC simulations used flexible protein and eps8. In simulation1, we used the reference energies obtained from extended peptide model. In simulation2, we used the reference energies obtained from empirical model with multiple proteins (Tiam1+Cask).

Simulation1&2		Position 0						Position -2					
		Bonded	Elec	αVdW	βSASA	SASA	Affinity	Bonded	Elec	αVdW	βSASA	SASA	Affinity
ALA	Complex	1047,3	-507,0	-45,1	38,5	7105,1	-12,4	1046,3	-507,2	-44,5	38,8	7156,5	-11,3
	Unbound	1047,3	-505,9	-39,4	44,2	8146,8		1046,3	-505,9	-39,5	43,8	8085,4	
ARG	Complex	1056,2	-530,4	-46,1	38,4	7081,6	-13,3	1050,4	-530,5	-45,9	38,3	7057,7	-12,3
	Unbound	1056,2	-530,3	-39,6	45,0	8298,1		1050,4	-530,0	-40,0	44,2	8150,0	
ASN	Complex	1057,0	-516,1	-45,4	38,4	7085,3	-12,0	1055,0	-516,5	-45,1	38,7	7135,6	-11,5
	Unbound	1057,0	-516,3	-39,3	44,4	8191,7		1055,0	-515,7	-39,8	44,0	8119,1	
ASP	Complex	1049,7	-509,5	-45,3	38,3	7065,2	-9,1	1048,8	-512,7	-44,8	38,7	7133,2	-10,8
	Unbound	1049,7	-512,5	-39,2	44,3	8169,3		1048,8	-512,3	-39,6	43,9	8094,4	
CYS	Complex	1049,7	-507,4	-45,4	38,4	7088,4	-12,3	1047,5	-506,4	-44,8	38,8	7157,5	-11,4
	Unbound	1049,7	-506,7	-39,5	44,2	8156,6		1047,5	-505,1	-39,7	43,9	8093,0	
GLN	Complex	1058,2	-515,9	-45,3	38,6	7114,6	-13,1	1054,5	-515,8	-45,2	38,6	7123,1	-11,8
	Unbound	1058,2	-515,4	-39,0	44,8	8273,1		1054,5	-514,7	-39,9	43,9	8104,1	
GLU	Complex	1052,4	-511,3	-46,0	38,4	7080,4	-11,7	1046,8	-511,5	-45,2	38,7	7135,6	-11,4
	Unbound	1052,4	-512,0	-39,7	44,6	8219,6		1046,8	-510,8	-39,9	44,0	8116,3	
HID	Complex	1082,9	-508,7	-43,1	38,5	7096,3	-10,8	1065,7	-508,1	-45,4	38,6	7116,8	-12,0
	Unbound	1082,9	-508,7	-38,6	44,6	8236,5		1065,7	-506,9	-40,0	44,0	8113,7	
HIE	Complex	1079,7	-508,1	-41,8	38,6	7125,2	-10,7	1065,0	-508,2	-45,4	38,5	7107,0	-11,9
	Unbound	1079,7	-507,7	-37,6	44,8	8259,6		1065,0	-507,1	-40,0	43,9	8100,0	
HIP	Complex	1083,5	-510,4	-43,1	38,4	7083,7	-12,1	1069,1	-508,3	-45,3	38,6	7124,5	-11,8
	Unbound	1083,5	-508,9	-38,5	44,5	8207,9		1069,1	-507,5	-39,8	44,1	8144,7	
ILE	Complex	1056,9	-509,5	-45,3	38,5	7094,4	-13,3	1055,4	-507,8	-44,7	38,8	7160,6	-11,9
	Unbound	1056,9	-508,6	-39,1	44,6	8232,2		1055,4	-506,4	-39,5	44,0	8115,8	
LEU	Complex	1062,3	-509,9	-44,5	38,4	7078,9	-12,8	1045,7	-510,1	-45,1	38,5	7106,8	-12,3
	Unbound	1062,3	-508,8	-38,9	44,5	8206,2		1045,7	-508,3	-39,9	43,8	8078,2	
LYS	Complex	1057,9	-509,6	-45,7	38,5	7100,2	-12,5	1050,2	-509,6	-45,3	38,7	7134,7	-11,4
	Unbound	1057,9	-510,0	-39,4	45,0	8300,8		1050,2	-509,3	-39,8	44,2	8154,0	
MET	Complex	1054,4	-507,6	-44,6	38,6	7119,4	-13,2	1048,4	-507,3	-45,4	38,5	7112,0	-12,1
	Unbound	1054,4	-506,5	-38,8	44,8	8265,8		1048,4	-506,1	-39,9	44,0	8121,8	
PHE	Complex	1076,9	-507,8	-41,3	38,4	7080,1	-10,9	1047,3	-507,1	-45,1	38,5	7105,0	-12,4
	Unbound	1076,9	-507,1	-37,3	44,7	8246,8		1047,3	-505,9	-39,4	44,2	8146,2	
SER	Complex	1047,0	-510,2	-45,6	38,3	7064,5	-12,1	1047,4	-508,5	-44,6	38,9	7185,7	-11,2
	Unbound	1047,0	-509,6	-39,8	44,0	8111,3		1047,4	-507,3	-39,5	43,9	8108,7	
THR	Complex	1053,4	-514,0	-45,4	38,3	7066,6	-12,2	1050,3	-510,7	-44,9	38,8	7167,8	-11,2
	Unbound	1053,4	-513,3	-39,6	44,0	8121,5		1050,3	-509,9	-39,7	44,0	8119,1	
TRP	Complex	1096,4	-507,0	-39,4	38,4	7081,1	-12,4	1043,6	-506,3	-45,9	38,4	7075,7	-13,1
	Unbound	1096,4	-506,1	-34,4	45,0	8297,1		1043,6	-504,8	-40,1	44,2	8152,0	
TYR	Complex	1086,3	-509,1	-37,3	38,4	7088,9	-8,1	1048,2	-509,6	-45,2	38,5	7097,3	-12,4
	Unbound	1086,3	-509,1	-35,5	44,8	8266,4		1048,2	-508,7	-39,4	44,2	8149,7	
VAL	Complex	1055,6	-511,9	-45,1	38,6	7123,5	-13,0	1046,4	-509,6	-44,5	38,7	7133,7	-11,7
	Unbound	1055,6	-510,9	-39,1	44,7	8241,0		1046,4	-507,6	-39,8	43,7	8065,8	

table8. xplor energy terms of Tiam1:peptide(mutant) complex at the bounded and unbounded states. The peptide mutant sequences are obtained from MC with Tiam1:Syndecan1 complex by single point mutations of peptide (at position 0 and -2), using flexible protein and eps8. The native sequence of Syndecan1 peptide is "TKQEEFYA". The structures of the 200 best conformation mutants (for each mutant type) are reconstructed by xplor, and minimized by 200 steps. The energy terms in the table are averaged over the 200 conformations of each aa. type. The simulation1 and simulation2 have the same results after rounded. The reference energies using in MC are from extended peptide model for simulation1 and from empirical model – version2 (using multiple protein Cask+Tiam1) for simulation2. The coefficient of VdW energy used α=0.11 and the coefficient for surface energy used β=0.00542.

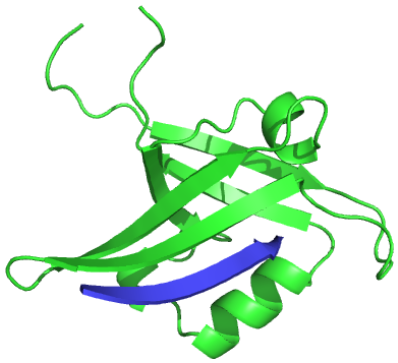


figure1. 3D structure of Tiam1PDZ (in green) binding Syndecan1 peptide (in blue).

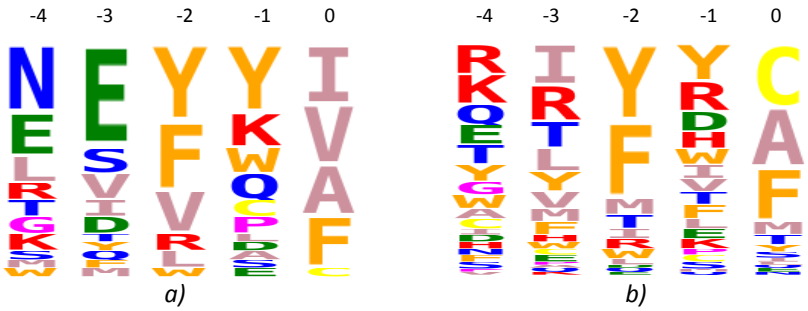


figure2. Sequence logos of the peptide ligands from experiments. a) Ligands of class II PDZ (26 in total) collecting from PDB website. b) Ligands are selected from a Tiam1-binding peptide library (69 in total)[11]. The complet sequences of these peptides are "YAAX$_{-4}$X$_{-3}$X$_{-2}$X$_{-1}$X$_{0}$", where X represents a given residue type.

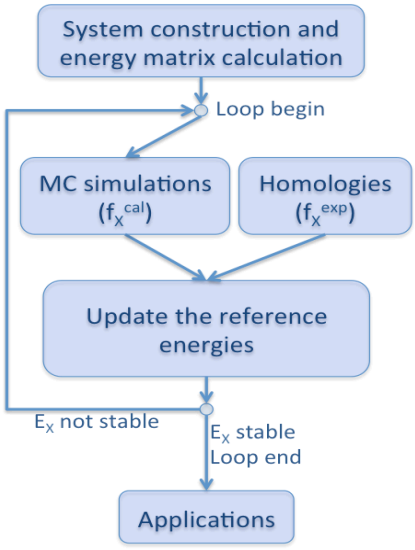


figure3. Flowchart of the general method for the reference energy optimization using empirical model.

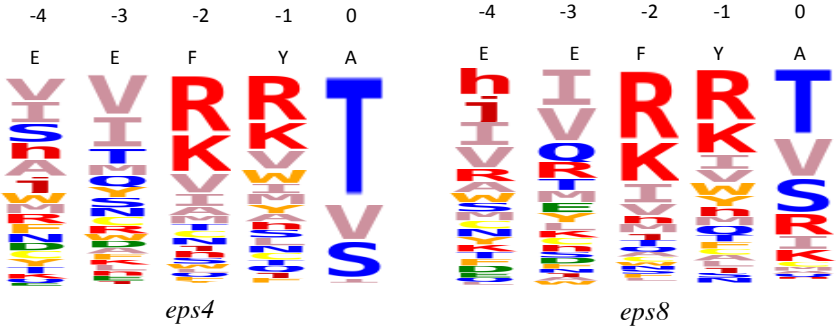


figure4. Sequence logos from MC by single point mutations at the last 5 active positions of peptide using extended peptide model. The complete sequence of peptide is "TKQEEFYA". The system is built with the extended peptide structure, which is extracted from Xray complex (4GVD_A_D) and minimized slightly.

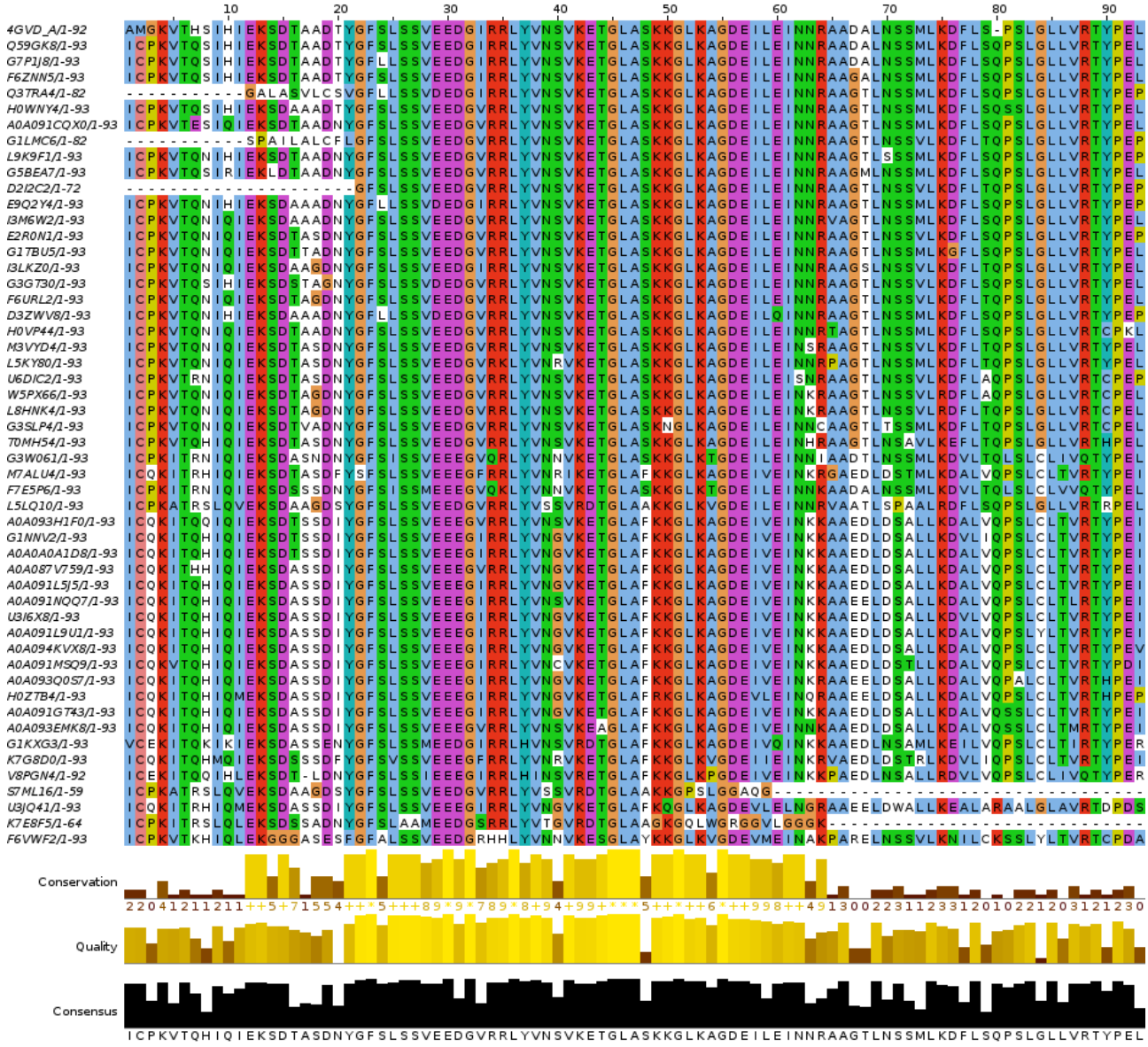


figure5. Sequence alignment with 50 homologues of PDZ Tiam1 protein (PDB: 4GVD_A). The selected homologues have a percentage of identification between 60% and 85%. The sequences are ordered by the percentage of identification, and the first sequence is the wild type sequence

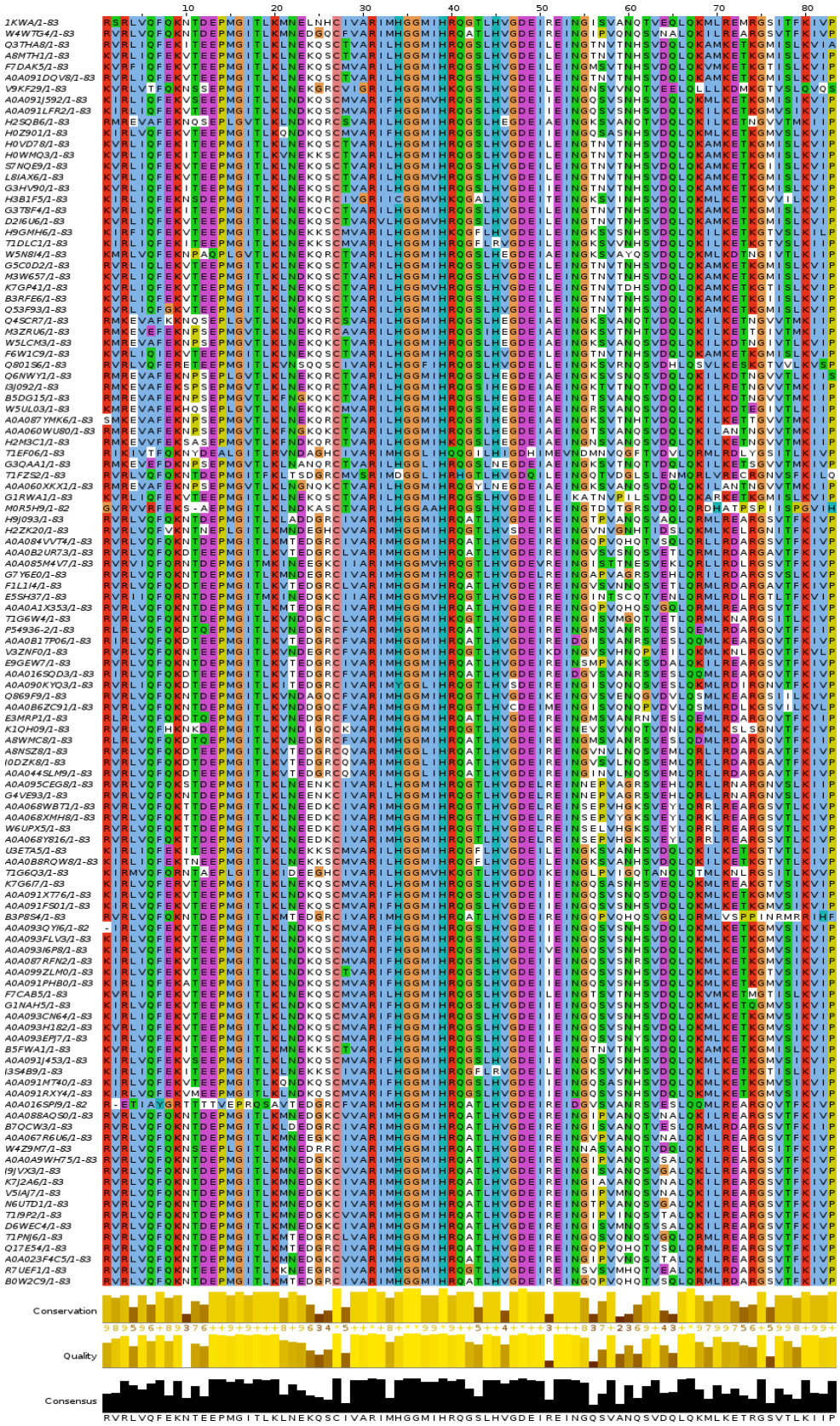


figure6. Sequence alignment with 126 homologues of PDZ hCask protein (PDB: 1KWA_A). The selected homologues have a percentage of identification between 60% and 85%. The sequences are ordered by the percentage of identification, and the first sequence is the wild type sequence.

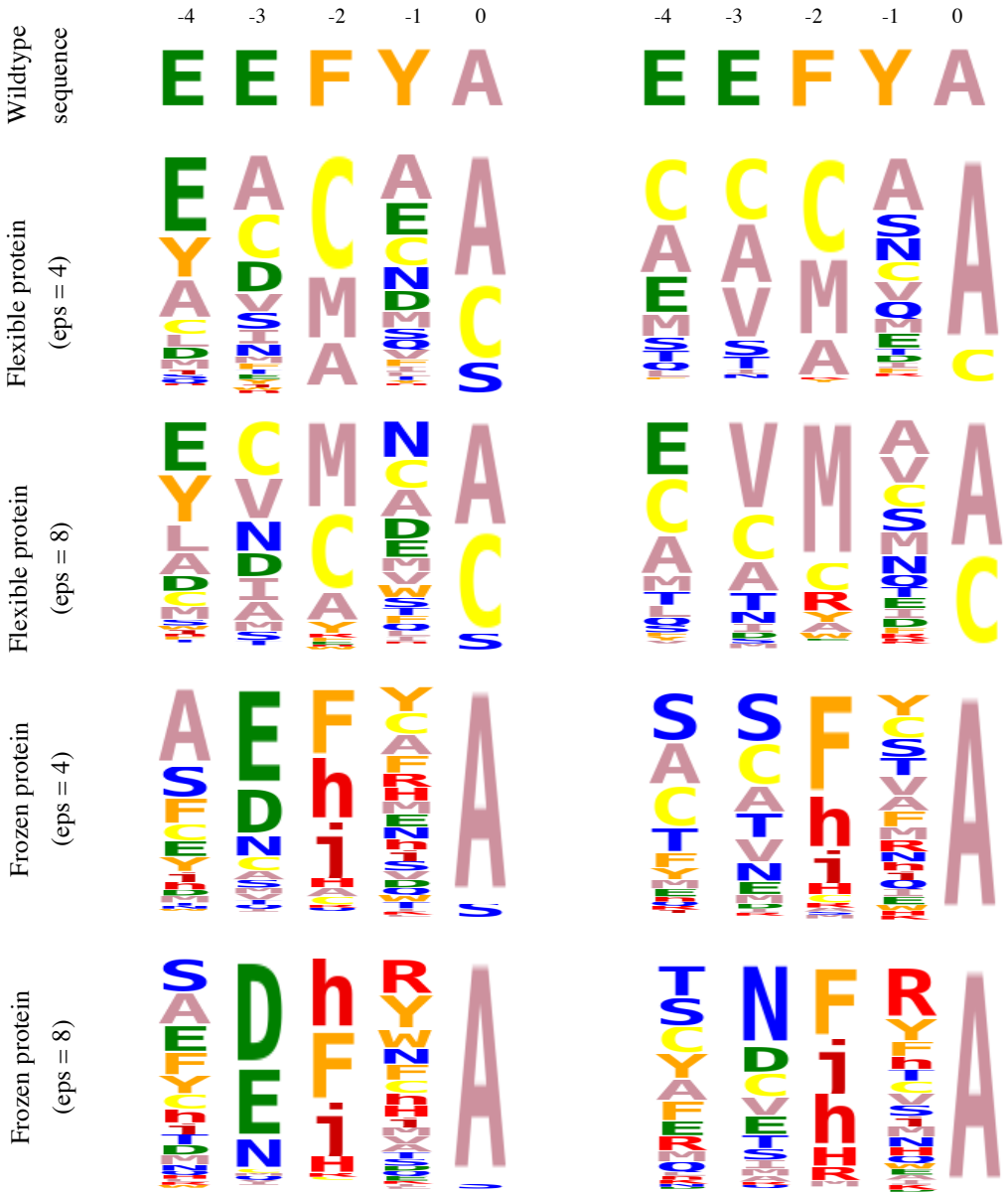


figure7. Sequence logo from MC by mutation explorations of peptide sequences with Tiam1:Syndecan1 complex. The system is built using the minimized structure from Xray (4GVD_A_D). The complete sequence of peptide is "TKQEEFYA". The reference energies are used the energies that obtained from the simulations of the syndecan peptide. For the MC simulations, we considered the protein as flexible or frozen, and used the dielectric constant equals 4 or 8. We explored mutations at the positions –4 to 0 of the peptide ligand. a): The multiple mutation results by 10 million step MC. b): The single point mutation results by 5 million step MC at each position.

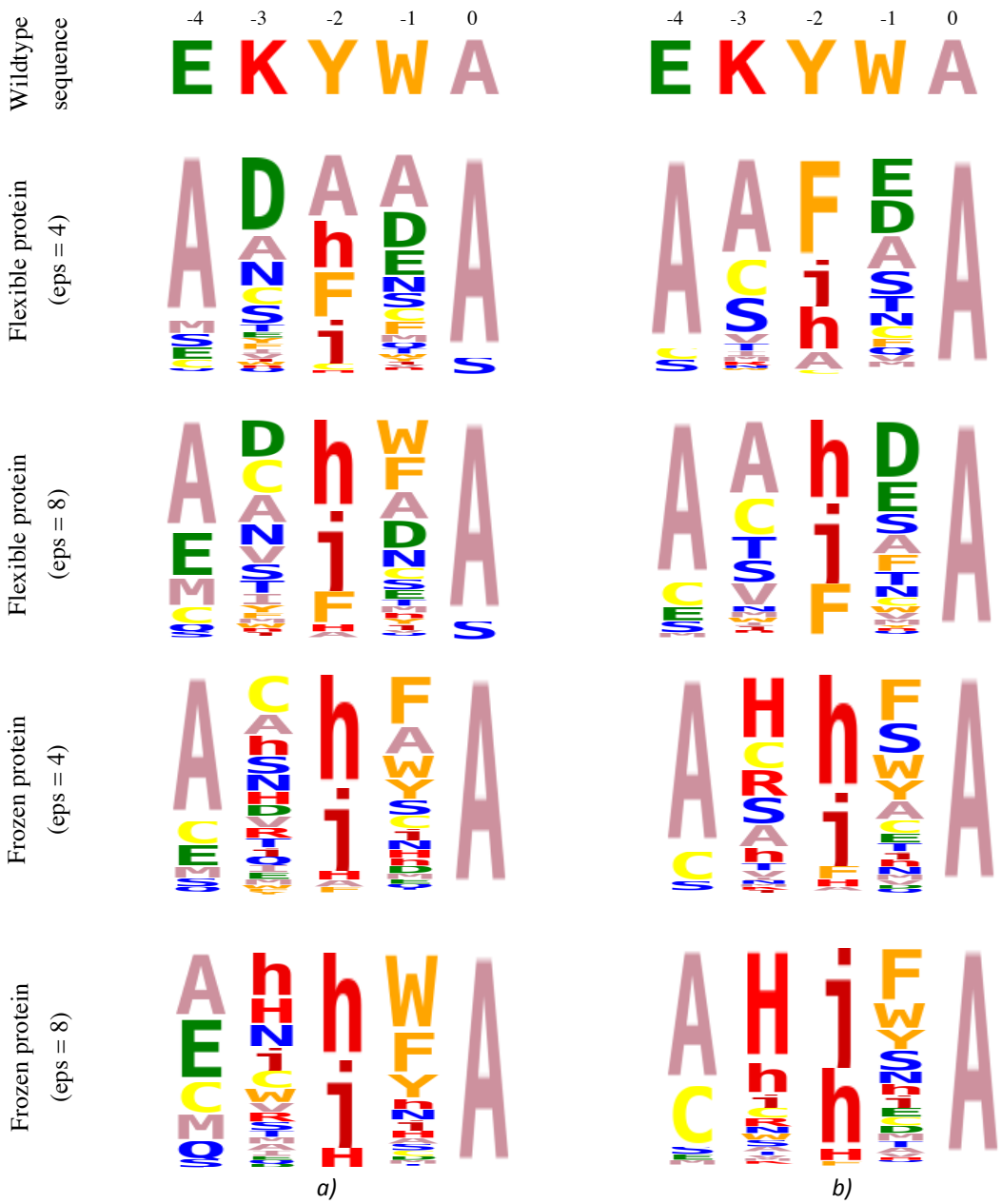


figure8. Sequence logo from MC simulation of Tiam1:peptide. The system is built using the structure obtained from 20ns MD. "Wildtype sequence"(above) is from the Tiam1-binding peptide library. The complete sequence is "YAAEKYWA"[11]. The reference energies are used the energies that obtained from the simulations of the syndecan peptide. For the MC simulations, we considered the protein as flexible or frozen, and used the dielectric constant equals 4 or 8. We explored mutations at the positions –4 to 0 of the peptide ligand. a): The multiple mutation results by 10 million step MC. b): The single point mutation results by 5 million step MC at each position.

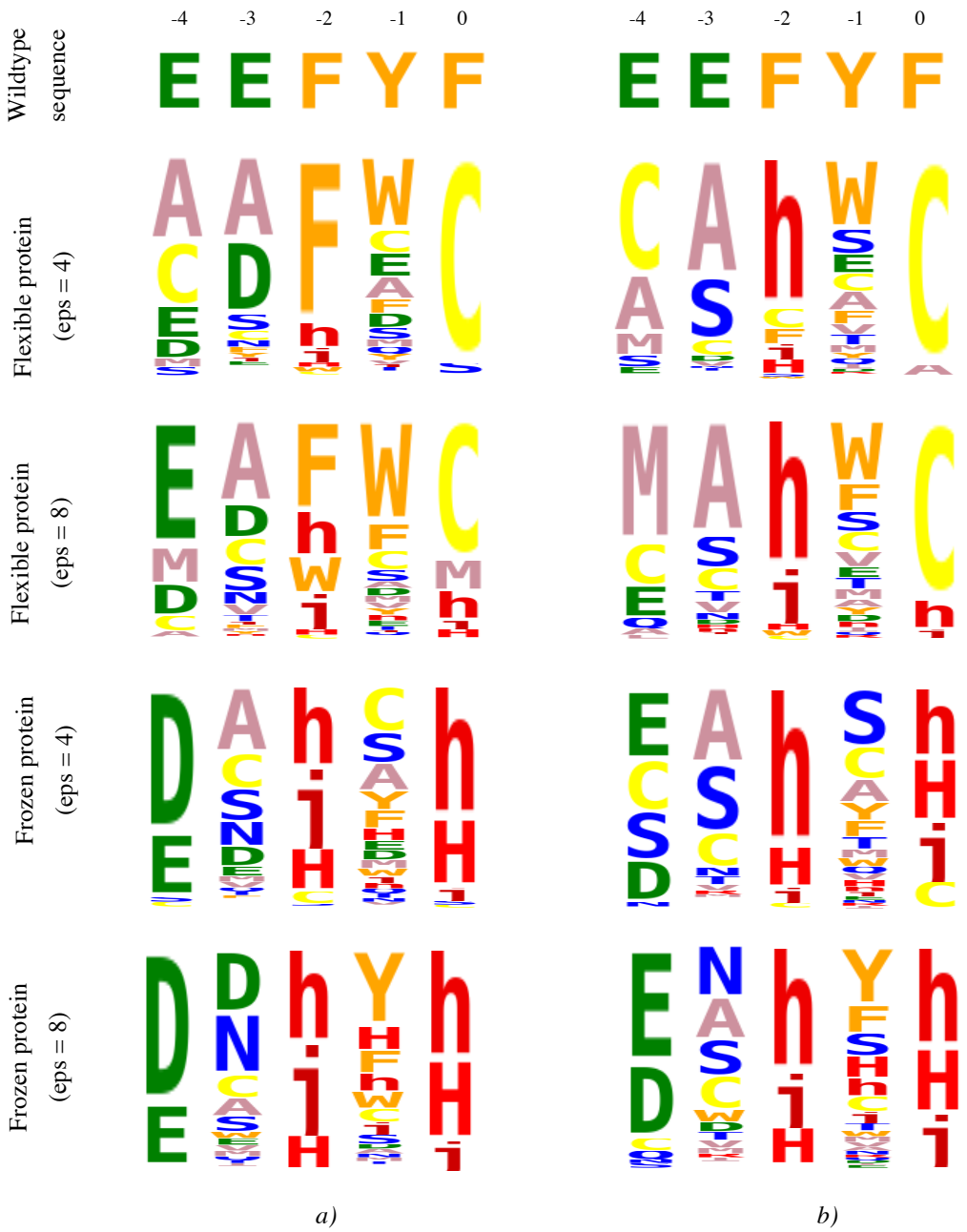
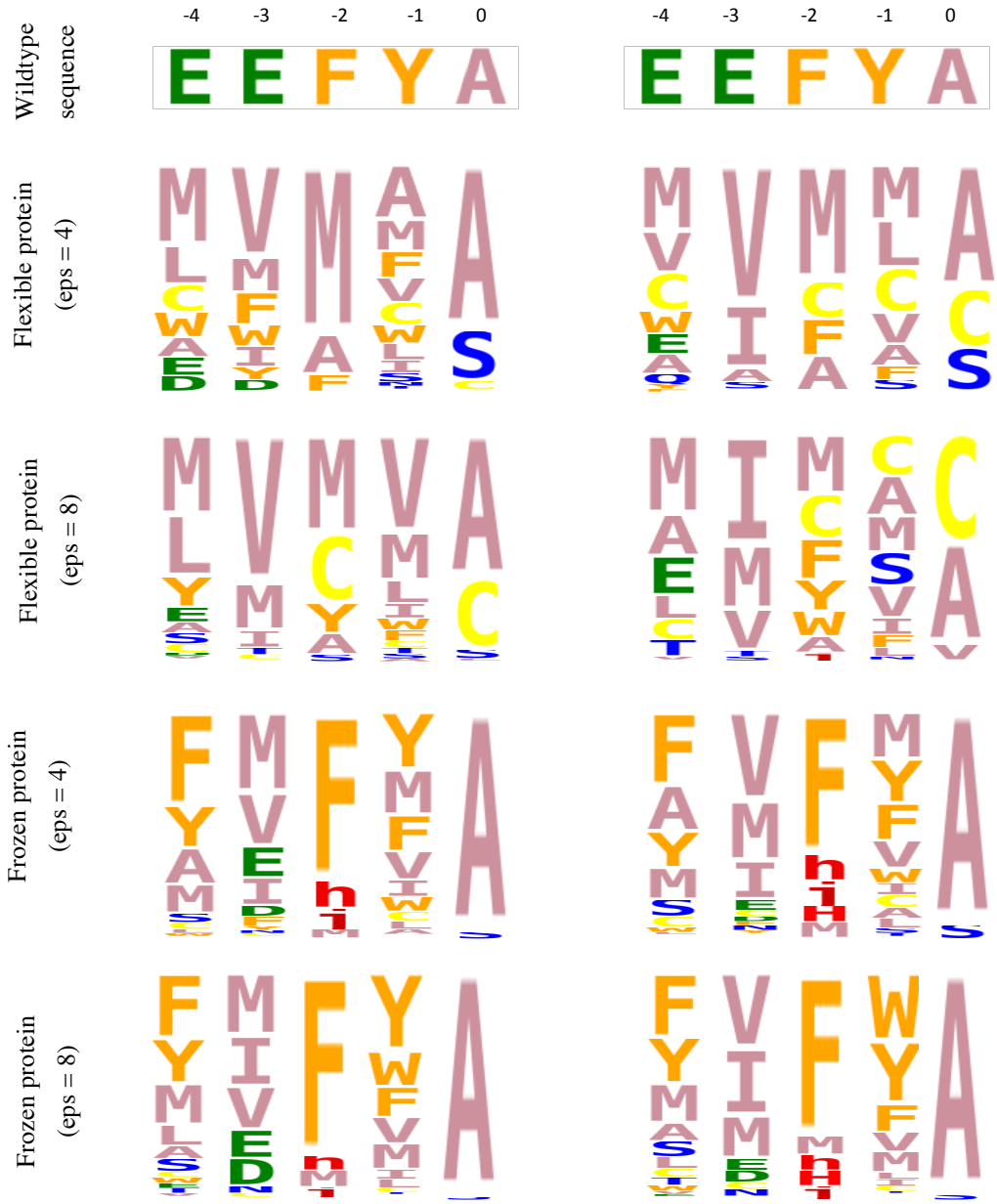
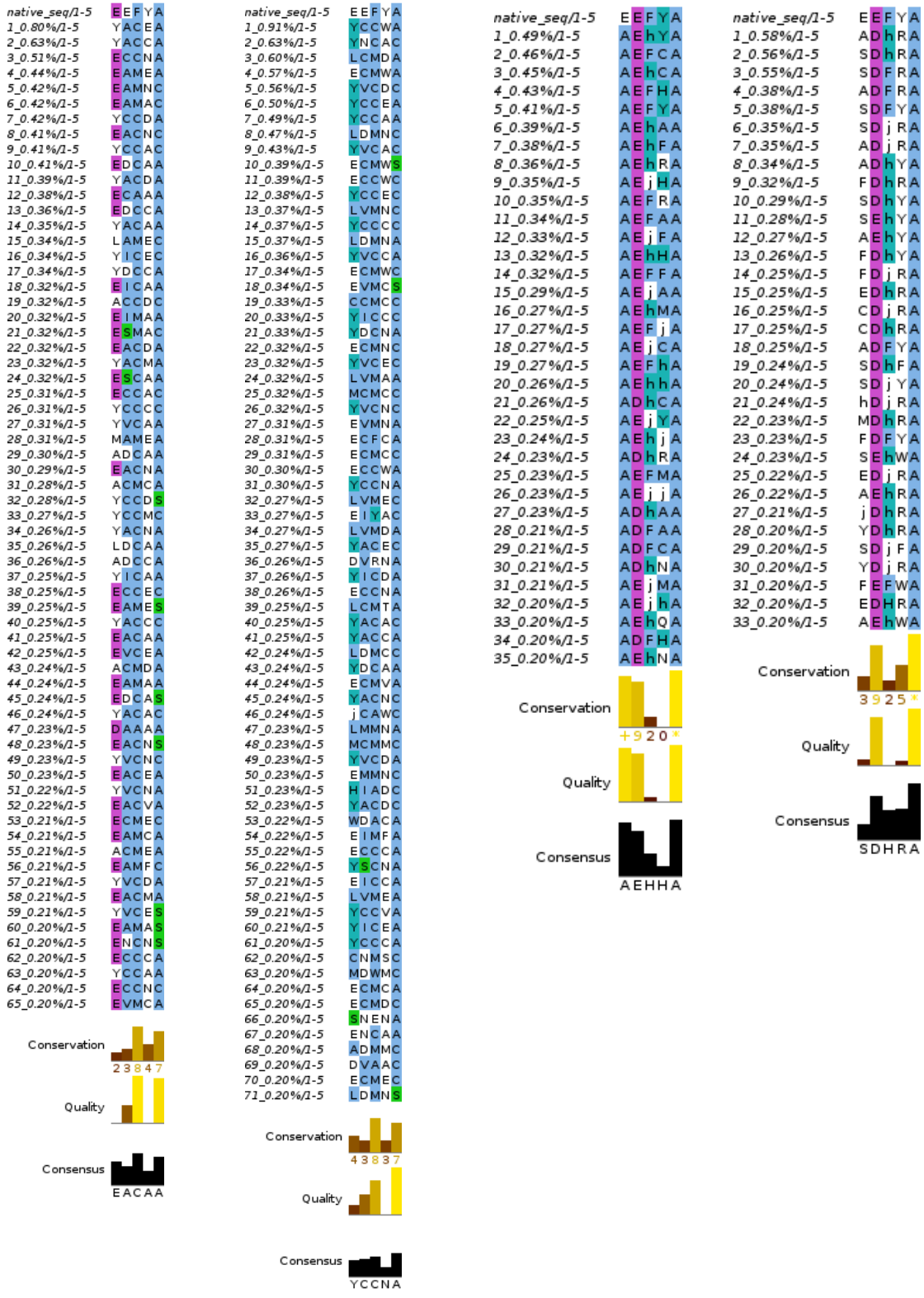


figure9. Sequence logo from MC simulation of Tiam1:peptide. The system is built using the averaged structure from 100ns MD. The "Wildtype sequence" (above) is from a mutant of Syndecan1 peptide. The complete sequence is "TKQEEFYF". The reference energies are used the energies that obtained from the simulations of the syndecan peptide. For the MC simulations, we considered the protein as flexible or frozen, and the dielectric constant equals to 8 or 4. We explored mutations at the positions –4 to 0 of the peptide ligand. a): The multiple mutation results by 10 million step MC. b): The single point mutation results by 5 million step MC at each position.



a) *b)*

figure10. Sequence logo from MC simulation of Tiam1:peptide. "wildtype" is the Syndecan1 peptide from the complex of the PDZ domain of Tiam1; the complete sequence is "TKQEEFYA". The reference energies are used the reference energies, which is optimized using multiple proteins (Cask and Tiam1) with 1/5 active positions. For the MC simulations, we considered the protein as flexible or frozen, and used the dielectric constant equals 4 or 8. We explored mutations at the positions –4 to 0 of the peptide ligand. a): The multiple mutation results by 10 million step MC. b): The single point mutation results by 5 million step MC at each position.

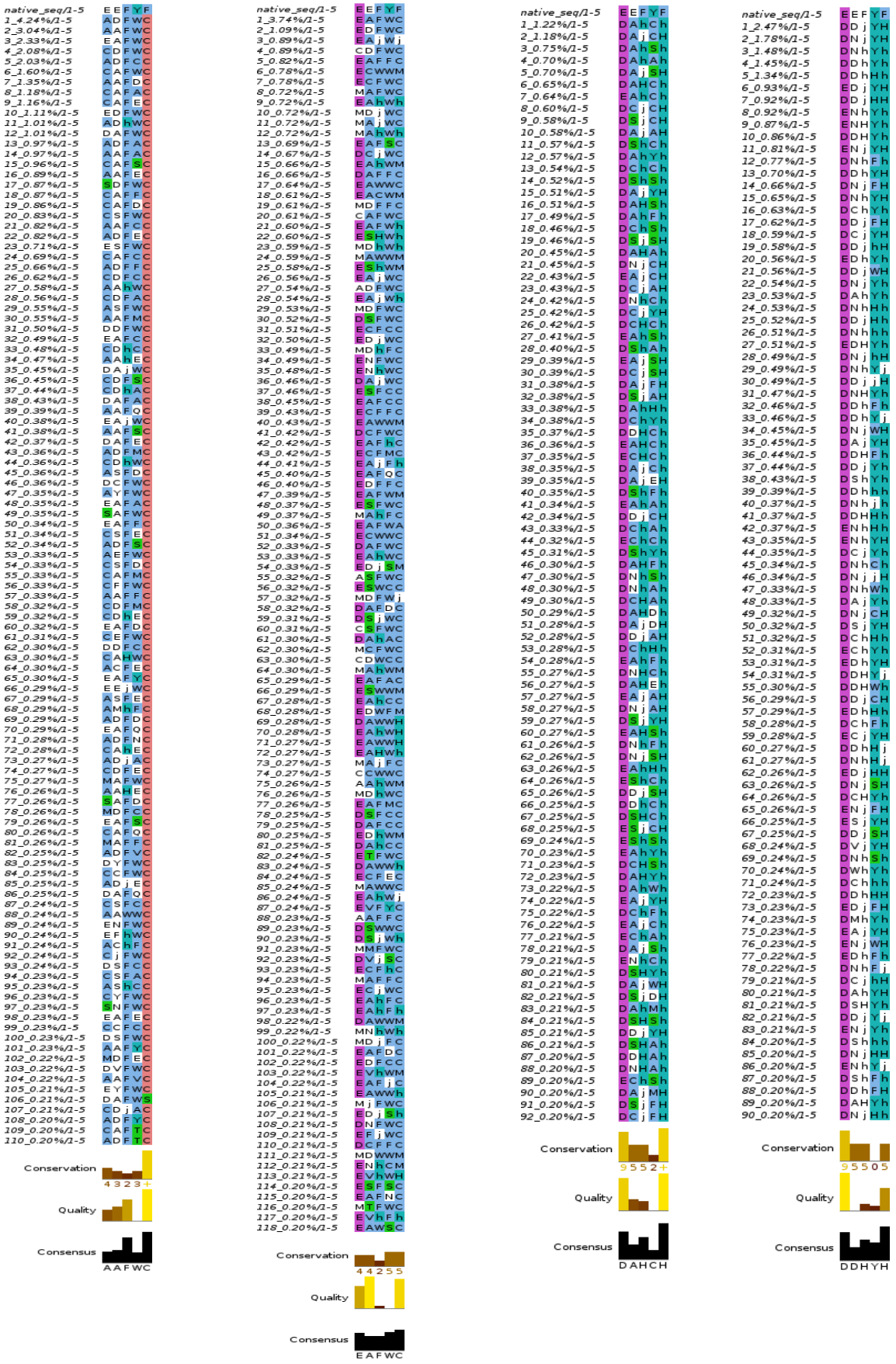


Flexible protein eps = 4 | Flexible protein eps = 8 | Frozen protein eps = 4 | Frozen protein eps = 8

figure11. Sequence alignments from MC by multiple mutation explorations of peptide sequences with Tiam1:Syndecan1 complex. The system is built using the minimized structure from Xray (4GVD_A_D). The complete sequence of peptide is "TKQEEFYA". For the MC simulations, we considered the protein as flexible or frozen, and used the dielectric constant equals 4 or 8. The reference energies are used the delta mutant energies from extended peptide. We explored mutations at the positions 0 to -4 of the peptide ligand. The occurrence threshold for selecting sequences is 0.2%. The first lines are the native sequence and the title of the other lines contains "seqid_occurences/1-5". The sequences are ordered by their occurrences.

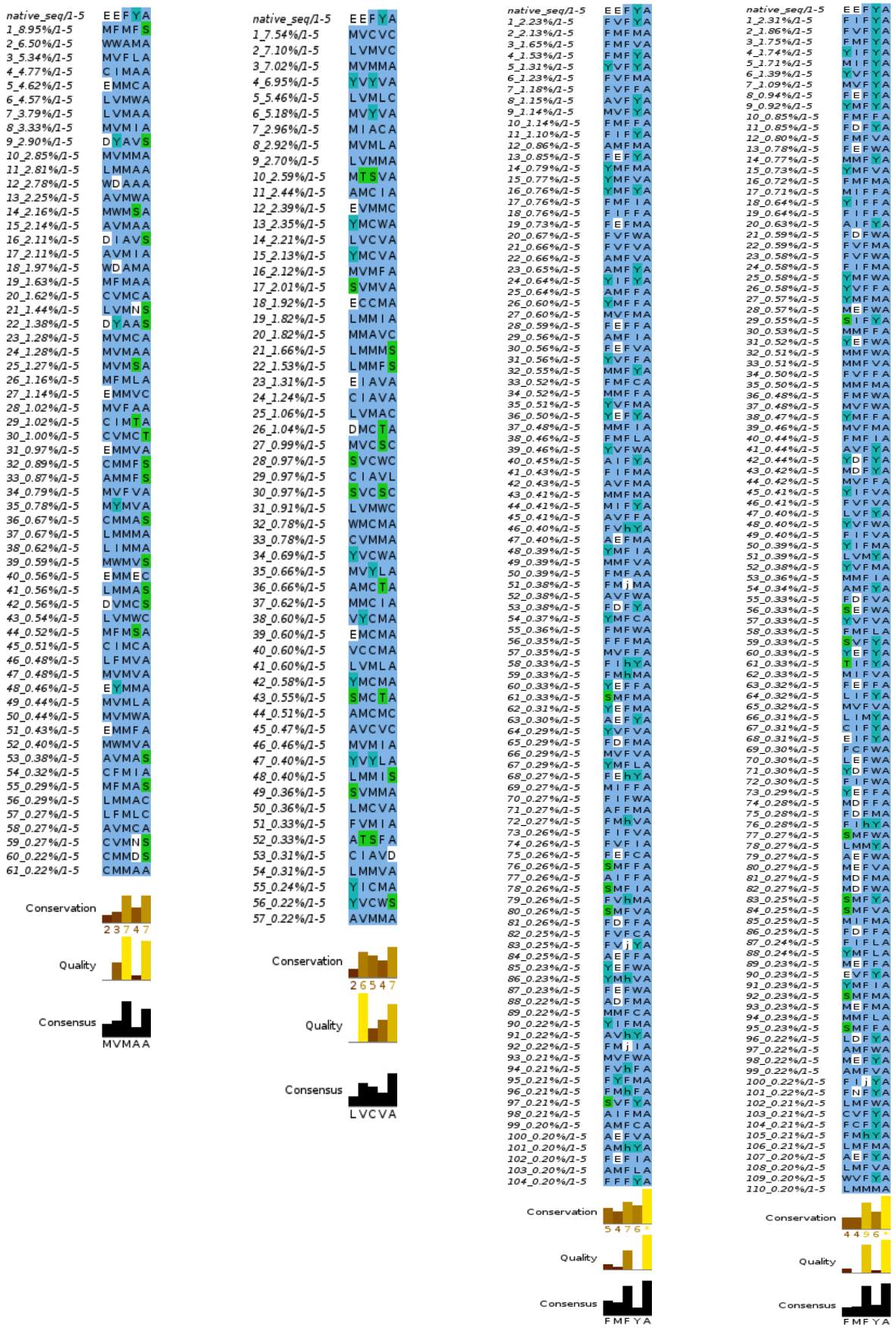
Flexible protein eps = 4 | Flexible protein eps = 8 | Frozen protein eps = 4 | Frozen protein eps = 8

figure12. Sequence alignments from MC by multiple mutation explorations of peptide sequences with Tiam1:peptide complex. The system is built using the structure obtained from 20ns MD. The peptide is selected from the Tiam1-binding peptide library with the complete sequence "YAAEKYWA"[11]. For the MC simulations, we considered the protein as flexible or frozen, and used the dielectric constant equals 4 or 8. The reference energies are used the delta mutant energies from extended peptide. We explored mutations at the positions 0 to -4 of the peptide ligand. The occurrence threshold to select sequences is 0.2%. The first lines are the native sequence and the title of the other lines contains "seqid_occurences/1-5". The sequences are ordered by their occurrences.



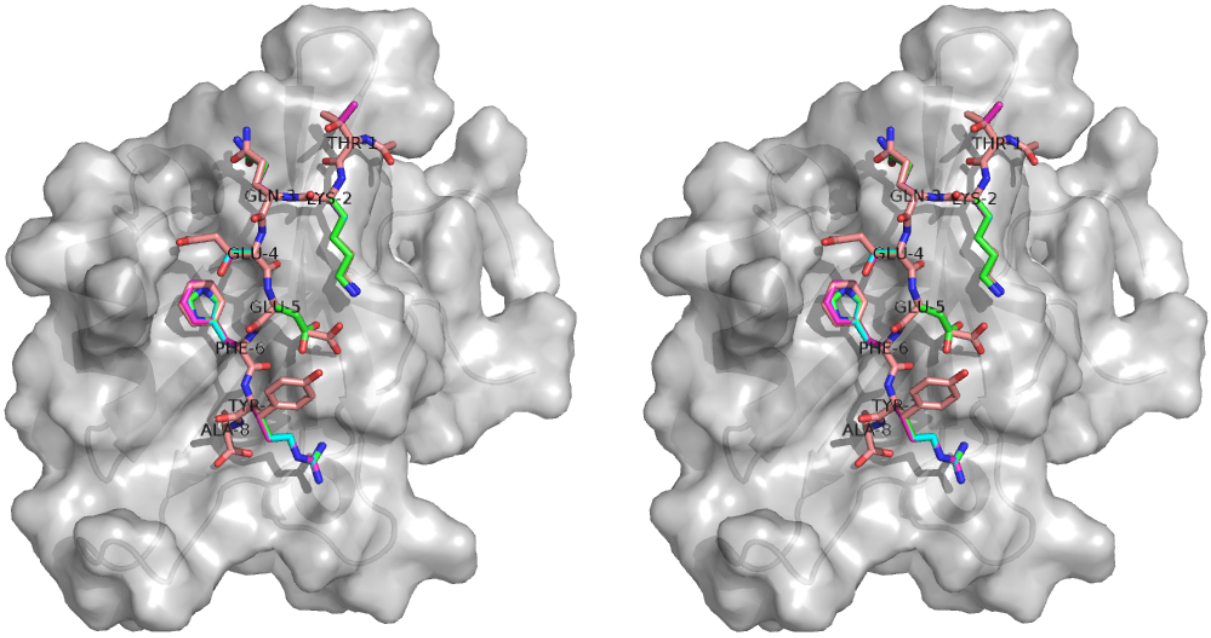
Flexible protein eps = 4 | Flexible protein eps = 8 | Frozen protein eps = 4 | Frozen protein eps = 8

figure13. Sequence alignments from MC by multiple mutation explorations of peptide sequences with Tiam1:peptide complex. The system is built using the averaged structure from 100ns MD. The peptide is from a mutant of Syndecan1 peptide with the complete sequence "TKQEEFYF". For the MC simulations, we considered the protein as flexible or frozen, and used the dielectric constant equals 4 or 8. The reference energies are used the delta mutant energies from extended peptide. We explored mutations at the positions 0 to -4 of the peptide ligand. The occurrence threshold for selecting sequences is 0.2%. The first lines are the native sequence and the title of the other lines contains "seqid_occurences/1-5". The sequences are ordered by their occurrences.

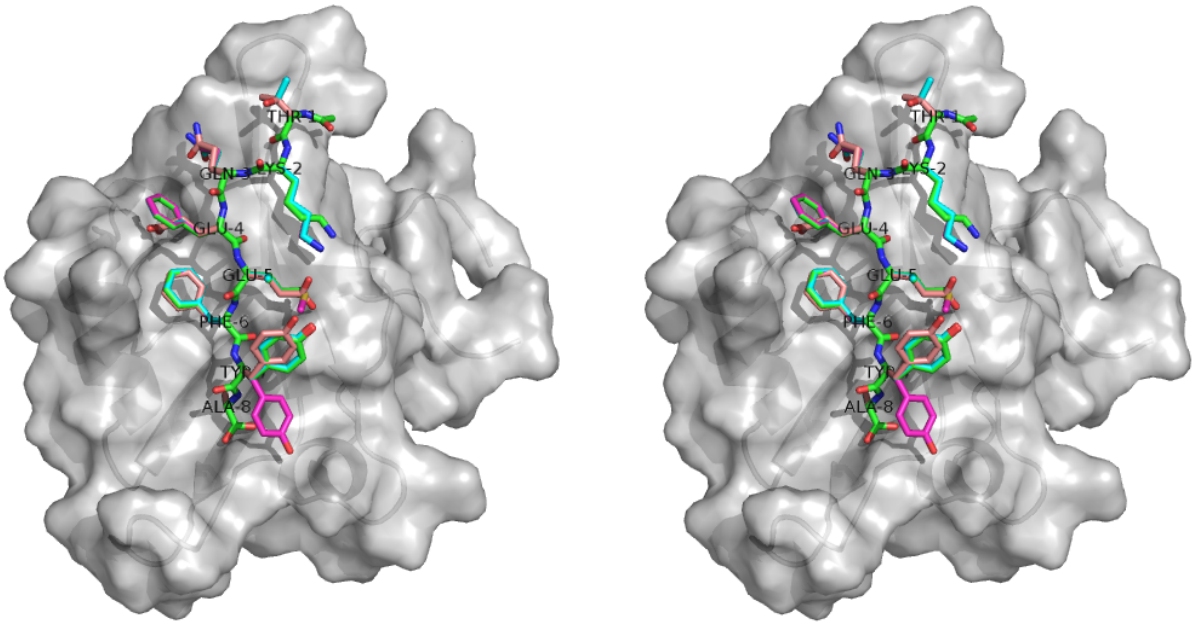


Flexible protein eps = 4 | Flexible protein eps = 8 | Frozen protein eps = 4 | Frozen protein eps = 8

figure14. . Sequence alignments from MC by multiple mutation explorations of peptide sequences with Tiam1:Syndecan1 complex. The system is built using the minimized structure from Xray (4GVD_A_D). The complete sequence of peptide is "TKQEEFYA". For the MC simulations, we considered the protein as flexible or frozen, and used ε= 4 or 8. The reference energies are optimized using multiple proteins (Tiam1+Cask) with 1/5 active positions. We explored mutations at the positions 0 to -4 of the peptide ligand. The occurrence threshold for selecting sequences is 0.2%. The first lines are the native sequence and the title of the other lines contains "seqid_occurences/1-5". The sequences are ordered by their occurrences.



a)



b)

figure15. Cross-eye stereo of the reconstructed structures and of the Xray structure (slightly minimized). The sequences of reconstructed structures are obtained from MC with Tiam1:Syndecan1 complex by multiple mutations, using frozen protein and eps8, which have top3 populations. The protein is represented in grey. The experimental structure of peptide is represented in salmon with the label. The reconstructed structures are represented in bleu, green and magenta. a) The MC simulation is used the reference energies from extended peptide model. The obtained sequences with top3 populations are: "TKQADhRA", "TKQSDhRA" and "TKQSDFRA". b) The MC simulation is used the reference energies from empirical model – version2 (using multiple protein Cask+Tiam1). The obtained sequences with top3 populations are: "TKQFIFYA", "TKQFVFYA" and "TKQFMFYA".

APPENDIX

The directories in the path /*home/xingyu/* are for the simulations with different structures:
/sdc_Proteus/ : Syndecan peptide alone (with first 3 inactive positions and last 5 active positions).
/PDZ_complexe_TKQ/ and /PDZ_complexe_TKQpf/ : Tiam1:Syndecan complex with flexible or frozen (pf) protein and the last 5 position flexible in peptide.
/PDZ_complexe_YAA/ and /PDZ_complexe_YAApf/ : Tiam1:peptide complex with flexible or frozen (pf) protein and the last 5 position flexible in peptide. The peptide is from the Tiam1-binding peptide library with sequence "YAAEKYWA"[11].
/PDZc_A0F/ and /PDZc_A0F_pf/ : Tiam1:peptide complex with flexible or frozen (pf) protein and the last 5 position flexible in peptide. The peptide is from a mutant of Syndecan1 peptide with sequence "TKQEEFYF".
/PDZ_Protein_act/ and /Cask_Pro_act/ : Tiam1 or Cask protein alone with all positions as active.
/ Proteus_2.0.1 / , Proteus package with some modified scripts:

1. The topology for N3 are added in the files *.str* and *.inp* :

```
parameters
DIHE N3   CT   C   N   MULT 4 0.00000000 4 0.0 ! four amplitudes and
                          0.55000000 3 180.0 ! phases for psi
                               1.58000000 2 180.0
                          0.45000000 1 180.0
end
```

2. A new version of analyze_proteus_sequences.pl

The reference energy optimization with Tiam1 and Cask can be achieved by scripts in cluster with the path (launch the job to 10 nodes in cluster):

```
/work/xingyu/PDZ_Protein_act/proteus/optEner_TC/
/work/xingyu/Cask_Pro_act/proteus/optEner_TC/
```

or by a version (calculate in local machine) with the path:

```
/home/xingyu/PDZ_Protein_act/proteus/optEner_v2/
```

A backup in the directory:

```
/home/xingyu/testcase/optiener_Tiam1_Cask/
```

We have scanned the 20 point mutants at position 0 and -2 of peptide with MC simulation in Tiam1:syndecan (flexible protein and eps8). The directory is:

```
/home/xingyu/PDZ_complexe_TKQ/proteus/run_impo.sh
```

A backup in the directory:

```
/home/xingyu/testcase/pt_mut_scan/
```

script1. The shell script for the reference energies optimization (in cluster).

```
#!/bin/bash
export CPD="/bioinfo/CPD/Proteus_2.0.1"
export MATRIX="/work/xingyu/PDZ_Protein_act/matrix"
export PROTEUSEXE="/work/xingyu/proteus/proteus"
export MYPATH="/work/xingyu/PDZ_Protein_act/proteus/optEner_TC"
export CASKPATH="/work/xingyu/Cask_Pro_act/proteus/optEner_TC"
export NODE=node$1
export MYPATHNODE=$MYPATH/$NODE
export i=`grep starting $MYPATH/iteration.log | awk '{print $3}'`
```

```
#copy the reference energies to the current directory
cp $MYPATH/refener_oldb.dat $MYPATHNODE/refener_oldb.dat
cp $MYPATH/refener_olde.dat $MYPATHNODE/refener_olde.dat

export prot=$1
echo iteration $i >> $MYPATHNODE.log
echo "protein: $prot" >> $MYPATHNODE.log
# select the active positions
awk j="$prot"'{if(NR%5 ==j%5) printf "%s,", $1}' $MATRIX/position_list.dat >
$MYPATHNODE/pos_act.dat

# Write the option Space_Constraints and Ref_Ener in the configure file of MC
cp $MYPATHNODE/MONTECARLO.conf $MYPATHNODE/MONTECARLO1.conf
python $MYPATHNODE/confMC.py $NODE

echo "<Space_Constraints>" >> $MYPATHNODE/MONTECARLO1.conf
sort -k1 $MYPATHNODE/space_constraints.dat >> $MYPATHNODE/MONTECARLO1.conf
echo "</Space_Constraints>" >> $MYPATHNODE/MONTECARLO1.conf

echo "<Ref_Ener>" >> $MYPATHNODE/MONTECARLO1.conf
cat $MYPATHNODE/refener_oldb.dat >> $MYPATHNODE/MONTECARLO1.conf
sort -k2 $MYPATHNODE/ref_ener.dat >> $MYPATHNODE/MONTECARLO1.conf
echo "</Ref_Ener>" >> $MYPATHNODE/MONTECARLO1.conf

# Active positions in array
active_resid_array=( `perl -nale 'if($F[1]eq"active"){push@a,$F[0]}END{print
join(q( ),sort{$a<=>$b}@a)}' $MATRIX/position_list.dat` )
# Number of active positions
nbactive=${#active_resid_array[@]}
echo "number of active positions: $nbactive" >> $MYPATHNODE.log
# Make a separated active resid list
active_resid_list=`perl -nale 'if($F[1]eq"active"){push@a,$F[0]}END{print
join(q(,),sort{$a<=>$b}@a)}' $MATRIX/position_list.dat`

# Run proteus multiples mutations
$PROTEUSEXE < $MYPATHNODE/MONTECARLO1.conf >&  $MYPATHNODE/MONTECARLO.log
echo Finished MONTECARLO of fixed backbone $NODE >> $MYPATHNODE.log
# Select the sequences with temperature 0.592
grep '0.592$' $MYPATHNODE/output.ener_* > $MYPATHNODE/output.ener
cat $MYPATHNODE/proteus_mult.seq_* > $MYPATHNODE/proteus_mult1.seq
python $MYPATHNODE/selcSeq.py $NODE

# Proteus post-processing
$PROTEUSEXE < $MYPATHNODE/POSTPROCESS.conf >& $MYPATHNODE/POSTPROCESS.log
echo Finished POSTPROCESS of fixed backbone $NODE >> $MYPATHNODE.log
```

```
# Analysis of sequences
$MYPATH/analyze_proteus_sequences.pl $MYPATHNODE/proteus_mult.seq
$MYPATHNODE/proteus_mult.rich $active_resid_list 500000000 >
$MYPATHNODE/proteus_mult.dat
cp $MYPATHNODE/proteus_mult.dat $MYPATH/proteus_mult_"$NODE".dat
mv $MYPATHNODE/proteus_mult.seq $MYPATHNODE/proteus_mult_$i.seq
# write the log file
echo $NODE finished > $MYPATH/log/ite_"$NODE"_"$i".log
# check if all 10 separated MC simulation have done
n=`ls $MYPATH/log/*.log | wc -l`
if [ $n -eq 10 ]; then
 for f in $MYPATH/proteus_mult_node1.dat \
         $MYPATH/proteus_mult_node2.dat \
         $MYPATH/proteus_mult_node3.dat \
         $MYPATH/proteus_mult_node4.dat \
         $MYPATH/proteus_mult_node5.dat \
         $CASKPATH/proteus_mult_node1.dat \
         $CASKPATH/proteus_mult_node2.dat \
         $CASKPATH/proteus_mult_node5.dat \
         $CASKPATH/proteus_mult_node4.dat \
         $CASKPATH/proteus_mult_node5.dat ; do
   cat $f >> $MYPATH/proteus_mult.dat
 done
 for f in $MYPATH/log/*.log ; do
  cat $f >> $MYPATH/iteration_$i.log
 done
 rm $MYPATH/log/*.log
 # Update refEner with python:
 python $MYPATH/refenerOpt1.py $MYPATH/proteus_mult.dat
 mv $MYPATH/proteus_mult.dat $MYPATH/proteus_mult_$i.dat
 cp $MYPATH/refener_newb.dat $MYPATH/refburied_$i.dat
 cp $MYPATH/refener_newe.dat $MYPATH/refexposed_$i.dat
 cp $MYPATH/refener_newb.dat $MYPATH/refener_oldb.dat
 cp $MYPATH/refener_newe.dat $MYPATH/refener_olde.dat
 cp $MYPATH/refener_newb.dat $CASKPATH/refener_oldb.dat
 cp $MYPATH/refener_newe.dat $CASKPATH/refener_olde.dat
 cp $MYPATH/dat/refdicb.dat $CASKPATH/dat/refdicb.dat
 cp $MYPATH/dat/refdice.dat $CASKPATH/dat/refdice.dat
 mv $MYPATH/freqCale.dat $MYPATH/freqCale_$i.dat
 mv $MYPATH/freqCalb.dat $MYPATH/freqCalb_$i.dat

 let i=i+1
 echo "starting iteration $i" > $MYPATH/iteration.log
 ssh master0 $MYPATH/optiEref.sh
fi
```

RESUME

Nous travaillions avec le PDZ domaine de la protéine Tiam1 *(T-cell lymphoma invasion and metastasis gene 1)*. Nous voulons tester la méthode Dessin Computationnel de Protéine (DCP) en utilisant différents paramètres : les deux constants diélectriques (4 et 8) et les différents énergies des références (E_X). Pour calculer les énergies des références, nous avons développé deux méthodes différents : La première méthode utilise un modèle de peptide déplié. Nous utilisons des Monte Carlo simulations pour calculer les énergies libres des mutants simples par rapport à Ala. La deuxième méthode utilise un modèle empirique, qui va améliorer les E_X itérativement. La boucle va s'arrêter jusqu'à toutes les compositions des acides aminés calculés convergent vers les expérimentales valeurs. Nous avons réalisé cette méthode en utilisant les protéines multiples (de Tiam1 et de Cask), en regroupant des acides aminés similaires ensemble, et en appliquant la conception de « position – dépendante ». Nous avons utilisé ces E_X améliorées pour explorer des mutations du peptide syndecan qui est associé avec Tiam1. Les résultats sont satisfaisants en comparant avec des séquences expérimentaux.

ABSTRACT

We work with the PDZ domain of Tiam1. We want to test the performance of Computational Protein Design (CPD) method using different parameters including two dielectric constants (4 and 8) and different reference energies (E_X). We used two methods to calculate the reference energies: In the first method, we use an unfolded peptide model and calculate the free energies of all single point mutants relative to Ala by Monte Carlo (MC) simulations. The second method uses an empirical model to iteratively update the E_X, until the amino acid compositions of calculated sequences are convergent to the experimental values. To implement this empirical model, we use multiple proteins (Tiam1 and Cask), regroup the similar amino acid types and apply position-dependent conception. We applied the improved E_X to explore mutations of the syndecan peptide bound to Tiam1. We have got satisfied results comparing with the experimental sequences.