CHAPTER 3

RESEARCH METHODOLOGY

The computational methods can be divided into two main parts, computational simulation and designing algorithm. The procedures are subdivided into eight parts:

(i) Searching and preparing for 3D structures,

(ii) Predicting the protein-protein complex,

(iii) Finding the intermolecular/interface neighbor,

(iv) Exacting considered CD4 residues,

(v) Identifying the hot spots,

(vi) Validating the hot spots,

Computational Simulation

Searching and preparing for 3D structures

The 3D template structures of human CD4 and DARPin 23.2 were explored in Protein Data Bank (PDB) (http://www.rcsb.org/pdb/home/home.do) based on sequence compatibility. Model of DARPin 23.2 structure was constructed by homology modeling method [109].

To prepare the protein structure, library of conformation along with configuration in term of typed forced field was utilized. The model was minimized in Discovery Studio (DS) 2.5. The DS 2.5 was used to generating reasonable 3D coordinates, correcting the missing residues, standardizing name of atom, eliminating alternate conformation, etc. CHARMm force field with the Momany-Rone partial charge estimation method was applied in this case. Before typing, the hydrogens were adjusted to comply with the force field and the bonds were localized. All the coordinations were minimized to relax the structure and eliminate the steric overlapping. The algorithm for minimization begins with 1000 steps of Steepest Descent (SD) with a RMS gradient tolerance of 3.0 Å, followed by Conjugate Gradient (CONJ). The SD uses first derivative information and coordinates that are adjusted in the negative direction using only current location. For CONJ, this method uses previous history of minimization steps and the current gradient to determine the next step of coordinate. During a cycle of minimization, the gradient was averaged and the change in total energy was calculated. The minimization routine exited when the average gradient and energy change are less than or equal to the tolerance. So, a tolerance of RMS gradient and energy change was defined as 0.1 Å and 0 kcal/mol. For this minimization algorithm, the RMS gradient or energy change controlled only the last CONJ minimization phase. In case of the average gradient or energy change were not less than or equal to the tolerance, the minimization routine exited when the last cycle is equal to tolerance of maximum step. Here, the maximum step of both proteins defined as 5000 minimizing steps. The solvent model and salt concentrations did not incorporate in this system. To calculate the non-bonded interaction, the distance cutoff value was defined for counting non-bonded interaction pairs. This

cutoff value was defined as 14.0 Å. Within cutoff distance, non-bonded interaction pairs were scaled smoothly using switching function which the starting and ending points of switching were 10.0 and 12.0 Å respectively. For calculating long range electrostatics, the spherical cutoff was specified. In the simulation, all bonds and hydrogens did not constraint. Moreover, the minimized structures were validated by assessing the stereochemistry quality by Ramachandran plot in PROCHECK program [110].

Predicting the protein-protein complex

The complexes were constructed from ZDOCK protocol and were refined by RDOCK protocol. In ZDOCK protocol, the optimized protein structures were put in this algorithm: CD4 was defined as receptor protein and DARPin 23.2 was defined as ligand protein as described in the previous work [109]. In this protocol, the molecular coordination of CD4 was fixed and the DARPin 23.2 relocated around CD4. The rotational sampling of the ligand orientations was specified by angular step size. In this case, the angular step size was defined as 6 leading to get 54000 poses for sampling. To filter the targeting area, the set of residues at the binding interface and the blocked residues were specified. Here, only DARPin 23.2 was performed to define the binding site or non-conserve residues which relied on AR consensus sequence reported by Binz et al (2003) [11]. In addition, the blocked residues of DARPin 23.2 also were defined as shown in Fig.11. The 54000 sampling poses were carried out to calculate the ZDock scoring within residues in binding interface. To select the binding residues of docked pose, the distance cutoff was specified as 10 Å.

In this case, not only pairwise shape complementarity (PSC) was used to calculate the ZDock scoring but also desolvation (DE) and electrostatic (ELEC) energies. Any sampling poses having positive ZDock score were selected to further make docked output. However, only top 2000 poses of positive ZDock score were kept. The positive ZDock poses or top 2000 poses were performed to calculate ZRank score based on energy calculation. Moreover, they were groped in to cluster based on same/different orientation of complex. The clustering interface cutoff which is interface region between receptor and ligand was specified to calculate the RMSD values for clustering. Any poses having RMSD less than RMSD cutoff was grouped in the same cluster. This experiment, The RMSD cutoff and interface cutoff were defined as 6.0 and 9.0 Å respectively. The maximum number of clusters was defined as100 clusters.

The outputs of ZDOCK protocol which are top 20 highest ZDock scored poses were used as inputs for refining in RDOCK protocol. Here the dielectric constant was defined as 4. The refined poses were carried out to calculate energy called E_RDock. Before running them in RDOCK, they were typed CHARMm polar H force field.

N-cap	DLGKKLLEAARAGQDDEVRILMANGADVNAT
1st domain	DTLGRTPLHMAAAWGHLEIVDVLLKHGADVNAI
2nd domain	EEVGMTPLHLAAFLGHLEIVEVLLKSGADVNAQ
C-cap	DKFGKTAFDISIDYGNEDLAEILQ

Figure 11 The binding and non-binding residue of DARPin 23.2 relied on AR consensus by Binz et al. The white residues in the black box are binding site and the others are blocked residues.

Finding the intermolecular/interface neighbor

The docking complexes obtained from the RDOCK protocol that show no binding of DARPin 23.2 onto CD4 domain 1 were excluded. Intermolecular/interface neighbor analysis of each CD4-DARPin complex was carried out, using DS 2.5, with distance threshold of 5.0 Å. Any atom pair (one from each protein) having distance between them is less than 5.0 Å is defined as intermolecular neighbor. The maximum distance between the hydrogen bond donor and the acceptor was defined at 2.5 Å and the donor proton-acceptor angles were defined in a range of 120 –180° to identify hydrogen bond interaction. Likewise, the CD4-gp120 [34] and CD4-MHCII [28] structures were identified interface neighbors.

Designing Algorithm for finding key residue

Extracting considered CD4 residues

To determine the hot spots, which were important residues for contributing binding complex, of CD4 to DARPin 23.2, we designed the two steps procedure to extract the information. First step was filtering the interface residue of CD4 binding to DARPin 23.2 at 5.0 Å, here called "considered CD4 residue" and the second was identifying the key binding residue called "key CD4 residue". Finally, the hot spots were analyzed by using key CD4 residues and bio-information. First of all, the intermolecular neighbor data in CD4-DARPin 23.2 complex were classified into five constructed criteria. First criterion was defined as the number of DARPin's amino acid positions those were bound to each CD4's amino acid. The second criterion was

the number of interactions in each CD4's amino acids bound to DARPin. Third criterion was the number of CD4's atom types in each CD4's amino acids those were bound to the DARPin's residues. The physic-chemical meaning of criterion 1, 2, and 3 were interaction between CD4's amino acid and DAPin's amino acid, CD4's atom and DARPin's atom, as well as CD4's atom and DARPin's amino acid respectively. The fourth criterion was defined as the percentage of CD4's atom types in each CD4's amino acids those were bound to the DARPin's residues. The fifth criterion was each hydrogen-bonded CD4's amino acids those are bound to the DARPin's residues. The other meaning in physic-chemical property was interaction between atoms making hydrogen bond.

The data of intermolecular neighbors, which were in a string form, were converted into a histogram value relying on four criteria (criterion 1-4). For each CD4's amino acid, the histogram value was counted by using the following equation:

$$x_i = \sum_{j=1}^{n} p(i,j), \quad p(i,j) = \begin{cases} 1 & \text{if } d(i,j) \le 5 \text{ Å} \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

, where x_i is a histogram value of each CD4's amino acid, $d(i, j)$ is the distance between two atom belonging to two residues (one from each chain), and $p(i, j)$ in case of criterion 1 is a frequency distribution of intermolecular neighbor of CD's residues i and DARPin's residue j. In case of criterion 2, $p(i,j)$ is a frequency distribution of intermolecular neighbor of CD4's residue i and DARPin's atom j. For $p(i,j)$ of criterion 3, it is a frequency distribution of CD4's atom j in CD's residue i.

The designed algorithm for criterion 2 started with reading the intermolecular neighbors and putting all amino acid of CD4 as typing in *input* and *CD4* variable,

respectively, in Fig. 12. Then the CD4's residue *j* was searched in intermolecular neighbor *i* as shown in variable *Find(j)(i)*. The histogram score was given when the CD4's amino acid *j* was found in intermolecular neighbor *j* which score *Find(j)(i)* equaled to 1. Note that the *strfind* function returned the strating index. Since the strating index of CD4's residue in intermolecular neighbor *j* was 1, this value was used to be tolerance for giving histogram value. The loop step was finished when the last residue of CD4 was searched in last intermolecular neighbor.

The designed algorithm of criterion 1 shows in Fig. 13. This algorithm was similar to algorithm for criterion 2. The different thing was that not only CD4's residue but also DARPin's residue were putted in algorithm. The histogram score was counted when meeting both of CD4's residue *j* and DARPin's residue *k* in intermolecular neighbor *i*. Note that the tolerance for giving histogram value was any values of 9-15 because these values were strating index of DARPin's residue in intermolecular neighbor *i*. For 3rd criterion, the algorithm designed similar to algorithm of 1st criterion. The differences were the atom of CD4 was searched, not DARPin's amino acid, and the tolerance for counting histogram score was 8 as shown in Fig. 14. The characters of atom were showed in Table 4.

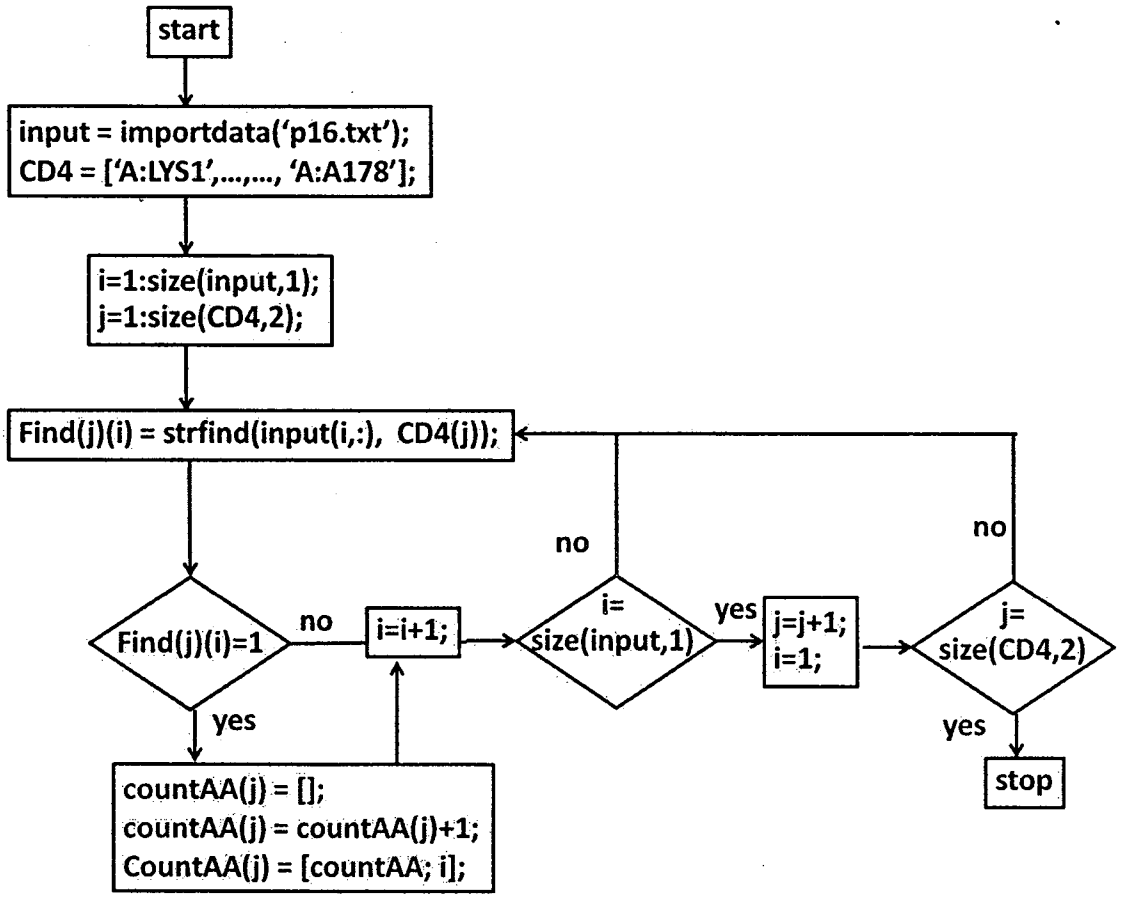

Figure 12 The designed algorithm to identify the number of interaction pair in each CD4's amino acid (criterion 2).

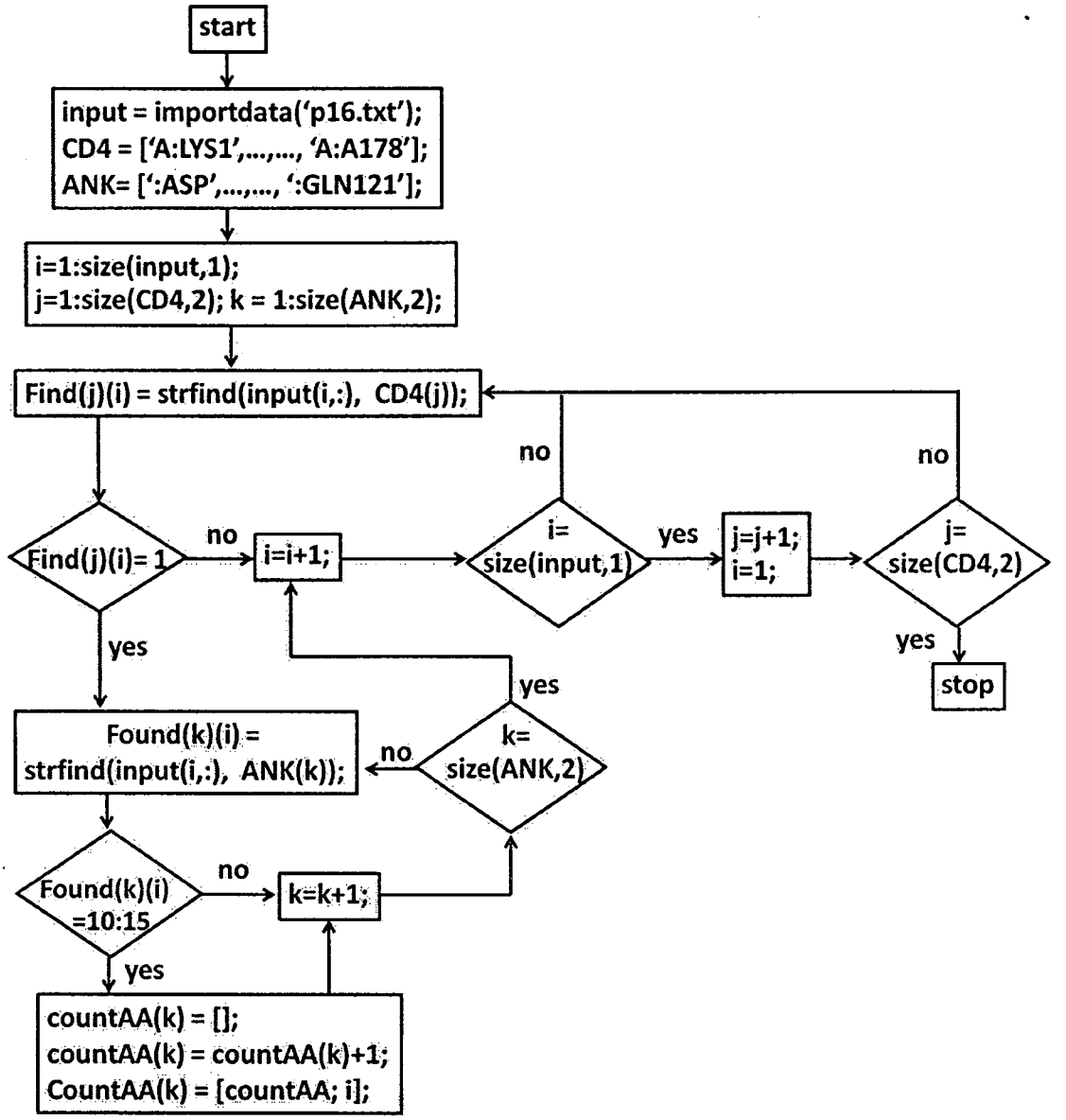

Figure 13 The designed algorithm to identify the number of DARPin's amino acid in each CD4's amino acid (criterion 1).

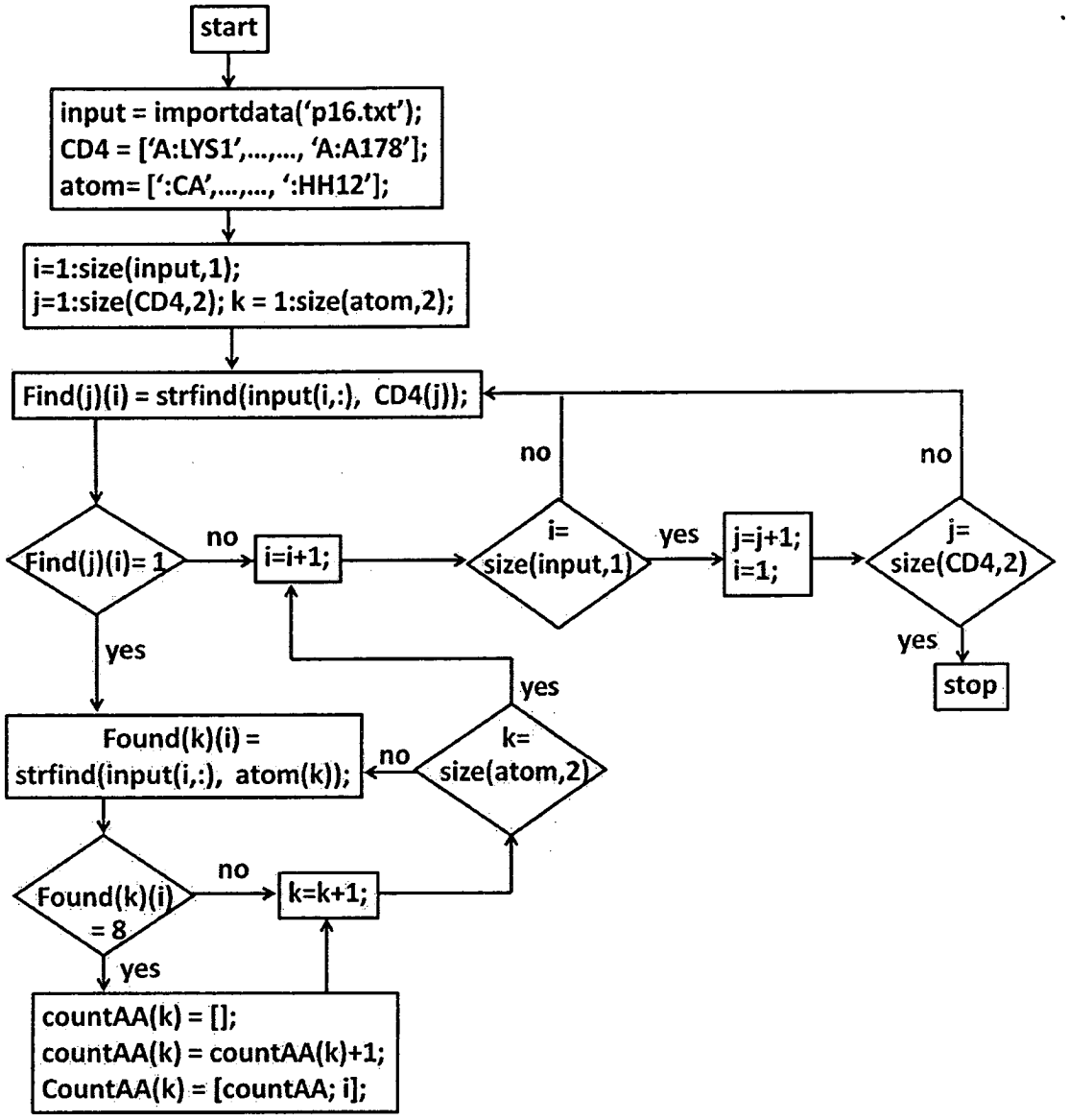

Figure 14 The designed algorithm to identify the number of CD4's atom in each CD4's amino acid (criterion 3).

For algorithm of criterion 4, it is mostly similar to criterion 3. To get the scoring of criterion 4, the scoring of criterion 3 was performed using followed equation:

$$\%atomType^{CD4} = \frac{atomType^{CD4} \times 100}{totalAtom} \tag{18}$$

, where $\%atomType^{CD4}$ is the score of criterion 4, $atomType^{CD4}$ is the number of CD4's atom type getting from criterion 3, *totalAtom* is the number of backbone and side chain atoms as shown in Table 4.

Table 4 Characters of atom using in criterion 3 and the number of backbone and side chain atoms using in criterion 4.

aa	Arg	His	Lys	Aps	Glu	Ser	Thr	Asn	Gln	Ala
atom	CB	CB	CB	CB	CB	CB	CB	CB	CB	CB
	CG	CG	CG	CG	CG	OG	OG1	CG	CG	
	CD	ND1	CD	OD1	CD	HG	HG1	OD1	CD	
	NE	HD1	CE	OD2	OE1		CG2	ND2	OE1	
	HE	CD2	NZ		OE2			HD21	NE2	
	CZ	NE2	HZ1					HD22	HE21	
	NH1	CE1	HZ2						HE22	
	HH11		HZ3							
	HH12									
	NH2									
	HH21									
	HH22									
#	17	12	13	9	10	8	9	11	12	6
aa	**Ile**	**Leu**	**Met**	**Phe**	**Trp**	**Tyr**	**Val**	**Pro**	**Gly**	**Cys**
atom	CB	CB	CB	CB	CB	CB	CB	CB		CB
	CG2	CG	CG	CG	CG	CG	CG1	CG		SG
	CG1	CD1	SD	CD1	CD2	CD1	CG2			
	CD1	CD2	CE	CD2	CE2	CE1				
				CE1	CE3	CD2				
				CE2	CD1	CE2				
				CZ	NE1	CZ				
					HE1	OH				
					CZ2	HH				
					CZ3					
					CH2					
#	9	9	9	12	16	14	8	7	5	7

Note: 1) The backbone atoms are N, HN, CA, C, and O which do not show in table

2) aa stands for amino acid; # is the number of backbone and side chain atom.

In each criterion excepting criterion 5, the CD4's amino acids with the top 10 highest values in the histogram were selected to be candidates for considering key CD4 residues. The value of 10 was sample size computing based on hypergeometric sampling as following equation:

$$n = \frac{NZ^2pq}{E^2(N-1)+Z^2pq} \tag{19}$$

, where n is the sample size to find candidate of considered CD4 residue, N is the population size of any CD4's residue having at least 1 interaction pair, p and q (q = 1 - p) are the population proportion, z is the value specifying the level of confidence, and E is the accuracy of sample proportions. Here, N, Z, p, and E were 23, 1.44 (at the confidence 85%), 30%, and 0.15 (at the confidence 85%) respectively. For population size N, 23 was average of the number of CD4's residue having at least one interaction pair in all 11 poses.

The candidate residues in four criteria were union together to get the considered CD4 residues. The frequency in considered CD4 residue was further performed to find key residues.

Identifying key binding residues

The frequency in considered CD4 residues in all five criteria was normalized by using the standardization as following equation:

$$z = \frac{x-\mu}{\sigma} \tag{20}$$

, where x is a histogram value of CD4's amino acid, μ is a CD4's criterion mean, and σ is a CD4's amino acid standard deviation.

Criteria combination was created in six patterns, i.e., patterns A, B, C, D, E and F. For each pattern, the normalized histogram values from each considered CD4 residue were combined. However, the criteria 1, 2, 3, 4, and 5 were used in pattern A; whereas the criteria 1, 2, 3, and 5 were used in pattern B. For pattern C, D, E, and F, the used criteria were; 1, 2, 4, and 5; 1, 2, 3, and 4; 1, 2, and 3; and 1, 2, and 4, respectively. Remarkably, the criterion 4 was derived from criterion 3, so, all six patterns consisted of either third or fourth criterion or both of them. Since, the pair potential interface between CD4's amino acid and DARPin' amino acid was important as between CD4's atom and DARPin's atom as well as between CD4's atom and DARPin amino acid, the criteria 1, 2, and 3 and/or 4 were involved in 6 patterns. For fifth criterion which was the subset of second criterion was considered to make combination but the importance was less than criteria 1-4. Therefore, the fifth criterion was involved in three patterns. In each pattern, the normalized histogram value of considered CD4 residues was combined and renormalized again using equation (20). Finally, the key CD4 residue decision, the maximum from all patterns was selected to be 1st key amino acid. Then the 2nd key residue was defined as the maximum value in six patterns without 1st key residue. As same identifying 2nd key residue, the 3rd key residue was identified in 6 patterns without 1st and 2nd key residues. Altogether, the designed algorithm for finding key residue of CD4 binding to DARPin 23.2 was shown in Fig. 15. Then the 1st-3rd key residues of 11 poses were analyzed to be hot spots.

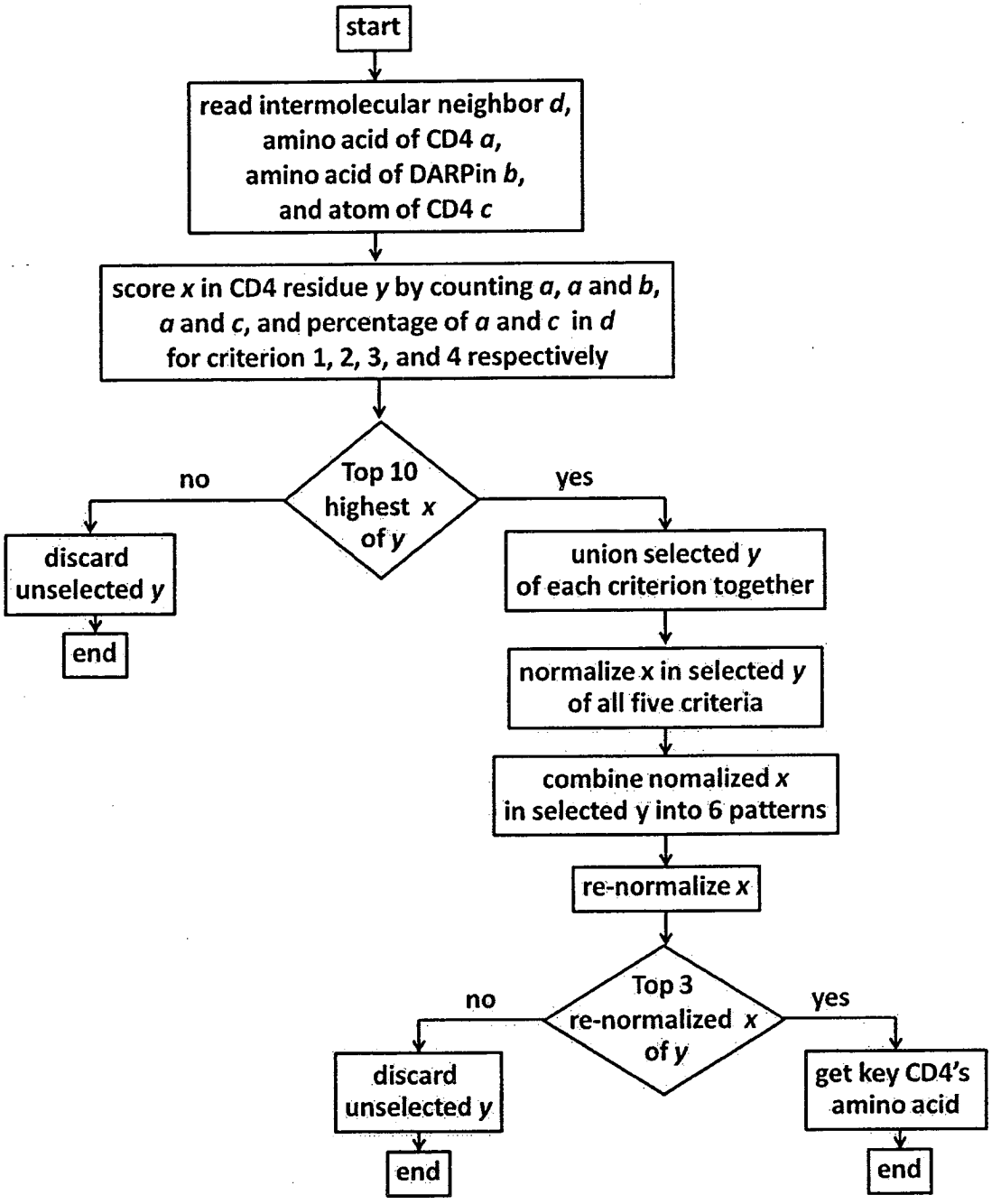

Figure 15 Flowchart of algorithm for finding key residues of CD4 binding DARPin 23.2. Definition of x is frequency and y is residue position of CD4.

Validating the predicted key residues

The 1^{st} – 3^{rd} key CD4 residues predicted by our constructed algorithm were validated with physicochemical properties of binding residue and HotPOINT as well as HSPred which were web server predicting hot spot that having difference methods. In case of validation, the 1^{st} – 3^{rd} key CD4 residues were called hot spots to compare with other programs. The physicochemical properties such as hydrophobic and hydrophilic residue as well as the propensity of hot spot were validated with our prediction. Here, the hot spot propensity analyzed by Bogan & Thron [58] was divided into 4 classed: high, moderate, low, and rare propensity. The high propensity was defined by suggesting the enriched hot spots by Bogan & Thorn. Likewise, the rare propensity was assigned with residue having frequency of hot spot percentage less than 3%. The moderate and rare propensities were set from the others which average of these data was used to be cut-off. The moderate propensity was residues that had percentage of hot spot more cut-off and also another was defined as frequency less than cut-off.

All 11 docking complexes were performed to identify the hot spots by HotPOINT [78] and HSPred [85, 86] web server. Before the complexes were submitted to two servers, these structures were converted to PDB file. The HotPoint web server is available at http://prism.ccbb.ku.edu.tr/hotpoint. In this server, here, any two atoms belonging to two residues (one from each chain), defined that the distance between them is less than the sum of their van der Waals radii plus a 0.5 Å, were extracted to be interface or interacting residues. Then, interacting residues were

calculated the solvent accessibilities and the pair potential. Finally, the hot spots were labeled with the criteria of the relative accessibility is $\leq$ 20% and the total contact potential is $\geq$ 18.0.

For HPRpred, the available server is http://bioinf.cs.ucl.ac.uk/hspred. Interface residues defined as those having at least one heavy atom within 5.0 Å of a heavy atom in the binding partner were extracted. Then the interface residues were mutated as ananine and were calculated the energy potentials in mutated complex. The positive value of calculated $\Delta\Delta G$ on any interface residue was considered to be hot spot.

Our predicted hot spots were validated with these servers by calculating percentage of identity prediction ($PID^{predict}$) using following equation:

$$PID^{predict} = \left(\frac{Identical\,\mathrm{Re}sidues}{TheNumberOfHotSpot} \right) \tag{21}$$

, where, *IdenticalResidues* is the number of two residues between our prediction and server that predicted in the same residue, *TheNumberOfHotSpot* is the maximum length of our prediction; because we identify key CD4 residue as three amino acid, in this case, this value is 3.