



Ratna Priya Moganti, Oskar Zinger and Boojala Vijay B. Reddy

Bioinformatics and *in Silico* Drug Design Lab, Computer Science Department, Queens College and
The Graduate Center – Computer Science Department of City University of New York

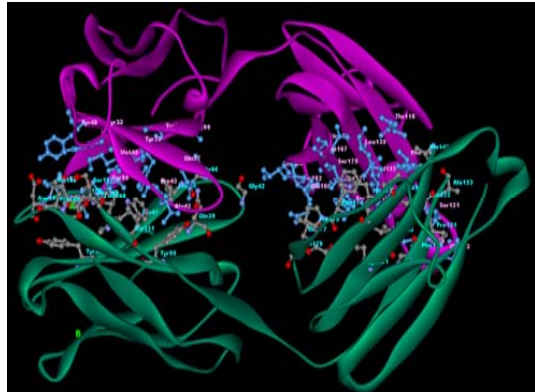
ABSTRACT

In order to systematically develop a more successful predictive scoring function for protein-protein docking we have undertaken analysis of 2038, 3Å or better resolved nonredundant (<25% sequence identity) protein dimer complex structures available in protein databank. We have identified interactively packed residues on the surface of each protein of the complex and computed a matrix of interactively packed amino acid pairs in the form of 20x20 data matrix. Similarly, to use as random model we have computed amino acids pairs of single domain proteins by assuming random complexes between them. We have used these two matrices to generate an amino acid interacting pair-potential matrix which we plan to use as one of the parameters in our proposed predictive scoring function.

INTRODUCTION

Biological function of gene products such as proteins mediated through interactions they make with one another. In humans, the larger number of proteins is expected to engage in hundreds of thousands of interactions, many of which involve large assemblies and play key roles in cellular function and disease. Such assemblies are, however, still rather poorly represented in the Protein Data Bank (PDB). Computational procedures capable of reliably generating structural models of multi-protein assemblies starting from the atomic coordinates of individual components, the so-called "docking" methods, should therefore play important role in bridging the gap. The 3D structures of protein complexes are pivotal for a full understanding of the mechanism of interactions because they provide specific interaction details at the atomic level. Such details are important for rational design of drug molecules to modulate protein interactions.

In molecular docking, transient complex structures are predicted by docking one monomeric structure on to the other. The docking consists of two major steps: generating multiple docking conformations and scoring to recognize the near native complex structure. To predict the interface regions one needs to know what distinguishes an interface region from the rest of the protein. Several attempts have been made using surface residue energy distributions, residue conservation and propensity information and salvation energies etc. However none of these calculations have given any successful results in rightly predicting the complex structures.

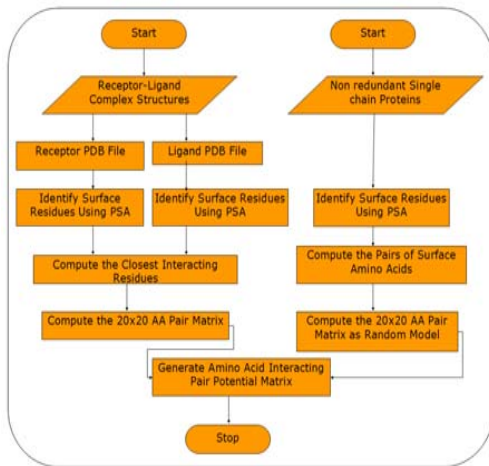


MATERIALS AND METHODS

- 2038 protein nonredundant (< 25% sequence identity) dimer complex structures resolved at 3Å or better were taken from the protein databank (PDB) and the files were separated into Receptor and Ligand pdb files.
- Identified residues involved in interactive packing with at least 10% of the accessible surface area of contact using the PSA (Protein Solvent Accessibility) Program.
- Identified pairs of closest interacting residues between receptor and ligand by calculating the distance between the C α coordinates and taking minimum distance between amino acids as the closest contact residues.
- Generated a 20x20 data matrix of total pair counts (P_{ij}) in all the complexes (shaded green in Table I).
- Used PSA on non-redundant single chain soluble proteins structures and identified surface residues with at least 10% accessible surface area.
- Computed the pairs of surface amino acids from the randomly generated complexes among all these proteins as a random model and generated a 20x20 data matrix of amino acid pairs similar to Table I.
- The random AA pair matrix was normalized to reflect same number of interacting pair counts as in real complexes (Q_{ij}) (shaded pink in Table I) and generated an amino acid interacting pair-potential (PP_{ij}) matrix (Table II) which can be used as one of the parameters in our proposed predictive scoring function.
- The amino acid pair potentials are calculated as follows:

$$PP_{ij} = 10 \log(P_{ij} / Q_{ij})$$

FLOW CHART OF THE METHOD



RESULTS

- Though there are 20 amino acids that code for proteins because of their differences in physicochemical properties some of them prefer to be present on the surface of the protein and others in the interior of protein core.
- Depending on the physicochemical properties certain amino acids on the receptor protein surface can interactively pack with a complementary amino acid present on the surface of ligand protein and *vice versa*.
- The green shaded values in Table I gives count of pairs of such interactively packed amino acids in the selected data set of protein complexes and the pink shaded value is count of pairs of amino acids on the surface of hypothetical complexes of random protein pair which represent noise level and used as random model to compute Table II.
- As can be seen from Table I, on average, each pair of amino acid is occurring about 231 times in the database. However, there are several pairs whose occurrence is significantly higher or lower than the average, reflecting their (i) general composition in the proteins, (ii) their preference to be present on the surface region and (iii) their presence as potential interacting partners on the surface.
- The amino acids C, H, I, L, M, F, W, Y, V presence on the surface of protein will increase its probability of interaction with other proteins. The amino acids E, D, Q, N presence on the surface regions decrease probability of its interaction with other proteins. However, the Table II gives over all pair potentials of interacting pairs of all the combinations of amino acid pairs.

FUTURE PLANS

- Our immediate plans are to characterize all the known interacting surfaces using the amino acid pair potentials in Table II to check how each of these score and helps us to set a cutoff scoring value for prediction of best interacting surface partners.
- Our next goal is to apply these pair potentials on the docked complex structures to score all the interacting surface and select the high scoring docked structure.
- We also plan to generate pair potential matrices using 2,3 and 4 near neighbor interactively packed residues and compare all the matrices to select best scoring matrix.
- We further plan to generate such statistical pair potential matrices for various combinations of atomic pairs to use as another variable in the scoring function.

REFERENCES

- Special Issue: Third Meeting on the Critical Assessment of Predicted Interactions (CAPRI), Proteins: Structure Function and Bioinformatics (2007) Volume 69 (4): 697-872.
- Special Issue: Second Meeting on the Critical Assessment of Predicted Interactions (CAPRI), Proteins: Structure Function and Bioinformatics (2005) Volume 60 (2): 149-323.

Table I: Total count of interactively packed amino acid pairs in the protein dimer complexes (shaded in green) and the normalized counts of pairs of amino acids on the surface of non-redundant single chain soluble proteins used as comparative random model (shaded in pink)

	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
ALA(A)	215	327	258	190	66	234	235	375	147	287	516	216	122	445	301	355	343	188	368	335
ARG(R)	239.4	106	272	559	69	463	271	433	108	194	372	161	136	239	292	443	234	127	329	233
ASN(N)	396.2	163.1	115	259	63	203	185	291	212	176	219	191	66	167	297	295	314	58	258	170
ASP(D)	470.6	389.5	230.1	70	50	150	223	267	201	139	420	398	84	168	233	369	296	129	209	207
CYS(C)	598.6	494.0	586.5	372.8	158	54	50	113	38	50	111	55	15	60	77	116	64	50	45	77
GLU(E)	27.2	22.5	26.7	34.0	0.8	91	220	221	142	212	315	473	101	245	296	344	299	59	290	223
GLN(Q)	661.2	545.3	648.1	824.0	37.5	454.4	109	325	95	169	391	216	72	205	213	327	233	121	211	272
ILE(I)	362.3	299.0	355.3	452.0	20.6	499.0	136.6	302	214	266	397	292	148	395	341	391	347	182	407	319
LEU(L)	583.1	481.6	571.9	727.6	33.1	803.8	440.6	354.3	23	87	152	74	65	90	133	324	152	62	143	124
HIS(H)	145.6	120.0	142.6	181.4	8.3	200.2	109.8	176.9	22.0	127	441	161	109	229	188	213	207	76	228	323
LYS(K)	141.0	116.3	138.1	175.7	8.0	194.1	163.5	171.3	42.7	20.6	399	273	223	932	353	554	439	231	439	793
VAL(V)	243.0	200.5	238.3	303.1	13.8	334.7	183.5	295.4	73.7	71.3	61.3	49	100	181	161	406	251	105	251	206
TRP(W)	723.7	597.3	708.7	901.6	41.1	995.5	546.1	879.6	219.2	212.4	266.5	543.9	37	95	115	146	96	98	124	121
MET(M)	74.9	61.7	73.3	93.3	4.2	103.0	56.5	91.0	22.7	22.0	37.8	112.8	5.8	127	305	647	472	136	291	369
PHE(F)	105.1	86.7	102.9	131.0	6.0	144.8	79.4	127.7	31.9	30.8	53.1	158.6	16.4	111.4	172	636	283	271	313	267
PRO(P)	412.6	240.4	404.5	514.6	23.4	568.2	311.5	501.6	125.0	121.1	208.9	621.9	64.4	90.4	177.1	382	634	227	524	461
SER(S)	515.8	426.2	505.6	547.9	29.3	711.3	389.8	627.0	156.5	151.6	261.4	778.1	80.5	113.1	443.7	276.9	207	74	226	284
THR(T)	455.8	376.7	446.7	189.9	25.9	628.5	344.5	554.2	138.3	134.0	231.0	687.5	71.2	99.9	392.1	489.9	216.2	72	179	100
TRP(W)	43.8	36.2	42.9	54.7	2.5	60.5	33.1	53.3	13.3	12.9	22.2	66.1	6.8	9.6	37.7	47.1	41.6	20	174	333
TYR(Y)	152.3	125.8	149.2	189.9	8.6	209.8	115.0	185.1	46.2	44.7	77.1	229.6	23.7	33.3	131.9	163.6	144.6	13.9	241	185
VAL(V)	235.0	194.0	230.4	293.1	13.3	323.7	177.5	285.7	71.3	69.0	119.0	354.4	36.6	51.4	202.1	252.7	223.3	21.5	74.6	57.4

Table II: Amino Acid Pair Potentials. The positive values indicates that those pair of amino acids influence for interactive packing of the proteins for complex formation if they are close to each other in protein-protein docking

	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
ALA(A)	-0.5	-0.8	-2.6	-5.0	3.8	-4.5	-1.9	-1.9	0.0	3.1	3.3	-5.3	2.1	6.3	-1.4	-1.6	-1.2	6.3	3.8	1.5
ARG(R)	-1.9	-1.5	0.5	4.9	-0.7	-0.4	-0.5	-0.5	2.2	2.7	-5.7	3.4	4.4	-0.7	0.2	-2.1	5.4	4.2	0.8	
ASN(N)	-3.0	-3.5	3.7	-5.0	-2.8	2.9	1.7	1.1	-0.4	-5.7	-0.5	2.6	1.3	-2.3	-1.1	1.2	4.1	-1.3		
ASP(D)	-7.4	-3.1	-4.4	0.4	-1.0	1.4	-3.6	0.5	1.1	-3.4	-2.4	-2.8	3.7	0.4	-1.5					
CYS(C)	23.2	1.6	3.9	5.3	6.6	8.0	9.1	1.3	5.5	10.0	5.2	6.0	3.9	13.0	7.2	7.6				
GLU(E)	-7.0	-3.6	-5.6	-1.5	0.4	-0.3	-3.2	-0.1	2.3	-3.8	-3.2	-3.2	-0.1	1.4	-1.6					
GLN(Q)	-1.0	-1.3	-0.6	2.0	3.3	-4.0	1.1	4.1	-1.7	0.8	1.7	5.6	2.5	1.9						
GLY(G)	-0.7	0.8	1.9	3.3	-4.8	2.1	4.9	-1.7	-2.1	-2.0	5.3	3.4	0.5							
HIS(H)	0.2	3.1	3.1	-4.7	4.6	4.5	0.3	3.2	0.4	6.7	4.9	2.4								
ILE(I)	7.9	7.9	-1.2	7.0	8.7	1.9	1.5	1.9	7.7	7.1	6.7									
LEU(L)	8.1	8.1	7.7	12.4	2.3	3.3	7.7	12.4	2.3	3.3										
LYS(K)	-10.5	-0.5	0.6	-5.9	-2.8	-4.4	2.0	0.4	-2.4											
MET(M)	8.0	7.6	2.5	2.6	1.3	11.6	7.2	5.2												
PHE(F)	10.5	5.3	7.6	6.7	11.5	9.4	8.5													
PRO(P)	-0.1	1.6	-1.4	8.6	3.8	1.2														
SER(S)	1.4	-0.2	2.5	1.9	1.0															
THR(T)	-0.2	2.5	1.9	1.0																
TRP(W)	15.6	11.1	6.7																	
TYR(Y)	8.6	6.5																		
VAL(V)	5.1																			