LRC: A new algorithm for prediction of conformational B-cell epitopes using statistical approach and clustering method

Mahnaz Habibi[1],*, Pooneh Khoda bakhsh[2], Rosa Aghdam[3]

1 Department of Mathematics, Islamic Azad University Branch of Qazvin, Iran
2 Department of Computer and IT, Islamic Azad University south Tehran Branch
3School of Biological Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
* Email: mhabibi@ipm.ir

1 Supplementary Material

In this section, more details of proposed algorithms are explained. In addition the PDB codes and chain IDs in training and testing dataset are presented.

1.1 Dataset

In this work, 90 Antibody-Antigen complexes are used. We choose randomly the 70 antigens as training and 20 antigens as testing dataset. To choose 70 antigens we use MATLAB Function "rand" which it uniformly distributed pseudorandom numbers. In Table S1, the PDB codes and chain IDs (antibody IDs and antigen ID) in training and testing dataset are presented. The testing dataset contains 3266 residues that 195 residues are reported as epitope residues in IEDB and the training dataset contains 13909 residues that 822 residues are known as epitope residues. Some details of training and testing dataset are presented in Table S2.

1.2 Methods

To clearly describe the LRC Algorithm, some notations need to be introduced. In this work, we use different criteria to predict epitope residues on antigen. In this subsection, some details of scales which used in this paper are presented.

1.2.1 Physicochemical Scale

The hydrophobicity of each amino acid is used as the one of the physicochemical scale to predict eitope residues. In general, we use different hydrophobicity scales with different rationales. In Table S3 we show some hydrophobicity scales which are considered in our work. The hydrophobicity scale that reported by Manavalan et al. with maximum DScore is chosen as the suitable scale to distinguish epitope and non-epitope residues. This scale shows the significant difference between epitope and non-epitope residues.

Another scale which used in our work is polarity scale. We consider two types of polarity scales which reported by Zimmermanet al. and Granthamet al. These scales for each amino acid are presented in Table S4. Also the polarity scale with maximum DScore is selected as the suitable scale.

The flexibility is another criterion to distinguish epitope and non-epitope residues on antigen. The scales of flexibility for each amino acid for two types of flexibility scales are presented in Table S5.

In this work, the flexibility scale which reported by Karpluset al. with maximum DScore is chosen as the suitable scale. Also we study two types of antigenicity scales which reported by Kolaskar. These scales for each amino acid are presented in Table S6.

Table S1: Ag-Ab complexes which they are included the PDB ID, Antibody heavy chain, Antibody light chain and Antigen chain.

Testing				
1AR1-CD-B	1CL7-HL-I	1JPS-HL-T	1JRH-HL-I	1MHH-DC-F
1OSP-HL-O	1OTS-CD-A	1TQB-BC-A	2FD6-HL-U	2Q8B-HL-A
2R29-HL-A	2UZI-HL-R	2VXT-HL-I	2XTJ-DB-C	3B9K-HL-B
3BN9-DC-B	3D85-BA-C	3KR3-HL-D	3KS0-HL-B	3NH7-HL-A
Training				
1BGX-HL-T	1CZ8-HL-W	1DEE-FE-H	1E6J-HL-P	1EGJ-HL-A
1EO8-HL-A	1EZV-XY-E	1FNS-HL-A	1FSK-IH-G	1H0D-BA-C
1IQD-BA-C	1LK3-HL-A	1N8Z-BA-C	1NFD-HG-D	1NL0-HL-G
1NMB-HL-N	1NSN-HL-S	1OAZ-HL-A	1OB1-BA-C	1ORS-BA-C
1PKQ-BA-E	1QKZ-HL-A	1R3J-BA-C	1RJL-BA-C	1V7M-HL-V
1W72-HL-A	1WEJ-HL-F	1YJD-HL-C	1ZTX-HL-E	2ADF-HL-A
2AEP-HL-A	2B2X-HL-A	2BDN-HL-A	2CMR-HL-A	2DD8-HL-S
2H9G-HL-S	2J4W-HL-D	2J5L-CB-A	2J88-HL-A	2JEL-HL-P
2NY1-DC-A	2NYY-DC-A	2QQK-HL-A	2QQN-HL-A	2R0L-HL-A
2R56-IM-B	2VXQ-HL-A	2VXS-IM-A	2XQB-HL-A	2XQY-GL-A
2XWT-AB-C	2YC1-AB-C	2ZCH-HL-P	3B2U-HL-A	3CVH-HL-A
3DVG-BA-Y	3GI9-HL-C	3GRW-HL-A	3H42-HL-B	3HI6-HL-A
3L5X-HL-A	3L95-BA-X	3LDB-CB-A	3LEV-HL-A	3LH2-HL-S
3LHP-HL-S	3LIZ-HL-A	3MJ9-HL-A	3MXW-HL-A	3NGB-HL-G

Table S2: The number of residues, epitope residues, surface residues and surface epitope residuesin training and testing dataset.

	Number of Residues	Number of Epitope Residues	Number of SurfaceResidues	Number of SurfaceEpitope Residues
Training dataset	13909	822	13855	822
Testing dataset	3266	195	3122	195

Table S3: The scales of hydrophobicity for each amino acid.

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile
Eisenberg et al.	0.62	-2.53	-0.78	-0.9	0.29	-0.85	-0.74	0.48	-0.4	1.38
Hopp et al.	-0.5	3	0.2	3	-1	0.2	3	0	-0.5	-1.8
Manavalan et al.	**12.97**	**11.72**	**11.42**	**10.85**	**14.63**	**11.76**	**11.89**	**12.43**	**12.16**	**15.67**
Black et al.	0.616	0	0.236	0.028	0.68	0.251	0.043	0.501	0.165	0.943
Fauchere et al.	0.31	-1.01	-0.6	-0.77	1.54	-0.22	-0.64	0	0.13	1.8
Argos et al.	0.3	-1.4	-0.5	-0.6	0.9	-0.7	-0.7	0.3	-0.1	0.7
Janin et al.	1.36	0.15	0.33	0.11	1.27	0.33	0.25	1.09	0.68	1.44
Tanford et al.	0.62	-2.53	-0.78	-0.09	0.29	-0.85	-0.74	0.48	-0.4	1.38
Parker et al.	2.1	4.2	7	10	1.4	6	7.8	5.7	2.1	-8
Rose et al.	0.74	0.64	0.63	0.62	0.91	0.62	0.62	0.72	0.78	0.88
Miyazawa et al.	5.33	4.18	3.71	3.59	7.93	3.87	3.65	4.48	5.1	8.83

	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
Eisenberg et al.	1.06	-1.5	0.64	1.19	0.12	-0.18	-0.05	0.81	0.26	1.08
Hopp et al.	-1.8	3	-1.3	-2.5	0	0.3	-0.4	-3.4	-2.3	-1.5
Manavalan et al.	**14.9**	**11.36**	**14.39**	**14**	**11.37**	**11.23**	**11.69**	**13.93**	**13.42**	**15.71**
Black et al.	0.943	0.283	0.738	1	0.711	0.359	0.45	0.878	0.88	0.825
Fauchere et al.	1.7	-0.99	1.23	1.79	0.72	-0.04	0.26	2.25	0.96	1.22
Argos et al.	0.5	-1.8	0.4	0.5	-0.3	-0.1	-0.2	0.3	-0.4	0.6
Janin et al.	1.47	0.09	1.42	1.57	0.54	0.97	1.08	1	0.83	1.37
Tanford et al.	1.53	-1.5	0.64	1.19	0.12	-0.18	-0.05	0.81	0.26	1.8
Parker et al.	-9.2	5.7	-4.2	-9.2	2.1	6.5	5.2	-10	-1.9	-3.7
Rose et al.	0.85	0.52	0.85	0.88	0.64	0.66	0.7	0.85	0.76	0.86
Miyazawa et al.	8.47	2.95	8.95	9.03	3.87	4.09	4.49	7.66	5.89	7.63

Table S4: Two types of polarity scales for each amino acid.

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile
Zimmerman et al.	0	5.2	3.38	40.7	1.48	3.53	49.91	0	51.6	0.15
Grantham et al.	8.1	10.5	11.6	13	5.5	10.5	10.3	9	10.4	5.2

	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
Zimmerman et al.	0.45	49.5	1.43	0.35	1.58	1.67	1.66	2.1	1.61	0.13
Grantham et al.	4.9	11.3	5.7	5.2	8	9.2	8.6	5.4	6.2	5.9

Table S5: Two types of flexibility scales for each amino acid.

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile
Karplus et al.	-1.27	2.79	1.77	1.42	-1.09	1.18	1.6	1.86	-0.82	-2.89
Bhaskaran et al.	0.36	0.53	0.46	0.51	0.35	0.49	0.5	0.54	0.32	0.46

	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
Karplus et al.	-2.29	2.88	-1.84	-2.14	0.52	3	1.18	-3.78	-3.3	-1.75
Bhaskaran et al.	0.37	0.47	0.3	0.31	0.51	0.51	0.44	0.31	0.42	0.39

Table S6: The antigencity scale for each amino acid.

ala	arg	asn	asp	cys	gln	glu	gly	his	ile
1.064	0.873	0.776	0.866	1.412	1.015	0.851	0.874	1.015	1.152

leu	lys	met	phe	pro	ser	thr	trp	tyr	val
1.25	0.93	0.826	1.091	1.064	1.012	0.909	0.893	1.161	1.383

1.2.2 Statistical Scale

In this work, we present the statistical criterion to calculate the probability occurrence of each amino acid in epitope region. To clearly describe statistical scale, some examples are presented. As example, in the protein with PDB code 1FSK, the probability of k=3 number of epitope residues in n=14 samples of GLU in the protein with N=122 surface residues that containsexactly K=13 epitope residues on surface is obtained by:

$$P(X_{GLU} = 3) = \frac{\binom{13}{3}\binom{122-13}{14-3}}{\binom{122}{14}} = 0.1283.$$

In this manner, the probability of k=2 number of epitope residues in n=7 samples of ASN, in the protein with N=122 residues that contain exactly K=13 epitope residues on surface is obtained by:

$$P(X_{ASN} = 2) = \frac{\binom{13}{2}\binom{122-13}{7-2}}{\binom{122}{7}} = 0.1360.$$

Comparing them shows that the probability occurrence of ASN in epitope region is higher than the probability occurrence of GLU in this special antigen. Table S7 shows the weighted mean related to statistical scale which obtained for each amino acid in training dataset.

Table S7: The statistical scale for each amino acid.

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile
Statistical Scale	0.428	0.330	0.357	0.779	0.310	0.376	0.283	0.329	0.467	0.461

	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
Statistical Scale	0.361	0.275	0.688	0.503	0.336	0.325	0.355	0.687	0.452	0.4282

1.2.3 Structural Scale

Protrusion Index from a three-dimensional protein structure is another criterion to help us to predict epitope residues. We use the server in http://hydra.icgeb.trieste.it/cx/to calculate the Protrusion Index from 3D antigen structures. The output of this server calculates Protrusion Index for each Atom ofantigen separately.The Figure S1 is shown part of an output of this server for an Antigen-Antibody complex to calculate the protrusion index (PDB ID 3NH7).

As you can see in figure S1, the output file contains the Atom number, atom name, residue name, chain name, residue number, the Atom coordinates and finally the protrusion index for each atom.

```
ATOM      1  N   PRO A  34       8.939  41.917  26.646  1.00  2.47
ATOM      2  CA  PRO A  34       8.003  40.804  26.503  1.00  1.71
ATOM      3  C   PRO A  34       6.537  41.252  26.703  1.00  1.85
ATOM      4  O   PRO A  34       6.247  42.056  27.598  1.00  2.36
ATOM      5  CB  PRO A  34       8.445  39.816  27.603  1.00  1.60
ATOM      6  CG  PRO A  34       9.804  40.253  28.021  1.00  2.11
ATOM      7  CD  PRO A  34       9.833  41.736  27.801  1.00  2.86
ATOM      8  N   PHE A  35       5.628  40.708  25.881  1.00  1.29
ATOM      9  CA  PHE A  35       4.247  41.214  25.769  1.00  1.51
ATOM     10  C   PHE A  35       3.254  40.172  25.250  1.00  1.02
ATOM     11  O   PHE A  35       2.144  40.520  24.870  1.00  1.04
ATOM     12  CB  PHE A  35       4.240  42.389  24.804  1.00  1.74
ATOM     13  CG  PHE A  35       4.749  42.031  23.437  1.00  1.48
ATOM     14  CD1 PHE A  35       4.673  42.941  22.380  1.00  1.82
ATOM     15  CD2 PHE A  35       5.293  40.772  23.207  1.00  0.95
ATOM     16  CE1 PHE A  35       5.159  42.600  21.112  1.00  1.74
ATOM     17  CE2 PHE A  35       5.772  40.406  21.957  1.00  0.88
ATOM     18  CZ  PHE A  35       5.708  41.311  20.900  1.00  1.32
ATOM     19  N   LEU A  36       3.663  38.909  25.194  1.00  0.74
ATOM     20  CA  LEU A  36       2.776  37.843  24.742  1.00  0.50
ATOM     21  C   LEU A  36       2.553  36.802  25.844  1.00  0.38
ATOM     22  O   LEU A  36       3.504  36.413  26.519  1.00  0.38
ATOM     23  CB  LEU A  36       3.380  37.158  23.524  1.00  0.30
ATOM     24  CG  LEU A  36       2.400  36.638  22.462  1.00  0.32
ATOM     25  CD1 LEU A  36       2.935  35.347  21.865  1.00  0.18
ATOM     26  CD2 LEU A  36       0.996  36.425  23.011  1.00  0.36
ATOM     27  N   LYS A  37       1.310  36.363  26.032  1.00  0.37
```

Figure S1: The part of an output of cx server an Antigen-Antibody complex

1.3 Logistic Regression

In this survey, we also use the structural and statistical criteria which are modeled by using logistic regression. The output of SPSS (version 20.0.0) gives us the significance of each criterion which used in our work.

1.4 Markov Clustering Algorithm

We next describe the Markov Clustering (MCL) algorithm for clustering graphs, proposed by Stijn van Dongen, The Markov Cluster Process (abbreviated MCL process) defines a sequence of stochastic matrices by alternation of two operators on a generating matrix.To see how this works, Figure S1, the MCL is applied on weighted graph (figure S2.a), it causes the flows promotein dense

region and demote otherwise;the result of applying the MCL on the weighted graph is the figure S1.b which it shows its clusters.

Figure S2: the output of MCL on a weighted graph (PDB ID 3NH7)

	62	63	64	65	66	67	68	69	70	71	72
62	0	0	0.281	0.133	0	0	0	0	0	0	0.182
63	0	0	0	0	0	0	0	0.154	0.190	0.352	0.267
64	0.281	0	0	0.253	0.113	0	0	0.104	0	0.112	0.341
65	0.133	0	0.253	0	0.275	0.146	0.128	0.231	0.297	0.111	0.139
66	0	0	0.113	0.275	0	0.419	0.206	0.275	0.352	0.150	0
67	0	0	0	0.146	0.419	0	0.574	0.165	0.155	0	0
68	0	0	0	0.128	0.206	0.574	0	0.331	0.1362	0.125	0
69	0	0.154	0.104	0.231	0.275	0.165	0.331	0	0.486	0.444	0.131
70	0	0.190	0	0.297	0.352	0.155	0.136	0.486	0	0.531	0.156
71	0	0.352	0.112	0.111	0.150	0	0.125	0.444	0.531	0	0.328
72	0.182	0.267	0.341	0.139	0	0	0	0.131	0.156	0.328	0

Figure S2 (a): The weighted Graph

	62	63	64	65	66	67	68	69	70	71	72
62	0	0	0	0	0	0	0	0	0	0	0
63	0	0	0	0	0	0	0	0	0	0	0
64	0	0	0	0	0	0	0	0	0	0	0
65	0	0	0	0	0	0	0	0	0	0	0
66	0	0	0	0	0	0	0	0	0	0	0
67	0	0	0	1	1	1	1	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0
71	0	1	0	0	0	0	0	1	1	1	1
72	0	0	0	0	0	0	0	0	0	0	0

Figure S2(b): the output of MCL algorithm on weighted graph

2 User Manual

2.1 Introduction

LRC algorithm is a Matlab (version 2007b) code for predicting epitope residues. It implements algorithm for predicting epitope reigns on an antigen using a logistic regression model and Markov CLustering algorithm (MCL).The LRC algorithm is applied to predict the epitope residues of clusters which resulted from MCL algorithm. Softwares are in the form of MATLAB codes.

2.2 Installation

The Matlab codes are ready for use. This package contains the following programs:

1/ Antigen_Surface_Graph4: Construct antigen-surface-graph based on accessible surface area and Delaunay triangulation.

2/ weigthed_Gragh: A method for constructing weighted graph based on some representative criteria and logistic regression model.

3/ MCL: Cluster weighted graph by alternation of two operators. See [].

4/LRC: Predict epitope residue of antigen from PDB file. The method is based on filtering the clusters obtained by MCL.

2.3 Credits

This package was implemented by M.Habibi and is supported in part bya grant from
Iran's Institute for Research in Fundamental Sciences and by Qazvin Islamic Azad University.

2.4 Usage and Examples

- **Antigen_Surface_Graph4:**

A method to construct antigen-surface-graph based on accessible surface area and Delaunay triangulation.

Usage of Antigen_Surface_Graph4:
Before running this program must be calculate accessible surface area using GetAreapakage andcall surface_residues.

Parameters:

1. Sur_Atom,Sur_Res; obtain from"surface_residues" program for the txt file resulted by GetArea package.
2. AntigenName,chainID; are the PDB file name and antigen chain.
3. Method:takes the following value as example;

```
[Sur_Res,Sur_Atom]  =  surface_residues ('GetArea-ASA3NH7.txt');
Protein = Antigen_Surface_Graph4(Sur_Atom,Sur_Res,'3NH7.pdb', 'A')
```

4. output: The Protein structure contains a filed for each surface residues
   Example:

```
    Protein =
                Atom_resSeq1: [212x1 double]
                Atom_resName1: {212x1 cell}
Atom_X: [1x212 double]
Atom_Y: [1x212 double]
Atom_Z: [1x212 double]
Atom_resSeq2: [1x212 double]
                Atom_resName2: {1x212 cell}
SurfaceRes_HorizonTal: [71x1 double]
SurfaceRes_VerTal: [1x71 double]
```

Matrix: [71x71 double]

- **weigthed_Gragh**

A method for constructing weighted graph based on some representative criteria and logistic regression model.

Usage of weigthed_Gragh:
This Program calls the “LR” program to calculate the weight of each vertex.

Parameters:

1. X: is a matrix containsrepresentative criteria. Each line of this matrix represents a set of criteria for each residue which containing hydropobicity, antigenicity, flexibility, polarity, Turn & helix, protrusion index and statistical criterion.Loading X_3NH7 presents the representative criteria which obtained for each surface residue of the antigen with PDB code 3NH7.
2. Un_W_graph: the surface-graph contains the adjacency matrix of un-weighted graph. It is obtained by Antigen_Surface_Graph4 program in the ‘Protein.Matrix’ field.
3. Output_file: contains the weight of each vertex (W) obtained by LC program and adjacency matrix (W_graph).

Example: [W,W_graph] = weigthed_Gragh(Protein.Matrix,X);

- **MCL**

A method to cluster weighted graph by alternation of two operators.This program calls the "Normalization", "Inflation2" and "MCLCluster" programs to divide the weighted graph.

Parameters:

1. W_graph: is resulted from weigthed_Gragh program
2. Vertices: is resulted from Antigen_Surface_Graph4 program
3. n: set it to 30
4. Method: takes the following value as example:

5. Output: The struct array with two fiels; cluster and Residues.

Cluster =
1x15 struct array with fields:
cluster
Residues

- **LRC**

A method to predict epitope residue of antigen from PDB file. The method is based on filtering the clusters obtained by MCL.

Usage of LRC:
This Program calls the Antigen_Surface_Graph4, weigthed_Gragh and MCL program to cluster the surface antigen.

Parameters:

1. ASA: the name of txt file resulted by GetArea package..
2. AntigenName, chainID; are the PDB file name and antigen chain.
3. Output_file: contains the weight of each vertex (W) obtained by LC program and adjacency matrix (W_graph).
4. X: is a matrix contains representative criteria.

   Example: [E]=LRC(ASA,AntigenName,chainID,X);
   Or [E]=LRC('GetArea-ASA3NH7.txt','3NH7.pdb', 'A',X);

5. Output: is the binary vector E (epitope residues are condidered as 1 and non-epitope residues are considered as 0)