SCIENTIFIC REPORTS

OPEN

# Network Analysis Reveals the Recognition Mechanism for Dimer Formation of Bulb-type Lectins

Received: 16 January 2017
Accepted: 20 April 2017
Published online: 06 June 2017

Yunjie Zhao[1,3], Yiren Jian[3], Zhichao Liu[3], Hang Liu[4], Qin Liu[5], Chanyou Chen[5], Zhangyong Li[2], Lu Wang[2], H. Howie Huang[4] & Chen Zeng[2,3,5]

**The bulb-type lectins are proteins consist of three sequential beta-sheet subdomains that bind to specific carbohydrates to perform certain biological functions. The active states of most bulb-type lectins are dimeric and it is thus important to elucidate the short- and long-range recognition mechanism for this dimer formation. To do so, we perform comparative sequence analysis for the single- and double-domain bulb-type lectins abundant in plant genomes. In contrast to the dimer complex of two single-domain lectins formed via protein-protein interactions, the double-domain lectin fuses two single-domain proteins into one protein with a short linker and requires only short-range interactions because its two single domains are always in close proximity. Sequence analysis demonstrates that the highly variable but coevolving polar residues at the interface of dimeric bulb-type lectins are largely absent in the double-domain bulb-type lectins. Moreover, network analysis on bulb-type lectin proteins show that these same polar residues have high closeness scores and thus serve as hubs with strong connections to all other residues. Taken together, we propose a potential mechanism for this lectin complex formation where coevolving polar residues of high closeness are responsible for long-range recognition.**

Plant lectins are carbohydrate-binding proteins that are abundant in seeds, flowers, leaves, roots, and other vegetative non-storage tissues. These lectins are recognized as plant defense proteins because they can specifically target the surface glycan of the epithelial cells lining the intestinal tract of insects and some herbivores[1–5]. The harmful and toxic effects of glycan binding vary from slight discomfort to even death. As our understanding of lectin-carbohydrate interaction grows, the biological applications of lectins also become much more diverse[6]. Besides the anti-insect activity, the plant lectins were used as molecular tools to study host-pathogen interactions, cell development and signaling, and many others in biomedical applications[7–15].

Plant lectins have been classified into 12 families based on their sequences, fold structures, and carbohydrate binding motifs[16–19]. The most general features of plant lectins are as follows. (1) The carbohydrate binding domains are evolutionarily related; and (2) lectins typically form dimer or oligomer for their biological activities. However, the recognition mechanism for lectin dimer or oligomer formation remains poorly understood and is the subject of our study.

Such study is becoming feasible given the rich data sources on plant lectin families. First, the known structures of most lectin complex deposited in Protein Data Bank (PDB) enable molecular dynamic simulations and the associated correlation network analysis[16, 17]. Graph theory concepts such as betweenness and closeness can be brought to bear in identifying critical residues for complex formation. Second, there happen to be abundant single-domain and double-domain lectins in plant genomes for us to distinguish these critical residues in terms of whether they are for short-range or long-range recognitions. In the crowded environment of a cell, one single-domain lectin may need to find the other single-domain interacting partner at a distance to form a dimer. However, such a long-range recognition is no longer required in double-domain lectin where two single-domain

[1]Institute of Biophysics and Department of Physics, Central China Normal University, Wuhan, 430079, China. [2]Research Center of Biomedical Engineering, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China. [3]Department of Physics, The George Washington University, Washington, DC, 20052, USA. [4]Department of Electrical and Computer Engineering, The George Washington University, Washington, DC, 20052, USA. [5]School of Life Sciences, Jianghan University, Wuhan, 430056, China. Yunjie Zhao and Yiren Jian contributed equally to this work. Correspondence and requests for materials should be addressed to C.Z. (email: chenz@gwu.edu)

lectins are fused into one protein with a short linker and are always in close proximity. We thus perform statistical inference in terms of direct coupling analysis (DCA) on sequence evolution for single- and double-domain lectins to probe the conservation as well as coevolution of the putative critical residue pairs from multiple sequence alignments. These methods together allow us to uncover the protein-protein recognition mechanism.

In this article, we select the bulb-type lectins for detailed computational analysis to gain insights into lectin recognition mechanism[18–28]. The bulb-type mannose-binding lectin is a beta-prism type II structure. The single-domain or monomer protein contains antiparallel beta-strands with 3-fold symmetry. Two monomers can assemble into a dimer structure by inserting their C-terminal beta-strand tails into each other to form beta-sheets. This particular lectin can also form a double-domain fusion protein via a short linker between two single domains.

Here, we utilize the dynamical network analysis to investigate the structural characteristic of the bulb-type mannose binding protein[29–34]. The network analysis reveals that the polar residues on the surface with high closeness are responsible for the long-range recognition of dimer formation. This observation is further supported by the direct coupling analysis that shows coevolution of these polar residues in dimer complex but not in double-domain construct. Taken together, these results suggest a new scheme to identify critical residues for bulb-type lectin complex formation that may be reengineered for novel biomedical applications.

## Methods

**Molecular dynamics simulations.** The MD simulations were carried out using the GROMACS software package[35]. The AMBER03 force field[36] and TIP3P[37] water solvation model were used for the simulations. A water solvent box of 12 Å was created between the outside of the protein and the edge of the box. All the structures were simulated at the room temperature (300 K). The initial structure was extracted from the PDB database (PDB code: 1KJ1)[18] and solvated with water molecules in a periodic rectangular box with a normal saline condition. The SHAKE algorithm was used to constrain all bond lengths[38]. The long-range electrostatic interactions were treated with the Particle Mesh Ewald method[39]. The non-bonded (electrostatic and VDW) cutoff range was 8 Å. A time step of 2 fs was used for numerical integration. Before the MD simulation, the entire system was first minimized by steepest descent calculation for 1000 steps followed by 300 ps equilibration. For each state, three 30 ns trajectories were generated. The solvent accessible surface areas were calculated by GETAREA[40]. The interface area is defined as the accessible surface on each of the two partners that subsequently become inaccessible to solvent in their dimer formation. The structures were visualized and analyzed by VMD and PyMOL[41].

**Network construction.** A dynamical network was constructed by Carma package from the final 20 ns portion of the entire 30 ns trajectories[33, 42]. A node in the network denotes a single amino acid residue. Two non-consecutive residues in sequence are connected by an edge if they contain a pair of heavy atoms, one from each residue, less than 4.5 Å apart for at least 75% of the times during the MD simulation. The weight of the edge between two connected nodes *i* and *j* is defined as:

$$W_{ij} = -\log(|C_{ij}|) \tag{1}$$

with $C_{ij}$ measuring the correlation of motions of nodes *i* and *j*:

$$C_{ij} = \frac{\left\langle \Delta\vec{r}_i(t) \cdot \Delta\vec{r}_j(t) \right\rangle}{\left( \langle \Delta\vec{r}_i(t)^2 \rangle \langle \Delta\vec{r}_j(t)^2 \rangle \right)^{1/2}} \textit{and } \Delta\vec{r}_i(t) = \vec{r}_i(t) - \langle \vec{r}_i(t) \rangle \tag{2}$$

where $\vec{r}_i(t)$ is the position vector of the $C_\alpha$ atom of the $i^{th}$ amino acid and the brackets indicate the time average. The values of $C_{ij}$ vary from $-1$ to 1. Since we focused on the nodes moving together in the same direction, we removed the edges if their correlations were from $-1$ to 0.

**Networks analysis.** We analyzed the closeness, betweenness, characteristic path length (CPL), and delta path length (DPL) of the coarse-grained dynamical network where only $C_\alpha$ atoms of amino acids are used to construct the network. The closeness of a node is defined as the inverse of the sum of its shortest distances to all other nodes as the following:

$$C(x) = \frac{n - 1}{\sum d(x, y)} \tag{3}$$

where $d(x, y)$ is the distance of the shortest path between the node *x* and any other node *y*[43]. The betweenness of a node *x* measures its contribution toward the network communication by counting the number of shortest paths between all pairs of nodes that also pass through the node *x*. The CPL is the average length of the shortest paths between all pairs of nodes. The DPL of a node *x* is the change of CPL induced by removing the node *x*. The shortest paths between all pairs of nodes are found using the Floyd-Warshall algorithm.

**Sequence evolution analysis.** The sequence evolution analysis measures the residue conservation by ConSurf program[44]. First, we obtained the alignment files (PDB code: 1KJ1, for chain A and chain D) from ConSurf-DB[45]. To focus on the plant lectins, we filtered the sequences by manually removing all non-plant entries. The final numbers of sequences were 79 for chain A (Table S1) and 82 for chain D (Table S2), respectively. Then, we used the program ClustalW2 to perform the sequence alignment on the filtered sequences[46, 47]. Lastly, we calculated the residue conservation scores by ConSurf[44]. The continuous conservation scores are divided into a discrete scale of 9 grades with grade 1–3 for the most variable positions and grade 7–9 for the most conserved positions. The 46 single-domain and 16 double-domain sequences were obtained from the annotations of the homology sequences in UniProt[48].

**Sequence coevolution analysis.** The Direct Coupling Analysis (DCA) was performed to infer the interacting residues by using information on sequence coevolution across different species[49–52]. The single- and double-domain sequence alignments were listed in two files: Supplementary Info File 2 (SI 2) and Supplementary Info File 3 (SI 3). The MUSCLE[53] program was used to perform the sequence alignment. The main steps of DCA are as follows.

Step 1: the columns in multiple sequence alignment (MSA) showing more than 50% gaps are removed.

Step 2: amino acid frequencies for single residue $f_i(A_i)$ and a pair of residues $f_{ij}(A_i, A_j)$ are computed by reweighting the $M$ sequences in MSA based on sequence identity as the following:

$$f_i(A_i) = \frac{1}{\lambda + M_{eff}}\left(\frac{\lambda}{21} + \sum_{a=1}^{M}\frac{1}{m_a}\delta_{A_i, A_i^a}\right) \tag{4}$$

$$f_{ij}\left(A_i, A_j\right) = \frac{1}{\lambda + M_{eff}}\left(\frac{\lambda}{21^2} + \sum_{a=1}^{M}\frac{1}{m_a}\delta_{A_i, A_i^a}\delta_{A_j, A_j^a}\right) \tag{5}$$

Here $A_i$ ($A_j$) denotes what is at the $i^{th}$ ($j^{th}$) location of the sequence of length L, which can be one of the 21 possible choices including 20 actual amino acid types or a gap insertion in MSA. A pseudo-count $\lambda = 0.5$ is introduced to treat possible finite sample effect. $M_{eff} = \sum_{a=1}^{M} 1/m_a$ is the effective number of sequences where $m_a$ counts the number of sequences with more than 80% sequence identity to the $a^{th}$ sequence $\{A_1^a, \ldots, A_L^a\}$ in MSA.

Step 3: the model statistical probabilities of a single residue and a pair of residues in MSA are,

$$P_i(A_i) = \sum_{\{A_k | k \neq i\}} P(A_1, \ldots, A_L) \tag{6}$$

$$P_{ij}\left(A_i, A_j\right) = \sum_{\{A_k | k \neq i, j\}} P(A_1, \ldots, A_L) \tag{7}$$

where $P(A_1, \ldots, A_L)$ is the global probability of the sequence $\{A_1, \ldots A_L\}$. Using the maximum-entropy model, the global probability involves the residue-pair interaction energy (pairwise couplings) $e_{ij}(A_i, A_j)$ and local energy $h_i(A_i)$,

$$P(A_1, \ldots, A_L) = \exp\left\{\sum_{i<j} e_{ij}\left(A_i, A_j\right) + \sum_i h_i(A_i)\right\}/Z \tag{8}$$

the normalization factor was defined as $Z = \sum_{A_1, \ldots, A_L} \exp\{\sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i)\}$ Applying the mean-field approximation, the residue-pair interaction energy can be estimated by the inverse of the covariance matrix,

$$e_{ij}\left(A_i, A_j\right) = -C_{ij}\left(A_i, A_j\right)^{-1} \tag{9}$$

where the covariance matrix is $C_{ij}(A_i, A_j) = P_{ij}(A_i, A_j) - P_i(A_i)P_j(A_j)$. Since we want to fit the one-site and two-site marginal of $P(A_1, \ldots, A_L)$ to the empirical reweighted counting frequency $f_i(A_i)$ and $f_{ij}(A_i, A_j)$, we substitute $C_{ij}(A_i, A_j)$ in the above equation with $C_{ij}^{emp}\left(A_i, A_j\right) = f_{ij}\left(A_i, A_j\right) - f_i(A_i) f_j\left(A_j\right)$.

Step 4: the direct couplings are defined as

$$DI_{ij} = \sum_{AB} P_{ij}^d(A, B) \ln \frac{P_{ij}^d(A, B)}{f_i(A) f_j(B)} \tag{10}$$

with the help of an isolated two-site model

$$P_{ij}^d(A, B) = \exp\left\{e_{ij}(A, B) + \tilde{h}_i(A) + \tilde{h}_j(B)\right\}/Z_{ij} \tag{11}$$

$\tilde{h}_i(A)$ and $\tilde{h}_j(B)$ are defined by the empirical single-residue frequency $f_i(A) = \sum_B P_{ij}^d(A, B)$ and $f_j(B) = \sum_A P_{ij}^d(A, B)$.

## Results

**Interface and mannose binding residues are more conserved than other surface residues.** The tertiary structure of the garlic bulb-type lectin protein was extracted from PDB database with a resolution of 2.2 Å (PDB code: 1KJ1)[18]. To understand the structural characteristic, we divide the protein into surface sites, interface sites, mannose-binding sites, and interior sites. A residue is defined as interface if it is solvent exposed in monomer but not solvent exposed in dimer complex. The surface residues are the solvent exposed residues both in monomer and dimer complex (Fig. 1A, Table S3). The solvent accessible surface area of a residue was calculated using the solvent accessible surface recognition program GETAREA (Table S5)[40]. The mannose binding sites were identified by protein-ligand interaction recognition program LIGPLOT[54] (Fig. 1B, Table S4, Figure S2). We then

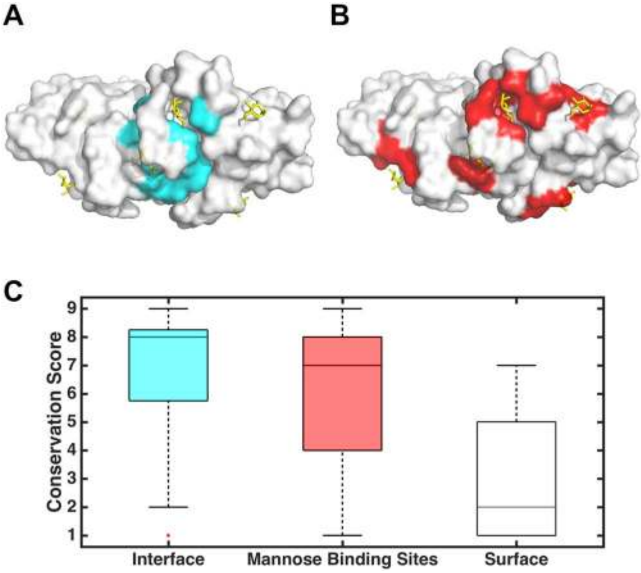

**Figure 1.** Classification of interface, mannose binding, and surface residues of the garlic mannose-binding lectin protein and their sequence conservation scores. (**A**) and (**B**) represent typical mapping of interface and mannose binding residues on a 3D structure, colored in cyan and red, respectively. (**C**) Distributions of average conservation scores of interface, mannose binding, and surface residues.

| Chain A | | Chain D | | Distance (Å) |
|---|---|---|---|---|
| Residue | Conservation | Residue | Conservation | |
| GLU91 | 2 | MET5 | 1 | 2.97 |
| ASN94 | 7 | THR105 | 9 | 3.34 |
| ASN94 | 7 | THR107 | 9 | 2.87 |
| TYR98 | 8 | TYR98 | 9 | 3.74 |
| TYR98 | 8 | ASP101 | 6 | 3.49 |
| GLY99 | 7 | GLY99 | 7 | 2.93 |
| GLY99 | 7 | ILE102 | 5 | 3.57 |
| ASP101 | 6 | TYR98 | 9 | 3.96 |
| ILE102 | 5 | GLY99 | 7 | 3.55 |
| SER104 | 8 | ASN94 | 8 | 3.99 |
| THR105 | 9 | ASN94 | 8 | 3.17 |
| THR107 | 9 | ASP92 | 8 | 3.80 |
| THR107 | 9 | ASN94 | 8 | 2.96 |

**Table 1.** Tertiary interactions within 4 Å at the interface. The interactions were calculated from the garlic mannose-binding lectin crystal structure (PDB code: 1KJ1).

performed sequence evolution analysis to investigate the sequence conservation of surface, interface, and mannose-binding residues in the dimer complex, respectively (see Materials and Methods for details). As shown in Fig. 1C, interface sites (average conservation score = 6.72, standard deviation = 2.46) and mannose binding sites (average conservation score = 6.18, standard deviation = 2.35) are significantly more conserved than the surface sites (average conservation score = 2.84, standard deviation = 1.90). It is believed that the interface residues tend to be more conserved across lectins for the stability of dimer formation while the slightly more varied mannose binding sites are for different mannose-binding specificity.

Table 1 lists all interacting pairs of the interface. A pair of amino acids across the interface is defined as interacting if the distance of any two heavy atoms, one from each amino acid, is less than 4 Å. As shown in Table 1, most interacting pairs are fairly conserved to maintain the interface stability. However, there is a highly variable pair of polar residues (A:GLU91-D:MET5). Ref. 53 already observed similar phenomena and suggested that such charged pairs should be for long-range steering effect in dimer formation. As discussed further below, a closeness score may provide a quantitative and practical measure to identify this polar pair. Moreover, our finding shows that this polar pair is largely absent in the double-domain fusion proteins. This offers the clearest evidence yet so far in support of this long-range recognition conjecture.

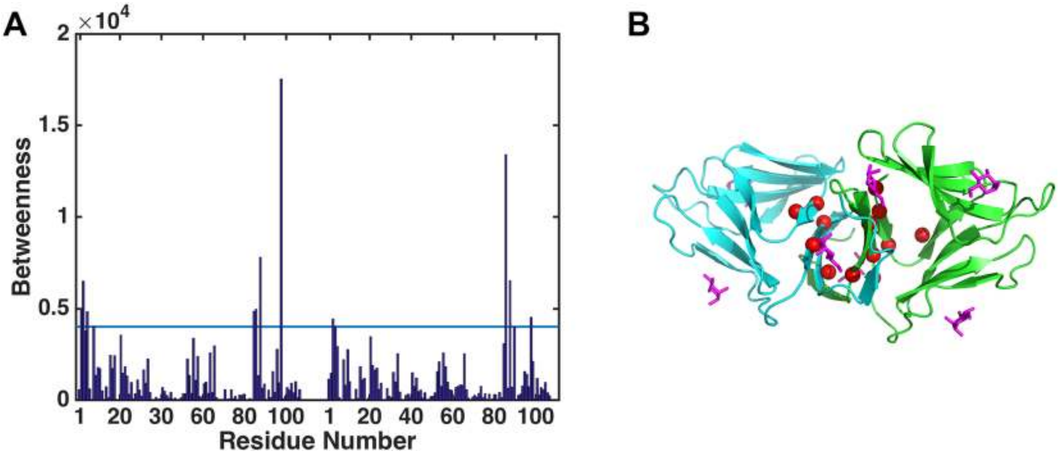

**Figure 2.** Betweenness centrality of residues in dynamical network. (**A**) ASN2, LEU3, THR5, GLU8, TYR85, VAL86, VAL88, TYR98 of chain A, and ILE3, LEU4, VAL86, VAL88, TYR98 of chain D have large betweenness with a Z-score greater than a cutoff value of 1.5 (light blue line). (**B**) The locations of the significant betweenness residues (red spheres). The chain A, chain D, and mannoses are colored in green, cyan, and magenta, respectively. The TYR98 with highest betweenness is the critical residue for intermolecular network communication.

**Network analysis reveals the critical residues for intermolecular communication.** The interface residues listed in Table 1 were obtained from the static structure of the dimer complex. It is necessary, however, to go beyond the static configuration to probe the importance of these residues in coordinating the dynamical motion of the entire complex. To this end, we performed the MD simulations and used the simulation trajectories of the complex to construct the dynamical network (see Materials and Methods for details). Given the network, graph theory concepts such as betweeness and delta path length (DPL) can be used to quantify the relative importance of each residue for the network communication between two monomers including some subtle allosteric effect.

First, we performed the betweenness calculation of the dynamic network. To probe the communication in this dynamical network, we identified the shortest path between each and every pair of nodes in the network and defined the betweenness of a node as the number of such shortest paths going through the node. Figure 2 shows the betweenness values in the entire dynamical network of the dimer complex with a Z-score value larger than 1.5. Most of the residues of high betweenness, especially TYR98, are located near interface indicating the importance of these residues in maintaining the correlated dynamics of the dimer complex. As such, it is no surprise that these residues are very conserved with conservation scores higher than 6.

Second, we performed the delta path length (DPL) calculation of the dynamical network. Since betweenness only considers the shortest path, it may overestimate the importance of a node in the network communication where there exist other paths of comparable length such as a very close but distinctive second shortest path. To overcome this potential pitfall, we also computed DPL of a node as the change of the average path length upon removal of the node (see Materials and Method). Figure 3 shows that most residues increase the path length upon their removal from the dynamic network. The values of betweenness and DPL share a high correlation of 0.845. The combined results of both metrics of betweenness and DPL suggest that the highly-conserved residue TYR98 is the critical residue for intermolecular network communication and dimer stability.

**Closeness analysis reveals the critical residues for long-range recognition.** In a connected graph, a node with small total distance to all other nodes acts as a hub for inter-network communication. To be precise, the closeness of a node is defined as the inverse of the sum of its shortest distances to all other nodes. It was proposed that the residues of high closeness are functional sites since, as network hubs, they can interact effectively with all other residues either directly or through a few intermediates. Indeed, previous benchmark tests showed that closeness scores successfully identified 70% of the protein active sites[43].

We further hypothesize that a pair of charged surface residues of high individual closeness value could best exert the long-range steering effect for dimer formation. Since each such residue forms a tighter connection with its own monomer, an attractive interaction for the pair can bring the two monomers together more effectively. To check this, we constructed the dynamic network from the MD simulations of the lectin monomer, and then computed the closeness values of all surface residues and classified them into three categories: (1) most likely recognition sites (high closeness values), (2) likely recognition sites (intermediate closeness values), and (3) unlikely recognition sites (small closeness values). Our results suggested that some residues (THR5, ASP17, and GLU19, colored in red) might be considered as the most likely recognition sites (Fig. 4). The crystal structure of garlic lectin showed that THR5 is responsible for dimer formation[18], and the crystal structure of snowdrop lectin indicated that ASP17/GLU19 might be responsible for tetramer formation[55]. Therefore, structural information on both monomer and complex and their associated network analysis supported our hypothesis that high closeness polar residues on surface might be responsible for long-range protein-protein recognition.

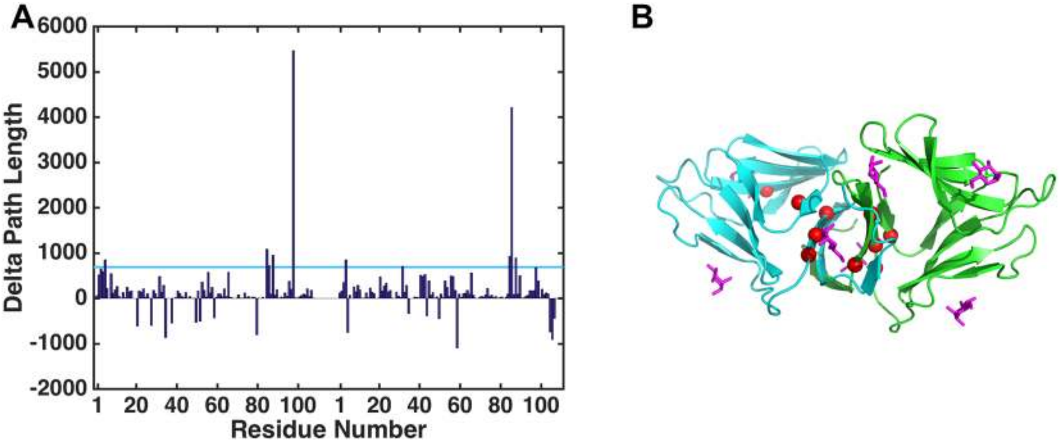

**Figure 3.** Delta path length of residues in dynamical network. (**A**) THR5, TYR85, VAL86, VAL88, TYR98 of chain A, and LEU4, VAL32, TYR85, VAL86, VAL88 of chain D have large delta path length with a Z-score greater than a cutoff value of 1.0 (light blue line). (**B**) The locations of the significant delta path length residues (red spheres). The chain A, chain D, and mannoses are colored in green, cyan, and magentas, respectively. The TYR98 with highest DPL is the critical residue for intermolecular network communication.

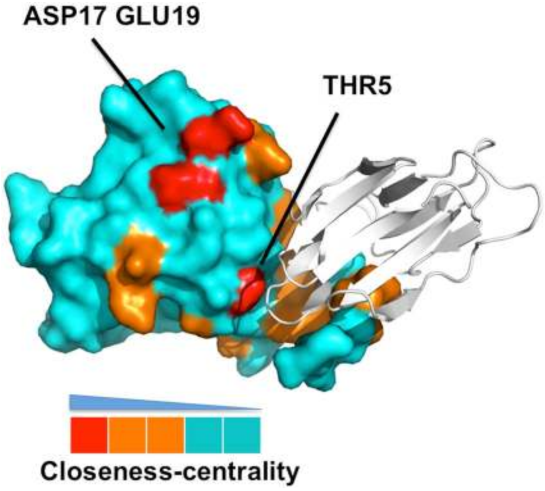

**Figure 4.** Surface representation of the garlic mannose-binding lectin protein (monomer, chain A). The residues are colored by closeness values with red, orange and cyan corresponding to the high (top 20%), intermediate (20~60%) and low (below than 60%) closeness values. The crystal structure of garlic lectin showed that THR5 is responsible for dimer formation and snowdrop lectin indicated that ASP17/GLU19 might be responsible for tetramer formation.

We analyzed additional representative plant homology sequences in both single- and double-domains of this specific kind of lectin (Supplementary Info 2 and 3). There are four such lectins with known crystal structures (Table S7)[56–58]. We compared the differences between single- and double-domain lectin proteins: (1) both single-domain lectin protein structures (PDB code: 1MSA and 3A0C) are similar to garlic lectin (PDB code: 1KJ1) with RMSDs around 1.5 Å, while double-domain protein structures (PDB code: 3MEZ and 3R0E) are different with RMSDs larger than 3 Å; (2) the residue-residue interaction of position 5 and position 91 is polar-polar interaction in single-domain lectin proteins but not in double-domain proteins; (3) the residues of position 5 and position 91 are surface residues in single-domain lectin proteins but not in double-domain lectin proteins; and (4) the residue closeness of position 5 is significantly high in single-domain but not in double-domain. These results supported our hypothesis that high closeness polar residues on surface of complex may responsible for long-range protein-protein recognition.

## Direct coupling analysis reveals that the polar pair is coevolving.

While it is possible to verify the importance of some residues toward the dimer formation via site mutagenesis, it is not clear how to measure if they are for short-range or long-range effect. Fortunately, the abundance of single- and double-domain lectins in plant genomes offers a unique opportunity to verify our hypothesis on long-range recognition mechanism.

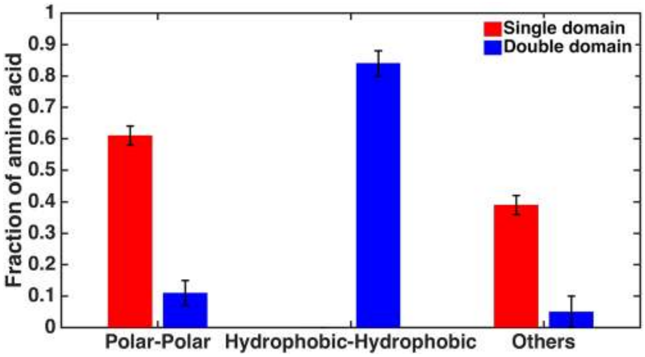

**Figure 5.** The analysis of interaction pattern between position 5 and 91 for single-domain and double-domain lectins in plant genomes. The red and blue colors are for the single-domain and double-domain lectins, respectively. We randomly select about 31 out of 46 single-domain sequences and 11 out of 16 double-domain sequences and analyze their interaction pattern, respectively. This is repeated five times. The residues at position 5 and 91 prefer to form polar-polar interaction pairs in single-domain lectins but hydrophobic-hydrophobic interaction pairs in double-domain lectins.

Unlike the dimeric lectin complex formed via protein-protein interaction of two single-domain lectins, the double-domain lectin is formed by fusing two single-domain lectins into one protein with a short linker and its two single domains are always in close proximity and thus do not require long-range recognition for its formation. Therefore, sequence features present in the single domain but absent in the double domain may be attributed to long-range effect. Specifically, we performed hydrophobic-polar pattern analysis and sequence coevolution analysis for both single- and double-domain lectins at position 91 and 5 to probe GLU91-MET5 pair.

For the hydrophobic-polar pattern analysis, we count the number of different types of interactions. In order to analyze the interaction pattern statistically, we randomly select 31 out of 46 single-domain sequences and 11 out of 16 double-domain sequences to compare the interaction patterns for single- and double-domain lectins, respectively. This calculation is repeated five times. The results indicate that the residues at position 5 and 91 prefer to form polar-polar interaction pairs in single-domain lectins but hydrophobic-hydrophobic interaction pairs in double-domain lectins. Figure 5 shows that dimer complex prefers polar interactions while such charged pair of GLU91-MET5 is largely absent in the double-domain lectins. This is strong evidence that GLU91-MET5 is indeed for long-range steering effect.

Direct Coupling Analysis (DCA) uses a global statistical model of multiple sequence alignments to infer direct interaction from coevolution of residue pairs[51, 59, 60]. We performed DCA for both single and double domains to identify the coevolving patterns for those residue pairs as displayed in Table 1. In previous coevolution analysis, the number of sequences used was typically comparable to the length of the target protein, and it was found that a DI score of 0.8 indicates a significant co-evolutionary signal[61, 62]. Therefore, we focused on the 18 residues at interface to perform the coevolution calculation due to the limited sequence information. The results indicate that less conserved residue pairs (with conservation scores less than 8 for both residues in the pair, Table S6) are coevolving in single-domain (with DI score greater than 0.8) but not in double-domain (with DI score less than 0.8). Specifically, the DI scores for GLU91-MET5 pair shows that there is strong correlation between GLU91 and MET5. The GLU91-MET5 pair is coevolved to maintain the interaction for long-range recognition in single-domain.

## Discussion and Conclusion

Protein-protein interactions are essential for carrying out various biological functions. Previous large-scale analysis of protein-protein interface of known complexes discovered a surprising pattern of highly variable and charged residue pairs at the interface. It was suggested that these pairs might provide the long-range steering force to bring together interacting proteins for dimer formation. Indeed, some mutagenesis experiments confirmed the importance of these charged pairs on protein surface for dimer formation and binding specificity[63–65].

The results from our case study on mannose-binding lectin complex are also consistent with this hypothesis. But beyond the qualitative description, we further proposed three practical and quantitative metrics to pinpoint such charged pairs for long-range recognition among a multitude of charged residues on protein surface without the complete structure of the dimer complex. Specifically, these charged pairs have the following unique features: (1) high closeness in the dynamical network of the monomer; (2) strong direct coupling indicating coevolution in the multiple sequence alignment; and (3) its absence in the double-domain construct. The last two measures above require sequence analysis only.

The identification of critical residue pairs for complex formation has many benefits. These pairs can serve as distance constraints to guide the structure modeling for much better accuracy. They can also facilitate drug design or protein engineering in order to regulate the complex formation for biological or medical applications.

In summary, we developed a hybrid approach of structure modeling, network analysis, and sequence statistical inference to identify critical residues for protein complex formation. Our results suggest that the coevolving polar residue pairs of high closeness initiate the long-range recognition of the bulb-type lectin complex formation that is further stabilized by short-range complementary interactions.

## References

1. Sharon, N. & Lis, H. History of lectins: from hemagglutinins to biological recognition molecules. *Glycobiology* **14**, 53R–62R (2004).
2. Kery, V. Lectin-carbohydrate interactions in immunoregulation. *Int J Biochem* **23**, 631–640 (1991).
3. Peumans, W. J. & Van Damme, E. J. Lectins as plant defense proteins. *Plant Physiol* **109**, 347–352 (1995).
4. Chrispeels, M. J. & Raikhel, N. V. Lectins, lectin genes, and their role in plant defense. *The Plant Cell* **3**, 1–9 (1991).
5. Vandenborre, G., Smagghe, G. & Van Damme, E. J. Plant lectins as defense proteins against phytophagous insects. *Phytochemistry* **72**, 1538–1550 (2011).
6. Lam, S. K. & Ng, T. B. Lectins: production and practical applications. *Appl Microbiol Biotechnol* **89**, 45–55 (2011).
7. Hirabayashi, J. Lectin-based structural glycomics: glycoproteomics and glycan profiling. *Glycoconj J* **21**, 35–40 (2004).
8. Paulson, J. C., Blixt, O. & Collins, B. E. Sweet spots in functional glycomics. *Nat Chem Biol* **2**, 238–248 (2006).
9. Fry, S., Afrough, B., Leathem, A. & Dwek, M. Lectin array-based strategies for identifying metastasis-associated changes in glycosylation. *Methods Mol Biol* **878**, 267–272 (2012).
10. Swanson, M. D., Winter, H. C., Goldstein, I. J. & Markovitz, D. M. A lectin isolated from bananas is a potent inhibitor of HIV replication. *J Biol Chem* **285**, 8646–8655 (2010).
11. Lam, S. K. & Ng, T. B. First report of a haemagglutinin-induced apoptotic pathway in breast cancer cells. *Biosci Rep* **30**, 307–317 (2010).
12. Souza, M. A., Carvalho, F. C., Ruas, L. P., Ricci-Azevedo, R. & Roque-Barreira, M. C. The immunomodulatory effect of plant lectins: a review with emphasis on ArtinM properties. *Glycoconj J* **30**, 641–657 (2013).
13. Liu, B., Bian, H. J. & Bao, J. K. Plant lectins: potential antineoplastic drugs from bench to clinic. *Cancer Lett* **287**, 1–12 (2010).
14. Jiang, Q. L. *et al.* Plant lectins, from ancient sugar-binding proteins to emerging anti-cancer drugs in apoptosis and autophagy. *Cell Prolif* **48**, 17–28 (2015).
15. Bies, C., Lehr, C. M. & Woodley, J. F. Lectin-mediated drug targeting: history and applications. *Adv Drug Deliv Rev* **56**, 425–435 (2004).
16. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
17. Rose, P. W. *et al.* The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res* **43**, D345–356 (2015).
18. Ramachandraiah, G., Chandra, N. R., Surolia, A. & Vijayan, M. Re-refinement using reprocessed data to improve the quality of the structure: a case study involving garlic lectin. *Acta Crystallogr D* **58**, 414–420 (2002).
19. Van Damme, E. J. *et al.* Phylogenetic and specificity studies of two-domain GNA-related lectins: generation of multispecificity through domain duplication and divergent evolution. *Biochem J* **404**, 51–61 (2007).
20. Ghequire, M. G., Loris, R. & De Mot, R. MMBL proteins: from lectin to bacteriocin. *Biochem Soc Trans* **40**, 1553–1559 (2012).
21. Milner, J. A. Garlic: its anticarcinogenic and antitumorigenic properties. *Nutr Rev* **54**, S82–86 (1996).
22. Raskin, I. *et al.* Plants and human health in the twenty-first century. *Trends Biotechnol* **20**, 522–531 (2002).
23. Banerjee, N. *et al.* Functional alteration of a dimeric insecticidal lectin to a monomeric antifungal protein correlated to its oligomeric status. *PloS one* **6**, e18593 (2011).
24. Clement, F. & Venkatesh, Y. P. Dietary garlic (Allium sativum) lectins, ASA I and ASA II, are highly stable and immunogenic. *Int Immunopharmacol* **10**, 1161–1169 (2010).
25. Clement, F., Pramod, S. N. & Venkatesh, Y. P. Identity of the immunomodulatory proteins from garlic (Allium sativum) with the major garlic lectins or agglutinins. *Int Immunopharmacol* **10**, 316–324 (2010).
26. Schafer, G. & Kaschula, C. H. The immunomodulation and anti-inflammatory effects of garlic organosulfur compounds in cancer chemoprevention. *Anticancer Agents Med Chem* **14**, 233–240 (2014).
27. Arreola, R. *et al.* Immunomodulation and anti-inflammatory effects of garlic compounds. *J Immunol Res* **2015**, 401630 (2015).
28. Karasaki, Y., Tsukamoto, S., Mizusaki, K., Sugiura, T. & Gotoh, S. A garlic lectin exerted an antitumor activity and induced apoptosis in human tumor cells. *Food Res Int* **34**, 7–13 (2001).
29. Zhao, Y. *et al.* Automated and fast building of three-dimensional RNA structures. *Sci Rep* **2**, 734 (2012).
30. Wang, J., Zhao, Y., Zhu, C. & Xiao, Y. 3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures. *Nucleic Acids Res* **43**, e63 (2015).
31. Zhao, Y., Zeng, C. & Massiah, M. A. Molecular dynamics simulation reveals insights into the mechanism of unfolding by the A130T/V mutations within the MID1 zinc-binding Bbox1 domain. *PloS one* **10**, e0124377 (2015).
32. Zhao, Y. *et al.* A new role for STAT3 as a regulator of chromatin topology. *Transcription* **4**, 227–231 (2013).
33. Sethi, A., Eargle, J., Black, A. A. & Luthey-Schulten, Z. Dynamical networks in tRNA:protein complexes. *Proc Natl Acad Sci USA* **106**, 6620–6625 (2009).
34. Chen, H. *et al.* Break CDK2/Cyclin E1 interface allosterically with small peptides. *PLoS One* **9**, e109154 (2014).
35. Van Der Spoel, D. *et al.* GROMACS: fast, flexible, and free. *Journal of computational chemistry* **26**, 1701–1718 (2005).
36. Duan, Y. *et al.* A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* **24**, 1999–2012 (2003).
37. Mahoney, M. W. & Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J Chem Phys* **112**, 8910–8922 (2000).
38. Jean-Paul Ryckaert, G. C. & Herman, J. C. B. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Chem. Phys.* **23**, 327–341 (1977).
39. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
40. Fraczkiewicz, R. & Braun, W. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *Journal of Computational Chemistry* **19**, 319–333 (1998).
41. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *Journal of molecular graphics* **14**, 33–38, 27–38 (1996).
42. Glykos, N. M. Software news and updates. Carma: a molecular dynamics analysis program. *Journal of computational chemistry* **27**, 1765–1768 (2006).
43. Amitai, G. *et al.* Network analysis of protein structures identifies functional residues. *J Mol Biol* **344**, 1135–1146 (2004).
44. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic acids research* **38**, W529–533 (2010).
45. Goldenberg, O., Erez, E., Nimrod, G. & Ben-Tal, N. The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic acids research* **37**, D323–327 (2009).
46. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
47. McWilliam, H. *et al.* Analysis Tool Web Services from the EMBL-EBI. *Nucleic acids research* **41**, W597–600 (2013).
48. UniProt, C. UniProt: a hub for protein information. *Nucleic acids research* **43**, D204–212 (2015).
49. Caleb Weinreb, A. J. R., John, B. I., Torsten, G. & Chris Sander, D. S. M. 3D RNA and Functional Interactions from Evolutionary Couplings. *Cell* **165**, 963–975 (2016).
50. De Leonardis, E. *et al.* Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res* **43**, 10444–10455 (2015).
51. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* **108**, E1293–1301 (2011).

52. Morcos, F., Hwa, T., Onuchic, J. N. & Weigt, M. Direct coupling analysis for protein contact prediction. *Methods in molecular biology* **1137**, 55–70 (2014).
53. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792–1797 (2004).
54. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng* **8**, 127–134 (1995).
55. Hester, G., Kaku, H., Goldstein, I. J. & Wright, C. S. Structure of mannose-specific snowdrop (Galanthus nivalis) lectin is representative of a new plant lectin family. *Nat Struct Biol* **2**, 472–479 (1995).
56. Wright, C. S., Kaku, H. & Goldstein, I. J. Crystallization and preliminary X-ray diffraction results of snowdrop (Galanthus nivalis) lectin. *J Biol Chem* **265**, 1676–1677 (1990).
57. Ding, J., Bao, J., Zhu, D., Zhang, Y. & Wang, D. C. Crystal structures of a novel anti-HIV mannose-binding lectin from Polygonatum cyrtonema Hua with unique ligand-binding property and super-structure. *J Struct Biol* **171**, 309–317 (2010).
58. Shetty, K. N., Bhat, G. G., Inamdar, S. R., Swamy, B. M. & Suguna, K. Crystal structure of a beta-prism II lectin from Remusatia vivipara. *Glycobiology* **22**, 56–69 (2012).
59. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci USA* **106**, 67–72 (2009).
60. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **3**, e02030 (2014).
61. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nat Biotechnol* **30**, 1072–1080 (2012).
62. Hopf, T. A. *et al*. Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3** (2014).
63. Zhang, Q., Zmasek, C. M. & Godzik, A. Domain architecture evolution of pattern-recognition receptors. *Immunogenetics* **62**, 263–272 (2010).
64. Wu, W., Ahlsen, G., Baker, D., Shapiro, L. & Zipursky, S. L. Complementary chimeric isoforms reveal Dscam1 binding specificity *in vivo*. *Neuron* **74**, 261–268 (2012).
65. Li, S. A., Cheng, L. N., Yu, Y. M. & Chen, Q. Structural basis of Dscam1 homodimerization: Insights into context constraint for protein recognition. *Sci Adv* **2** (2016).

## Acknowledgements

This work was partially supported by National Science Foundation (NSF) grant 0941228 (C.Z.), CAREER 1350766, grant 1618706 (H.H.) and Scientific Research Foundation of Central China Normal University 20205170045 (Y.Z.).

## Author Contributions

Y.Z. performed most computational analysis; Y.J. Z.L. performed network analysis with the help pf H.L. and H.H.; Q.L. C.C. Z.L. and L.W. helped with mannose-binding lectins sequences; Y.Z. and C.Z. supervised the overall study, analyzed the data and wrote the paper. All authors edited the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at doi:10.1038/s41598-017-03003-5

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.