

Discovering Interacting Domains and Motifs in Protein–Protein Interactions

Willy Hugo, Wing-Kin Sung, and See-Kiong Ng

Abstract

Many important biological processes, such as the signaling pathways, require protein–protein interactions (PPIs) that are designed for fast response to stimuli. These interactions are usually transient, easily formed, and disrupted, yet specific. Many of these transient interactions involve the binding of a protein domain to a short stretch (3–10) of amino acid residues, which can be characterized by a sequence pattern, i.e., *a short linear motif (SLiM)*. We call these interacting domains and motifs *domain–SLiM* interactions. Existing methods have focused on discovering SLiMs in the interacting proteins’ sequence data. With the recent increase in protein structures, we have a new opportunity to detect SLiMs directly from the proteins’ 3D structures instead of their linear sequences. In this chapter, we describe a computational method called SLiMDIet to directly detect SLiMs on domain interfaces extracted from 3D structures of PPIs. SLiMDIet comprises two steps: (1) interaction interfaces belonging to the same domain are extracted and grouped together using structural clustering and (2) the extracted interaction interfaces in each cluster are structurally aligned to extract the corresponding SLiM. Using SLiMDIet, de novo SLiMs interacting with protein domains can be computationally detected from structurally clustered domain–SLiM interactions for PFAM domains which have available 3D structures in the PDB database.

Key words: Short linear motifs, Protein structural mining, Domain–SLiM interactions, Protein–protein interactions

1. Introduction

Many protein–protein interactions (PPIs), such as those in signal transductions pathways, require fast response to stimuli. These interactions, also known as transient interactions, are designed to be easily formed and disrupted, and specific. While other PPIs are mediated by the binding of two large globular domain interfaces (domain–domain interactions), these transient interactions typically involve the binding of a protein domain to a short stretch (3–10) of amino acids (domain–motif interactions).

Many bioinformatics researchers have worked on discovering domain–domain interactions computationally. One of the earlier works is the InterDom database (1) created by detecting interacting domains from overabundant domain pairs in the protein sequence data of PPIs. With the increase in the availability of protein 3D structure data, researchers are able to detect domain–domain interactions directly from the PDB structural database (2); the databases in this line include iPFAM (3), 3DID (4), SCOPPI (5), and SNAPPI-DB (6).

Recently, researchers have found that in addition to domain–domain interactions, protein domains can recognize a second type of interaction motifs on other proteins called *short linear motifs* (SLiMs) (7–12). The SLiMs are short and degenerate, typically only 3–20 residues long containing just a few conserved positions. The SLiMs are often found to mediate PPIs that are specific but, at the same time, easily formed and disrupted. This type of interaction is found in many key biological pathways such as the signal transduction. Because of their small sizes, SLiM-based PPIs are good targets for small-molecule drug therapy, for it is easier to design drugs to mimic the SLiMs (13) than the larger structural motifs like domains. The listing of all known SLiMs to date can be found in databases like ELM (14) and MiniMotif (MnM) (15).

Experimental methods to find SLiMs include site-directed mutagenesis and phage display. These are tedious and expensive methods to apply on whole interactomes. As such, bioinformatics researchers have developed a number of computational methods for predicting SLiMs from other biological data. The current methods can be broadly classified into two approaches. The first approach mines motifs from a given set of *related* protein sequences, with the relations being established by prior biological knowledge such as similarity in known biological functions, similar localization to a certain cell compartment, or sharing of interaction partners. Methods in this class include DILIMOT (16), SLiMDisc (17), and SLiMfinder (18). The second approach is to mine SLiMs that are overrepresented in the available PPI data. Methods in this class include D-STAR (19), MotifCluster (20), and SLIDER (21). There are several drawbacks with these two approaches. First, the motifs identified via these sequence-based approaches are not guaranteed to occur on the binding interface. Such atomic level of details can only come from high-resolution 3D structures (22). Second, because SLiMs are highly degenerative, most of these algorithms masked conserved structured regions (which are assumed not to have many SLiMs) such as globular domains to reduce false positives. However, it was found that such filtering has caused true motifs to be missed (18). Third, the algorithms are highly dependent on the accuracy of the interaction identification experiments, but high-throughput PPI data are well known to be noisy (23).

Just as the development of domain–domain interaction detection methods has progressed from sequence-based into

structure-based approaches, the rapid increase of protein structure data in the PDB database also offers an excellent opportunity for detecting SLiMs directly from 3D structures instead of the proteins' sequences. The atomic level of details available in the high-resolution 3D structures are much richer than the linear protein sequences for discovering the weak signals of SLiMs, and the detected SLiMs are guaranteed to occur on the binding surfaces. In this chapter, we describe a method called SLiMDIet (24) to find SLiMs solely from 3D structure data. From the protein structure dataset downloaded from PDB on August 24, 2009, SLiMDIet was able to detect 452 distinct SLiMs on the domain interaction interfaces. One hundred and fifty-five of them were validated using the literatures, available structures, or statistical enrichment in the high-throughput PPI data. In addition, 198 SLiMs have been detected on domain–domain interaction interfaces (we call these *domain–domain SLiMs*), suggesting that the common belief that SLiMs occur outside the globular domain regions is not completely accurate, and that some of the apparent domain–domain interactions could in fact be mediated by domain–SLiM interactions.

2. Materials

1. *Protein 3D structure data.* The protein structure dataset can be downloaded from public databases such as the PDB. In the running example of this chapter, we used a protein 3D structure dataset downloaded from PDB on August 24, 2009, containing 57,559 structures. We chose structures with at least one protein chain and whose crystallographic resolution is 3.0 Å or better. We also included all NMR structures. In total, we have a working dataset of 54,981 structures with 130,488 protein chains.
2. *Protein domain annotations.* We compute PFAM domain annotations on each PDB chain using the *hmmpfam* program from the HMMER library version 2.3.2 (25) with the PFAM 23.0 library (26). We use PFAM as our choice of protein domain definition instead of the structurally defined SCOP (27) or CATH (28) because of the relatively better coverage of the former (see Note 1). However, PFAM domain does have its own limitation. It currently does not define structural domains that are formed by multiple protein chains. Nevertheless, SLiMDIet can also be applied on SCOP/CATH domain definitions if needed.
3. *PPIs.* To compute the statistical significance of the SLiMs detected by SLiMDIet, we compute their enrichment within a large nonhomologous PPI dataset which can be downloaded from online databases such as the BioGRID (29) (see Note 2).

3. Methods

SLiMDIet consists of two main steps: a DIet step (*Domain Interface extraction and clustering step*), followed by a SLiM step (*SLiM extraction step*):

1. The DIet step takes a set of protein structures from PDB as input, finds all known domains within the input structures, and extracts the domain interfaces associated with each of them. A domain interface comprises two sets of amino acid residues: one found within a protein domain (the set is called the *domain face*) while the other on a partner chain (the *partner face*) that are in close vicinity of each other. The interaction interfaces of each domain are then clustered based on their structural similarity.
2. In the SLiM step, we conduct an approximate structural multiple alignment to align the domain faces and the partner faces in each cluster. We then check if the alignment of the partner faces contains any conserved linear region (called a “block”) of an appropriate length. If so, we construct a (linear) gapped position-specific scoring matrix (PSSM) from the block to represent the detected SLiM.

The details of each step are given as follows (see also Fig. 1).

3.1. DIet Step: Domain Interface Extraction and Clustering

3.1.1. Domain Interface Extraction

1. For each PDB structure, we use the HMMER program to identify the PFAM domains in its chains. Regions with overlapping domain annotations are resolved by choosing the PFAM domain with the best HMMER *E*-value.
2. For all possible domain interfaces, we retain those in which each amino acid on the domain face is within a distance threshold of 5 Å (as done in PSIMAP (30)) from some amino acid on the partner face and vice versa. We define the distance between two amino acid residues to be the nearest distance between any pair of non-hydrogen atoms in the two residues.
3. To curb possible nonbiological (crystal) interfaces, which generally have smaller interface area, we set a threshold of having domain interfaces involving a minimum of *eight amino acids on the domain face* and *four amino acids on the partner face*. This lower bound corresponds to a binding area larger than 800 Å²—the average size of a domain interface as given in (5).
4. For intrachain domain interfaces, we also require that the residues on the partner face are at least *ten residues* from the ends of the domain (see Note 3).

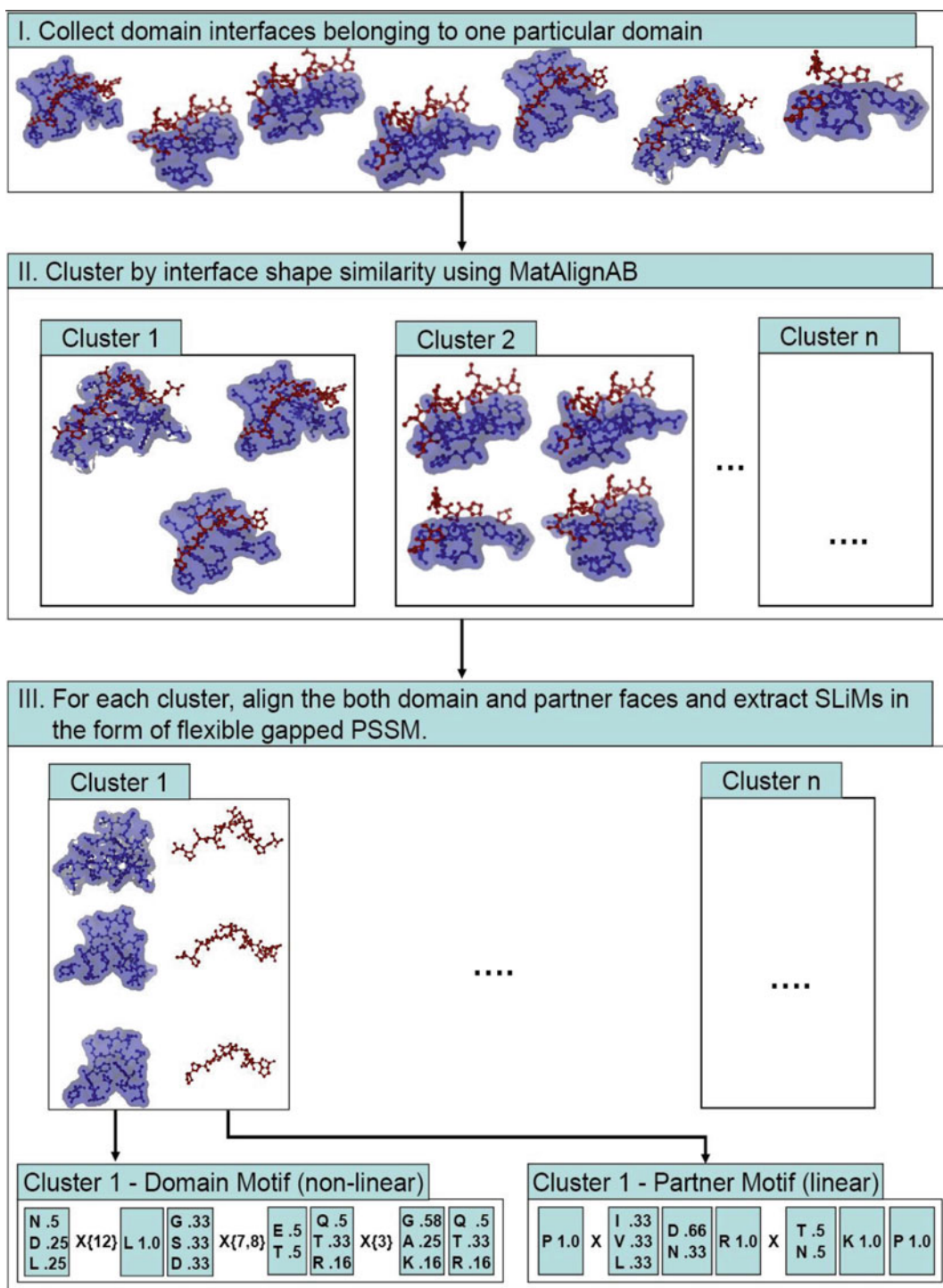


Fig. 1. SLiMDlet's overview. The domain interfaces of each PFAM domain are clustered by their structural similarity. Next, from each cluster, the domain and partner faces are structurally aligned and we build a gapped PSSM based on the contacts on the partner faces. The gapped PSSM has flexible gaps defined by the minimum and maximum gaps observed between two PSSM positions. We define a gapped PSSM as linear when the total length of its non-gap positions is 3–20 residues with gaps of at most four residues between any consecutive residue positions. To detect domain–SLiM interfaces, we collect domain interface clusters whose partner faces are covered by a linear gapped PSSM.

3.1.2. Pairwise Structural Alignment

1. Next, we compute the similarity scores and pairwise alignments among all pairs of domain interfaces of each PFAM domain in our dataset.
2. Alignment of two domain interfaces is done by treating each interface (both the domain and partner face) as one rigid body. Moreover, we enforce the alignment of the domain face residues on one interface against the domain face residues on the other, and do the same for the partner face residues (see Note 4).
3. We define the similarity of two interfaces using the modified S-score function by Alexandrov and Fischer (31) as follows: $S_{\text{norm}} = \frac{1}{(1+\Delta)} \frac{N}{\min(|A|, |B|)}$ where Δ is the root mean square distance (RMSD) between the two structures being aligned, N is the number of aligned residues between the two interfaces, and $|A|$ and $|B|$ are the sizes of the aligned interfaces, respectively. The similarity of two interfaces is measured using both the backbone and side chain conformation of the residues on each interface (see Note 5). To this end, we designed MatAlignAB for comparing domain interfaces' both C α and C β atoms, based on the MatAlign algorithm (32).

3.1.3. Hierarchical Agglomerative Clustering of the Domain Interfaces

1. For every domain, we cluster its interfaces using a hierarchical agglomerative clustering algorithm using average linkage (see Note 6) as follows.
2. We start by setting every domain interface as a trivial cluster with one member.
3. Next, we pick the pair of clusters which has the highest similarity and combine them into a new cluster. We compute the average similarity of the newly combined cluster.
4. We repeat the above step until the similarity score between every possible pair of the clusters is below a certain threshold (for threshold setting, see Note 7).

3.2. SLiM Step: SLiM Extraction

3.2.1. Multiple Alignment of the Partner Faces in a Cluster

1. For each resulting domain interface cluster, we choose the interface with the least average distance to all other cluster member as the cluster center.
2. To generate an approximate multiple alignment of the partner faces, we align the partner faces from the interfaces in the cluster to the cluster center's partner face. We keep only the alignments that contain at least four nonhomologous partner faces. A face f_a is defined as homologous to f_b when (1) f_a 's and f_b 's aligned residues in the alignment are exactly the same and (2) their full protein chains share more than 50% sequence similarity.
3. We also make sure that each alignment column has at least 50% occupancy, i.e., the number of nonempty residues aligned in each column (see Note 8) must be at least half of the number of nonhomologous interfaces aligned.

3.2.2. SLiM Extraction from the Longest Linear Block

1. We identify the longest linear block in each of the above alignments of the nonhomologous faces (see also Fig. 2 for an example). A linear block is defined as a set of 3–12 consecutive alignment positions with gaps of at most four residues.
2. We also require that the block must cover at least half of the partner faces in the alignment. A block is said to cover a partner face f_a when it includes at least half of the contact residues in f_a .
3. From the longest linear block thus identified, we construct a gapped PSSM (i.e., a PSSM with flexible gaps) to represent the SLiM recognized in the particular domain interface cluster. The (flexible) gap in between each alignment column is computed by taking the minimum and maximum gap observed between two residue positions.
4. Given that our multiple alignment is derived from limited structural data, we do not directly score a residue with its observed frequency in the alignment. Instead, we define the score of a residue X on the alignment column i by

$$\text{GappedPSSM}(i, X) = \ln \left(\sum_{AA \in \text{Res}(i)} \text{freq}_i(AA) \cdot e^{\text{BLOSUM}(X, AA)} \right),$$

where $\text{Res}(i)$ is the set of amino acids seen in the column i of the alignment and $\text{freq}_i(AA)$ is the frequency of residue AA in column i . Basically, the equation computes the weighted combination of the BLOSUM62 substitution score (33) of any residue X against the residues observed in the alignment—it extrapolates the feasibility of having other residues in that position based on the BLOSUM62 substitution matrix. An illustration of gapped PSSM construction can be seen in Fig. 3.

3.2.3. Computing the Statistical Significance of the SLiM Using PPI Data

1. For each SLiM extracted from an interface cluster, we also verify whether the motif occurs significantly more in the interaction partners of the domain as compared to random PPIs.
2. We define the gapped PSSM score of a particular position j in a given protein sequence S as the maximum sum of the gapped PSSM's residue scores starting at j over all possible gap value in the PSSM (see Note 9 for an example).
3. We define a position j in a protein with a gapped PSSM score s as an *occurrence* of the PSSM if the probability of scoring j with s or higher in a set of random protein sequence is at most 0.0001 (see Note 10).
4. Given a SLiM's gapped PSSM and a set of PPI data, the probability of observing a certain number of occurrences in the interaction partners of a protein domain by random can be

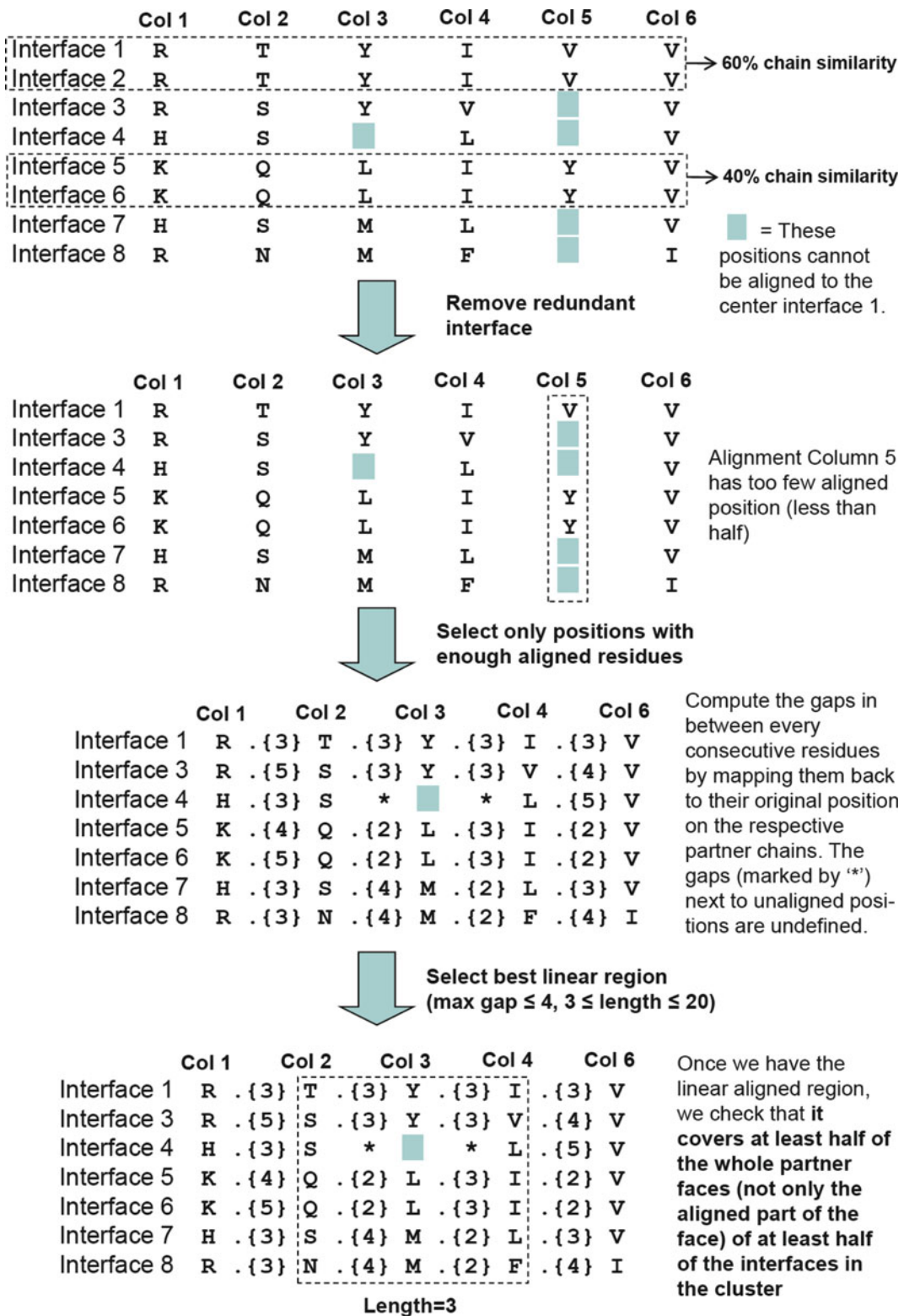


Fig. 2. Partner face alignment steps for finding the longest linear block. The latter is where we extract the SLiM from.

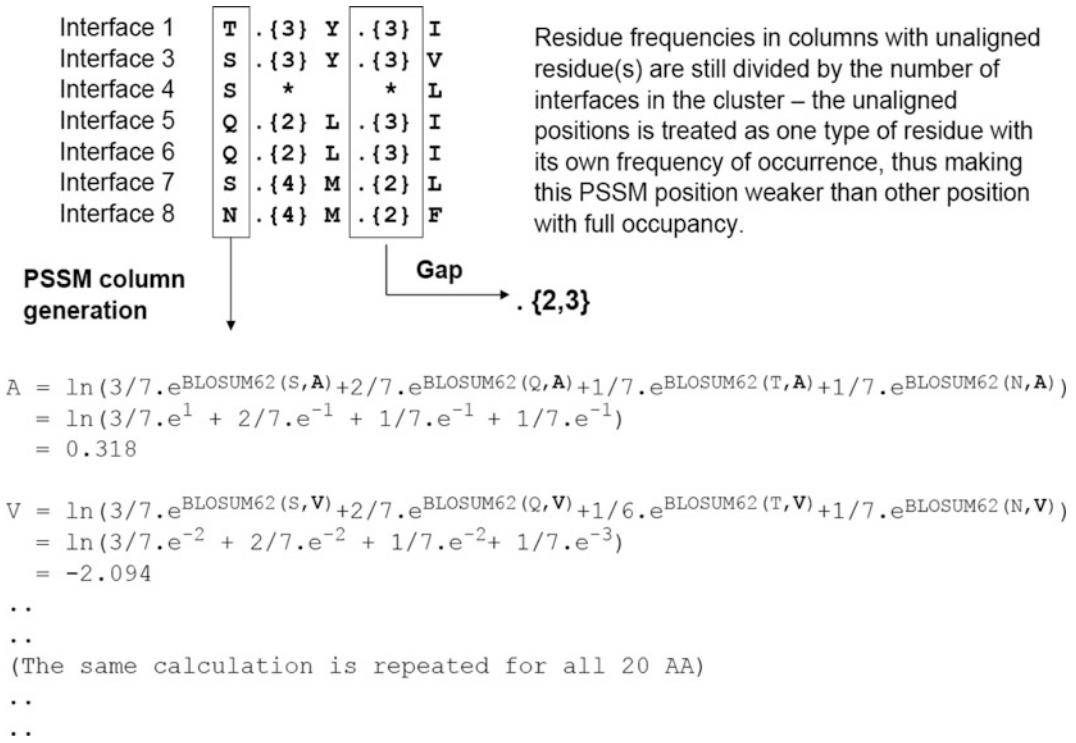


Fig. 3. An illustration of SLiMDlet's gapped PSSM generation from a linear block computed from the multiple interface alignment.

computed by the standard hypergeometric distribution function:

$$P\text{-Value}(I, I_D, I_M, I_{DM}) = \frac{\binom{|I_M|}{|I_{DM}|} \binom{(|I|-|I_M|)}{(|I_D|-|I_{DM}|)}}{\binom{|I|}{|I_D|}},$$

where I is the whole set of the high-throughput PPI data, I_M is the subset of I which contains an occurrence of the gapped PSSM M , I_D is the subset of I containing the domain D , and I_{DM} is the subset of I_D which contains an instance of M in the domain D 's interaction partners. SLiMs with P -value ≤ 0.05 are considered to be enriched in the PPI data and reported as detected SLiMs.

4. Notes

1. PFAM has higher PDB chain coverage on the current dataset [it covers 112,424 chains (86.16% coverage)] as compared to SCOP [version 1.75, dated June 2009, covering 87,064 chains (66.72% coverage)] and CATH [version 3.2.0, dated July 2008, covering 86,105 chains (65.99% coverage)].

2. We collected a set of 181,997 nonhomologous PPI data from the BioGRID interaction database version 2.0.58. We removed genetic (nonphysical) interactions (as defined by BioGRID) and those derived directly from structural data (to avoid self-discovery). Non-homology is enforced by keeping only one interaction among those whose both proteins are at least 70% homologous to another pair(s) of interacting proteins. The PPIs are collected as ordered tuples, i.e., for a given pair of interacting protein A and B , the tuple (A, B) is distinct from (B, A) . From each tuple, we collect the domain face from the left element and the partner face from the right one.
3. The requirement is important in order to avoid recognizing local secondary structures at the end of a domain as interface contacts. A similar filtering is also used by Stein and Aloy (34) (published shortly after SLiMDIet).
4. We do not align the structures of entire domains as done in SCOPPI (5) and SNAPPI-DB (6), which greatly reduces computation time. By considering both domain and partner face as one rigid body, we also avoid the need of considering the relative orientation between the domain and partner face.
5. Usually, the RMSD between two proteins is approximated only by the RMSD of their backbone's $C\alpha$ atoms. Since SLiMDIet's domain interfaces only consist of the contact residues (instead of the whole domain), the $C\alpha$ representation is rather inadequate. We use the $C\beta$ atom position as a first-order approximation of the side chain with respect to its backbone $C\alpha$ (a similar $C\beta$ approximation was mentioned in (35)).
6. The similarity of two clusters is the average pairwise similarity between all the members of the two clusters (as done in (5)).
7. We use the multiple thresholds, 0.15, 0.2, 0.25, and 0.3, to generate different sets of (possibly overlapping) domain interface clusters. Those clusters which originate from different thresholds but have more than 70% overlapping cluster member are grouped and SLiMDIet only reports the one with the most stringent cutoff as the representative of the group.
8. Some alignment column may have empty residues because the pairwise structural alignment may not align a residue from a particular partner face to the cluster center's residue when these residues' 3D positions are too different.
9. For example, the best score of position 0 in the string *FSDTK* based on the gapped PSSM $\begin{bmatrix} L : 4.62 \\ F : 1.38 \end{bmatrix} \cdot \{1, 2\} \begin{bmatrix} T : 2.4 \\ D : -0.12 \end{bmatrix}$ would be $\max \begin{cases} 1.38 + (-0.12) & (\text{gap} = 1), \\ 1.38 + 2.4 & (\text{gap} = 2) \end{cases}$.

Note that this is a mini-version of a gapped PSSM for exemplary purpose; the real gapped PSSM would have entries for all 20 amino acids.

10. We created a set of 10,000 random protein sequences, each of length 500, whose amino acid distribution follows the distribution observed in our PPI data (BioGRID 2.0.58). For each gapped PSSM, we compute its scores on all protein positions in the random dataset (of approximately five million positions) and sort the scores in nonincreasing order. The 500th score on the sorted score list would have an empirical *P*-value of 0.0001 and is chosen as the cutoff score for the gapped PSSM's occurrence.

References

1. Ng SK et al (2003) InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res* 31:251–254
2. Berman HM et al (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235–242
3. Finn RD, Marshall M, Bateman A (2005) iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* 21(3):410–412
4. Stein A, Céol A, Aloy P (2011) 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res* 39:D718–D723
5. Kim WK et al (2006) The many faces of protein-protein interactions: a compendium of interface geometry. *PLoS Comput Biol* 2(9):e124
6. Jefferson ER et al (2007) SNAPPI-DB: a database and API of structures, iNterfaces and alignments for protein-protein interactions. *Nucleic Acids Res* 35(Database Issue):D580–D589
7. Pawson T, Scott JD (1997) Signaling through scaffold, anchoring, and adaptor proteins. *Science* 278(5346):2075–2080
8. Sudol M (1998) From Src homology domains to other signaling modules: proposal of the ‘protein recognition code’. *Oncogene* 17:1469–1474
9. Neduva V, Russell RB (2005) Linear motifs: evolutionary interaction switches. *FEBS Lett* 579(15):3342–3345
10. Neduva V, Russell RB (2006) Peptides mediating interaction networks: new leads at last. *Curr Opin Biotechnol* 17(5):465–471
11. Diella F et al (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front Biosci* 13:6580–6603
12. Fox-Erlich S, Schiller MR, Gryk MR (2009) Structural conservation of a short, functional, peptide-sequence motif. *Front Biosci* 14:1143–1151
13. Vagner J, Qu H, Hrubby VJ (2008) Peptidomimetics, a synthetic tool of drug discovery. *Curr Opin Chem Biol* 12:1–5
14. Puntervoll P et al (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res* 31(13):3625–3630
15. Rajasekaran S et al (2009) Minimoto miner 2nd release: a database and web system for motif search. *Nucleic Acids Res* 37(Database Issue):D185–D190
16. Neduva V et al (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol* 3(12):e405
17. Davey NE, Shields DC, Edwards RJ (2006) SLIMDisc: short, linear motif discovery, correcting for common evolutionary descent. *Nucleic Acids Res* 34(12):3546–3554
18. Edwards RJ, Davey NE, Shields DC (2007) SLIMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS One* 2(10):e967
19. Tan SH et al (2006) A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC Bioinformatics* 7:502
20. Leung HC et al (2009) Clustering-based approach for predicting motif pairs from protein interaction data. *J Bioinform Comput Biol* 7(4):701–716
21. Boyen P et al (2009) SLIDER: mining correlated motifs in protein-protein interaction networks. In: *Proceedings of the 2009 ninth IEEE*

- international conference on data mining (ICDM) Miami, FL, USA, on December 6–9, pp. 716–721
22. Aloy P, Russell RB (2006) Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol* 7:188–197
 23. von Mering C et al (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417(6887):399–403
 24. Hugo W et al (2010) SLiM on Diet: finding short linear motifs on domain interaction interfaces in Protein Data Bank. *Bioinformatics* 26(8):1036–1042
 25. Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14:755–763
 26. Finn RD et al (2008) The Pfam protein families database. *Nucleic Acids Res* 36(Database Issue):D281–D288
 27. Andreeva A et al (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* 36(Database Issue):419–425
 28. Cuff AL et al (2009) The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in super-families. *Nucleic Acids Res* 37(Database Issue): D310–D314
 29. Stark C et al (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res* 39(Database Issue):D698–D704
 30. Dafas P et al (2004) Using convex hulls to extract interaction interfaces from known structures. *Bioinformatics* 20(10):1486–1490
 31. Alexandrov NN, Fischer D (1996) Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures. *Proteins* 25(3):354–365
 32. Aung Z, Tan K (2006) MatAlign: precise protein structure comparison by matrix alignment. *J Bioinform Comput Biol* 4(6):1197–1216
 33. Henikoff S, Henikoff JG (2005) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89(22):10915–10919
 34. Stein A, Aloy P (2010) Novel peptide-mediated interactions derived from high-resolution 3-dimensional structures. *PLoS Comput Biol* 6(5):e1000789
 35. Torrance JW et al (2005) Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *J Mol Biol* 347(3):565–581

Data Mining for Systems Biology

Methods and Protocols

Mamitsuka, H.; DeLisi, C.; Kanehisa, M. (Eds.)

2013, XII, 279 p., Hardcover

ISBN: 978-1-62703-106-6

A product of Humana Press