BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 8 2010, pages 1036–1042
doi:10.1093/bioinformatics/btq065

Structural bioinformatics

Advance Access publication February 18, 2010

SLiM on Diet: finding short linear motifs on domain interaction interfaces in Protein Data Bank

Willy Hugo[1,2], Fushan Song[1], Zeyar Aung[2], See-Kiong Ng[2] and Wing-Kin Sung[1,3,*]

[1]Department of Computer Science, National University of Singapore, 13 Computing Drive S(117417), [2]Data Mining Department, Institute for Infocomm Research, 1 Fusionopolis Way, S(138632) and [3]Department of Information and Mathematical Science, Genome Institute of Singapore, 60 Biopolis Street, S(138672), Singapore

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: An important class of protein interactions involves the binding of a protein's domain to a short linear motif (SLiM) on its interacting partner. Extracting such motifs, either experimentally or computationally, is challenging because of their weak binding and high degree of degeneracy. Recent rapid increase of available protein structures provides an excellent opportunity to study SLiMs directly from their 3D structures.

Results: Using domain interface extraction (Diet), we characterized 452 distinct SLiMs from the Protein Data Bank (PDB), of which 155 are validated in varying degrees—40 have literature validation, 54 are supported by at least one domain–peptide structural instance, and another 61 have overrepresentation in high-throughput PPI data. We further observed that the lacklustre coverage of existing computational SLiM detection methods could be due to the common assumption that most SLiMs occur outside globular domain regions. 198 of 452 SLiM that we reported are actually found on domain–domain interface; some of them are implicated in autoimmune and neurodegenerative diseases. We suggest that these SLiMs would be useful for designing inhibitors against the pathogenic protein complexes underlying these diseases. Our findings show that 3D structure-based SLiM detection algorithms can provide a more complete coverage of SLiM-mediated protein interactions than current sequence-based approaches.

Contact: ksung@comp.nus.edu.sg

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 11, 2009; revised on February 6, 2010; accepted on February 13, 2010

1 INTRODUCTION

Protein–protein interactions are vital to virtually every biological process. Some important biological processes, such as the signaling pathways, require protein–protein interactions that are designed for fast response to stimuli. These interactions are usually transient, easily formed and disrupted, and specific. While other protein–protein interactions 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 acid residues, which is usually characterized by a simple sequence pattern, i.e. *a short linear motif (SLiM)*. We call these *domain–SLiM* interactions. Numerous well-known and biologically important domains such as SH2, SH3, WW, 14-3-3, FHA and PDZ (Puntervoll *et al.*, 2003) have been found to interact with their partners via domain–SLiM interactions. The small binding areas on the SLiMs result in weak binding affinity (Neduva and Russell, 2005) that makes them suitable for mediating the transient interactions (Pawson and Nash, 2003). Compared with the larger domain–domain interaction interfaces, domain–SLiM interfaces are also better candidates for intervention by small molecules (Neduva and Russell, 2006).

However, current wet lab experiments for detecting SLiMs are laborious and time consuming. It is also a challenge to detect these motifs *in silico*, due to their short length and highly degenerative nature (Neduva and Russell, 2005). One popular approach is to mine SLiMs that are overrepresented in a set of non-homologous proteins known to be interacting with a particular protein/domain or known to share a similar biological function [e.g. DILIMOT (Neduva *et al.*, 2005) and SLiMFinder (Edwards *et al.*, 2007)]. Another line of computational approach finds SLiMs from sets of densely interacting protein pairs, for example, the work of Li *et al.* (2006) and D-STAR (Tan *et al.*, 2006). There are several drawbacks with these approaches. First, as the SLiMs are highly degenerative, most of these algorithms mask conserved structured regions (which are assumed not to have many SLiMs) such as globular domains to reduce false positives. Recently, it was found that such filtering has caused some true motifs to be missed (Edwards *et al.*, 2007). Second, the motifs identified via the 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 (Aloy and Russell, 2006). Third, the algorithms are highly dependent on the accuracy of the interaction identification experiments. However, these interaction data are well known to be noisy (von Mering *et al.*, 2002).

The rapid increase of protein structure data in the Protein Data Bank (PDB) database (Berman *et al.*, 2000) offers an excellent opportunity to detect SLiMs directly from 3D structures instead of the proteins' sequences. Some researchers have begun to exploit the structural data by using the structures as templates to find seed binding motifs, which are subsequently enriched using the available Protein-Protein Interaction (PPI) data (Betel *et al.*, 2007). They therefore suffer from the accuracy and coverage limitations of the PPI data like the previous methods. In this work, we directly find *de*

*To whom correspondence should be addressed.

novo SLiMs on domain interfaces extracted from 3D structures of protein–protein interactions (domain interface extraction or Diet). The SLiMs are extracted from structurally clustered domain–SLiM interaction classes for all PFAM domains that have available structures in the PDB database.

Our SLiMDIet method comprises two steps: (i) *Domain interface clustering:* interaction interfaces belonging to the same domain are grouped together and classified using structural clustering; and (ii) *SLiM extraction:* interaction interfaces in each domain interface cluster are structurally aligned and the corresponding SLiM is extracted from the alignment. We reported 452 distinct SLiMs found on the domain interaction interfaces where 40 of them are known in the literature, 54 have at least one supporting *domain–short peptide structure* (a PDB structure which shows that a single short peptide instance of the SLiM is sufficient for binding the protein domain) and another 61 SLiMs are found to be overrepresented in the PPI data collected from the BioGRID (Breitkreutz *et al.*, 2008).

Our data also revealed that the common assumption that SLiMs occur outside the globular domain regions could be a cause for the lacklustre coverage of current SLiM detection methods (Edwards *et al.*, 2007; Neduva *et al.*, 2005). Among the 452 distinct SLiMs that we reported, 198 of them have been detected on domain–domain interaction interfaces (we call these *domain–domain SLiMs*). Current SLiM detection methods are not amenable to mining these domain–domain SLiMs since they rely on a motif's overrepresentation over a set of non-homologous protein sequences. It is virtually impossible to detect the overrepresentation of a domain–domain SLiM using sequence-based methods since the domain's homology would overwhelm the SLiM's much weaker similarity.

We conducted a further study on four novel domain–domain SLiMs, which we have found. The first one is a domain–domain SLiM bound by the tumor necrosis factor (TNF; ID: PF00229) domain on the BAFF proteins that have been implicated in B-cell hyperplasia and development of severe autoimmune diseases (Gross *et al.*, 2000; Khare *et al.*, 2000). A previous experiment reported in the literature has showed that an instance of our predicted SLiM (a short peptide DLLVRHWV) can prevent the pathogenic condition from BAFF overexpression (Gordon *et al.*, 2003). Another domain–domain SLiM of interest is a novel SLiM found on the dimer interfaces of the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) enzyme, which is associated with neurodegenerative disorders such as Huntington's disease, Alzheimer's disease, Parkinson's disease and Machado–Joseph disease (Berry and Boulton, 2000; Tatton *et al.*, 2003). We also discovered two SLiMs that are implicated in amyloid fibril formation implicated in several debilitating human diseases such as Alzheimer's disease, prion-based encephalopathies, liver cirrhosis and lung emphysema (Carrell and Gooptu, 1998). The class of domain–domain SLiMs could, therefore, be particularly useful for designing inhibitors to disrupt the domain–domain interactions, which underlie the formation of pathogenic protein complexes.

In conclusion, the fine atomic details offered by structural data made them an attractive data source for discovering SLiMs that are beyond the coverage of existing sequence-based methods. With the number of available protein structures expecting to grow rapidly, we can expect to discover even more biologically significant novel SLiMs in the near future.

2 METHODS

2.1 SLiMDIet's workflow

In this study, we devised a method named SLiMDIet, a *de novo* SLiM discovery method by Diet from 3D protein structure data. SLiMDIet consists of two steps: a Diet step, followed by a SLiM step. 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 along a domain chain (the set is called *the domain face*) while the other on a partner chain (*partner face*), which are in close vicinity of each other. The interaction interfaces of each domain are then clustered based on structural similarity. The resulting domain interface clusters represent various modes of interactions for the domain. 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 length 3–12 residues. To ensure robustness, we require that a block is constructed only from non-homologous partner chains and we require at least four of them. Finally, we construct a (linear) gapped position specific scoring matrix (PSSM) from the block to represent the predicted SLiMs. An illustration of SLiMDIet algorithm can be seen in Figure 1.

2.2 Domain identification

A structural dataset was downloaded from PDB on August 24, 2009, containing 57 559 structures. We chose structures containing at least one protein chain and whose resolution is 3.0 Å or better, giving a total of 54 981 legible structures with 130 488 protein chains. PFAM domain annotations on each PDB chain are computed by running the *hmmpfam* program from the HMMER library version 2.3.2 (Eddy, 1998) using the latest PFAM 23.0 library (Bateman *et al.*, 2004).

We use PFAM (Bateman *et al.*, 2004) as our choice of protein domain definition as opposed to SCOP (Andreeva *et al.*, 2008) or CATH (Cuff *et al.*, 2009) because of the relatively better coverage of PFAM. PFAM was previously reported to have 57% coverage on SWISSPROT+TREMBL sequences, while SCOP covers 31% (Elofsson and Sonnhammer, 1999). PFAM also has higher PDB chain coverage on the current dataset [version 23.0, released July 2008, covering 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)]. However, PFAM domain does have its own limitation. It currently does not define structural domains that are formed by multiple protein chains. Nevertheless, one can always apply SLiMDIet on SCOP/CATH domain definition without major change on the program.

2.3 Interface extraction

For each PDB structure, we find the PFAM domains in its chains. For each domain, we computed the domain interfaces as follows. First, we define the distance between two amino acid residues to be the nearest distance between any pair of non-hydrogen atoms between the two residues. As done in PSIMAP (Dafas *et al.*, 2004), we also use a contact distance cutoff of 5 Å here.

A domain interface comprises two sets of amino acid residues: the domain and partner faces. Each amino acid on one face must be within the defined contact distance from some amino acid on the other face. The residues on each face must originate from a single protein chain (named domain and partner chain, respectively). However, they need not be located consecutively in their respective chains. For the domain face, the residues must also be within a single protein domain region of the domain chain.

To curb possible non-biological (crystal) interfaces, which are generally of smaller 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

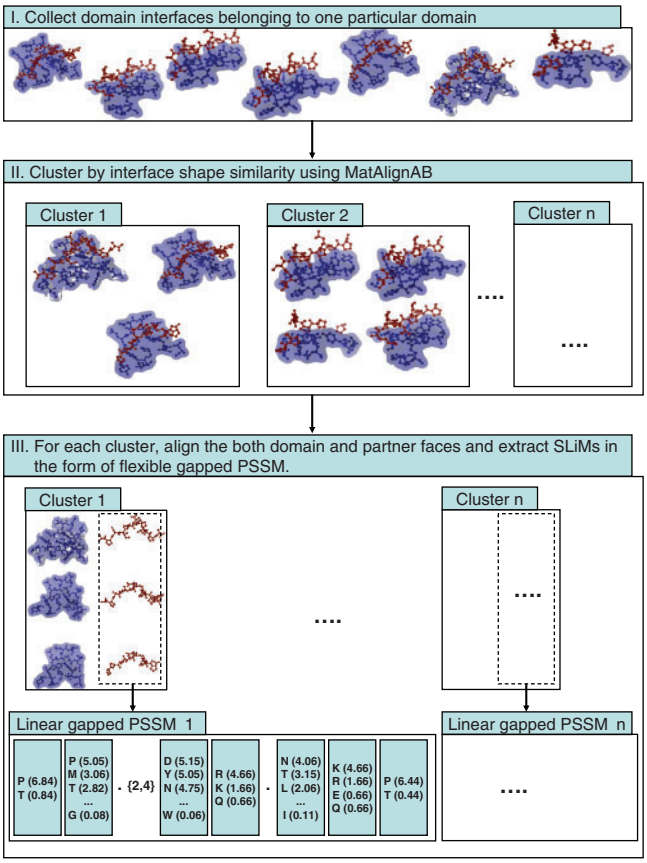

Fig. 1. SLiMDIet's overview. The domain interfaces of each PFAM domain are clustered by their structural similarity. Next, from each cluster, the domain and partner faces are structurally aligned and we build a gapped PSSM based on the contacts on the partner faces. The gapped PSSM has flexible gaps defined by the minimum and maximum gaps observed between two consecutive PSSM positions. We define a gapped PSSM as linear when the total length of its non-gap positions is 3–12 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.

the partner face. This lower bound corresponds to a binding area >800 Å^2—which is roughly the average size of a domain interface (Kim *et al.*, 2006). For intrachain domain interfaces, we also require that the residues on the partner face are not within 10 residues from the ends of the domain, to avoid recognizing local contacts as interaction interfaces. This resulted in 270 739 domain interfaces involving 4780 PFAM domains.

2.4 Pairwise structural alignment within each domain interface group

To classify similar interfaces that correspond to the same domain interaction class, we define the similarity of two interfaces using the modified[1] *S*-score function from (Alexandrov and Fischer, 1996) as follows:

$$S_{\text{norm}} = \frac{1}{(1+\Delta)} \cdot \frac{N}{\min(|A|,|B|)}$$

[1]The function is normalized by the size of the interface and scaled to yield similarity score between 0 and 1.

where Δ is the root mean square distance (RMSD) between the two structures being aligned, N the number of aligned residues between the two interfaces, and $|A|$ and $|B|$ the sizes of the aligned interfaces.

Usually, the RMSD between two proteins is approximated by the RMSD of their backbone's C_α atoms. Since SLiMDIet's domain interfaces only consist of the contact residues (instead of the whole protein or domain), the C_α representation is rather inadequate. To capture the similarity better, we measure the similarity of two interfaces using the backbone and side chain conformation of the residues on each interface. We use the C_β atom position to represent the direction of the side chain with respect to its backbone C_α (a similar C_β approximation was mentioned in Torrance *et al.*, 2005).

When comparing two interfaces, we treat both domain and partner faces of each domain interface as one rigid continuous structure. We designed MatAlignAB for comparing domain interfaces, a modified algorithm of MatAlign (Aung and Tan, 2006), which only aligns residues from the same face type (i.e. residues from domain face in one interface can only be aligned to residues in the domain face of the other) and aligns atoms of the same atom type [i.e. C_α (C_β, respectively) to C_α (C_β, respectively)]. As with the original algorithm, MatAlignAB produces alignments that follow the sequential ordering of the residues within their respective domain and partner sequences. The final results of this step consist of the similarity scores and pairwise alignments among all pairs of domain interfaces of each domain.

2.5 Hierarchical agglomerative clustering on the domain interfaces using average linkage

For every domain, we cluster its interfaces into domain interface clusters by following the steps of hierarchical agglomerative clustering algorithm using average linkage, where the similarity of two clusters is defined to be the average pairwise similarity between all the members of the two clusters (as done in Kim *et al.*, 2006). The algorithm starts by setting every domain interface as a cluster with one member. Next, it picks the pair of clusters that has the highest pairwise similarity and combine the pair. Then, it computes the average similarity of the combined cluster with the rest of the cluster. The latter two steps are repeated until the similarity score between every possible pair of the clusters is below a certain threshold. In SLiMDIet, we use the following range of thresholds 0.15, 0.2, 0.25 and 0.3 to generate sets of (possibly overlapping) clusters each under the corresponding threshold level. For those clusters that have >70% overlap, we group them together and report one of the clusters as the representative (see Supplementary Material 2 for details).

2.6 Quantification of the clustering performance

Suppose C is a cluster of domain interfaces computed by a particular algorithm and R the reference cluster [in our case, R is the set of domain interfaces (manually) grouped according to the literature]. We use the *F-score*, which is the harmonic mean of the sensitivity and specificity scores (Rijsbergen, 1979), to quantify the similarity of the predicted cluster C and the reference cluster R.

$$F\text{-score}_{(C,R)} = \frac{2 \times \text{Spec}_{(C,R)} \times \text{Sens}_{(C,R)}}{\text{Spec}_{(C,R)} + \text{Sens}_{(C,R)}}$$

where $\text{Spec}_{(C,R)}$ is the specificity of the cluster C with respect to a reference cluster R, which is computed by $\text{Spec}_{(C,R)}=|C \cap R|/|C|$. $\text{Sens}_{(C,R)}$ is the sensitivity of the cluster C with respect to R, which is computed by $\text{Sens}_{(C,R)}=|C \cap R|/|R|$.

The F-score of an algorithm for a particular reference cluster R is the best score among its computed cluster C. The F-score measure is used to compare the clustering performance of SLiMDIet to SCOWLP's on the benchmark data.

2.7 SLiM extraction from the interface clusters

We employ a PSSM with flexible gaps, called *gapped PSSM* to define the binding motif on the interaction interfaces. The gaps are defined between any two consecutive positions in the PSSM.

Given a cluster of domain interfaces, the construction of a gapped PSSM is performed in two steps. First, interfaces from the same cluster are aligned to the *cluster center*, which is the domain interface with the best average similarity to the rest of the interfaces in a cluster, to generate an approximate multiple interface alignment of the interfaces. Then, we ensure that the alignment contains four non-homologous interfaces. An interface I_a is defined as homologous to I_b when I_a and I_b's aligned residues in the alignment are exactly the same and their full partner chains share >50% sequence similarity. This means that two interfaces whose partner chains share high sequence similarity can still be defined as non-homologous as long as their aligned interface residues differs.

In the alignment of the non-homologous interfaces, a block is defined as a set of 3–12 consecutive alignment positions with gaps of at most four residues in between. The SLiM corresponding to an interface alignment is computed from the longest block in it. A SLiM is said to be covering an interface when it covers at least half of the contact residues on the partner face of that interface. To make sure that the computed SLiM represents the interfaces of a domain interface cluster, we require it to cover at least half of the non-redundant interfaces in it.

With a block that satisfies the coverage constraint, we construct a gapped PSSM by extrapolating the score of all 20 amino acids based on the BLOSUM62 substitution score (Henikoff and Henikoff, 2005) of all observed amino acids in each column in the block. The detailed steps of the gapped PSSM construction is included in Supplementary Material 1 (Section 2). From 39 170 domain clusters with at least four members, SLiMDIet found 7473 with at least four non-homologous interfaces. Out of these, only 1592 met the coverage constraint. We then grouped interface clusters from different similarity cutoffs when they have at least 70% member overlap. The grouping yields 452 distinct gapped PSSMs involving 280 PFAM domains. The full listing of SLiMDIet's predicted SLiMs and their gapped PSSM are listed in Supplementary Material 2.

2.8 Computing the statistical significance of the SLiM using PPI data

When a SLiM is extracted from a particular domain–SLiM interface clusters, we conduct statistical tests to see if the motif occurs significantly more in the interaction partners of the domain as compared to any random interaction.

Given a protein sequence S, the gapped PSSM score of one particular position j in S is just the maximum sum of the gapped PSSM's residue scores starting at j over all possible gap combination in the PSSM. For example, the best score of position 0 in the string FSDTK based on the gapped PSSM[2] $\begin{bmatrix} L:4.62 \\ F:1.38 \end{bmatrix} \cdot \{1,2\} \begin{bmatrix} T:2.4 \\ D:-0.12 \end{bmatrix}$ would be

$$\max \begin{cases} 1.38+(-0.12) & (\text{gap}=1), \\ 1.38+2.4 & (\text{gap}=2). \end{cases}$$

For a position in a protein with a gapped PSSM score s, it is defined as an *occurrence* of the PSSM if the probability of scoring s or better by random is at most equal to 10^{-4}. To this end, we created 10 000 random protein sequences, each of length 500, with their amino acid distribution following the one observed in our PPI data from BioGRID (Breitkreutz *et al.*, 2008). For each gapped PSSM, we computed the scores of all positions in the random dataset (of approximately 5 million positions) and sorted the scores in non-increasing order. The 500th score on the sorted score list would have an empirical P-value of 10^{-4} and is chosen as the cutoff score for the occurrence of the gapped PSSM.

[2]This is a mini-version of gapped PSSM for exemplary purpose, the real gapped PSSM would have scores for all 20 amino acids.

Given a SLiM's gapped PSSM, the probability of observing a certain number of occurrences in the partners of a protein domain by random can be computed by the standard hypergeometric distribution function

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

where I is the whole set of the high-throughput PPI data, I_M the subset of I that contain an occurrence of the motif M, I_D the subset of I containing the domain D and I_{DM} the subset of I_D that contain an instance of M. The correctness of the hypergeometric scoring function on the PPI data is presented in Supplementary Material 1 (Section 5).

3 RESULTS

3.1 Both known and novel SLiMs are discovered

SLiMDIet detected 452 distinct SLiMs from the whole PDB dataset (dated August 2009). Forty of that are known in the literature. Amongst the remaining 412 candidate novel SLiMs, 54 have at least an instance of a domain–short peptide structure in their respective domain–SLiM clusters. The presence of such a domain–short peptide structure is a strong indicator that the domain is capable of binding a linear peptide defined by the predicted SLiM. Indeed, all of the literature-backed SLiMs have at least one domain–short peptide structure.

From the remaining 358 candidate novel SLiMs, we found 61 are overrepresented in the interaction partners of their respective domains within the high-throughput PPI data (P-value ≤ 0.05). It is important to note that SLiMs with poor P-value are not necessarily erroneous since the PPI data is far from complete. Indeed, as many as 145 of the remaining 297 SLiMs (those with P-value > 0.05) have <10 distinct interaction data—99 of them have no PPI data support at all. This shows the limitation of SLiM detection methods that relied solely or heavily on PPI data.

3.2 SLiMs with validations from the literature

We compared our predicted SLiMs with those listed in the ELM (Puntervoll *et al.*, 2003) and MiniMotif database (Balla *et al.*, 2006). SLiMDIet reported 40 SLiMs with strong similarity with the known SLiMs in literature. Since there is a significant overlap in the entries of ELM and MiniMotif, most of our SLiMs correspond to more than one database entry in both databases. In summary, our SLiMs covered 30 out of 136 known ELM SLiMs and 72 of 524 MiniMotif SLiMs (from the publicly available MiniMotif version 1). The coverage is significant considering that the SLiMs are solely computed from a more limited structural data source. The detailed listing of the 40 SLiMs with their corresponding literature SLiM is given in Supplementary Material 4.

As a comparison, we also checked the discovery of these literature-backed SLiMs in the profiles collected by D-MIST (Betel *et al.*, 2007). D-MIST, like SLiMDIet, constructs binding profiles of different domains based on the structural data. However, it relies on the high-throughput PPI data to refine their predicted motifs. Out of the 40 literature-backed SLiMs found by SLiMDIet, we could only find the corresponding D-MIST profiles for nine of them. For the missing 31 SLiMs, D-MIST did not have any profile related to the SLiM's domain for 24 of them and for the remaining seven SliMs, D-MIST's profiles are too divergent from the literature SLiMs.

Such poor coverage could be due to the fact that D-MIST was collected from a subset of PDB (10064 structures). However, we observe that even the older, well-studied SLiMs recognized by domains like SH2(Grb2), WW, FHA, PDZ and PID(PTB) were also missing. We present the detailed listing of matched D-MIST profiles in the Supplementary Material 5.

4 DISCUSSION

4.1 Different SLiM classes have different interface geometries

It has been known that some SLiM-recognizing domains can bind multiple classes of SLiMs. The SH3 domain, for example, is known to recognize two classes of SLiMs; [KRY]..P..P (SH3 class 1 SLiM) and P..P.[KR] (SH3 class 2 SLiM) (Puntervoll *et al.*, 2003). We hypothesize that the existence of such different classes of SLiM that can bind to the same domain is due to observable differences in their corresponding domain interface geometries. In other words, one can differentiate domain–SLiM interfaces belonging to different classes of SLiMs through geometric comparison.

To verify our conjecture, we hand-curated a benchmark set of 230 domain–SLiM interfaces from three well-studied domains—SH2 (123 interfaces), SH3 (80 interfaces) and 14-3-3 (27 interfaces)—whose interaction classes are well annotated in the literature. The detailed listing of the benchmark interfaces is given in the Supplementary Material 3.

We compare the structural clustering of SLiMDIet with an existing domain interface clustering method SCOWLP (Teyra *et al.*, 2008) on the benchmark clusters. Table 1 shows that SLiMDIet's clustering has better average specificity, sensitivity and *F*-score for all three domains in the benchmark. However, we should note that SCOWLP was not designed specifically for clustering domain–SLiM interfaces. We compared with SCOWLP because it was the only existing method we found suitable for clustering domain–SLiM interfaces. The discussion on the performance of both methods on each benchmark domain is presented in Supplementary Material 3.

The overall high correspondence of SLiMDIet's structural clusters with the literature reference clusters indicates that different classes of domain–SLiM interfaces indeed are associated with different domain interface geometries. It also shows that SLiMDIet, which took in account the interface geometries of the SLiMs, is more suitable for domain–SLiM interface clustering as compared with such an existing domain interface clustering method as SCOWLP.

Table 1. Clustering performance comparison of SLiMDIet and SCOWLP

Interaction class	SLiMDIet			SCOWLP		
	Sens	Spec	F-Scr	Sens	Spec	F-Scr
SH3-class 1	0.97	1.00	**0.98**	0.71	0.55	0.62
SH3-class 2	0.98	0.92	**0.95**	0.88	0.54	0.67
SH3 P.[VI][DN]R..KP	1.00	1.00	**1.00**	0.25	1.00	0.40
SH2-(class 1A)	0.75	0.86	**0.80**	0.62	0.67	0.65
SH2-(class 1B)	0.67	1.00	0.80	0.75	1.00	**0.86**
SH2-(class 1C)	1.00	1.00	**1.00**	0.83	0.59	0.69
SH2-(class 2A)	0.25	1.00	0.40	0.50	1.00	**0.67**
SH2-(class 2B)	0.67	0.86	0.75	0.67	1.00	**0.80**
14-3-3 Mode 1	1.00	0.50	0.67	0.50	1.00	0.67
14-3-3 Mode 2	1.00	1.00	**1.00**	0.67	1.00	0.80
14-3-3 Mode 3	0.50	1.00	**0.67**	1.00	0.33	0.50

We collected the interfaces of the SH2, SH3 and 14-3-3 domains whose domain–SLiM interaction class is defined in their respective reference papers. The grouping from the literature constitutes the reference clusters, against which the accuracy of both SLiMDIet and SCOWLP are computed. The cases where one method outperforms the other are printed in bold.

4.2 Known and Novel SLiMs are found on domain–domain interaction interfaces

Interestingly, we observed that 198 of the total 452 predicted SLiMs are domain–domain SLiMs. A domain–domain SLiM is a SLiM found in an interface cluster with at least four non-homologous interfaces whose partner faces occur within some (not necessarily the same) PFAM domain. We found two of our 198 reported domain–domain SLiMs have literature support. They are the SH3_1 and ubiquitin domain. All of the instances of our predicted SLiM for ubiquitin are found within a domain called ubiquitin interacting motif (UIM, ID: PF02809). We also found another six domain–domain SLiMs with supporting domain–short peptide structures and 35 with overrepresentation in the PPI data. We also observe that 143 out of 198 domain–domain SLiMs that we found are overrepresented in their respective PFAM domains. The complete listing of the domain–domain SLiMs and the procedure to compute their overrepresentation within the PFAM domain in which they are found are given in Supplementary Material 6.

Finding such domain–domain SLiMs can be an important discovery since it is commonly believed that SLiMs occur outside the globular domain regions (Puntervoll *et al.*, 2003). In fact, most of the current SLiM detection methods remove domain regions from the search space (Edwards *et al.*, 2007; Neduva *et al.*, 2005) because of this belief. The discovery of such domain–domain SLiMs also indicates that many of the apparent domain–domain interactions could be mediated by domain–SLiM interactions. Indeed, a recent study had actually found genuine occurrences of ELM SLiMs on the accessible parts of a globular domain (Via *et al.*, 2009).

One particularly interesting novel domain–domain SLiM found by SLiMDIet is a SLiM that is bound by the GAPDH, C-terminal (Gp_dh_C) domain (ID: PF02800). The Gp_dh_C domain is the C-terminal domain of GAPDH enzyme. The enzyme exists as a tetramer of identical chains, each containing two conserved functional domains, the Gp_dh_N (ID: PF00044) and Gp_dh_C (ID: PF02800) domain. Figure 2A shows the structure of half of the tetramer, comprising of two chains of GAPDH (one chain on the left and one on the right). Figure 2B and C illustrates only the Gp_dh_C domain surfaces with the linear peptide regions of Gp_dh_N on them.

GAPDH has an important role in glycolysis and gluconeogenesis, and it is also involved in the signaling mechanism for programmed cell death (apoptosis; Berry and Boulton, 2000). Several studies associated the enzyme with neurodegenerative disorders such as Huntington's disease, Alzheimer's disease, Parkinson's disease and Machado–Joseph disease (Berry and Boulton, 2000; Tatton *et al.*, 2003). The SLiM computed by SLiMDIet for Gp_dh_C is [YH]..[KRQ][YH]D[ST], which is found within the Gp_dh_N domain. The predicted SLiM is found within nine non-homologous

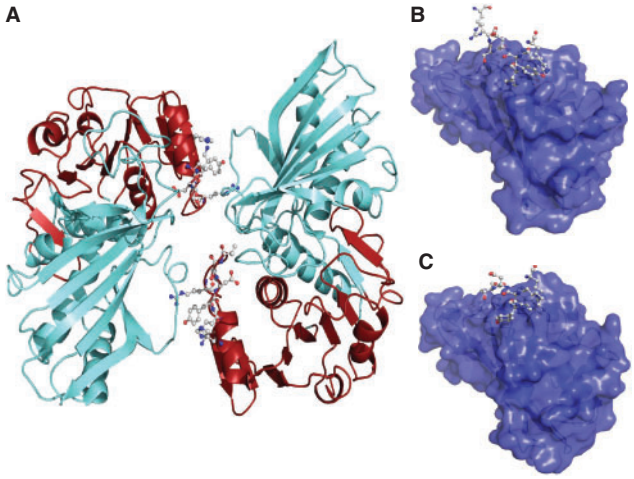

Fig. 2. Domain–SLiM interface between Gp_dh_C domain (ID: PF02800) and Gp_dh_N (ID: PF00044). (**A**) The dimer of the GAPDH complex (PDB ID:1gd1). The cyan (light) part is the C-terminal domain and the red (dark) part mark the N-terminal domain. The C-terminal domain binds to a linear region on the N-terminal domain of the opposite chain (highlighted in ball-and-stick mode). SLiMDIet's predicted SLiM for this region is [YH]..[KRQ][YH]D[ST]. (**B**) The surface representation of the Gp_dh_C domain of Holo-GAPDH from *Bacillus stearothermophilus* (PDB ID:1gdl). The linear region **H**LL**KYDS**VHGR of the opposite N-terminal domain bound to the domain is shown in ball-and-stick representation. (**C**) The structure of linear sequence **Y**QM**KHDT**VHGR bound to the Gp_dh_C domain of Leishmania mexicana's glycosomal GAPDH (PDB ID:1a7k).

GAPDH dimers. It was reported in an earlier study that inhibition on the formation of GADPH tetramer protects against neuronal-induced cell-death (Fukuhara *et al.*, 2001), a phenomenon frequently seen in many neurogenerative diseases. Interestingly, the dimeric and monomeric form of the enzyme retain its glycolysis and gluconeogenesis functionality and research had shown that they have higher catalytic activity (Minton and Wilf, 1981). We suggest that our domain–domain SLiM could be used as a template for designing inhibitors to disrupt the enzyme's complex formation and keep it in its monomeric form.

Another notable example of domain–domain SLiMs is a SLiM interacting with the TNF domain (ID: PF00229) of BAFF proteins. SLiMDIet predicted that it binds a SLiM D[LHS]L[LV][RH]..[IV] on its domain partners [BaffR-Tall_bind (ID: PF09256), BCMA-Tall_bind (ID: PF09257) and TACI-CRD2 (ID: PF09305)]. BAFF protein overexpression was previously shown to result in B-cell hyperplasia and development of severe autoimmune diseases (Gross *et al.*, 2000; Khare *et al.*, 2000). In fact, it has already been reported that an instance of the SLiM can confer BAFF binding and block the signaling pathway leading to the pathogenic condition from BAFF overexpression (Gordon *et al.*, 2003). However, there were no TNF-binding SLiM for BAFF reported in the literature and SLiMDIet managed to predict one. The predicted SLiM could provide further insights for designing more effective treatments. Figure 3 shows two PDB structures in which two TNF domains are binding a short peptide and a full partner domain, respectively; both containing our predicted SLiMs.

A third domain–domain SLiM is found on the dimer interface of RNaseA domains (ID: PF00074) of Ribonuclease protein. The

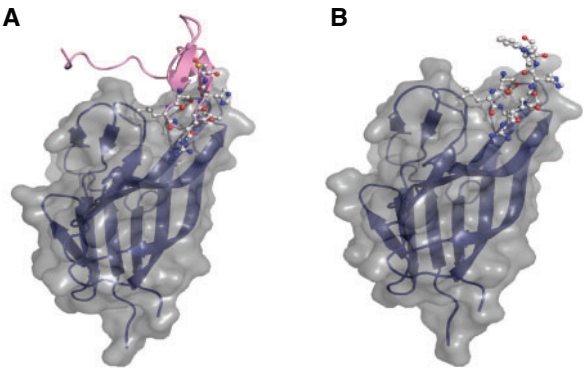

Fig. 3. Domain–SLiM interfaces of TNF domain of BAFF proteins recognizing the SLiM D[LHS]L[LV][RH]..[IV]. (**A**) The TNF interface from BAFF with a part of BAFF receptor protein (PDB ID:1oqe). The linear region is shown in ball-and-stick display, comprising the residues DLLVRHCV. (**B**) The structure between the TNF domain of BAFF complexed with only the minimal peptide DLLVRHWV (shown in ball-and-stick, PDB ID:1osg)

protein is known to form dimers using two modes of domain swapping. The major mode swaps the C-terminal beta sheets (Liu *et al.*, 2001) while the minor mode swaps the N-terminal helix (Liu *et al.*, 1998). Previous experiments have shown that a peptide instance of the N-terminal helix could compete with the minor mode of the domain swapping and disrupt dimer formation (Liu *et al.*, 1998). It has also been reported that domain swapping is one possible mechanism of amyloid fibril formation (Carrell and Gooptu, 1998; Liu *et al.*, 2001) and based on the domain swapping observed in RNase, Liu *et al.* proposed a model of amyloid fibril formation, which is stabilized by the swapped domain binding (Liu *et al.*, 2001). The formation of amyloid is associated with a variety of neurodegenerative diseases such as Alzheimer's disease, Huntington's disease and the new variant Creutzfeldt–Jakob disease (nvCJD). It is also implicated in other diseases such as the sickle cell anemia, α-antitrypsin related liver cirrhosis and emphysema (Carrell and Gooptu, 1998). In such a model, knowing the SLiM bound by the domain would enable one to design an inhibitor to destabilize and prevent the amyloid formation. SLiMDIet predicted two distinct novel SLiMs that correspond to the two swapping modes of the RNase domain, namely YVPVH[FYL][DAN]AS (major mode) and AA..[FAM]ERQH.DS (minor mode).

5 CONCLUSIONS

SLiMs are important mediators of protein–protein interactions but they are difficult to detect experimentally and computationally. In this work, we showed that it is possible to systematically detect *de novo* SLiMs on domain interaction interfaces extracted directly from structural data. The atomic level of details available in the high-resolution 3D structures provide a rich source of data for discovering SLiMs that are guaranteed to occur on the binding surfaces. In fact, by mining the different domain–SLiM interaction classes from the PDB database, our SLiMDIet method detected many novel SLiMs, including the domain–domain SLiMs.

The discovery of domain–domain SLiMs uncovered a limitation in the current SLiM detection approaches. These SLiMs are located in regions that are routinely masked out by the current SLiM

detection methods. They cannot be detected simply by turning off the masking step—the strong similarity of the domain regions would bury the weak signal of the degenerate SLiM(s) in them. This class of SLiM is, therefore, currently under-represented in the known databases and literature, and they present real opportunities for domain–domain interaction inhibitor design.

Current SLiM detection methods also rely heavily on PPI data and are thus affected by its accuracy. An earlier study (Neduva *et al.*, 2005) has reported that some of the known SLiMs were not detected in the PPI due to noisy and incomplete interaction data. In our study, we also observed a similar problem where as many as 111 SLiMs do not have any PPI data containing their binding domains. Among them, two are known in the literature, namely the Toxin_1 (Scherf *et al.*, 1997) and fn1 domains (Bingham *et al.*, 2008) and 10 have domain-short peptide evidences.

As the structural genomic initiatives continue to make more and more high-quality structural data available, we can have a viable chance of detecting the SLiMs that mediate many of our important protein–protein interactions directly from 3D structural data. As future work, we plan to continue to improve SLiMDIet's capability by refining the notion of interface similarity to take into account the interface residues' chemical properties and their connectivity within the domain interfaces.

Funding: Ministry of Education, Singapore (AcRF grant R-252-000-326-112 to H.W. and W.-K.S); Agency for Science, Technology and Research (A*STAR) of Singapore.

Conflict of Interest: none declared.

REFERENCES

Alexandrov,N.N. and Fischer,D. (1996) Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures. *Proteins*, **25**, 354–365.

Aloy,P. and Russell,R.B. (2006) Structural systems biology: modelling protein interactions. *Nat. Rev. Mol. Cell Biol.*, **7**, 188–197.

Andreeva,A. *et al.* (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.

Aung,Z. and Tan,K.L. (2006) MatAlign: precise protein structure comparison by matrix alignment. *J. Bioinform. Comput. Biol.*, **4**, 1197–1216.

Balla,S. *et al.* (2006) MiniMotif miner: a tool for investigating protein function. *Nat. Methods*, **3**, D175–D177.

Bateman,A. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Berry,M.D. and Boulton,A.A. (2000) Glyceraldehyde-3-phosphate dehydrogenase and apoptosis. *J. Neurosci. Res.*, **60**, 150–154.

Betel,D. *et al.* (2007) Structure-templated predictions of novel protein interactions from sequence information. *PLoS Comput. Biol.*, **3**, e182.

Bingham,R.J. *et al.* (2008) Crystal structures of fibronectin-binding sites from Staphylococcus aureus FnBPA in complex with fibronectin domains. *Proc. Natl Acad. Sci. USA*, **105**, 12254–12258.

Breitkreutz,B. *et al.* (2008) The BioGRID interaction database: 2008 update. *Nucleic Acids Res.*, **36**, D637–640.

Carrell,R.W. and Gooptu,B. (1998) Conformational changes and disease-serpins, prions and Alzheimer's. *Curr. Opin. Struct. Biol.*, **8**, 799–809.

Cuff,A.L. *et al.* (2009) The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.*, **37**, D310–D314.

Dafas,P. *et al.* (2004) Using convex hulls to extract interaction interfaces from known structures. *Bioinformatics*, **20**, 1486–1490.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Edwards,R. *et al.* (2007) SLiMFinder: a probabilistic method for identifying overrepresented, convergently evolved, short linear motifs in proteins. *PLoS ONE*, **2**, e967.

Elofsson,A. and Sonnhammer,E.L. (1999) A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics*, **15**, 480–500.

Fukuhara,Y. *et al.* (2001) GAPDH knockdown rescues mesencephalic dopaminergic neurons from MPP+ induced apoptosis. *Neuroreport*, **42**, 2049–2052.

Gordon,N.C. *et al.* (2003) BAFF/BLyS receptor 3 comprises a minimal TNF receptor-like module that encodes a highly focused ligand-binding site. *Biochemistry*, **42**, 5977–5983.

Gross,J.A. *et al.* (2000) TACI and BCMA are receptors for a TNF homologue implicated in B-cell autoimmune disease. *Nature*, **404**, 949–950.

Henikoff,S. and Henikoff,J.G. (2005) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Khare,S.D. *et al.* (2000) Severe B cell hyperplasia and autoimmune disease in TALL-1 transgenic mice. *Proc. Natl Acad. Sci. USA*, **97**, 3370–3375.

Kim,W.K. *et al.* (2006) The many faces of protein–protein interactions: a compendium of interface geometry. *PLoS Comput. Biol.*, **2**, e124.

Li,H. *et al.* (2006) Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics*, **22**, 314–324.

Liu,Y. *et al.* (1998) The crystal structure of a 3D domain-swapped dimer of RNase A at a 2.1-Å resolution. *Proc. Natl Acad. Sci. USA*, **95**, 3437–3442.

Liu,Y. *et al.* (2001)A domain-swapped RNase a dimer with implications for amyloid formation. *Nat. Struct. Biol.*, **8**, 989–996.

von Mering,C. *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.

Minton,A. and Wilf,J. (1981) Effect of macromolecular crowding upon the structure and function of an enzyme: glyceraldehyde-3-phosphate dehydrogenase. *Biochemistry*, **20**, 4821–4826.

Neduva,V. *et al.* (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol.*, **3**, e405.

Neduva,V. and Russell,R.B. (2005) Linear motifs: evolutionary interaction switches. *FEBS Lett.*, **579**, 3342–3345.

Neduva,V. and Russell,R.B. (2006) Peptides mediating interaction networks: new leads at last. *Curr. Opin. Biotechnol.*, **17**, 465–471.

Pawson,T. and Nash,P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445–452.

Puntervoll,P. *et al.* (2003) ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, **31**, 3625–3630.

Rijsbergen,C.J.V. (1979) *Information Retrieval*. Butterworth-Heinemann, Newton, MA.

Scherf,T. *et al.* (1997) Three-dimensional solution structure of the complex of alpha-bungarotoxin with a library-derived peptide. *Proc. Natl Acad. Sci. USA*, **94**, 6059–6064.

Tan,S.H. *et al.* (2006) A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC Bioinformatics*, **7**, 502.

Tatton,W. (2003) Neuroprotection by deprenyl and other propargylamines: glyceraldehyde-3-phosphate dehydrogenase rather than monoamine oxidase B. *J. Neural Transm.*, **110**, 509–515.

Teyra,J. *et al.* (2008) SCOWLP classification: structural comparison and analysis of protein binding regions. *BMC Bioinformatics*, **9**, 9.

Torrance,J.W. *et al.* (2005) Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *J. Mol. Biol.*, **347**, 565–581.

Via,A. *et al.* (2009) A structure filter for the Eukaryotic Linear Motif Resource. *BMC Bioinformatics*, **10**, 351.