Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach

Lei Li¹, Chenggang Wu¹, Haiming Huang¹, Kaizhong Zhang², Jacob Gan³ and Shawn S.-C. Li¹,*

¹Department of Biochemistry and the Siebens-Drake Medical Research Institute, Schulich School of Medicine and Dentistry, ²Department of Computer Science, University of Western Ontario, London, Ontario N6A 5C1, Canada and ³School of Mechanical & Aerospace Engineering, Nanyang Technological University, Singapore

Received February 26, 2008; Revised March 19, 2008; Accepted March 20, 2008

ABSTRACT

Systematic identification of binding partners for modular domains such as Src homology 2 (SH2) is important for understanding the biological function of the corresponding SH2 proteins. We have developed a worldwide web-accessible computer program dubbed SMALI for scoring matrix-assisted ligand identification for SH2 domains and other signaling modules. The current version of SMALI harbors 76 unique scoring matrices for SH2 domains derived from screening oriented peptide array libraries. These scoring matrices are used to search a protein database for short peptides preferred by an SH2 domain. An experimentally determined cut-off value is used to normalize an SMALI score, therefore allowing for direct comparison in peptide-binding potential for different SH2 domains. SMALI employs distinct scoring matrices from Scansite, a popular motif-scanning program. Moreover, SMALI contains built-in filters for phosphoproteins, Gene Ontology (GO) correlation and colocalization of subject and query proteins. Compared to Scansite, SMALI exhibited improved accuracy in identifying binding peptides for SH2 domains. Applying SMALI to a group of SH2 domains identified hundreds of interactions that overlap significantly with known interactions mediated by the corresponding SH2 proteins, suggesting SMALI is a useful tool for facile identification of signaling networks mediated by modular domains that recognize short linear peptide motifs.

INTRODUCTION

Phosphorylation by protein kinases is a central paradigm in signal transduction and it regulates almost all essential cellular functions such as proliferation, differentiation, migration and survival (1). Deregulated phosphorylation of proteins is often associated with an abnormal state of a cell and can result in malignant transformation (2). The human genome encodes ~518 protein kinases, of which 90 are tyrosine kinases and another 43 are tyrosine kinase-like (3). By adding a phosphate moiety to the hydroxyl group of a Tyr residue, protein-tyrosine kinases can directly modulate the activity of the target protein, alter its subcellular localization and/or promote the formation of specific signaling complexes. The latter function of tyrosine phosphorylation is mediated by protein modules, such as the Src homology 2 (SH2) and phosphotyrosine-binding (PTB) domains, which recognize pTyr-containing peptides (4,5). Binding of an SH2 or a PTB domain to a phosphotyrosyl sequence provides a general mechanism for the formation of specific protein complexes in intracellular signal transduction, which serves to propagate and regulate a signal emanated from a protein-tyrosine kinase.

The importance of tyrosine phosphorylation in normal cellular function is also highlighted by the great number of SH2 and PTB domains identified in metazoa (6,7). The human genome encodes 120 SH2 domains distributed in 110 distinct proteins, which constitutes the largest family of modular domains capable of recognizing a phosphotyrosine (7). Although the pTyr residue is indispensable for SH2-binding in the majority of cases (8), the specificity of a given SH2 domain is typically determined by a few residues C-terminal to the pTyr (5). Identifying the specific
phosphotyrosyl peptide motif recognized by an SH2 domain is a key to understand the function of the corresponding SH2-containing protein. On a larger scale, comprehensive knowledge about the specificity of all mammalian SH2 and PTB domains would make it possible to gauge, in principle, the phosphotyrosine cellular signaling network mediated by these domains. As a first step towards this lofty goal, we recently determined the phosphotyrosyl motifs selected, respectively, by 76 human SH2 domains using an oriented peptide array library (OPAL) approach (13). The parent library consisted of the degenerated sequence XX-pY-XXXX, where X denotes a mixture of 19 naturally occurring amino acids except Cys, and screening of the OPAL yielded selectivity for positions −2 to +4 with respect to the pTyr. This specificity information is necessary for future exploration of SH2 domain function and for the identification of SH2-mediated protein–protein interactions. To take advantage of the OPAL screen data, we generated position-specific scoring matrices (PSSM) for 76 SH2 domains and developed a world-wide web-based (WWW) computer program called scoring matrix-assisted ligand identification (SMALI) for facile identification of linear peptides preferred by an SH2 domain from searching a protein database. Although SMALI is similar to the motif-scanning method, Scansite, developed by Yaffe, Cantley and colleagues (9), SMALI contains PSSMs for 76 SH2 domains in contrast to 14 employed by the latter. Moreover, an SMALI PSSM incorporates selectivity information for six positions (from −2 to +4 relative to the pTyr) of a peptide, whereas most of the PSSMs for SH2 domains used in Scansite were derived from earlier studies that addressed the selectivity from pTyr +1 through pTyr +3 (10,11). To restrict the return from a search to target proteins that have a high probability to be physiologically relevant, SMALI contains an optional filter for phosphorylated peptides. The physiological relevance of a predicted interaction may be further enhanced by applying two additional filters namely signal transduction and subcellular colocalizations (of the query and subject proteins). These novel features make SMALI a useful approach besides Scansite to identify phosphotyrosine-mediated binding events. Here, we describe the usage of the SMALI program and an experimental approach by which to determine the cut-off value for a prediction. We evaluated the performance of SMALI against Scansite for predicting binding peptides for the NCK, CRK and FGR SH2 domains, and applied SMALI to a representative group of 12 SH2 domains in order to identify the corresponding protein–protein interaction (PPI) network. The SMALI-derived PPI network overlaps significantly with known interactions for these SH2-containing proteins, suggesting that SMALI can recapitulate known interactions and identify novel PPIs. The SMALI program, accessible via http://liblab.uwo.ca/SMALI.htm, is frequently updated to include more modular interaction domains and the corresponding PSSMs. To maximize the usage of these matrices, we are also making them available to other bioinformatic programs such as the Scansite and NetPhorest (Linding et al. unpublished results) that aim at identifying protein-binding events and/or signaling pathways according to the principle of domain-short linear motif recognition.

MATERIALS AND METHODS

Derivation of position-specific scoring matrices

The OPAL membrane was scanned and quantified on a BioRad FluoroImager. A selectivity value \( X_{i,p} \) is assigned to each amino acid \( i \) at position \( p \) in the peptide based on an OPAL result, by subtracting the background signal of the membrane from each data spot. The \( X_{i,p} \) is used to calculate a score \( S_{i,p} \) defined as an element of the query SH2 domain scoring matrix, by the formula \( S_{i,p} = R_{i,p}(\log_2 N - (\sum_{p=1}^{N} R_{p} \log_2 R_{p})) \), where \( N \) is the number of residue types in the OPAL array (\( N = 19 \) except for Cys) and \( R_{i,p} = X_{i,p}/\sum_{i=1}^{N} X_{i,p} \). In this formula, the term \( \log_2 N - (\sum_{p=1}^{N} R_{p} \log_2 R_{p}) \) represents information content of all residues at position \( p \) and \( S_{i,p} \) denotes information content of residue \( i \) at this position. Information content of Cys, which was not included in the OPAL, is set equal to the mean \( S_{i,p} \) value at a given position. A peptide score \( S_m \), or SMALI score, is calculated using the formula \( S_m = \sum_{n=3}^{N} S_{i,p} (P \neq 0) \), assuming entropy independence between positions. A peptide with a larger SMALI score is considered to have a greater propensity for binding to the query SH2 domain. A relative score is defined as the ratio of SMALI score over a cut-off value, corresponding to the score at that separates the top 4.5% of peptides from the remaining Tyr-containing peptides taken from all human proteins in the Swiss-Prot database, with the exception of BRDG1 SH2 (3.5%) and GRB2 SH2 (5.5%).

Peptide array synthesis and probing

Peptide arrays were synthesized following established protocols (12). To determine the ability of the peptides on the array to bind an SH2 domain, the SH2 domain was expressed as GST-fusion and purified to homogeneity on a glutathione affinity column and fast-performance liquid chromatography (FPLC) column. The same procedures used for OPAL screening (13) was used to probe the array for binding to the GST-SH2 protein (applied at 1 \( \mu \)M). Finally, the peptide array was scanned and quantified on a BioRad FluorolImager and the background signal was subtracted from each peptide spot.

Differentiation of binding and nonbinding peptides in an array

While in most cases the spot value will provide the quantitative information about the strength of binding for a peptide on the array, the line between binding and nonbinding peptides becomes blurred when the binding signal is weak. We used the distribution pattern on spot values on an array to determine a cut-off value by which to differentiate binding from nonbinding peptides. When the numbers of binding and nonbinding peptides are comparable, the distribution of spot values follows a bimodal pattern where the peak at a large spot value represents binding, while the peak at the small value represents
non-binding peptides. In this case, the transition point between the two peaks is selected as the cut-off. When the signals are extremely biased, the distribution of spot values can be unimodal, and therefore no apparent transition is detected. This is the case with the BRDG1 SH2 peptide array for which an overwhelming number of peptides showed binding. In this case, we define a non-binding peptide as one with a spot value smaller than the average spot value across the entire array subtracted by 1.5 × SD. Based on the earlier definition, peptides with spot values > 1.3 are considered binding peptides for the BRDG1 SH2 domain (Table S1), > 1.8 for the GRB2 SH2 (Table S2), > 0.8 for NCK SH2 (Table S3), > 0.7 for CRK SH2 (Table S4) and > 0.4 for FGR SH2 (Table S5). The five peptide arrays together contained 16 known binding peptides for different SH2 domains. All are correctly classified, suggesting the classification scheme outlined above is a reasonable representation of the true binding data.

RESULTS

Overview of the SMALI program

The derivation of PSSM based on the experimental data from OPAL screens was described elsewhere (13). Briefly, the OPAL-binding profile for an SH2 domain was obtained and quantified for signal strength at each peptide spot on the array (Figure 1A). The information-entropy algorithm was applied to the signals to generate the corresponding scoring matrix (Figure 1B). The current version of the SMALI program includes two modules, peptide scan and domain scan. The peptide scan module is used to identify short peptides that have a high propensity to bind a modular interaction domain such as SH2. In contrast, the domain scan module is used to identify domains that are preferred by a query protein. To predict peptide ligands for a query SH2 domain, all Tyr-containing peptides in the Swiss-Prot database (14) are retrieved and scored using PSSM for that SH2 domain. Peptides are ranked in a descending order based on the SMALI scores, and a peptide with a larger score is considered to have a greater tendency to bind the query SH2 domain. Inside the peptide scan module, a user could select one of the 76 SH2 domains currently covered by the SMALI site. After selecting a protein database (the Swiss-Prot database is used as a default in SMALI), one can choose to run the program without filters or with filters to restrict the proteins to be included in the output file (Figure 1D). Because the Swiss-Prot database contains over 200,000 tyrosines from human proteins, it is necessary to limit the output size of a SMALI prediction by parsing the output through a number of filters.

Three filters were therefore implemented that may be used individually or in combination. The ‘phosphorylation potential’ filter selects only peptides whose phosphorylation has been experimentally verified. This information is taken directly from the databases PhosphoSite (15) and Phospho.ELM (16). Because SH2 domains bind specifically to pTyr-containing sequences, those that are not phosphorylated on Tyr are unlikely to be of physiological relevance even when they produce large SMALI scores. The application of phosphorylation filter reduces the candidate peptides from over 200,000 to ~8000 (15). The second filter, signaling transduction, limits proteins returned from a search to those involved in signal transduction processes. Because most SH2 domains are involved in cellular signal transduction, the identification of signaling proteins that bind to SH2 domains may have a greater potential to be physiologically relevant. Signaling proteins are identified according to the PFAM domain database and Gene Ontology (GO) terms (17–19). Specifically, a subject is classified as a signaling protein if it contains one or more of the 116 signaling domains defined in the PFAM and/or SMART databases (20,21), or if it is annotated with one or more of the following GO terms or their child terms: signal transduction, signal transducer activity, protein kinase activity, phosphoprotein phosphatase activity and protein amino acid dephosphorylation. The third filter is created to keep in an output only those proteins that colocalize with the query SH2 protein in specific subcellular compartments as annotated in Swiss-Prot. The following compartments are used with this filter: (i) cytoplasm, (ii) nucleus, (iii) mitochondrion, (iv) golgi apparatus, (v) endoplasmic reticulum and (vi) endosome. Approximately 34% of human proteins in Swiss-Prot are annotated with a role in signal transduction, while 71% are assigned to specific cellular compartments. To date, 63 SH2 domain-containing proteins have been annotated by subcellular localization, some of which are identified in more than one cellular compartment (e.g. ABL1 exists in either cytoplasm or nucleus). In cases where different regions of a protein are assigned to distinct subcellular locations (i.e. membrane proteins), the region containing the putative binding site (s) for the query SH2 is considered. For instance, the cytoplasmic region (residues 323–428) of the membrane protein NACHR alpha 10 (Swiss-Prot ID: Q9GZZ6) is scanned if a query SH2 is annotated with cytoplasmic localization.

Typical output format of the peptide scan module is shown in Figure 1E. The output size can be set by a user to 100, 250 or 500. The first two columns of the output file report a SMALI score of the peptide target and its relative score calculated by normalizing the raw SMALI score against a cut-off value (defined as the score corresponding to the top 4.5% of peptides ranked by SMALI, see subsequent sections for detail). A relative score of > 1.0 suggests a strong potential for binding. The output file also includes information about the peptide sequence, the position of Tyr residue in the subject protein, gene name, protein name, GeneBank identification (ID), Swiss-Prot ID, molecular weight of the subject protein and localization if available. To match the prediction with known interactions, the last two columns of the output list interactions between the query and subject proteins that have been curated in PPI databases or in domain-peptide interaction databases such as Phospho.Elm (16). Two PPI databases are currently linked to SMALI: the IntAct database where interactions are derived from experiments (22,23), and the I2D database that combines literature-derived human PPIs with those inferred from other species (24). IntAct contains over 400 interactions that may
involve SH2 domains and the I2D collects ~2000 potential SH2-mediated interactions. The confidence level of an SH2 domain-ligand interaction predicted by SMALI is greater if the corresponding PPI is also listed in a database.

In contrast to the peptide scan module that identify peptide targets for a query SH2 domain/protein, the domain scan module of SMALI is used to identify SH2 domains preferred by a query protein that harbors one or more Tyr phosphorylation sites. A query protein can be specified by its Swiss-Prot/TrEMBL ID or its complete or partial sequence entered in FASTA format in the space provided (Figure 2A). Prior to activate a search, the user has the option of selecting one, a subgroup or all SH2 domains (default). The output file of a domain scan lists the query protein sequence with all tyrosine residues highlighted. In a separate panel, the Tyr-containing peptides are listed along with a group of SH2 domains preferred by them (Figure 2B). The numbers in the parenthesis besides an SH2 domain denotes its relative SMALI score for a given Tyr site. An SH2 domain with a larger relative score has a greater tendency to bind to a Tyr site. The output file lists only those SH2 domains that have a relative SMALI score > 1.0 (see next section for the derivation of relative SMALI score).
Experimental determination of SMALI cut-off values

While it is reasonable to assume that a peptide with a larger SMALI score has a greater tendency to bind a query SH2 domain, this assumption has to be verified experimentally. In addition, a cut-off value is needed to limit the size of the output file and to identify interactions that have a high probability to occur. Moreover, a given peptide may produce different SMALI scores for different SH2 domains, and it would be impossible to determine which SH2 domain is preferred by the peptide based on the raw SMALI scores. Therefore, it is necessary to derive a relative SMALI score that allows for direct comparison between SH2 domains. To this end, we applied SMALI to predict peptide ligands for the BRDG1 and GRB2 SH2 domains, respectively and synthesized these peptides in an array format to test their binding to the purified BRDG1 SH2 domain (25), whereas a dozen or so have been characterized for the GRB2 SH2 domain. To gauge the repertoire of peptides that potentially bind the BRDG1 SH2 domain, we searched the Swiss-Prot human protein database and retrieved 1488 peptides ranked in the top 5% by SMALI (Table S1). These peptides were then synthesized as an array and screened for binding to the purified BRDG1 SH2 domain following established procedures (12, 13). As shown in Figure 3A, while the majority of peptides belonging to the top two-thirds of list displayed binding to the BRDG1 SH2 domain, only a small fraction of the bottom third exhibited binding, suggesting that the ability of a peptide to bind BRDG1 SH2 domain correlates grossly with the raw SMALI score. Because only a small fraction of all Tyr residues contained in the Swiss-Prot database is expected to be phosphorylated in vivo, we performed a more targeted binding assay for the GRB2 SH2 domain on a set of peptides selected from the Phosphosite database. We selected a total of 720 peptides of which 360 corresponded to the peptides with large SMALI scores (upper half in Figure 3B) and the remaining 360 were taken randomly from the Phosphosite database (Table S2). While most peptides predicted by SMALI indeed exhibited binding to the GRB2 SH2 domain, only a small fraction of the randomly chosen peptides (lower half in Figure 3B) showed detectable binding.

To correlate the peptide array results with the SMALI score, we calculated the experimentally observed ‘hit-rates’ of peptide-domain interactions and graphed them against the corresponding SMALI scores (at 0.1 or 0.2 intervals). It is apparent from Figure 3C and D that a larger SMALI
Figure 3. Validation of SMALI predicted interactions by peptide array and derivation of cut-off SMALI values. (A) Binding profile of the BRDG1 SH2 domain to an array of 1488 top-ranked phosphotyrosine-containing peptides selected by SMALI from the Swiss-Prot human protein database. (B) Binding of the GRB2 SH2 domain to 720 phosphopeptides taken from the Phosphosite database (15). The first 360 peptides (upper portion) was based on SMALI prediction, whereas the second half (lower portion) was randomly chosen from the database. Dark spots indicate positive binding. (C and D) Distribution of binding peptides over SMALI scores for the BRDG1 (C) and GRB2 SH2 (D) domains. The histograms show ‘hit rate’, defined as the percentage of binding peptides, at a given SMALI score range (in increments of 0.1 and 0.2, respectively for C and D). (E and F) An optimal SMALI cut-off value is arbitrarily defined as the SMALI score that produces the greatest F-measure. F-measure = 2 × precision × recall / (precision + recall), where precision = binding peptides correctly predicted/binding peptides predicted and recall = binding peptides correctly predicted/real binding peptides. For the BRDG1 SH2 domain, the SMALI score 1.4 produced the largest F-measure 0.84 (E). Coincidently, this SMALI value corresponds to a hit-rate of ~50%. For the GRB2 SH2 domain, the cut-off SMALI score is 1.6. (F and G) Distribution of all Tyr-containing peptides (total 203 494) in Swiss-Prot human database according to SMALI scores calculated using PSSM for BRDG1 (G) or the GRB2 SH2 (H) domain. The SMALI cut-off of 1.4 for the BRDG1 SH2 domain corresponds to the top 3.5% scoring peptides located to the right of the cut-off value (G). For GRB2 SH2, the cut-off corresponds to the top 5.5% peptides ranked according to SMALI.
Table 1. Known GRB2 SH2-peptide interactions re-examined in the peptide array experiment

<table>
<thead>
<tr>
<th>SH2 Protein (Alias)</th>
<th>Description</th>
<th>Peptide</th>
<th>pY site</th>
<th>SMALI score</th>
<th>Peptide array</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCR_HUMAN (Bcr)</td>
<td>Breakpoint cluster region protein</td>
<td>177</td>
<td>KPFpYVNVEF</td>
<td>2.67</td>
<td>+</td>
<td>(34)</td>
</tr>
<tr>
<td>IRS1_RAT (Irs1)</td>
<td>Insulin receptor substrate 1</td>
<td>895</td>
<td>PGEpYVNIIF</td>
<td>2.61</td>
<td>+ (35,36)</td>
<td></td>
</tr>
<tr>
<td>FAK2_HUMAN (PYK2)</td>
<td>Focal adhesion kinase 2</td>
<td>881</td>
<td>DLVPYLNME</td>
<td>2.53</td>
<td>+ (37)</td>
<td></td>
</tr>
<tr>
<td>ERBB2_HUMAN (ErbB2)</td>
<td>Receptor tyrosine-protein kinase erbB-2</td>
<td>1139</td>
<td>QPEpYVNQPD</td>
<td>2.51</td>
<td>(38)</td>
<td></td>
</tr>
<tr>
<td>FAK1_HUMAN (FAK)</td>
<td>Focal adhesion kinase</td>
<td>925</td>
<td>DKVpYVENVTG</td>
<td>2.43</td>
<td>+ (39)</td>
<td></td>
</tr>
<tr>
<td>SHC1_HUMAN (Shc)</td>
<td>SHC-transforming protein 1</td>
<td>427</td>
<td>DPsypYVNYQN</td>
<td>2.42</td>
<td>+ (40)</td>
<td></td>
</tr>
<tr>
<td>VGFRI_HUMAN (VEGFR-1)</td>
<td>Vascular endothelial growth factor receptor 1</td>
<td>1213</td>
<td>DVRpYVNAFK</td>
<td>2.41</td>
<td>+ (41)</td>
<td></td>
</tr>
<tr>
<td>PFGFR_HUMAN (PDGFR-β)</td>
<td>Beta-type platelet-derived growth factor receptor</td>
<td>716</td>
<td>AEIPYSNALP</td>
<td>2.40</td>
<td>+ (42)</td>
<td></td>
</tr>
<tr>
<td>LAT_MOUSE (LAT)</td>
<td>Linker for activation of T-cells family member 1</td>
<td>175</td>
<td>DDdpYVNVPE</td>
<td>2.38</td>
<td>+ (43)</td>
<td></td>
</tr>
<tr>
<td>TIE2_HUMAN (TIE2)</td>
<td>Angiopoietin-1 receptor</td>
<td>1102</td>
<td>RKTPYVNTTL</td>
<td>2.35</td>
<td>+ (44)</td>
<td></td>
</tr>
<tr>
<td>LAT_MOUSE (LAT)</td>
<td>Linker for activation of T-cells family member 1</td>
<td>235</td>
<td>APhpYENLQE</td>
<td>2.24</td>
<td>+ (45)</td>
<td></td>
</tr>
<tr>
<td>PTN1_HUMAN (Ptn11)</td>
<td>Tyrosine-protein phosphatase non-receptor type 11</td>
<td>546</td>
<td>GHPpYTNIKY</td>
<td>1.94</td>
<td>+ (46)</td>
<td></td>
</tr>
<tr>
<td>SHC1_HUMAN (Shc)</td>
<td>SHC-transforming protein 1</td>
<td>349</td>
<td>DHQpYYNDFP</td>
<td>1.86</td>
<td>+ (47)</td>
<td></td>
</tr>
</tbody>
</table>

a) Protein names are according to Swiss-Prot convention with the commonly used alias given in parenthesis.

b) Peptides showing positive binding in the array (Figure 3B) are identified with ‘+’. See Methods section for details of experimentation.

score generally corresponds to a greater hit rate for either the BRDG1 or the GRB2 SH2 domain. To generate a cut-off value for SMALI prediction, we next calculated the F-measure and plotted it against the SMALI score (Figure 3E and F). We arbitrarily defined a SMALI cut-off as the score corresponding to the greatest F-measure value, which represents the best compromise between precision of prediction and the rate of recall. For the BRDG1 SH2 domain, the cut-off of 1.4 corresponds to peptides ranked in the top 3.5% by SMALI (Figure 3G).

BRDG1 SH2 domain, the cut-off of 1.4 corresponds to 82% of the peptides with the pY site. In a previous study, we synthesized 22 peptides and measured their respective dissociation constants (Kd) for the BRDG1 SH2 domain in solution (13). Half of these peptides have a SMALI score >1.4, whereas the remaining half has scores below the cut-off. For the first half, 10 (or 91%) displayed strong binding in solution. In contrast, 9 (or 82%) of the second group of peptides exhibited weak or no binding to the BRDG1 SH2 domain. These results suggest that the cut-off is suitable for identifying authentic binding partners for BRDG1.

Analysis of the F-measure led to a SMALI cut-off value of 1.65 for the GRB2 SH2 domain, which corresponds to the top 5.5% of all Tyr-containing peptides collected in the Swiss-Prot human protein database (Figure 3H). Interestingly, all 13 known ligands of the GRB2 SH2 domain have scores greater than the cut-off, were correctly identified by SMALI, and showed strong binding in the peptide array screen (Table 1). Therefore, the experimentally determined cut-off value is suitable for identifying physiological binding partners for GRB2.

While in principle one could carry out similar experiments for the remaining SH2 domains in order to determine the corresponding cut-off values, the amount of work involved would be enormous. Nevertheless, from the binding data obtained for the BRDG1 and GRB2 SH2 domains, it is reasonable to assume that the top 4.5% (average cut-off value for the BRDG1 and GRB2 SH2 domains) of peptides ranked by SMALI have a high probability to bind a query SH2 domain. We therefore set the SMALI score that separate the top 4.5% of peptides from the remainder (except for the BRDG1 and GRB2 SH2s) as the reference point for an SMALI prediction. The cut-off value was used as a common denominator to normalize the raw SMALI score. This produces the relative SMALI score listed in Figure 1E, which serves as a measure of propensity for a peptide to bind a query SH2 domain. A relative score of >1.0 indicates high potential, whereas a score smaller than 1.0 indicates a low potential for binding. The assignment of a relative SMALI score also makes it possible to compare and rank different SH2 domains for their propensity to bind a given peptide ligand in the ‘domain scan’ module of the SMALI program.

Comparison between SMALI and Scansite

Scansite is a web-based program capable of identifying domain-binding peptides or kinase substrates using PSSMs derived from screening peptide libraries synthesized chemically or displayed on bacteriophages (26,27). Scansite incorporates three threshold values—‘high’, ‘medium’ or ‘low’ stringency—to determine the accuracy of prediction. For instance, a peptide is reported as a ‘high stringency’ hit if its f score falls within the top 0.2% of all peptides from the same group (i.e. Tyr-containing). Scansite currently incorporates PSSMs for 14 SH2 domains from ABL1, CRK, FGR, FYN, GRB2, ITK, LCK, NCK, SRC, SHIP, SHIP, PIK3R1, PLCG1_N and PLCG1_C, respectively. All matrices have counterparts in SMALI except for the PLCG1_N SH2 domain.

Since both SMALI and Scansite can be used to predict SH2–ligand interactions, we next compared their performance in predicting targets for SH2 domains from NCK, CRK and FGR. For each SH2 domain, the top 336 candidate peptides selected by either Scansite or SMALI were synthesized on a membrane and tested for binding to the SH2 domain. The sequences and ranking orders of the peptides by either SMALI or Scansite are listed in Tables S3–S5. Results of screening the peptide arrays with the corresponding SH2 domain are shown in Figure 4. For peptides predicted by SMALI to bind an SH2 domain,
also Figure 4 and Tables S3–S5 for experimental data).

array experiment (see Materials and Methods section for details; see FGR SH2 domain, based on the distribution of spot values in a peptide

<table>
<thead>
<tr>
<th>SH2 domain cut-off</th>
<th>SMALI score</th>
<th>SMALI</th>
<th>Scansite</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hit (average, SD) rate (%)</td>
<td>Hit (average, SD) rate (%)</td>
<td></td>
</tr>
<tr>
<td>NCK 1.40</td>
<td>2.02, 0.11</td>
<td>40</td>
<td>1.73, 0.29</td>
</tr>
<tr>
<td>CRK 1.65</td>
<td>2.19, 0.08</td>
<td>90</td>
<td>1.64, 0.28</td>
</tr>
<tr>
<td>FGR 1.35</td>
<td>1.84, 0.10</td>
<td>98</td>
<td>1.49, 0.30</td>
</tr>
</tbody>
</table>

*Peptides with spot values >0.8 are defined as binding peptides for the NCK1 SH2 domain, >0.7 for the CRK SH2 domain and >0.4 for FGR SH2 domain, based on the distribution of spot values in a peptide array experiment (see Materials and Methods section for details; see also Figure 4 and Tables S3–S5 for experimental data).

40% are found real for NCK, 90% for CRK and 98% for FGR. In contrast, 15% of peptides identified by Scansite as binders for NCK were real, while 32 and 87% were real for CRK and FGR SH2-binding, respectively (Table 2). Interestingly, neither program predicted NCK SH2 ligands with a >50% accuracy. We speculate that other factors, such as negative selection and position-dependence, which are not accounted for in a PSSM, may play a ’dominant negative’ role in some SH2–ligand interactions. We calculated the average SMALI score of the Scansite-predicted peptides and found it to be smaller than the average score for the SMALI-predicted peptides. This agrees with our observation that peptides with larger SMALI scores have greater propensities to bind a query SH2 domain (Table 2). Taken together, SMALI exhibited improved accuracy than Scansite in identifying peptide ligands for the three SH2 domains examined herein. Nevertheless, we also observed that the combination of the two programs identified more binding peptides for an SH2 domain than did either alone. Therefore, the integration of SMALI and Scansite should facilitate the identification of SH2 domain–ligand interactions.

Predicting SH2 signaling network by SMALI

The determination of specificity of two-thirds of human SH2 domains makes it possible to gauge the signaling space involving the SH2 domain. To interrogate whether SMALI can aid in the identification of authentic SH2–ligand interactions in a larger scale than described earlier, we applied it to a group of 12 SH2 domains with the phosphorylation filter and identified all peptides with a relative SMALI score >1.0. These SH2 domains were selected to represent the major specificity groups I (motif poYξϕ, where ξ denotes a hydrophilic residue and ϕ is a hydrophobic residue) and II (motif poYxxϕ, where x denotes any residue) (13). The corresponding SH2-containing proteins have also been studied extensively by either conventional or proteomic approaches such that a number of interactions involving them have been reported in the literature. As seen in Table 3, each SH2 domain could potentially interact with hundreds of target proteins, suggesting that other factors such as protein expression and localization must play a role in dictating which interactions occur in vivo. To assess the accuracy of the prediction, we examined the overlap between the predicted and known interactions ranging from 20.3% for Fyn to 49.3% for PIK3. This overlap is significantly greater than expected by chance (P < 0.006), confirming that SMALI is an efficient method to recapitulate authentic SH2–ligand interactions. The overlap would have been more extensive if we had knowledge on which interactions listed in a PPI database indeed involve an SH2 domain and discounted those that are not directly mediated by the query SH2 domain. It should also be noted that the intersection between the PPI space and corresponding SMALI space for a given SH2 protein is rather small (with the exception of Grb2, Table 3), suggesting that many authentic SH2–target interactions await identification or experimental validation.

DISCUSSION

It is clear from comparative genomic analyses that signaling domains have undergone a drastic expansion in multicellular organisms (28,29). Taking the SH2 domain for example, in contrast to yeast that contains no functional SH2 domain, a human cell harbors over a hundred such domains. The same pattern of domain expansion is also observed for other signaling modules such as SH3, PTB and PDZ, to name just a few. The abundance of these interaction modules in the human genome suggests that they play important roles in regulating normal cellular function (30). Because a number of prevalent signaling domains promote PPIs by binding to short linear motifs present in other proteins, delineating the specificity of these
domains provides an effective means to decipher the multitude of protein interactions mediated by them. Additionally, the specificity information allows for ready identification of potential binding partners for an interaction domain. In this regard, Scansite was developed to capitalize on the knowledge of domain and kinase specificity, and has become an essential tool in the toolbox of signal transduction (26,27). The SMALI method described herein utilizes the same principles as those guided the Scansite, but is distinguishable from the latter in the following. First, the current version of SMALI contains specificity information and the corresponding scoring matrices for 76 human SH2 domains, making it possible now to predict phosphorysineptide–SH2 domain interactions at or near proteome scale. The origin of the PSSMs (13) dictates that SMALI is dedicated to the prediction of human or other PPIs. Second, the scoring matrices employed by SMALI contain experimentally derived selectivity information for positions −2 through +4 with respect to the invariant pTyr. In comparison, most SH2 matrices employed by Scansite contain experimentally derived selectivity information on the C-terminal residues only (10,11). Although we have not subjected it to rigorous tests yet, the inclusion of N-terminal selectivity may enhance the accuracy of prediction since it allows distinctions to be made between two peptides that may contain an identical C-terminal sequence. Moreover, some SH2 domains, including those from BRDG1, PLCγ1 and SHP2, have shown selectivity beyond P+3. Third, SMALI is imbedded with several filters to limit the return from a search to proteins that are most likely to be of physiologically relevance. Of particular usage is the phosphorylation filter, since it limits the output to proteins whose phosphorylation has been experimentally verified.

Fourth, the threshold value for a SMALI prediction is inferred from experiments and the resulting normalized SMALI score can be used as a direct measure of binding propensity of a peptide to a query SH2 domain. The normalized propensity score also eliminates the difference in the range of SMALI scores for different SH2 domains and allow for direct comparison of two SH2 domains for propensity to bind a given peptide.

A useful bioinformatic program should not only be capable of recapitulating known knowledge but also predict novel biology. We have put SMALI to rigorous tests on both functions. SMALI faithfully recapitulated all known interactions mediated by the GRB2 SH2 domain and predicted novel interactions involving the BRDG1 SH2 domain (12). Our network analysis on a set of 12 SH2 domains also revealed a significant overlap between SMALI predicted SH2–ligand interactions and known interactions mediated by the GRB2 SH2 domain (32). A web-based program that was developed recently to identify phosphorylation sites and the corresponding kinases based on linear motif-recognition and network context (32,33). The predicted PTK–substrate–SH2 network not only recapitulates many known interactions, but reveals a number of novel signaling pathways (Li and Li, unpublished data). This exercise suggests that by combining SMALI with existing programs on kinase specificity and/or network analysis, novel signaling pathways can be uncovered.

---

**Table 3. Overlap between SMALI-predicted SH2-ligand interactions and those listed in PPI databases**

<table>
<thead>
<tr>
<th>SH2 domain classification</th>
<th>SH2-containing proteins</th>
<th>SH2-interacting proteins predicted by SMALI</th>
<th>SH2-interacting proteins included in PPI databases</th>
<th>Intersection between SMALI and PPI space</th>
<th>Overlap of SMALI network with PPI databases (%)</th>
<th>Statistical significance of overlap (P-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA</td>
<td>SRC</td>
<td>298</td>
<td>104</td>
<td>63</td>
<td>15 (23.8)</td>
<td>&lt;0.00004</td>
</tr>
<tr>
<td>IA</td>
<td>LYN</td>
<td>204</td>
<td>69</td>
<td>50</td>
<td>13 (26.0)</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>IA</td>
<td>ABL1</td>
<td>253</td>
<td>63</td>
<td>46</td>
<td>11 (23.9)</td>
<td>&lt;0.00006</td>
</tr>
<tr>
<td>IA</td>
<td>FYN</td>
<td>313</td>
<td>99</td>
<td>69</td>
<td>14 (20.3)</td>
<td>&lt;0.0006</td>
</tr>
<tr>
<td>IB</td>
<td>CRK</td>
<td>395</td>
<td>44</td>
<td>35</td>
<td>14 (40.0)</td>
<td>&lt;0.000005</td>
</tr>
<tr>
<td>IB</td>
<td>CRKL</td>
<td>274</td>
<td>40</td>
<td>30</td>
<td>9 (30)</td>
<td>&lt;0.0006</td>
</tr>
<tr>
<td>IC</td>
<td>GRB2</td>
<td>420</td>
<td>383</td>
<td>250</td>
<td>68 (27.2)</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>IC</td>
<td>GRAP2</td>
<td>308</td>
<td>27</td>
<td>18</td>
<td>7 (38.9)</td>
<td>&lt;0.0009</td>
</tr>
<tr>
<td>IIA</td>
<td>PIK3R1</td>
<td>317</td>
<td>98</td>
<td>73</td>
<td>35 (49.3)</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>IIA</td>
<td>PTPN11</td>
<td>288</td>
<td>67</td>
<td>59</td>
<td>20 (33.9)</td>
<td>&lt;0.00001</td>
</tr>
<tr>
<td>IIA</td>
<td>VAV1</td>
<td>170</td>
<td>54</td>
<td>45</td>
<td>13 (28.9)</td>
<td>&lt;0.000005</td>
</tr>
<tr>
<td>IIB</td>
<td>SHC1</td>
<td>275</td>
<td>98</td>
<td>77</td>
<td>22 (28.6)</td>
<td>&lt;0.00001</td>
</tr>
</tbody>
</table>

*The SH2 domain classification is based on (13). Group IA has a common motif poY—φ, IB has poY:xφ, IC has poY:xN, IIA has poY:φ:xφ and IIB has the motif poY[E/D]φ:xφ, where ‘—’ denotes a negatively charged residue, φ denotes a hydrophobic residue and x is any type of residues. Number of proteins predicted to bind to a specific SH2 domain-containing protein in the table by SMALI with a relative score >1.0. The Phosphorylation filter is applied with the prediction. Number of binding proteins for a specific SH2-containing protein according to PPI databases I2D (24) and IntAct (22). SMALI space is the number of proteins (3253) used in the prediction. These include all proteins listed in the PhosphoSite and Phospho.ELM databases that contain a pTyr. Intersection is defined as the protein space covered by both SMALI and the PPI databases. Number of common interactions shared between the PPI databases and SMALI prediction for a given SH2-containing protein. The percentage of overlap (in parenthesis) is calculated by dividing this number by the intersected space between PPI and SMALI. Observed overlap over that expected by chance.
To make full use of the OPAL-derived scoring matrices, we have made them available to other bioinformatic programs such as NetPhorest (Linding et al. unpublished data). We will also make our matrices available to Scansite and related programs that predict PPIs based on linear motifs. Moreover, the SMALI site will be updated regularly to include more scoring matrices derived from OPAL or other experiments. Because the OPAL approach can be applied in principle to any modular domains, including kinases and phosphatases that recognize short linear peptide motifs, we anticipate SMALI will be expanded to the prediction of interactions mediated by a variety of interaction domains and for the identification of kinase substrates in a similar manner as described here. Despite the usefulness of SMALI or Scansite in identifying peptide ligands for an SH2 domain, it should be realized that the physiological relevance of a prediction remains to be established by experiments. An in vitro binding event does not always correspond to an in vivo interaction because other factors such as protein expression, phosphorylation, localization and/or scaffolding may dictate whether a given interaction will indeed occur in a cell.

**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online.

**ACKNOWLEDGEMENTS**

This work was supported by grants from Genome Canada (to S.S.-C.L.) through the Ontario Genome Institute, the Canadian Institute of Health Research (to S.S.-C.L.) and the Canadian Cancer Society (to S.S.-C.L.). S.S.-C.L. holds a Canada Research Chair in Functional Genomics and Cellular Proteomics. Funding to pay the Open Access publication charges for this article was provided by Genome Canada.

**Conflict of interest statement.** None declared.

**REFERENCES**


