doi:10.1016/j.jmb.2006.01.005 J. Mol. Biol. (2006) 357, 1322–1334

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Prediction of Binding Sites of Peptide Recognition Domains: An Application on Grb2 and SAP SH2 Domains

William A. McLaughlin, Tingjun Hou and Wei Wang*

Department of Chemistry and Biochemistry, Center for Theoretical Biological Physics University of California, San Diego, 9500 Gilman Drive La Jolla, CA 92093-0359, USA

*Corresponding author

Determination of the binding motif and identification of interaction partners of the modular domains such as SH2 domains can enhance our understanding of the regulatory mechanism of protein–protein interactions. We propose here a new computational method to achieve this goal by integrating the orthogonal information obtained from binding free energy estimation and peptide sequence analysis. We performed a proof-of-concept study on the SH2 domains of SAP and Grb2 proteins. The method involves the following steps: (1) estimating the binding free energy of a set of randomly selected peptides along with a sample of known binders; (2) clustering all these peptides using sequence and energy characteristics; (3) extracting a sequence motif, which is represented by a hidden Markov model (HMM), from the cluster of peptides containing the sample of known binders; and (4) scanning the human proteome to identify binding sites of the domain. The binding motifs of the SAP and Grb2 SH2 domains derived by the method agree well with those determined through experimental studies. Using the derived binding motifs, we have predicted new possible interaction partners for the Grb2 and SAP SH2 domains as well as possible interaction sites for interaction partners already known. We also suggested novel roles for the proteins by reviewing their top interaction candidates.

Keywords: binding site motif; MM/PBSA; protein–protein interactions; SH2 domain

Introduction

The Src homology 2 (SH2) domain functions as a protein-binding module that is used in the control of cellular signal transduction.[1–3] It can serve as an adapter molecule that coordinates the assembly of intracellular signaling proteins in response to an extracellular signal.[2,4–6] Signals mediated by SH2 domains ultimately lead to alterations of the cellular processes such as growth, differentiation, and metabolism.[7] Malfunctions in SH2 domains can lead to a host of human diseases.[8]

Two examples of the SH2 domains are found in the Grb2 and the SLAM-associated proteins (SAP). The Grb2 protein is composed of an SH2 domain and two flanking SH3 domains.[9] A primary function of the Grb2 protein is to bind to protein receptors at the cell surface *via* its SH2 domain and to bind to the SOS protein through its SH3 domains, thereby coupling SOS protein to the membrane where it can activate the Ras protein to initiate a kinase signaling cascade that ultimately leads to modifications in transcription (reviewed by Schlessinger).[10] The SAP protein consists solely of the SH2 domain and is a regulator of signaling events induced by members of the SLAM-related protein receptors found on the surface of T and NK cells.[11–14]

A common feature of all SH2 domains is that binding to an interaction partner is regulated, in part, by the phosphorylation state of a tyrosine residue within that partner.[7] While phosphorylation is required for binding for most SH2 domains,[15] there are SH2 domains that bind to their partners in the absence of phosphorylation, albeit with lower affinity.[3,11,16] Using the method of

Abbreviations used: SH2, Src homology 2; SAP, SLAM-associated protein; HMM, hidden Markov model.

E-mail address of the corresponding author: wei-wang@ucsd.edu

peptide library screening (finding the sequence binding preference based on affinity measurements of oriented degenerate set of random peptide sequences), it has been demonstrated that the sequence determinants of binding or binding specificity of an SH2 domain depend partly on the sequences flanking the tyrosine phosphorylation site.[17,18] For the majority of SH2 domains characterized to date, that specificity is dictated by residues C-terminal to the phosphorylation site, but in some cases, such as the SAP SH2 and the EAT2 SH2 domain, amino acid residues at positions C and N-terminal to the phosphorylation site have been demonstrated to play a role in binding.[19,20]

Peptide library screening has been applied to a number of SH2 domains and there has been accumulation of binding sequence motifs corresponding to each of the domains that have been studied. Many of these binding motifs have been compiled into a web resource called SCANSITE.[21,22] A searching script at that website can be used to predict interaction partners of particular protein domains and thus provides a starting point for identifying candidate interaction partners. Since SCANSITE searching can be conducted only for those domains that have been experimentally characterized, and the strong binding peptides present in the random library but not in the human genome may bias the binding motif determined by the peptide library experiment, alternative techniques are currently needed.

With the aim of addressing these limitations, we have developed a computational method for identifying binding candidates of the modular domains and applied it to the SH2 domains in the proteins of Grb2 and SAP. The method does not rely on peptide library experiments and combines information obtained from binding affinity estimation and sequence motif preference, which is different from the previous approaches using only either type of the information.[22,23] We first created three-dimensional models of known binding peptides as well as randomly selected peptides from the human proteome in complex with the SH2 domain. We next estimated their binding free energies using the molecular mechanics/Poisson–Boltzmann solvent-accessible surface area (MM/PBSA) method.[24,25] These peptides were then clustered based on the binding energy and sequence characteristics and a binding motif was extracted from the cluster of peptides containing those known to interact. The resulting motifs were represented by hidden Markov models (HMMs)[26] and utilized to scan a representative set of human protein sequences in the SWISS-PROT database for likely interaction partners,[27] among which experimental documentation for an interaction with the associated SH2 domain was identified. Possible sites of interaction with the associated SH2 domain were also identified for each interaction candidate. Moreover, based on a literature review of the candidate proteins, new biological roles for the SAP and Grb2 SH2 domains were inferred.

Results

Energy measurements separate known binders from random peptides

The known binding peptides should, on average, have more negative or more favorable binding free energies than peptides selected at random from the background. Two energy measurement protocols were examined with respect to how well the known binding peptides could be separated from the background set of peptides using MM/PBSA. One protocol had the peptides in a phosphorylated state while the other had the peptides in an unphosphorylated state. Student's *t*-test was used to evaluate the significance of the difference between the means.

For peptides in the phosphorylated state, the *p*-value associated with the difference in the mean binding energies of the 15 known peptide binders *versus* the 1400 randomly selected peptides in the Grb2 SH2 domain dataset was 6.41×10^{-5}. For the peptides in the unphosphorylated state, the *p*-value associated with the separation of the two means was lower at 2.31×10^{-9}, which indicated a better separation. Similarly, the *p*-value associated with the separation of the mean binding energy for 11 known binders and the 1799 other peptides in the SAP SH2 domain dataset was lower for the unphosphorylated peptides (7.37×10^{-6}) than for the phosphorylated peptides (3.48×10^{-5}). Figure 1 illustrates the separation between the known binders and peptide candidates in an unphosphorylated state by a histogram plot.

To predict binding motifs and interacting partners of SH2 domains, we chose to use unphosphorylated peptides for the following reasons. First, in our method, rather than to calculate the binding free energy for each binding or non-binding peptide accurately, we only need to establish two distinctive distributions for binders and non-binders. We assume that excluding phosphate does not distort these two distributions, which seems reasonable based on the comparison between the distributions of energy calculations using phosphorylated and unphosphorylated peptides. Second, the binding energy contribution by the phosphate moiety was similar for the known binding peptides and the background set of peptides, and there was a relatively high error associated with its calculation. The binding energy contributed by the phosphate moiety was estimated by subtracting the binding energy of the phosphorylated peptides from the binding energy of the unphosphorylated peptides. For the known binding peptide the average energy contribution due to phosphate binding was 64.78 ($\pm$13.01) kcal/mol for Grb2 and 75.28($\pm$8.83) kcal/mol for SAP. For the background set of peptides, the average energy contribution due to phosphate was estimated to be 71.41($\pm$17.80) kcal/mol for Grb2 and 75.77($\pm$24.34) kcal/mol for SAP. Therefore, the average contribution of phosphate to binding and