J|A|C|S
COMMUNICATIONS

Published on Web 09/29/2006

Circumventing the Problems Caused by Protein Diversity in Microarrays: Implications for Protein Interaction Networks

Andrew Gordus and Gavin MacBeath*

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138

Received July 26, 2006; E-mail: macbeath@chemistry.harvard.edu

Over the past decade, much effort has been applied to defining protein−protein interactions on a large scale. The hope is that the resulting networks will reveal new biology and provide insight into how cells and organisms are organized at a systems level. Various assays have been used for these studies, including those based on the yeast two-hybrid system,[1,2] affinity purification of protein complexes,[3] and protein microarrays.[4−6] One concern with all of these approaches is that proteins vary considerably with respect to their physical properties. Some proteins express well in heterologous cells, while others do not. Some proteins are soluble and monomeric, while others are sticky and tend to aggregate. How, then, can we expect to obtain accurate data from any standardized assay performed in high-throughput?

One way to alleviate this problem is to adopt a domain-oriented approach.[2,4,6,7] Most eukaryotic proteins are modular, comprising both interaction and catalytic domains.[8] By focusing on families of interaction domains, it is more likely that the proteins under investigation will exhibit similar properties. Protein microarrays provide a system in which environmental conditions can be defined and concentrations specified.[9] Yet even with this control, can we reasonably expect different domains to behave similarly enough to obtain accurate information? To answer this question, we assembled a collection of seven Src homology 2 (SH2) domains, which mediate protein−protein interactions by binding to sites of tyrosine phosphorylation on their target proteins. We have previously cloned, expressed, and purified virtually every human SH2 domain.[6] From this set, we selected seven that express well in *Escherichia coli* and are >88% monomeric. Each domain was produced with an amino-terminal thioredoxin tag to favor solubility and an amino-terminal His_6 tag to facilitate purification.

To investigate how uniformly the domains behave in a microarray format, we labeled each domain with cyanine-5 (Cy5), yielding a dye/protein ratio of 0.27−0.53. We then arrayed each domain in quadruplicate on aldehyde-displaying glass substrates at five concentrations, ranging from 10 to 100 μM. Since we used a piezoelectric microarrayer, we were able to measure the volume of the droplets being arrayed (350 pL). By scanning the arrays before and after quenching them, we calculated the fraction of surface area covered by each protein (see Supporting Information). Even when the seven proteins were printed at the same concentration, they varied by up to 3-fold with respect to the fraction of surface area they covered (Figure 1a). More importantly, at concentrations below ~30 μM, the fraction of surface covered by each protein varied in proportion to its concentration. Since the signal intensity in an array experiment depends on the surface density of the immobilized protein, this result underscores the importance of normalizing protein concentrations before printing.

The experiment described above provides a measure of how much total protein is immobilized in each spot. To determine how much active protein is immobilized, we probed the same arrays with EGFR-1016, a 5(6)-carboxytetramethylrhodamine (5(6)-TAMRA)-labeled phosphopeptide derived from the epidermal growth factor receptor that is recognized by five of the domains.[6] Since specific binding is saturable, we probed the arrays in duplicate with eight concentrations of the peptide, ranging from 10 nM to 5 μM. This resulted in 50 independent saturation binding curves performed in quadruplicate: two for each of five domains at each of five concentrations (see, for example, Figure 1b,c). Assuming the system reaches equilibrium during the incubation step, the mean fluorescence of replicate spots (F_{obs}) can be described by

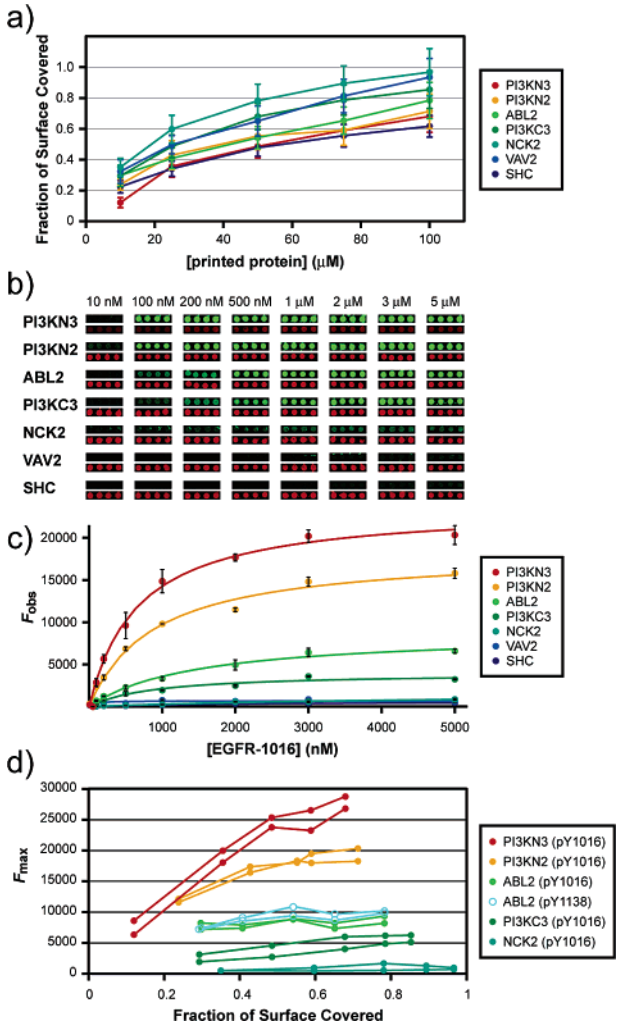

Figure 1. (a) Fraction of surface area covered by SH2 domains when printed at different concentrations; (b) microarray images of SH2 domains printed at 50 μM and probed with phosphopeptide EGFR-1016 at the indicated concentrations; red, Cy5; green, 5(6)-TAMRA; (c) quantified 5(6)-TAMRA fluorescence for the arrays shown in panel b; (d) observed F_{max} for SH2 domains printed at different concentrations. F_{max} was obtained by fitting binding curves for the indicated domain−peptide pairs to eq 1. All error bars indicate the SEM of replicate measurements.

10.1021/ja065381g CCC: $33.50

$$F_{obs} = (F_{max}[pep])/(K_D + [pep]) \quad (1)$$

where F_{max} is the maximum fluorescence, [pep] is the total peptide concentration, and K_D is the equilibrium dissociation constant. Importantly, F_{max} provides a measure of the amount of active protein on the surface. When F_{max} is plotted as a function of the total fraction of surface covered by each domain (Figure 1d), we find that the percentage of active protein varies considerably from one domain to the next. For example, at 70% coverage, the amino-terminal SH2 domain of PI3 kinase-γ (PI3KN3) is 50-fold more active than the SH2 domain of Nck2, even though they both behave well in solution. Curiously, the amount of active protein increases in proportion to the fraction of surface covered for some domains (PI3KN3, PI3KN2, and PI3KC3), but remains constant for the others. Apparently, as the surface becomes increasingly saturated, the fraction of immobilized protein that is active decreases for ABL2 and NCK2. To verify that this behavior is a property of the immobilized domains and is not particular to EGFR-1016, we obtained saturation binding curves using EGFR-1138, a different phosphopeptide that is also recognized by ABL2.[6] Although the peptide binds with a different K_D, the values for F_{max} obtained using this peptide are virtually indistinguishable from those obtained using EGFR-1016 (Figure 1d).

What implications do these findings have for protein microarray experiments? Even when closely related proteins are studied under idealized conditions, they vary with respect to the surface density of active protein (F_{max}). Since the intensity of a spot depends both on K_D, which is biologically relevant, and F_{max}, which is not, the information obtained by probing an array with a single concentration of ligand can be misleading. To illustrate this point, we probed our arrays with eight concentrations of phosphopeptide ErbB3-1054, which is recognized by ABL2, PI3KC3, and PI3KN3.[6] These domains vary by less than 2.5-fold with respect to F_{max} when printed under identical conditions (Figure 2a), yet there is no single concentration of peptide that correctly orders the domains according to their affinities for ErbB3-1054.[6] At low peptide concentration (<700 nM), the order of spot intensities is ABL2 > PI3KN3 > PI3KC3; at higher concentrations, the order is PI3KN3 > ABL2 > PI3KC3 (Figure 2a). When the data are normalized with respect to F_{max}, however, the correct order of affinities emerges: ABL2 > PI3KC3 > PI3KN3 (Figure 2b).

We submit that the way to avoid collecting misinformation in microarray experiments is to obtain saturation binding curves for every protein−ligand interaction. If this is to serve as a general strategy, it is important that the K_D values determined in this way are independent of F_{max}. To investigate this issue, we selected the domain from our studies that exhibits the greatest variation when printed at different concentrations: PI3KN3 (Figure 1d). We then overlaid the binding curves obtained by probing PI3KN3 with peptide EGFR-1016 (Figure 2c). Although F_{max} varies by up to 5-fold across the different concentrations of PI3KN3, the K_D values obtained from these independent curves, as well as from replicate experiments, are independent of F_{max} and are narrowly distributed (mean K_D = 630 nM; s.d. = 88 nM; Figure 2d). We conclude that obtaining K_D values from binding curves provides a way to circumvent the problems caused by protein diversity on microarrays.

What implications do these results have for protein interaction networks? It is likely that the problems highlighted by this study are not unique to protein microarrays. The results of any assay that detects protein−protein interactions, including those based on the yeast two-hybrid system and on affinity purification, depend on the concentrations and activities of the proteins being investigated. To date, most protein interaction networks are binary; proteins are reported either to "interact" or "not interact". At best, this represents a single slice through the underlying quantitative network at a single affinity threshold. More realistically, the threshold varies from one protein to the next based on how well each protein behaves in the assay. As a result, binary networks determined using high-throughput methods may be very misleading.[10] Although protein microarrays are harder to implement than screens that do not require protein purification, their greatest value may lie in the control they offer over ligand and receptor concentrations. Here, we have shown that a threshold-based approach, if not accompanied by quantitative measurements, is inaccurate. How much these inaccuracies affect system-level insights derived from binary networks remains to be determined. We submit, however, that an increased emphasis on obtaining quantitative information should drive future efforts to define large-scale protein interaction networks.

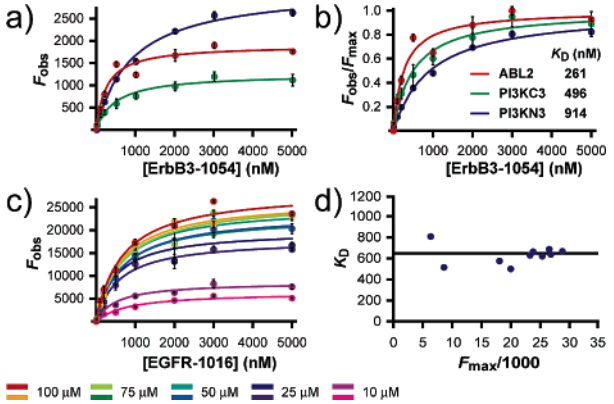

Figure 2. (a) Saturation binding curves for SH2 domains probed with ErbB3−1054: red, ABL2; green, PI3KC3; blue, PI3KN3. (b) The binding curves of panel a normalized to F_{max} and (c) binding curves for PI3KN3 printed at different concentrations are shown. The curves were obtained using EGFR-1016. (d) Data are the K_D values obtained by fitting the curves in panel c to eq 1, plotted as a function of F_{max}. All error bars indicate the SEM of quadruplicate spots.

Acknowledgment. This work was supported by awards from the W. M. Keck Foundation and the Arnold and Mabel Beckman Foundation. A.G. is the recipient of an NSF Graduate Fellowship.

Supporting Information Available: Materials and methods, peptide sequences, calculations, and complete lists of authors for refs 1, 3, and 5. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) (a) Uetz, P.; et al. *Nature (London)* **2000**, *403*, 623−627. (b) Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori, M.; Sakaki, Y. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 4569−4574. (c) Li, S.; et al. *Science* **2004**, *303*, 540−543.
(2) Newman, J. R.; Wolf, E.; Kim, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 13203−13208.
(3) (a) Gavin, A. C.; et al. *Nature (London)* **2002**, *415*, 141−147. (b) Ho, Y.; et al. *Nature* **2002**, *415*, 180−183. (c) Gavin, A. C.; et al. *Nature*, **2006**, *440*, 631−636.
(4) Newman, J. R.; Keating, A. E. *Science* **2003**, *300*, 2097−2101.
(5) Ptacek, J.; et al. *Nature (London)* **2005**, *438*, 679−684.
(6) Jones, R. B.; Gordus, A.; Krall, J. A.; MacBeath, G. *Nature* **2006**, *439*, 168−174.
(7) (a) Espejo, A.; Cote, J.; Bednarek, A.; Richard, S.; Bedford, M. T. *Biochem. J.* **2002**, *367*, 697−702. (b) Stiffler, M. A.; Grantcharova, V. P.; Sevecka, M.; MacBeath, G. *J. Am. Chem. Soc.* **2006**, *128*, 5913−5922.
(8) Pawson, T.; Nash, P. *Science* **2003**, *300*, 445−452.
(9) (a) MacBeath, G.; Schreiber, S. L. *Science* **2000**, *289*, 1760−1763. (b) Zhu, H.; Bilgin, M.; Bangham, R.; Hall, D.; Casamayor, A.; Bertone, P.; Lan, N.; Jansen, R.; Bidlingmaier, S.; Houfek, T.; Mitchell, T.; Miller, P.; Dean, R. A.; Gerstein, M.; Snyder, M. *Science* **2001**, *293*, 2101−2105.
(10) Deeds, E. J.; Ashenberg, O.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 311−316.

JA065381G