Determination of the Substrate Specificities of Protein Kinases by Prediction Algorithms and Substrate Peptide Microarrays

Sommerfeld, D., Lai, S., Safaei, J., Shi, J., Zhang, H. and Pelech, S.

KiNEXUS · Brain Research Centre · UBC

1. Overview

Peptide microarrays were employed to experimentally elucidate the optimal substrate amino acid sequence information for 112 human protein kinases to compare with predictions from the Kinase Predictor 1.0 algorithm. These findings will be used to generate training data to improve Kinase Predictor 2.0, to identify novel kinase-substrate interactions, and to map novel phosphorylation-dependent signalling pathways.

2. Introduction

Protein kinases play a virtually universal role in the regulation of eukaryotic cellular processes by phosphorylating a plethora of protein (and lipid) substrates. Over two thirds of the proteins encoded by the human genome are subjected to phosphorylation on multiple sites, and there may be in excess of 650,000 phosphorylation sites (P-sites) in the human proteome. To ensure signalling fidelity amidst the vast complement of potential substrate P-sites, protein kinases must exert specificity for their substrates and act only on appropriate cellular targets. Substrate recognition by a particular kinase is largely determined by molecular recognition of the amino acid sequence surrounding a target P-site. An important aspect of substrate peptide specificity is that the target P-site falls within a consensus amino acid sequence motif, which exhibits structural and chemical complementarity to the kinase active site. Consensus P-site motifs for individual kinases have been determined through alignment of known substrate P-sites, screening of substrate peptide libraries in solution and arrays, and systematic mutation of substrate sequences. However, the substrate consensus sequences for most of the 516 human protein kinases remain unknown. To fill this gap in knowledge, there have been several attempts to predict kinase specificities for substrate recognition in the last decade since the sequencing of the human genome.

3. Experimental Approach

Figure 1: Overview of input and output of Kinase Predictor 1.0

INPUT
(a) Aligned amino acid sequences for P-sites from known substrates for 229 'typical' protein kinases (from 9,125 confirmed kinase-substrate pairs)

(b) Aligned amino acid sequences for 488 catalytic domains from 478 'typical' kinases

Algorithm 1 Computing SDRs
Determine SDRs and their interaction probabilities for each amino acid position in the P-site region of optimal substrates using deduced consensus P-sites and kinase catalytic domain information.

Algorithm 2 Prediction of PSSM matrices for all kinases
Generate Profile and PSSM matrices for each kinase catalytic domain using SDR's, predicted amino acid interactions, and the unique amino acid sequence of each of the 488 catalytic domains.

OUTPUT
PSSM matrices of all kinases. PSSMs are used to derive an optimal P-site consensus sequence.

We previously developed computer algorithms to predict the specificities for all of the human typical protein kinases (Kinase Predictor 1.0). These algorithms utilize mutual information and charge information theory to identify Specificity Determining Residues (SDRs) in the catalytic domains of kinases and predict the preferred substrate sequences (13 amino acid residues in length) for 488 typical human protein kinase catalytic domains. They generate a frequency table or Position Specific Scoring Matrix (PSSM) that provides the probability of occurrence for each of the 20 amino acids at each of the 12 positions surrounding a target P-site for a specific protein kinase (Figure 1).

While the accuracy of our algorithms have been tested in silico by comparing the computer-predicted consensus sequences with deduced consensus sequences by alignment of known protein substrates, they have yet to be validated empirically by in vitro assays. To test the accuracy of our algorithms, we used the PSSM for each of 488 typical human protein kinase catalytic domains to generate a list of 'optimal' substrate peptide sequences. We selected 445 peptide sequences (omitting nearly identical consensus sequences for highly related kinases) and printed them onto glass slide microarrays (Figure 2).

Peptide microarrays were incubated with ATP and purified recombinant active kinases (SignalChem Pharmaceuticals). The phosphorylation of the peptides was detected using the fluorescent Pro-Q Diamond® phosphoprotein/peptide stain (Figure 3). Slides were scanned and the images were analyzed using advanced microarray image analysis software (ImaGene® 8.0).
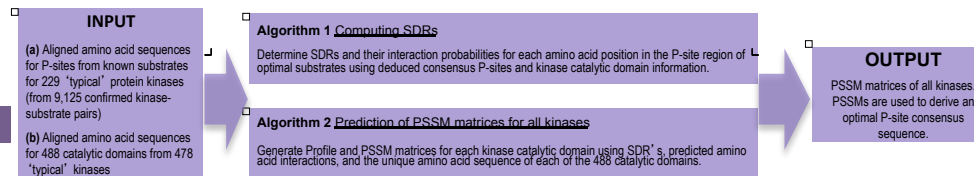
Figure 2: Synthesis and printing of optimal peptide sequences onto glass slide microarrays

PEPTIDE MICROARRAY PRINTING
Peptides printed onto epoxysilane-coated glass slides. Epoxysilane coating provides an epoxy ring that reacts with and covalently binds amine groups on the spotted peptides.
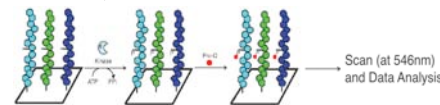Image below is modified from SCHOTT Nexterion® SlideE Product Information Sheet.
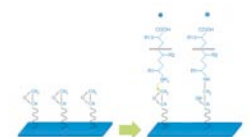
MICROARRAY SET-UP
Every peptide microarray has 4 fields that each contain 445 peptides (plus control spots and orientation markers) printed in triplicate. This setup allows for the testing of 4 different kinases per slide.

Figure 3: Detection of phosphorylation with Pro-Q Diamond phosphoprotein stain

Pro-Q Diamond® (Invitrogen) is a proprietary fluorescent stain that binds directly to the highly negatively charged phosphate group on phosphorylated serine, threonine and tyrosine amino acid residues, and allows for the detection of phosphopeptides (or phosphoproteins) without the requirement for antibodies or radioisotopes. Image below is modified from Gao et al., 2010.*

Scan (at 546nm) and Data Analysis

*Gao, L., Sun, H. and Yao, S.Q. (2010). Activity-based high-throughput determination of PTPs substrate specificity using a phosphopeptide microarray. Peptide Science 94 (6): 810-819.

4. Results

We are aiming to test 204 purified recombinant human protein kinases in high-throughput phosphorylation assays against 445 peptide substrates. To date we have evaluated 112 kinases, and these different kinases exhibit distinct substrate fingerprints/phosphorylation profiles (Figure 4).

The Pro-Q Diamond stain binds directly to the phosphate groups on phosphorylated peptides. Therefore, the signal intensity from peptide spots is proportional to the degree of phosphorylation of the respective peptide substrate, and we can use the signal intensity as a measure of the 'efficiency' of a peptide as a substrate for a particular kinase. Peptide substrates are ranked based on signal intensities (or degree of phosphorylation), and the sequences of the top ranking peptides are aligned and used to derive consensus sequences for target kinases. Table 1 demonstrates the alignment of top ranking peptide substrates for the AMPKα1 kinase and the comparison between the consensus sequences obtained from our peptide arrays (Peptide Consensus), alignment of known substrates P-sites (Protein Consensus), and our Kinase Predictor 1.0 algorithm (Predicted). Comparison of the consensus sequences shows a close match between the different methods, indicating that our algorithms are capable of predicting specificities of kinases, and that our peptide array can be used to test kinase specificities and validate our algorithms.

Our preliminary results indicate that the first generation kinase specificity algorithms are capable of predicting optimal substrate sequences with relatively high accuracy. Table 2 illustrates the comparison of consensus sequences derived from our peptide arrays, computer prediction, and alignment of confirmed substrate proteins for a variety of protein kinases .
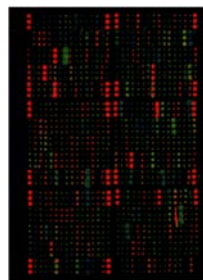
Figure 4: Phosphorylation profile of the kinase AMPKa1
Different kinases display distinct substrate fingerprints. Red indicates high signal intensity corresponding to higher phosphorylation, while green indicates low signal intensity.

Table 1: AMPKα1 peptide substrate ranking and alignment.
Peptides exhibiting signal intensities above 20% of the highest signal detected are aligned and used to derive a consensus sequence (Peptide Consensus).

	Substrate Peptide	Signal
F L S R R G S F G N G G H	SIKpep	4177
G R S R R G S F C H K T G	YANK1 pep	4000
G L G R R G S F A N L G H	AMPKa1 pep	3638
V L S R R L S F G N G G H	MARK3pep	3601
F R S R R L S F G N G G H	MARK2pep	2992
G L G R R T L Y F V V R G	MEK1 pep	2364
G L S R R G S F G N G A H	QIKpep	2125
V L S R R L S F C N F G H	MARK4pep	2035
G L G R R G S F C V F G G	TRIOpep	1813
G L G R R G S F V V F G H	TSSK1 pep	1681
G R S R R P L Y F Y V R G Y	MEK2 pep	1668
G L G R P G S F G N F G H	TSSK3 pep	1566
G R S R R G S F C N G G H	BRSK1/2 pep	1561
K L S R R G S F R F N F H	QSKpep	1458
G L D R R G S L L G K G I	TESK2pep	1444
G L G R R G L Y F Y V G G G	SuRTK106 pep	1377
G L G R R G S L F F F G	PKD1/2/3 pep	1365

f L s R R X S F X nv X X h	Peptide Consensus
X l r R X S F X a n L X h	Protein Consensus
X lm R X S F a n L X h	Predicted

Table 2: Comparison of consensus sequences obtained from peptide arrays (Peptide Consensus), Kinase Predictor 1.0 algorithm (Predicted), and alignment of known substrates P-sites (Protein Consensus)
Shown are the comparisons of derived consensus sequences for protein kinases of various groups and families. Our results also demonstrate the similarities in substrate specificities between members of the same kinase groups.

Group	Kinase	Peptide Consensus	Predicted Consensus	Protein Consensus
AGC	AKT3			
AGC	PKAca			
AGC	PKAcg			
CAMK	AMPKa1			
CAMK	PIM2			
CAMK	CAMK1			
CMGC	GSK3a			
CMGC	JNK1			
CMGC	p38d			
Other	AuroraB			
STE	MEK1(MU)			
STE	PAK3			
TK	Abl1			
TK	JAK2			
TK	HER4			
TK	FRK			
TKL	IRAK4			
TKL	Raf1			

5. Significance and Future Directions

The substrate phosphorylation data obtained from testing all of 204 protein kinases against our peptide microarrays will be used to expand the input datasets for training of our algorithms. In addition, the acquisition of 3D structural information of kinase catalytic domains, the incorporation of a larger number of SDRs, and the adoption of an empirically-derived amino acid interaction table will also be included in the development of second generation Kinase Predictor algorithms with improved accuracy and predictive power for mapping phosphoproteome interactions. Importantly, the in silico prediction of kinase specificities and substrate interactions can significantly facilitate the identification of kinase-substrate interactions in vivo.

Understanding the molecular determinants that guide substrate recognition by a kinase is essential to understanding the role of individual protein kinases in particular cellular processes. The knowledge of substrate specificities that can be gained through our work has significant implications to the mapping of phosphorylation sites in the human proteome, the identification of previously uncharacterized substrates, and the generation of model substrates for small-molecule inhibitor design for drug development. The identification of novel P-sites and substrates for kinases may ultimately lead to a high-resolution map of kinase-dependent phosphorylation signalling networks inside the cell, and a greater understanding of their role in the pathology of many human diseases, including cancer, diabetes, neurodegenerative and immune disorders.