PhosphoBlast, a Computational Tool for Comparing Phosphoprotein Signatures among Large Datasets* S

Yingchun Wang and Richard L. Klemke‡

Identification of specific protein phosphorylation sites provides predicative signatures of cellular activity and specific disease states such as cancer, diabetes, Alzheimer disease, and rheumatoid arthritis. Recent progress in phosphopeptide isolation technology and tandem mass spectrometry has provided the means to identify thousands of phosphorylation sites from a single biological sample. These advances now make it possible to profile global changes in the phosphoproteome at an unprecedented level. However, although this technology is generating a wealth of information, there is currently no efficient means to identify phosphoprotein signatures shared among large phosphoprotein databases. Identification of common phosphoprotein signatures found in biologically relevant systems and their conservation throughout evolution would provide valuable insight into mechanisms of signal transduction and cell function. Here we describe the development of a computational program (PhosphoBlast) that can rapidly match thousands of phosphopeptides that share phosphorylation sites within and across species. PhosphoBlast analysis of several large phosphoprotein datasets from the literature revealed common phosphorylation signatures shared across diverse experimental platforms and species. Moreover PhosphoBlast is a powerful analysis tool to identify specific phosphosite mutations. Comparison of the mouse and human phosphoproteomes revealed more than 130 specific phosphoamino acid mutations, some of which are predicted to alter protein function. Further analysis revealed that known phosphorylated amino acids are more evolutionarily conserved than the Ser/Thr/Tyr amino acids not known to be phosphorylated. Together our results demonstrate that PhosphoBlast is a versatile mining tool capable of identifying related phosphorylation signatures and phosphoamino acid mutations among complex proteomics datasets in a highly efficient and accurate manner. PhosphoBlast will aid in the informatics analysis of the phosphoproteome and the identification of phosphoprotein biomarkers of disease. Molecular & Cellular Proteomics 7:145–162, 2008.

Identification of physiologically important phosphorylation sites is crucial to understanding mechanisms of signal transduction as they provide functional signatures that are predicative of cellular activity as well as specific disease states such as cancer, diabetes, Alzheimer disease, and rheumatoid arthritis. Reversible protein phosphorylation on serine, threonine, and tyrosine residues has been widely studied and shown to control numerous cellular functions. Current estimates indicate that one-third of all proteins in the eukaryotic cell are phosphorylated at any given time by ~518 kinases encoded by the human genome (1–3). Phosphorylation can control protein function by changing its three-dimensional structure, activity, protein-protein binding properties, diffusion, spatial organization, and stability (1). Phosphoproteins and the kinases/phosphatases that regulate their phosphorylation are under intense investigation and remain attractive drug targets in a number of disease paradigms (4). Recent progress in phosphopeptide isolation technology and tandem mass spectrometry has provided the means to identify thousands of phosphorylation sites from a single biological sample (5–7). These advances now make it possible to profile global changes in the phosphoproteome at an unprecedented scale. Although the stoichiometry of phosphorylation needs to be quantified and functionally validated, this technology has provided new insight into how signaling networks control growth factor signaling (8–10), cell chemotaxis (11), and yeast mating responses (12). However, although a tremendous number of phosphorylation sites have been and will continue to be identified and validated from diverse experimental systems, no computational tool is available to compare phosphoproteomics datasets across studies and laboratories. Such a tool could provide a meaningful comparison of phosphorylation results obtained across diverse experimental platforms, species, laboratories, and cellular systems. This in silico information could provide valuable insight into conserved mechanisms of phosphorylation at the systems level and provide an important resource for biomarker discovery and prognostic indicators of disease.

Previously we developed an automated blast program (BlastPro) that allows the comparison of large proteomics and genomics datasets based on nucleotide and amino acid identity (13). BlastPro provides data compatibility across all platforms including nucleotide to protein and species to species with a high level of accuracy. In this report, we describe a...
In Silico Comparison of Phosphorylation Sites

revised version of the BlastPro program that facilitates analysis of complex phosphopeptide datasets. This new blast program called PhosphoBlast facilitates the comparison of thousands of phosphorylation sites among large phosphoproteomics datasets within and across species. Within minutes PhosphoBlast can match thousands of phosphopeptides based on similarity in amino acid composition and phosphorylation status with a high level of accuracy. Using PhosphoBlast, we compared several large phosphoproteomics datasets published in the literature. Our analysis of these datasets uncovered common phosphoprotein signatures and phylogenetically conserved phosphorylation sites between mouse and human and revealed phosphoamino acid mutations that may alter protein function and biological activity. Our findings demonstrate that PhosphoBlast is a useful mining tool to identify and uncover the relationship of phosphorylation signatures shared among large complex phosphoprotein datasets.

EXPERIMENTAL PROCEDURES

Software—The PhosphoBlast was modified and automated for large scale phosphopeptide comparison from stand-alone Blast (14), which was downloaded from ftp.ncbi.nlm.nih.gov/blast/executables/. The user interface and the automation of stand-alone Blast were programmed using html, Perl, and Javascript. An Apache server was used as the World Wide Web server.

Retrieval of Phosphopeptide Datasets—All phosphopeptide datasets were retrieved on line from published studies as indicated in the text. For each datum, only the accession number, annotation of the protein (protein name), and the corresponding phosphopeptide sequence were retrieved and saved in a tab-delimited text file, which served as the input for PhosphoBlast.

Phosphosite Analysis for Aligned Phosphopeptide Pairs—A matched phosphopeptide pair in the output of PhosphoBlast contains two aligned sequences with the same length, one from query and the other from database. Each phosphopeptide sequence contains not only the characters representing the amino acid residues but also the symbol "*" representing the phosphate moiety and the "-" representing the gap in sequence alignments. Both of the paired phosphopeptide sequences were split into single characters or symbols, which were sequentially saved into two arrays, A and B, respectively. For example, the peptide AS′ET-R can be split and saved into the array A where \( A_1 = \text{Ala}, A_2 = \text{Ser}, A_3 = *, A_4 = \text{Glu}, A_5 = \text{Thr}, A_6 = - \), and \( A_7 = \text{Arg} \), whereas its counterpart peptide AS′ET′R can be split and saved into the array B where \( B_1 = \text{Ala}, B_2 = \text{Ser}, B_3 = - , B_4 = \text{Glu}, B_5 = \text{Thr}, B_6 = - , \) and \( B_7 = \text{Arg} \). In this approach, the sequence information and the phosphosite positions can be saved and tracked. In this particular example, because \( A_3 = * , B_3 = - , A_5 = \text{Ser}, \) and \( B_5 = \text{Ser} \), we can find that both peptides have a Ser residue at the second position, but it is only phosphorylated on the first peptide. The rationale of the algorithm for counting phosphosites on the aligned pairwise phosphopeptides and site-specific analysis is the extrapolation of the example, which can be described as follows. 1) If two phosphosites are matched at sites \( i \), then \( A_i = B_i \), and \( A_{i+1} = B_{i+1} = * \). 2) If one peptide A is phosphorylated at site \( i \) and its counterpart B has the same amino acid residue at the matched position but is not phos-

phorylated, then \( A_i = B_i, A_{i+1} = *, \) and \( B_{i+1} = - \). 3) If two homologous peptides have a Ser/Thr/Tyr exchange at the same position and both are phosphorylated, then \( A_i = B_i, A_{i+1} = *, A_i = \text{Ser/Thr/Tyr}, \) and \( B_i = \text{Ser/Thr/Tyr} \), but \( A_i \neq B_i, 4) \) If two homologous peptides A and B have a Ser/Thr/Tyr exchange at the same position and only A is phosphorylated, then \( A_i = \text{Ser/Thr/Tyr} \) and \( B_i = \text{Ser/Thr/Tyr} \), but \( A_i \neq B_i, 5) \) If there is a point mutation that changes the phosphorylated amino acid in A to a non-phosphoamino acid in B, then \( A_i \neq B_i, A_{i+1} = *, \) and \( B_{i+1} = - \). In this way, we can count and group all phosphosites on the aligned phosphopeptide sequences. The algorithm was implemented into an in-house software to facilitate the analysis and will be included in the PhosphoBlast software package for download.

Determination of the Mutation Rate of Phosphorylated and Unphosphorylated Ser/Thr/Tyr Residues from Human to Mouse—The phosphosites identified by Olsen et al. (6) were mapped to corresponding human protein sequence using the International Protein Index (IPI) database ipl.HUMAN.v3.30 using an in-house software. The protein sequences containing mapped phosphosites were subsequently aligned to their mouse orthologous sequences by searching the mouse proteome sequence database ipl.MOUSE.v3.30 using the software BlastPro (13). The aligned sequences were scanned using an in-house software that utilizes a 10-character frame. Only the frames containing sequences with similarity higher than 60% were used to count the number of phospho- and non-phospho-Ser/Thr/Tyr residues. The phosphosite analyses were performed in the same way as described under “Phosphosite Analysis for Aligned Phosphopeptide Pairs.” The rate of each type of mutations for phosphorylated (or unphosphorylated) Ser/Thr/Tyr was calculated using the following equation: \( \left( n/m \right) \times 100\% \) where \( n \) represents the total number of a specific mutation on a phosphorylated (or unphosphorylated) Ser/Thr/Tyr residue and \( m \) represents the total number of the counted phosphorylated (or unphosphorylated) form of the particular amino acid.

Statistics—The t test was used to evaluate the significance of the difference of the mutation rate of the phosphorylated and unphosphorylated Ser/Thr/Tyr residues from human to mouse. Each particular amino acid substitution (e.g., Ser → Ala) was treated as an observation. The t test was performed using the means and variance from the paired substitution rates of all observations. The analysis was performed using the computer software Excel.

RESULTS

Design of the PhosphoBlast Program—PhosphoBlast uses the fasta format for storing the query and the phosphosite database sequences. The data to be compared (query and database) can be copied from published studies or databases and saved in tab-delimited text files (Fig. 1). The data are then converted to fasta format by the format converter of PhosphoBlast prior to comparison (Fig. 1, step 1). Each datum in the query or database contains a description line and a sequence line. The description line includes the protein accession number and the protein name, and the sequence line contains the amino acid sequence of the experimentally identified peptide and its phosphorylation sites (Fig. 1). Because the query or the database may contain multiple phosphopeptides that are identified from a single protein, different data may have the same description line, which will introduce redundancy to the database causing a problem because the formatdb program in the stand-alone Blast package is unable to format the database if the peptides are not distinctly in-

1 The abbreviations used are: html, hypertext markup language; PS, peptide size; E-value, expect value; IPI, International Protein Index.
files. The format converter transforms the two text files into a fasta format. The converted files can be copied from published sources and saved as tab-delimited text files. The format converter is designed to identify matched peptides (step 3). The raw output of PhosphoBlast is then saved in an html file (step 4), which is scanned by the parser to retrieve the first hit for each queried peptide based on the appropriate E-value and similarity cutoffs set by the user. The data are then formatted into a table that contains the gene accession numbers for the query peptides, names of the protein for each matched query/database peptide, the score (bits), E-value, the similarity value, and sequence alignments (step 5). The accession numbers are hyperlinked to the raw output file, which displays all possible peptide matches in the database and query datasets.

Because researchers may use different nomenclature to designate phosphorylation sites, the phosphopeptide sequences in the query or database also need to be preformatted by PhosphoBlast. For example, a phosphotyrosine residue can be described as pY, Yp, “Y,” and Y* by different researchers. PhosphoBlast can read all these formats and transform them to a single format Y*. p is not used because it may be confused with proline.

Inherited from the Blast program (14), PhosphoBlast uses sequence similarity and E-value to describe the homology of two peptide sequences. Because a high similarity score and low E-value indicate high homology of two peptide sequences, we can use these two values as criteria to filter out false-positive hits. However, it is important to recognize that in general the longer the peptide pair the lower its E-value because the probability of a random match (false-positive hit) between peptides increases with decreasing peptide length. Therefore, a false-positive hit for two long peptides may have a lower E-value compared with a positive hit for two short peptides. If we use a low E-value as a cutoff to filter out the false-positive hits for long peptides, then we may also filter out the positive hits for short peptides. On the other hand, if we use a high E-value cutoff to include the positive hits for short peptides, we may also include some false-positive hits for long peptides. A typical large complex phosphopeptide dataset may contain over a thousand peptides that vary in size from 5 to over 30 amino acid residues in length. If we compare two large datasets using a high E-value to include all positive hits, it would be quite cumbersome to remove the false-positive hits. To resolve this problem, we developed a module, Query separator, to partition the query dataset according to peptide size (Fig. 1, step 2). For example, we can separate a query file into four subquery files in which all peptides have a peptide size (PS) ≥ 20, 15 ≤ PS < 20, 10 ≤ PS < 15, and PS < 10, respectively. Each subquery file can then be blasted against the database using optimal criteria cutoffs to remove the majority of false-positive hits (Fig. 1, step 3). This is particularly important for analysis of small phosphopeptides as the probability of random matching is high due to low sequence complexity. For example, a phosphopeptide AS*DR from protein A may be nonspecifically matched to the partial sequence of KAS*DR in protein B. We analyzed three non-redundant large datasets (5–7) and found that less than 10% of the phosphopeptides were shorter than 10 characters (data not shown). For the long peptides we found that more than 96% of the hits were positive. The remaining 4% were protein isoforms or paralogs of the annotated proteins or some true false-positive hits that are caused by random match of two long peptides containing the same low complexity peptide repeats.

Because phosphopeptide sequences are of low complexity per se, the SEG filter, which filters low complexity or repeat sequences, is turned off for the stand-alone Blast. The scoring algorithm is designed to identify shared and differential phosphorylation sites among large datasets using PhosphoBlast.

Inherited from the Blast program (14), PhosphoBlast uses sequence similarity and E-value to describe the homology of two peptide sequences. Because a high similarity score and low E-value indicate high homology of two peptide sequences, we can use these two values as criteria to filter out false-positive hits. However, it is important to recognize that in general the longer the peptide pair the lower its E-value because the probability of a random match (false-positive hit) between peptides increases with decreasing peptide length. Therefore, a false-positive hit for two long peptides may have a lower E-value compared with a positive hit for two short peptides. If we use a low E-value as a cutoff to filter out the false-positive hits for long peptides, then we may also filter out the positive hits for short peptides. On the other hand, if we use a high E-value cutoff to include the positive hits for short peptides, we may also include some false-positive hits for long peptides. A typical large complex phosphopeptide dataset may contain over a thousand peptides that vary in size from 5 to over 30 amino acid residues in length. If we compare two large datasets using a high E-value to include all positive hits, it would be quite cumbersome to remove the false-positive hits. To resolve this problem, we developed a module, Query separator, to partition the query dataset according to peptide size (Fig. 1, step 2). For example, we can separate a query file into four subquery files in which all peptides have a peptide size (PS) ≥ 20, 15 ≤ PS < 20, 10 ≤ PS < 15, and PS < 10, respectively. Each subquery file can then be blasted against the database using optimal criteria cutoffs to remove the majority of false-positive hits (Fig. 1, step 3). This is particularly important for analysis of small phosphopeptides as the probability of random matching is high due to low sequence complexity. For example, a phosphopeptide AS*DR from protein A may be nonspecifically matched to the partial sequence of KAS*DR in protein B. We analyzed three non-redundant large datasets (5–7) and found that less than 10% of the phosphopeptides were shorter than 10 characters (data not shown). For the long peptides we found that more than 96% of the hits were positive. The remaining 4% were protein isoforms or paralogs of the annotated proteins or some true false-positive hits that are caused by random match of two long peptides containing the same low complexity peptide repeats.

Because phosphopeptide sequences are of low complexity per se, the SEG filter, which filters low complexity or repeat sequences, is turned off for the stand-alone Blast. The scoring algorithm is designed to identify shared and differential phosphorylation sites among large datasets using PhosphoBlast.
matrix PAM30 is used for peptide-peptide comparison instead of BLOSUM62, which is used for protein-protein comparison. The other arguments used for peptide-peptide comparison are the same as recommended by the National Center for Biotechnology Information (NCBI), including the default word size and expect value, which are 2 and 20,000, respectively.

The query peptides have now been formatted and are ready to be compared with peptides in the database using the PhosphoBlast program. Each peptide will be blasted against the database, and the first five significant hits that match the query peptide will be included in the raw output file (an html file) for analysis by the parser (Fig. 1, step 4). The first hit that meets the user-defined similarity and E-value cutoffs will be retrieved and outputted to a table (Fig. 1, step 5). The table contains accession numbers for queried peptides, the protein names assigned to the peptide pairs in the query and database, respectively, the score (bits), the E-value, the percentage of sequence similarity of two matched peptides, and the sequence alignments (Fig. 1). It is notable that the sequence alignment and the protein name assigned to a particular peptide pair provide an important means to validate a positive hit. Also each accession number is hyperlinked to the raw output of PhosphoBlast, which provides a trace-back function that can be used to display all hits for a queried peptide. This is important because a query peptide may match multiple peptides in the database (the same peptide or its homologs) (Fig. 1) and may require further analysis for proper assignment.

Currently the PhosphoBlast is installed on a computer with a 1.6-GHz Pentium M central processing unit. The command-line version of the software will be available through the Cell Migration Consortium gateway as is the BlastPro (13). The source code of PhosphoBlast and all other accessory software described in this study are also available upon request.

**PhosphoMerger Removes Redundancy from Query and Database Datasets**—Many of the published phosphopeptide datasets contain redundant phosphorylation sites due to the repeated identification of the same phosphopeptides or peptides generated by trypsin miscleavage. The latter can result in the generation of peptides of various lengths from the same protein that partially share the same amino acid sequence and phosphosites. The redundancy itself does not affect the PhosphoBlast analysis, but it does make it difficult to count the number of distinct phosphosites in each dataset and in the matched peptide pairs created by PhosphoBlast. To remove the redundancy, we developed the PhosphoMerger program to facilitate PhosphoBlast analysis; it is used to create a non-redundant phosphopeptide dataset for the query or database. The redundancy in a dataset can be removed in two steps. First, PhosphoMerger removes multiple copies of the repeated phosphopeptides. Second, PhosphoMerger reduces the redundancy by assigning the different phosphosites identified for each peptide of a given protein that fully or partially overlapped to an *in silico* merged peptide (Fig. 2). In this way overlapping peptides generated by tryptic miscleavage from a given protein can be merged into a single long peptide containing all identified phosphosites (Fig. 2). This function greatly reduces the size and complexity of the datasets and allows one to accurately determine the number of phosphorylation sites identified in a particular phosphopeptide dataset. In addition, this function allows integration of short phosphopeptides into a single long one and hence dramatically reduces the number of false-positive hits by preventing random matches between short peptides. However, it is important to note that merging phosphosites from multiple peptides that share the same amino acid sequence but dif-

![Fig. 2. Schematic showing the strategy to remove phosphosite redundancy from phosphoprotein datasets using PhosphoMerger.](image-url)
different phosphosites will result in the loss of the differential phosphorylation information. Therefore, to retain the site-specific information of peptide phosphorylation one should not use PhosphoMerger but should rather use the primary mass spectrometry-derived datasets for comparison.

**Criteria for Determining Peptide Similarity**—The similarity and E-value are two critical criteria used to evaluate whether two phosphopeptides share amino acid sequence homology and phosphorylation status. Low E-value and high similarity score indicate high homology between a matched peptide pair. To determine the optimal similarity cutoff that allows the parser of PhosphoBlast to retrieve two phosphopeptides that are identically or differentially phosphorylated as a positive hit, we first need to consider the relationship of its value with the size of the peptides and their phosphorylation status (singly, doubly, or multiply phosphorylated). MS-based approaches primarily identify phosphopeptides equal to or larger than 5 amino acids because the number of fragment ions generated from smaller peptides by MS is usually not sufficient to identify the phosphopeptides. In fact, the smallest phosphopeptide found in our analyses from diverse species (i.e., human, mouse, and yeast) was 5 amino acids in length (5–7, 11, 15–17). Therefore, we chose to use a 5-amino acid peptide to illustrate the rationale for determining the similarity cutoff. If a peptide in a database dataset has a single phosphorylation site (the peptide sequence now contains six characters, and the size of the peptide becomes 6) but has two phosphorylation sites in the query dataset, one of which overlaps with the site in the database (the size of the sequence becomes 7), then the similarity of the two sequences is equal to 85.7% (6 divided by 7). In some rare cases, a five-residue peptide pair may have three phosphorylation sites. If two of the sites are phosphorylated in the database and the third is phosphorylated in the query, then the similarity of the two sequences becomes 71.4% (5 divided by 7). Likewise, we can determine the similarity of differentially phosphorylated six-residue peptide and seven-residue peptide pairs of which the similarity between the doubly phosphorylated form and their singly phosphorylated counterpart is 87.5% (75% if all three sites in the two datasets are different) and 88.9% (77.8% if all sites in the two datasets are different), respectively. For all peptides with equal to or more than 8 amino acid residues, the lowest similarity between their doubly phosphorylated and singly phosphorylated counterparts is 90% if one site is overlapping (80% if all three sites in the two datasets are different). Therefore, a cutoff of 70% can be used to compare all phosphopeptides with at least 5 amino acid residues and at most two phosphorylation sites. In a few cases, a peptide pair may share more than three different phosphosites, but this usually happens with long peptides. Nevertheless their similarity is still high due to the significant number of aligned amino acid sequences. In our experience, 70% is the optimal similarity cutoff for all peptide comparisons within the same species.

To compare peptides from different species, it will be necessary to decrease the similarity cutoff because the similarity score decreases due to differential phosphorylation as well as inherent differences in amino acid sequences between orthologous peptides. In this case, we arbitrarily chose a similarity cutoff of 60% because it is difficult to determine whether a pair of sites is correctly aligned between two matched peptides that share less than 60% homology even though we know that the two peptides are from orthologous proteins (18). Therefore, peptides with sequence similarity below the 60% cutoff were excluded from further analysis.

To determine the E-value cutoff, we analyzed its value in relation to the database and peptide size, respectively. We first created a large mixed database from multiple published phosphosites datasets (5–7, 19). We then retrieved 125, 250, 500, 1000, 2000, 4000, 8000, and 16,000 peptides from the mixed database to create new subdatabases. A 5-residue peptide was blasted against the subdatabases that contain this known 5-residue peptide. The E-value of the positive hit in each blast increased linearly as the database increased in size (Fig. 3A, left). This correlation is also applicable to peptides with different sizes. To evaluate the relationship between the E-value and peptide size, we blasted six different peptides ranging from 5 to 10 amino acids against a database containing 125 peptides that includes the particular six different peptides. The \(-\log_{10}(E\text{-value})\) of the positive hit of each peptide increased linearly with the peptide size (Fig. 3A, right). This correlation is also applicable to different size databases.

Using this approach, we generated a two-dimensional matrix of the E-value correlations associated with different size databases and peptide lengths (Fig. 3B). The matrix can be used to determine the E-value cutoff for the corresponding peptide.
and database size. For example, if the smallest peptide in the query file is 7 residues and the database contains 500 peptides, then the E-value 7.00e−04 should be chosen as the cutoff (Fig. 3B). In this case, E-values higher than 7.00e−04 will introduce false-positive hits, whereas values lower than 7.00e−04 may exclude positive hits. Again one must be aware that the size of a peptide and its phosphorylated counterpart is different because there is an additional character representing the phosphate moiety in the peptide sequence. Therefore, one must increase the size of a phosphorylated peptide by one and then choose the appropriate E-value cutoff from the matrix.

It is important to note that the E-value cutoffs in the matrix are for peptide pairs that are fully matched in amino acid sequence and phosphorylation sites. But in reality, many phosphopeptides may be differentially phosphorylated or contain mismatched amino acid residues (cross-species comparison) and thereby contain gaps and mismatches in the sequence alignment resulting in a higher E-value score. Therefore, to compare differentially phosphorylated peptides or phosphopeptides across species, we must use E-value cutoffs higher than the one given in the matrix to allow for more hits to be retrieved and then manually filter out the false-positive hits by looking at the sequence alignment in the formatted output table.

**Analysis of PhosphoBlast Efficiency and Specificity**—To evaluate the efficiency of PhosphoBlast, we first performed a self-comparison of the large phosphopeptide dataset generated by Olsen et al. (6) that contains 4928 phosphopeptides after removing redundant peptides and phosphosites by PhosphoMerger. The similarity cutoff is 100% because all phosphopeptides should be unambiguously matched. The E-value cutoff is set at 0.5 to allow all positive hits with different peptide length to be retrieved by PhosphoBlast. We did not partition the query dataset into multiple datasets because there is no need to filter out false-positive hits in this particular case. Under these conditions each peptide was correctly matched (ratio, 100%), and the analysis proved to be highly efficient taking less than 2 min to complete.

To further evaluate the efficiency and specificity of PhosphoBlast we compared two large phosphopeptide datasets generated from the same cell line but by different research groups. The query dataset was generated by Beausoleil et al. (5) and contains 1777 phosphopeptides identified from purified nuclei of HeLa cells. The database dataset was generated by Olsen et al. (6) and contains 18,958 phosphopeptides (before removing redundancy) identified from whole HeLa cell extracts. Because the query and database contain many redundant phosphosites, we used PhosphoMerger to create two non-redundant phosphosite datasets from the original query and database, respectively. After filtering, the query dataset contained 1564 phosphopeptides bearing 1965 distinct phosphosites, whereas the database contained 4928 phosphopeptides bearing 9704 distinct phosphosites. It is important to note that of the 9704 sites only 5568 belong to class I sites that were assigned with a confidence level higher than 75%; the rest of the sites were assigned with a confidence level lower than 75% (6). Therefore, this will increase the probability that a peptide in the query may not be fully matched to its counterpart in the database (similarity score <100%) due to the low confidence in phosphosite assignment. The query dataset was first separated by peptide size (20, 15, 10, 8, and 5) into five subquery files. Each subquery file was then blasted against the database using high stringent E-value cutoffs as shown in Fig. 4A. This approach maximally excluded false-positive hits while maintaining the majority of positive hits. A similarity cutoff of 70% was chosen based on the rationale discussed above. The computation took ~10 min to complete. We then removed the false-positive hits by manually examining the sequence alignments of the peptides and their corresponding protein names; this took ~20 min. The ratios (numbers) of positive hits obtained for each of the subquery files (20, 15, 10, 8, and 5) were 98.6 (208), 100 (197), 97.7 (168), 67.3 (37), and 46.2% (12), respectively (Fig. 4A). A comparison of the first three subquery datasets consisting of the longer phosphopeptides (20, 15, and 10) was then repeated using higher E-value cutoffs as shown in Fig. 4A. As expected, the ratio of positive hits decreased slightly to 96.4, 97.1, and 94.1%, but the number of positive hits increased slightly to 212, 204, and 191. The new positive hits retrieved by the higher E-value cutoffs were found to be peptide pairs that either have large differences in the number of phosphosites (and hence lower similarity) or contain longer peptides in the query that were matched to their short counterpart in the database due to peptide misalignment. If we use even higher E-value cutoffs, no new positive hits are retrieved. The stringency of the E-value cutoff used for phosphopeptides smaller than eight characters was sufficient to allow all positive E-value cutoffs; therefore, we did not repeat the comparison using higher cutoffs. These findings indicate that a high stringent E-value cutoff can increase the ratio of positive hits retrieved by PhosphoBlast but may exclude some positive hits. On the other hand, a low stringent E-value cutoff can increase the number of total hits but may introduce some false-positive hits. Therefore, when comparing two phosphopeptide datasets for the first time, it is advisable to use a low stringent E-value cutoff to include all possible peptide pairs and then increase the stringency to filter out the false-positive hits during subsequent runs. The protein annotation, sequence information, and phosphorylation signatures assigned to each phosphopeptide pair are included in the formatted table making it straightforward to manually filter out the false-positive hits (Fig. 1).

In total, 655 peptide pairs from the database and query were matched (supplemental Table S1), whereas 909 peptides bearing 1070 phosphosites are unique in the query, and 4273 peptides bearing 8324 phosphosites are unique in the database (Fig. 4B, Venn diagram). The matched peptide pairs
In Silico Comparison of Phosphorylation Sites

Fig. 4. Determination of PhosphoBlast efficiency. A, a HeLa cell nuclear phosphoproteome dataset (5) containing 1965 phosphosites (query) was separated into five subquery datasets according to PS. Each subquery dataset was then compared with the phosphoproteome of whole HeLa cell extracts containing 9704 distinct phosphosites (database) (6) using the indicated range of E-value cutoffs under high or low stringency. The bars represent the percentage of positive hits from each comparison. B, the Venn diagram shows the number of unique phosphopeptides in the query, database, and the phosphopeptides matched between the two datasets in A. The pie chart shows the number of phosphosites on matched phosphopeptide pairs, including sites unique in query, database, and sites matched between the two datasets. C, the ratio of completely matched phosphopeptides (similarity score = 100%) in three different comparisons including the comparisons between the same datasets as shown in A (a), the same datasets as shown in A but without merging the phosphosites using PhosphoMerger (b), and the same query dataset as shown in A and a subset of the database dataset in A containing only class I phosphosites that were assigned with high confidence (confidence >75%) (6) (c).

contain a total of 1495 phosphosites including the 780 (52.2%) matched, 115 (7%) query-unique, and 600 (40.1%) database-unique sites (Fig. 4B, pie chart, and supplemental Table S1). It is notable that the vast majority of phosphosites identified in each dataset were assigned to peptides other than the 655 matched peptide pairs (Fig. 4B, Venn diagram). In total, 1185 sites are unique in the query dataset including the 1070 sites on the 909 unmatched peptides (Fig. 4B, Venn diagram) and the 115 query-unique sites on the matched peptides (Fig. 4B, pie chart), whereas 8924 sites are unique in the database including the 8324 sites on the 4273 unmatched peptides (Fig. 4B, Venn diagram) and the 600 database-unique sites on matched peptides (Fig. 4B, pie chart). Thus, analysis of these independent experiments uncovered a specific subset of the phosphoproteome in HeLa cells.

It is notable that of the 655 matched peptide pairs only 281 pairs (42.9%) are completely identical in phosphorylation (supplemental Table S1 and Fig. 4C, a). The low ratio of completely matched phosphopeptides is likely due to not only the experimental error resulting from different criteria used to identify and assign phosphorylation sites but also the merging of phosphosites from multiple copies of a peptide bearing different combinations of phosphosites onto a single in silico peptide. For example, if the query contains a singly phosphorylated peptide and the database also contains the same singly phosphorylated peptide and a doubly phosphorylated form, merging the sites using PhosphoMerger will make no change to the singly phosphorylated peptide in the query but will merge the two peptides in the database into a single one containing all phosphosites on the original two peptides. In this case, PhosphoBlast will reveal that the similarity between the singly phosphorylated peptide in the query and the merged phosphopeptide in the database is less than 100%. To maintain the differential phosphorylation signatures from the original datasets, we can compare the query and the database datasets without using the PhosphoMerger program because PhosphoBlast is a stand-alone tool. To demonstrate this ability, we tried to compare the query and database datasets using PhosphoBlast without completely premerging the redundancy using PhosphoMerger. Instead we only partially filtered the redundancy of the original datasets prior to PhosphoBlast analysis by only removing the extra copies of the peptides that have more than one copy bearing identical phosphosites. The comparison revealed 791 matched peptide pairs that do not contain merged peptides of which 452 pairs (57.1%) were identical (supplemental Table S2 and Fig. 4C, b). The increase in identically matched phosphopeptides suggests that many of the peptides in the original datasets have multiple copies bearing a different combination of phosphosites. The comparison without premerging the phosphosites allowed more peptides in the two datasets that share the same amino acid sequence and phosphosites to be matched with similarity score 100%. It is important to note that a phosphopeptide in query may not only be matched to its identical counterpart in database (similarity score = 100%) but also to differentially phosphorylated forms if the counterpart of the query peptide has multiple copies bearing a different combination of phosphosites in the database (similarity score <100%). However, only the hit with the highest
similarity score will be retrieved by the PhosphoBlast parser and outputted to the formatted table. The hits with lower similarity score are also positive but can only be seen through the hyperlink function (Fig. 1). To further evaluate these findings, we removed all non-class I phosphosites in the database that were assigned with low confidence to decrease the possibility of false-positive phosphosite assignments. Comparison of the non-merged query and database using PhosphoBlast revealed 607 matched peptides of which 379 (62.4%) were completely identical (supplemental Table S3 and Fig. 4C, c). Thus, as expected, the percentage of completely matched phosphopeptides increased from 57.1 to 62.4% (379 of 607).

**Comparing Phosphopeptides among Diverse Species**—One of the most important functions of PhosphoBlast is its ability to identify phosphorylation sites shared among diverse species. The conservation of phosphorylation sites among species can reflect important biological functions and helps to validate specific phosphorylation signatures. To compare phosphosites among different species, we need to use lower stringent similarity and E-value cutoffs because a phosphopeptide and its cross-species homolog may differ not only in phosphorylation status but also in amino acid composition. To demonstrate this, we compared two large datasets from mouse (query) and human (database). The query dataset was generated from mouse liver and contains 8529 phosphopeptides (7). After removing the phosphopeptide redundancy and merging the partially overlapping peptides, 3310 distinct phosphopeptides were generated bearing 5813 unique phosphosites. This dataset was separated into five subqueries based on peptide length (20, 15, 10, 8, and 5) as described above. The database is the same dataset obtained from the HeLa cell extracts described above (6). PhosphoBlast matched 918 phosphopeptides between the two datasets that are recognized as orthologous peptides bearing the same or different phosphosites (supplemental Table S4 and Fig. 5A, Venn diagram). On the 918 matched phosphopeptide pairs, 1390 phosphosites were matched, whereas 377 and 593 phosphosites were unique in the query and database, respectively (Fig. 5A, pie chart). Therefore, the matched 918 peptide pairs contain 1767 (1390 + 377) phosphosites representing the query and 1983 (1390 + 593) phosphosites representing the database. In addition to the matched phosphopeptides, 2392 peptides bearing 4046 phosphosites are unique in the query, and 4010 peptides bearing 7721 phosphosites are unique in the database (Fig. 5A, Venn diagram), illustrating the significant difference in phosphosites uncovered in these two studies. To keep track of the original identified differential phosphorylation information, we also compared the two datasets without being preprocessed by PhosphoMerger. In total, 1884 peptide pairs were matched of which 518 are completely conserved between human and mouse in both amino acid sequence and phosphorylation (supplemental Table S5).

**Identification of Mutations That Target Known Phosphorylation Sites**—The differences in phosphosignatures observed between matched human-mouse peptide pairs may be due to differences in phosphorylation or result from mutations occurring during evolution that target Ser/Thr/Tyr amino acid resi-
Substitution of phosphosites with a negatively charged amino acid could act as a natural phosphomimetic altering protein function, whereas neutral or basic amino acids could be silencing mutations. Understanding the divergence of specific phosphosignatures could be informative about the evolution of specific signaling processes and cellular functions. Therefore, we performed a detailed analysis of the human-mouse peptides that showed high homology at the amino acid level but displayed different phosphorylation signatures.

Of the 377 unique phosphosites identified in the query dataset (mouse), 317 (84.1%) of these sites have the same amino acid residue in the same position as their matched counterpart in the human database but lack a phosphate moiety (Fig. 5, A, pie chart, and B, a). Of the 593 phosphosites identified in the human database, 520 (87.7%) sites have the same amino acid residue in the same position as their counterparts in the query dataset but lack phosphorylation (Fig. 5, A, pie chart, and B, a). Thus, the majority of the unmatched phosphosites are caused by observed differential phosphorylation on conserved amino acid residues. Importantly, however, we did discover among the two datasets 10 phosphosites that were phosphorylated but have Ser/Thr exchange mutations (Fig. 5B, b, and Table I). We also uncovered five and 11 phosphosites in the query and database, respectively, that were phosphorylated but have Ser/Thr exchange mutations in the same position of the counterpart peptide, which is not phosphorylated (Fig. 5B, c, and Table II). Finally and most importantly, we identified 45 and 52 phosphosites in the query and database datasets, respectively, that display amino acid substitutions other than Ser, Thr, or Tyr (Fig. 5B, d, and Table III).

The most frequent phosphosite mutations were substitutions with neutral amino acids (Thr → Ala, Ser → Asn, Ser → Ala, and Ser → Gly) (Fig. 5C and Table III). Interestingly three phosphoproteins displayed acidic amino acid (Glu and Asp) substitutions (Table III). In some cases, these mutations can serve as natural phosphomimetics altering protein function. Surprisingly there was also a high frequency of phosphosite mutations involving proline substitutions (Table III). Although these mutations would be expected to prevent phosphorylation, they could also introduce significant changes in protein structure. These data highlight the usefulness of PhosphoBlast to uncover phylogenetically important mutations affecting protein phosphorylation. Overall our findings indicate that PhosphoBlast can be used to uncover conserved and divergent phospho-
rylation signatures among large complex phosphoprotein datasets within and across species.

Analysis of Overall Mutation Rates in Phosphorylated and Unphosphorylated Ser/Thr/Tyr Residues—To further investigate whether known phosphorylated sites (denoted as phosphorylated) are more conserved than those not known to be phosphorylated (denoted as unphosphorylated for convenience) between human and mouse, we chose a representative subset of human proteins of which the phosphosites were identified by Olsen et al. (6) because these proteins contain the most comprehensive set of identified phosphosites. To map these phosphosites to corresponding human proteins, we retrieved protein sequences from the IPI human database (ipi.HUMAN.v3.30). For simplicity, we only mapped the sites

**TABLE II**

Identification of Ser/Thr exchange mutations that show differential phosphorylation between human and mouse orthologs

| Accession (mouse) | Name in Query (Mouse) | Name in Database (Human) | Sequence Alignments |
|-------------------|-----------------------|--------------------------|---------------------|
| IPI00319973       | Membrane associated progestosterone receptor component 1. | Membrane associated progestosterone receptor component 1 | Mouse: 1 EGE8P*VY*S*DEE5P*DEE*AR 22  
|                  |                       |                          | Human: 1 EGE8P*VY*S*DEE5P*KDEE*AR 24 |
| IPI00154109       | Protein c9orf42 homolog. | Protein C9orf42 | Mouse: 1 SNSAPL1HGLS*DS*SPVFQAEAPSAR 26  
|                  |                       |                          | Human: 1 SNSAPL1HGLS*DT*SPVFQAEAPSAR 26 |
| IPI00226265       | Red51-associated protein 1. | Splice Isoform 1 of RAD51-associated protein 1 | Mouse: 1 FINYQ*QEDS*GND*S*DFF1SS*T-TVPNK 29  
|                  |                       |                          | Human: 1 FYNYQ*QEDS*DS-DSSF1SSAT*VPLNK 27 |
| IPI00112339       | Epithelial protein lost in neoplasms. | Splice Isoform 3 of Epithelial protein lost in neoplasms | Mouse: 18 ES*PHS*PGVDAP1K 34  
|                  |                       |                          | Human: 1 ET*PHS*PGVDAP1K 16 |
| IPI00311534       | Poly a-specific ribonuclease parn. | Poly(A)-specific ribonuclease PARN | Mouse: 1 EAALEDRESVEVS*DS*ELEQTDCTCPLPEGR 33  
|                  |                       |                          | Human: 10 EALEDGVSE8S*DT*ELEQTDCAEPLSEGR 40 |
| IPI00123624       | Predicted: p30 dBC protein. | P30 DBC | Mouse: 1 LEET*EVRS*VAM*NS*EMEY*SSLQOMPK 30  
|                  |                       |                          | Human: 1 LEO*EVRS*VAM*NS*EMEY*SSLQOMPK 30 |
| IPI00172210       | Proliferation potential-related protein. | Retinoblastoma binding protein 6 isoform 1 | Mouse: 1 VACG*EGSSSTLVDYTS*T*STGCGS*PVR 29  
|                  |                       |                          | Human: 1 VTG*EGSSSTLVDYTS*T*STGCGS*PVR 27 |
| IPI00321978       | Ran-specific gtpase-activating protein fragment. | Ran-specific GTPase-activating protein | Mouse: 1 DS*HECDHTST*ENADES*NHPQDFEPVS*LPEGQIK 36  
|                  |                       |                          | Human: 1 DT*HECDHTST*ENTIDES*NHPQDFEPVS*LPEGQIK 37 |
| IPI00274030       | Splicing factor, arginine/serine-rich 8 isoform 1. | Splicing factor, arginine/serine-rich 8 isoform 1. | Mouse: 1 SGT*DS*DBE6DDEEDGSYLPSLFK 28  
|                  |                       |                          | Human: 1 SGVS*DS*DBE6DDEEDGNYLPSLFK 29 |
| IPI00321860       | Phosducin-like protein. | Phosducin-like | Mouse: 1 LQYY*S*DS*EDEDS*DHEKD 21  
|                  |                       |                          | Human: 1 LQYY*S*DS*EDEDS*DHEKD 22 |
| IPI00319933       | Protein kinase c and casein kinase ii substrate protein 3. | Protein kinase c and casein kinase in neurons protein 3 | Mouse: 1 DGT*APPGQ*PSS*PCS*OQEDWS*DHEEPR 29  
|                  |                       |                          | Human: 1 DGT*APPGQ*PSS*PCS*OQEDWS*DHEEPR 32 |
| IPI00110265       | Splice isoform 1 of nad-dependent deacetylase siru2-2. | Splice Isoform 4 of NAD-dependent deacetylase siru2-2 | Mouse: 1 VQEAQDS*DS*DT*EGGATGGAEMDFLR 29  
|                  |                       |                          | Human: 1 VQEAQDS*DS*DS*EGGAGAEGDFLR 26 |
| IPI00454123       | 4930535b03rik protein. | Similar to Cytokinesis protein sepA | Mouse: 1 NTGVSASRPS*PGI*TPPSNL*SGSLO 28  
|                  |                       |                          | Human: 1 NTGVSPASRPS*PGI*TPPSNL*T-SLO 27 |
| IPI00453839       | Mksa1064 protein. | PREDICTED: chromosome 19 open reading frame 7 | Mouse: 1 T*GT*GS*PFAAGN*PAREQDGASL 28  
|                  |                       |                          | Human: 1 T*GS*GS*PFAAGN*PAREQDGASL 28 |
| IPI00130065       | Ras-gtpase-activating protein binding protein 1. | Ras-GTPase-activating protein binding protein 1 | Mouse: 1 S-T-S*PAPADVAPA-GQDLR 18  
|                  |                       |                          | Human: 1 S*S*S*PAPADVAPA-GQDLR 18 |
| IPI00280015       | Chloride channel CLICKLIKE 1. | Mid-1-related chloride channel 1 | Mouse: 1 FHS*GNKS*PEVLR 14  
|                  |                       |                          | Human: 1 FCT*GNKS*PEVLR 13 |
### TABLE III
Identification of mutations that replace phosphorylated mouse or human Ser/Thr/Tyr residues with non-Ser/Thr/Tyr amino acids
Phosphoamino acids substituted with an acidic amino acid residue are inversely displayed. Phosphoamino acids substituted with a proline residue are lightly shaded.

| Accession (Mouse) | Name in Query (Mouse) | Name in Database (Human) | Sequence Alignments |
|-------------------|-----------------------|--------------------------|---------------------|
| IPI00131130       | Splice isoform long of transcription intermediary factor 1-alpha. | Splice Isoform Long of Transcription intermediary factor 1-alpha | Mouse: 1 HEDKCRS*DDDDDDDYQPR 22  
Human: 1 NES*EDKCRS*DDDDDDDYQPR 23 |
| IPI000224200      | Hypothetical acyl-coa dehydrogenase/glutamic acid-rich region containing protein. | Hypothetical protein FLJ10006 | Mouse: 1 VDS*DDDS*DGVS*DK 19  
Human: 3 VS*DDDS*DGVS*DK 22 |
| IPI00227539       | Pogo transposable element with zinc domain. | Pogo transposable element with ZNF domain | Mouse: 1 SLDAEFVS*FAC*APR*SF*PEK 21  
Human: 1 SDGAEFS*VARS*APR*SF*PEK 20 |
| IPI00316133       | Predicted: signal recognition particle 72. | Signal recognition particle 72 kDa protein | Mouse: 1 AVS*S*FPT*S*FPPGS*APAT*SSS*HSSSHPFPR 33  
Human: 1 TVS*F*FPT*S*FPPGS*APAT*SSS*HSSSHPFPR 30 |
| IPI00396671       | ATP-binding cassette sub-family F member 1. | Splice Isoform 2 of ATP-binding cassette sub-family F member 1 | Mouse: 2 LS*VPS*SDDEEFPAP 13  
Human: 1 LS*VPS*SDDEEFPAP 18 |
| IPI00224570       | Camp-dependent protein kinase type II-beta regulatory subunit. | PRKAR2A protein | Mouse: 2 S*YCAE*Y*NPDEEDEDDAER 21  
Human: 2 S*YCAE*Y*NPDEEDEDDAER 18 |
| IPI00360737       | Nuclear envelope pore membrane protein pom 121. | Nuclear envelope pore membrane protein PCM2 121 | Mouse: 1 HDS*G*GS*HSS*AFPELHAN*GYPAAAFYPKPSGLK 34  
Human: 1 HDS*G*GS*HSS*AFPELHAN*GYPAAAFYPKPSGLK 36 |
| IPI00228693       | Hypothetical arm repeat structure containing protein. | Splice isoform of protein WD-repeat protein 75 | Mouse: 1 AAMAPGSS*GSSGGGVHFLS*DS*DEDDDEVQAT*EDR 37  
Human: 1 AAMAPGSS*GSSGGGVHFLS*DS*DEDDDEVQAT*EDR 36 |
| IPI00136689       | Locus46723 protein. | Protein phosphatase inhibitor 2 | Mouse: 3 EOE*S*S*GEED*LYS*PEER 21  
Human: 5 EOE*S*S*GEED*LYS*PEER 24 |
| IPI00172174       | Similar to zinc finger protein 183. | Zinc finger protein 183 | Mouse: 1 ATY*Y*GDL*S*EDS*HEPGLOVYK 26  
Human: 1 AA*Y*GDL*S*EEHNEEPEPGLOVYK 26 |
| IPI00116294       | Tacc2 protein. | Splice Isoform 1 of Transforming acidic coiled-coil-containing protein 2. | Mouse: 1 LONTPAS*FPSR*PETFS*DTPIAK 25  
Human: 1 LONTPAS*FPSR*PETFS*DTPIAK 24 |
| IPI00111169       | Cytosolic phospholipase a2. | Cytosolic phospholipase A2 | Mouse: 1 HIVNSNS*DS*DDA*GFP 20  
Human: 1 HIVNSNS*DS*DDA*GFP 22 |
| IPI00308560       | Wdr44 protein. | Hypothetical protein FLJ11116 | Mouse: 1 EYSVAAT*QS*DQEK 17  
Human: 1 EYSVAAT*QS*DQEK 16 |
| IPI00331289       | Splice isoform B of ras-GTPase activating protein binding 2. | Splice Isoform B of Ras-GTPase activating protein binding protein 2 | Mouse: 2 LEEELEKS*T*EFPFEEAPFQPEK 28  
Human: 2 LEEELEKS*T*EFPFEEAPFQPEK 29 |
| IPI000312128      | Splice isoform 1 of transcription intermediary factor 1-beta. | Splice Isoform 1 of Transcription intermediary factor 1-beta. | Mouse: 1 AAS*AAAAATGAS*AAS*GSSGEGSAGGGEK 36  
Human: 1 AAS*AAAAATGAS*AAS*GSSGEGSAGGGEK 36 |
| IPI00115660       | Treacle protein. | Treacher Collins-Franceschetti syndrome 1, isoform c | Mouse: 1 SATA*ANTVL*ETS*EDNHQALQTPK 30  
Human: 1 SAT*ANTVL*ETS*EDNHQALQTPK 30 |
| IPI001161546      | Camp-dependent protein kinase type II-alpha regulatory subunit. | PRKAR2A protein | Mouse: 1 VS*VCAET*F*NPDEEEDNPPR 21  
Human: 1 VS*VCAET*F*NPDEEEDNPPR 23 |
| IPI001119795      | Splice isoform 1 of ralbp1 associated eps domain containing protein 1. | Splice Isoform 1 of RALBP1 associated Eps domain containing 1 | Mouse: 1 T*S*SDHTNFSPLLWES*DLSENNK 26  
Human: 1 T*S*SDHTNFSPLLWES*DLSENNK 26 |
| IPI00320422       | Periodic tryptophan protein 1 homolog. | Periodic tryptophan protein 1 homolog | Mouse: 1 LOR.gov*E*EKEAG*P3*EDGNGS 24  
Human: 1 LOR.gov*E*EKEAG*P3*EDGNGS 25 |
| IPI00122174       | Predicted: scaffold. | Similar to Scaffold | Mouse: 1 SEPVKEEG*S*BEEQPFAQAT*S-VGDP 27  
Human: 1 SEPVKEEG*S*BEEQPFAQAT*S-VGDP 27 |
| Attachment factor a. | Attachment factor b. | Human | Mouse | Other |
|---------------------|----------------------|-------|-------|-------|
| **Splice isoform alpha of nuclear receptor corepressor 2** | **Splice isoform TRAC-1 of nuclear receptor corepressor 2** | 1 | 1 | 1 |
| IPI00123871 | | 1 | 1 | 1 |
| **Weakly similar to at2g19370 protein, full insertion sequence.** | Hypothetical protein | 3 | 3 | 3 |
| IPI00454028 | | 3 | 3 | 3 |
| **Yap1 protein.** | **65 kDa Yes-associated protein** | 1 | 1 | 1 |
| IPI00108989 | | 1 | 1 | 1 |
| **Similar to Cytokinesis protein sepA** | **Protein tyrosine phosphatase, non-receptor type 12.** | 1 | 1 | 1 |
| IPI00454123 | | 1 | 1 | 1 |
| **Tuftelin-interacting protein 11.** | **Splice isoform 1 of Tuftelin-interacting protein 11** | 1 | 1 | 1 |
| IPI00112101 | | 1 | 1 | 1 |
| **Fatty acid elongase 1 homolog.** | **Long chain polyunsaturated fatty acid elongation enzyme** | 2 | 2 | 2 |
| IPI00308278 | | 2 | 2 | 2 |
| **Splice isoform 3 of protein c2orf17 homolog.** | **Protein C2orf17** | 1 | 1 | 1 |
| IPI00421012 | | 1 | 1 | 1 |
| **Src substrate cortacin.** | **Src substrate cortacin** | 1 | 1 | 1 |
| IPI00118143 | | 1 | 1 | 1 |
| **Novel protein.** | **NSPSbeta3alpha** | 1 | 1 | 1 |
| IPI00022577 | | 1 | 1 | 1 |
| **Novel protein.** | **DKFZP434K1421 protein** | 1 | 1 | 1 |
| IPI00345960 | | 1 | 1 | 1 |
| **RNA binding protein homolog.** | **Splicing coactivator subunit 5Rm300** | 1 | 1 | 1 |
| IPI00225062 | | 1 | 1 | 1 |
| **Rho-interacting protein 3.** | **Myosin phosphatase-Rho interacting protein isoform 1** | 1 | 1 | 1 |
| IPI00125348 | | 1 | 1 | 1 |
| **Predicted: p30 dbc protein.** | **P30 DBC** | 1 | 1 | 1 |
| IPI00123624 | | 1 | 1 | 1 |
| **Histone h1.5.** | **Histone H1.5** | 1 | 1 | 1 |
| IPI00230133 | | 1 | 1 | 1 |
| **Splice isoform 1 of protein aatf.** | **Protein AATF** | 1 | 1 | 1 |
| IPI00123565 | | 1 | 1 | 1 |
| **Predicted: ran binding protein 3.** | **Splice isoform 1 of Ran-binding protein 3** | 1 | 1 | 1 |
| IPI00135190 | | 1 | 1 | 1 |
| **Epithelial protein lost in neoplasm.** | **Splice isoform Alpha of Epithelial protein lost in neoplasm** | 2 | 2 | 2 |
| IPI00112339 | | 2 | 2 | 2 |
| **CCAT/enhancer binding protein zeta.** | **CCAT/enhancer binding protein zeta** | 1 | 1 | 1 |
| IPI00132439 | | 1 | 1 | 1 |
| **Hypothetical protein bc021438.** | **Hypothetical protein FLJ12949** | 1 | 1 | 1 |
| IPI00311761 | | 1 | 1 | 1 |
| **Rest corepressor 1.** | **CoREST protein** | 1 | 1 | 1 |
| IPI00285217 | | 1 | 1 | 1 |
| Human: 1 | EES*EDELEEEA-NGNNPIDIEVDQDK 25 |
| Mouse: 1 | GT-GGQQLQLQDSLSS*S*DDATATQ-NKPSATK 34 |
| Human: 1 | GT-GGQQLQLQDSLSS*S*DDGAAQSTNKPSATK 37 |
| Mouse: 1 | INS*S*GKS*PGS*EEFQAKS-HK 27 |
| Human: 1 | VNS-N-GKS*PGS*EEFQAKS-HK 26 |
| Mouse: 1 | EGINPGY-GY-DY-ADS*DEQDQKAT-LE 26 |
| Human: 1 | EGMNP*K*T*DY-ADS*DEQDQKAT-LE 28 |
| Mouse: 1 | GGFGGEAFAAFAACGGPS*D*EMVEFDRS*PGK 37 |
| Human: 1 | GGFGGEAFAAFAACGGPS*D*EMVEFDRS*PGK 36 |
| Mouse: 1 | IQQQACTYPKS*EDEED*VDEED 24 |
| Mouse: 1 | VQQPTAGGHA-QEDEG*QDEED 30 |
| Mouse: 1 | DGSEY-DPS-DF--EAEQMPQVHIT-PK 25 |
| Mouse: 1 | DGDSL*DPS-DF--EAEQMPQVHIT-PK 29 |
| Mouse: 1 | VEDTTGSSGAA-KDQDDDDILLG*DDDEEES-EKAK 40 |
| Human: 1 | VEDTTGSSGAA-KDQDDDDILLG*DDDEEES-EKAK 42 |
| Mouse: 1 | HLGSGFS-PFQAQQDEEFK 21 |
| Mouse: 1 | HLGSGFS-PFQAQQDEEFK 26 |
| Mouse: 1 | IFDEINGCO-DPQGGS*FFPEBEGRTLIYDE 34 |
| Mouse: 1 | IFDEINGCO-DPQGGS*FFPEBEGRTLIYDE 34 |
| Mouse: 1 | T*S*FCQSFTVPF--SAPSLPFIRSD 26 |
| Mouse: 1 | T*S*FCQSFTVPF--SAPSLPFIRSD 25 |
| Mouse: 1 | EESKEFLS*DEEED-NODEPILFEDER 28 |
| Mouse: 1 | EESKEFLS*DEEED-NODEPILFEDER 29 |
| Mouse: 1 | HGTL-DIDWMSNS--APS*H*PEK 22 |
| Human: 1 | HGTL-DIDWMSNS--APS*H*PEK 24 |
| Mouse: 1 | DDEWKS*S*FPKPTPVIFISVIGAR 26 |
| Mouse: 1 | DDEWKS*S*FPKPTPVIFISVIGAR 25 |
| Mouse: 1 | S*RS*RS*FT*PPS*A-AGLGSPS*APPFD 32 |
| Mouse: 1 | S*RS*RS*FT*PPS*A-AGLGSPS*APPFD 31 |
| Mouse: 1 | LEEEKS*A-T*EPPELAQSLPEP 28 |
| Mouse: 1 | LEEEKS*A-T*EPPELAQSLPEP 29 |
| Mouse: 1 | LPELWS*WRS*DEEDEDS-DFGEER 24 |
| Mouse: 1 | LPELWS*WRS*DEEDEDS-DFGEER 25 |
| Mouse: 1 | AT*DS*HS*DS*EDEAEACAPQLQDLSVAR 31 |
| Human: 1 | AM-DN*HS*DS*EDEAEACAPQLQDLSVAR 28 |
| Mouse: 1 | DB*HSPQGAT+ET-HLYLGPDQEESTVLQDK 29 |
| Mouse: 1 | DB*HSPQGAT+ET-HLYLGPDQEESTVLQDK 30 |
| Mouse: 1 | A-S*DDGGS*PDEHA-DIYDFDQLVQNT*LR 31 |
| Mouse: 1 | T*S*DDGGS*PDEHA-DIYDFDQLVQNT*LR 34 |
| Mouse: 1 | GCCGG*CQDMGLOGL 16 |
| Mouse: 1 | GCCGG*CQDMGLOGL 14 |
| Mouse: 1 | AAVGELS*G*S*ILZ*G*EDPEER 23 |
| Human: 1 | AAVGELS*G*S*ILZ*G*EDPEER 22 |
| Mouse: 1 | SPPDQAVP*TPPS*T-PVK 20 |
| Mouse: 1 | SPPDQAVP*TPPS*T-PVK 20 |
| Mouse: 1 | LADDEGDS*G*ESE*VQYQR 20 |
to proteins whose IPI accession numbers are active and ignored those with obsolete accession numbers. In total, 8367 phosphosites were mapped to 2532 proteins. The protein sequences were then aligned with their mouse orthologs by searching the IPI mouse database (ipi.MOUSE.v3.30) using BlastPro (13). In total, 2511 protein sequences were successfully aligned with their mouse orthologs, whereas 21 sequences did not have a ortholog, or the homology was too low (<50%) (supplemental protein sequence alignments).

Sequence similarities are not always homogeneous across the whole length of many aligned protein sequences, but rather they can show discontinuous regions of highly conserved or divergent amino acid sequences. This makes it difficult to decide whether a specific Ser/Thr/Tyr residue present in a divergent region of a human protein is present in its mouse ortholog. To solve this problem, we scanned the aligned sequences using a 10-character frame as described recently (18) where only the Ser/Thr/Tyr residues that were in frames with sequence similarity higher than 60% were counted and analyzed (18). The numbers of mutations from human to mouse for phosphorylated or unphosphorylated Ser/Thr/Tyr residues were counted as described under “Experimental Procedures.” In total, 6141 phosphorylated Ser/Thr/Tyr residues in the human protein sequences were

### Table III—continued

| Protein ID | Description                                      | Human Phospho Peptide | Mouse Phospho Peptide |
|------------|--------------------------------------------------|-----------------------|-----------------------|
| IPI00330331| Predicted: similar to Brd1 protein.              | LADREDGAS...KSGMIKISI | VHGERP...KSGMIKISI    |
| IPI00340294| elf5b protein fragment.                         |                      |                       |
| IPI00330272| Arginine glutamic acid dipeptide re repeats.     |                      |                       |
| IPI00357145| Similar to interferon regulatory factor 2 binding protein 2 |                      |                       |
| IPI0122559 | Kinectin.                                        |                      |                       |
| IPI00153833| Cutaneous T-cell lymphoma-associated antigen 5 homolog. |                      |                       |
| IPI00340815| Similar to rbm15 protein.                        |                      |                       |
| IPI00323820| Minichromosome maintenance deficient 2 mitotin.  |                      |                       |
| IPI00407613| Mitochondrial tumor suppressor gene 1, isoform 3 |                      |                       |
| IPI00123537| Nuclear transcription factor, x-box binding 1.   |                      |                       |
| IPI00337844| 341 kda protein.                                |                      |                       |
| IPI00118438| Splice isoform 2 of serine/arginine repetitive matrix protein 1 |                      |                       |
| IPI00169477| Splice isoform 2 of bcl-2-associated transcription factor 1 |                      |                       |

Phosphoamino acids substituted with an acid amino acid residue are inversely displayed.

Phosphoamino acids substituted with a proline residue are lightly shaded.
counted of which 370 sites have mutations in mouse (6.4%). A total of 282,180 unphosphorylated Ser/Thr/Tyr residues were counted in human of which 30,037 have mutations in mouse (10.6%) (supplemental Table S6). These data suggest that phosphorylated amino acids are more conserved than those not known to be phosphorylated. Similar findings were obtained when we aligned the human protein sequences containing only class I phosphosites identified by Olsen et al. (6) with the mouse protein sequences (data not shown).

Interestingly similar to the distribution of the mutations shown in Fig. 5C, the most frequent mutations from human to mouse in Ser/Thr/Tyr residues are Thr → Ala, Ser → Asn, Ser → Gly, Ser → Ala, Ser → Pro, and the exchange of Ser and Thr (Fig. 6A). Two major reasons can explain this distribution: First, the total number of analyzed Ser, Thr, and Tyr amino acids are different where the ratio of the total counted Ser, Thr, and Tyr is 3.6:2.1:1 (154,952:90,485:42,878). Therefore, the total numbers of mutations observed for each amino acid (#Ser, #Thr, and #Tyr) are exhibited as #Ser > #Thr > #Tyr both in the phosphorylated and unphosphorylated forms. Second, these mutations require only a single nucleotide change in the genetic code, which occurs relatively frequently. To investigate the potential influence of selection pressure of amino acid phosphorylation on each mutation, we calculated the mutation rate for Ser/Thr/Tyr residues as described under “Experimental Procedures” (Fig. 6B). Several interesting observations were made from the results. First, although the highest mutation rate is for Thr → Ala, there is no significant difference in mutation rate between the phosphorylated and unphosphorylated forms indicating that this mutation is the most frequent random mutation that occurs independently of phosphorylation. Second, for the majority of the amino acids with high mutation rates including Ser → Pro, Ser → Asn, Ser → Gly, Ser → Ala, and Tyr → His, the phosphorylated form shows a reduced mutation rate indicating that phosphoamino acids are more conserved evolutionally. Finally a few mutations have higher mutation rates in the phosphorylated form including Tyr → Val, Tyr → Pro, and Tyr → Asp. The significance of the difference in the mutation rates in phosphorylated/unphosphorylated Ser/Thr/Tyr residues was further evaluated by t test for paired samples. The p value was 0.0026, suggesting that phosphorylation significantly affects the mutation rate of Ser/Thr/Tyr residues. Again the same conclusions can be drawn if we calculate the mutation rate from mouse to human (data not shown).

**DISCUSSION**

Although recent advances in mass spectrometry allow the identification of thousands of sites of protein phosphorylation...
In Silico Comparison of Phosphorylation Sites

(5–7, 11) there are currently no automated computational tools available for biologists to compare intra- or interlaboratory phosphoprotein datasets. The PhosphoBlast program described here provides a versatile mining tool that can be used to uncover similarities and differences in phosphorylation signatures within and across species from large complex phosphoprotein datasets in a highly efficient and user-friendly manner.

PhosphoBlast has several important features and possible utilities. First, it facilitates the identification of evolutionally conserved phosphorylation signatures. These signatures are likely to play important roles in various cell functions and homeostasis. Second, identification of common phosphorylation signatures among independent studies and their conservation across species aid in the validation of specific phosphorylation sites. These putative phosphosites would be important candidates to validate using molecular biology and functional biological assays. Finally PhosphoBlast can be used to understand mutational events that have targeted functionally important phosphoamino acid residues. Identification of such mutations could provide valuable insight into the evolution of signaling processes associated with phosphoprotein function as well as provide important biomarkers of disease including cancer. In this regard, a mutation that introduces an acidic amino acid into a key phosphoamino acid residue could serve as a phosphomimetic altering its biological activity. In some cases, phosphomimetics engineered in the laboratory can recapitulate constitutive phosphorylation states that promote activation/deactivation of enzymatic function, protein conformational changes, and protein interactions (20, 21). However, in some other cases this type of substitution does not functionally substitute for the actual phosphorylation event (22–26). Conversely introduction of a basic or neutral amino acid into a functional phosphosite may have a silencing effect on protein function. Following functional validation, these sites would provide valuable biomarkers and potential targets for therapeutic intervention.

Although we demonstrate here that PhosphoBlast can be used to identify mutations between divergent datasets, it can also be used to examine a single dataset for phosphosite mutations that occur across multiple species. Indeed in our recent work we used a pilot form of PhosphoBlast (BlastPro) (11) to evaluate the amino acid conservation of identified phosphorylation sites from monkey COS-7 cell proteins (using a human proteome database) compared with that of mouse, rat, and zebrafish. Interestingly we found that 44, 80, and 88% of the identified phosphosites were conserved in zebrafish, rat, and mouse, respectively.

Although PhosphoBlast is highly efficient at identifying phosphosites shared among phylogenically related species, the comparison of phylogenetically distant species such as yeast and human requires user interpolation to confidently assign a matched phosphopeptide pair to a particular protein. The parser of the PhosphoBlast can retrieve positive hits with similarity scores lower than 100%. This feature enables the retrieval of identical but differentially phosphorylated peptides as well as peptides from cross-species, which may not be 100% identical in amino acid sequence. However, in some cases false-positive hits will be generated. For example, the peptide T*GGGSGGGGGGSGGSDVK in the query dataset (5) was matched to the peptide GDGGGASGGGGGGGSGGSGGGGGGGSS*R in the database (6) using a low stringent E-value cutoff (e~04). This is because the two peptides contain a 14-amino acid residue repeat even though they are from distinct proteins. This type of false positive is rare but can be removed using a more stringent E-value cutoff or by manual interpolation. In some cases, protein isoforms also generate false-positives hits. For instance, we observed that the peptide SAS*EPSLHR from A-RAF, which has only one mismatched amino acid residue. Although this false-positive hit can be manually removed, it may also be informative about phosphorylation events shared among related proteins or families of proteins.

The PhosphoMerger program was developed in house to create a non-redundant phosphosite dataset prior to comparative analyses. The program increases efficiency and allows for a more accurate determination of the number of phosphorylation sites identified in a particular dataset. Initially we attempted an alternative approach to remove redundancy by performing in silico full tryptic digestion of all peptides containing miscleaved tryptic sites (Split algorithm). Using this approach it is possible to integrate the phosphosites of all peptides with identical amino acid sequences from a given protein into a single peptide. Using this method, we removed the redundancy and counted the phosphosites from the above phosphopeptide datasets. As expected, this approach counted the same number of phosphopeptides as counted by the PhosphoMerger program. However, this approach can generate many small phosphopeptides due to in silico digestion. This is a particular problem as these small peptides can be randomly matched with other small peptides or fragments of larger peptides, making it difficult to filter out the false-positive hits. Therefore, we chose the Merge algorithm instead of the Split algorithm to develop the PhosphoMerger module.

The PhosphoBlast program compares phosphosites by aligning sequences at the peptide level instead of the whole protein. This provides a straightforward and efficient means to track combinations of phosphorylation events that occur between independent studies. For instance, a tryptic peptide may be singly and multiply phosphorylated at the same time in a given cell, and a protein may contain many such peptides. To map these sites to a given protein sequence and to track all of the specific phosphosite combinations on this protein would require repeated mapping of the phosphorylation sites present on each identified peptide to each copy of the same
protein sequence. This would be quite cumbersome when comparing thousands of whole protein sequences containing multiple phosphosite modifications because it would require significantly more computational time and memory than comparing thousands of phosphopeptides. Moreover, following PhosphoBlast analysis, it is much easier to manage and display the different combinations of phosphosite modifications identified on each peptide in two different datasets. This is because the matched phosphopeptides can be saved in a formatted table and the differential phosphorylation states can be tracked by rapidly scanning the table (supplemental Tables S2, S3, and S5). This is not possible if the phosphosites are mapped to aligned protein sequences because of the large protein sequences. Also, the peptide alignment approach is highly accurate as shown by the fact that the number of total matched peptides between the datasets generated by Olsen et al. (6) and Beausoleil et al. (5) that was calculated by PhosphoBlast (Fig. 4B) is the same as that obtained using the protein sequence alignment approach.

Although PhosphoBlast is a useful tool to compare phosphosignatures among large datasets, it is important to note that the presence or absence of a particular phosphoprotein signature can result not only from differential phosphorylation but also from differences in protein abundance as well as failed detection due to the inherent error in the experimental methods and instrumentation. Therefore, the phosphoprotein datasets analyzed here are neither comprehensive nor quantitative in nature. Identifying a true phosphoprotein signature with confidence will require a detailed understanding of the stoichiometry of protein phosphorylation as well as its functional validation. Nevertheless, sensitive mass spectrometry methods for quantitative phosphoproteomics are rapidly evolving along with powerful informatics programs (6, 7, 11). Future work in this area will undoubtedly provide a wealth of information on the phosphoproteome of normal and diseased cells and provide a valuable resource for comparative analyses using PhosphoBlast and other informatics programs.

In summary, PhosphoBlast is a highly efficient and versatile data mining tool to compare phosphorylation sites between large phosphoproteomics datasets. Our data indicate that the software can be used to reveal putative phosphorylation sites shared across diverse experimental platforms and species and to identify specific phosphosite mutations. This system also displayed an excellent run time matching thousands of peptide pairs singly or multiply phosphorylated with a high level of accuracy with minimal user input. With the generation of more and more biologically relevant phosphoprotein datasets, PhosphoBlast will provide a valuable tool to uncover the relationship among these datasets and help identify critical signaling events that contribute to different biological functions. Future work using quantitative phosphoproteomics combined with PhosphoBlast and other informatics tools could provide comprehensive assignment of phosphoprotein signatures or networks of signatures to specific biological functions. Such an approach could provide a powerful means to develop reliable prognostic indicators of health and disease.

Acknowledgments—We thank Wei Wang for the manual validation of the PhosphoBlast results and Konstantin Stoletov, Oliver Pertz, and Laurie Gay for critical comments. We also thank Dr. Cizhong Jiang for constructive advice.

* This work was supported by Susan G. Komen Foundation Grant PDF0503999 (to Y. W.), National Institutes of Health Grants GM068487 and CA097022 (to R. L. K.), and by Cell Migration Consortium Grant GM064346. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ To whom correspondence should be addressed: Dept. of Pathology and Moores Cancer Center, University of California, San Diego, Basic Science Bldg. 1040, 9500 Gilman Dr., MC 0612, La Jolla, CA 92039. Tel.: 858-822-5610; Fax: 858-822-4566; E-mail: rklemke@ucsd.edu.

REFERENCES

1. Cohen, P. (2000) The regulation of protein function by multistate phosphorylation—a 25 year update. Trends Biochem. Sci. 25, 596–601
2. Mann, M., Ong, S. E., Gronborg, M., Steen, H., Jensen, O. N., and Pandey, A. (2002) Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. Trends Biotechnol. 20, 261–268
3. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002) The protein kinase complement of the human genome. Science 298, 1912–1934
4. Blume-Jensen, P., and Hunter, T. (2001) Oncogenic kinase signalling. Nature 411, 355–365
5. Beausoleil, S. A., Jedrychowski, M., Schwartz, D., Elias, J. E., Villen, J., Li, J., Cohn, M. A., Cantley, L. G., and Gygi, S. P. (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. Proc. Natl. Acad. Sci. U. S. A. 101, 12130–12135
6. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 127, 635–648
7. Villen, J., Beausoleil, S. A., Gerber, S. A., and Gygi, S. P. (2007) Large-scale phosphorylation analysis of mouse liver. Proc. Natl. Acad. Sci. U. S. A. 104, 1488–1493
8. Kratchmarova, I., Blagoev, B., Haack-Sorensen, M., Kassem, M., and Mann, M. (2005) Mechanism of divergent growth factor effects in mesenchymal stem cell differentiation. Science 308, 1472–1477
9. Thelenmann, A., Petti, F., Griffin, G., Iwata, K., Hunt, T., Settinari, T., Fenyo, D., Gibson, N., and Halevy, J. D. (2005) Phosphotyrosine signaling networks in epidermal growth factor receptor overexpressing squamous carcinoma cells. Mol. Cell. Proteomics 4, 356–376
10. Wolf-Yadlin, A., Hautaniemi, S., Laufnburger, D. A., and White, F. M. (2007) Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. Proc. Natl. Acad. Sci. U. S. A. 104, 5860–5865
11. Wang, Y., Ding, S. J., Wang, W., Jacobs, J. M., Qian, W. J., Moore, R. J., Yang, F., Camp, D. G., II, Smith, R. D., and Klemke, R. L. (2007) Profiling signaling polarity in chemotactic cells. Proc. Natl. Acad. Sci. U. S. A.
12. Gruhler, A., Olsen, J. V., Mohammed, S., Mortensen, P., Faergeman, N. J., Mann, M., and Jensen, O. N. (2005) Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. Mol. Cell. Proteomics 4, 310–327
13. Wang, Y., Hanley, R., and Klemke, R. L. (2006) Computational methods for comparison of large genomic and proteomic datasets reveal protein markers of metastatic cancer. J. Proteome Res. 5, 907–915
14. Attasch, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new
In Silico Comparison of Phosphorylation Sites

162 Molecular & Cellular Proteomics 7.1

generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
15. Diella, F., Cameron, S., Gemund, C., Linding, R., Via, A., Kuster, B., Sicheritz-Ponten, T., Blom, N., and Gibson, T. J. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* 5, 79
16. Ficarro, S. B., McCleland, M. L., Stukenberg, P. T., Burke, D. J., Ross, M. M., Shabanowitz, J., Hunt, D. F., and White, F. M. (2002) Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *BMC Bioinformatics* 5, 79
17. Rush, J., Moritz, A., Lee, K. A., Guo, A., Goss, V. L., Spek, E. J., Zhang, H., Zha, X. M., Polakiewicz, R. D., and Comb, M. J. (2005) Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nature* 433, 633–638
18. Nousiainen, M., Sillje, H. H., Sauer, G., Nigg, E. A., and Korner, R. (2006) Phosphorylation causes a conformational change in the R domain of the cystic fibrosis transmembrane conductance regulator. *Biochemistry* 33, 4072–4079
19. Gomez-del Arco, P., Maki, K., and Georgopoulos, K. (2004) Phosphorylation controls Ikaros’s ability to negatively regulate the G1-S transition. *Mol. Cell. Biol.* 24, 2797–2807
20. Huang, W., and Erikson, R. L. (1994) Constitutive activation of Mek1 by mutation of serine phosphorylation sites. *Proc. Natl. Acad. Sci. U. S. A.* 91, 8960–8963
21. Yang, Y., Craig, T. J., Chen, X., Ciuflo, L. F., Takahashi, M., Morgan, A., and Gillis, K. D. (2007) Phosphomimetic mutation of Ser-187 of SNAP-25 increases both syntaxin binding and highly Ca^{2+}-sensitive exocytosis. *J. Gen. Physiol.* 129, 233–244
22. Dulhanty, A. M., and Riordan, J. R. (1994) Phosphorylation by cAMP-dependent protein kinase causes a conformational change in the R domain of the cystic fibrosis transmembrane conductance regulator. *Biochemistry* 33, 4072–4079
23. Germann, U. A., Chambers, T. C., Ambudkar, S. V., Licht, T., Cardarelli, C. O., Pastan, I., and Gottesman, M. M. (1996) Characterization of phosphorylation-defective mutants of human P-glycoprotein expressed in mammalian cells. *J. Biol. Chem.* 271, 1708–1716
24. Sun, K., Montana, V., Chellappa, K., Brelivet, Y., Moras, D., Maeda, Y., Parpura, V., Paschal, B. M., and Sladek, F. M. (2007) Phosphorylation of a conserved serine in the deoxyribonucleic acid binding domain of nuclear receptors alters intracellular localization. *Mol. Endocrinol.* 21, 1297–1311