Phosphoproteomic Analysis of the Mouse Brain Cytosol Reveals a Predominance of Protein Phosphorylation in Regions of Intrinsic Sequence Disorder*‡

Mark O. Collins†‡¶§, Lu Yu†¶§, Iain Campuzano‖, Seth G. N. Grant**‡‡, and Jyoti S. Choudhary†¶§§

We analyzed the mouse forebrain cytosolic phosphoproteome using sequential (protein and peptide) IMAC purifications, enzymatic dephosphorylation, and targeted tandem mass spectrometry analysis strategies. In total, using complementary phosphoenrichment and LC-MS/MS strategies, 512 phosphorylation sites on 540 non-redundant phosphopeptides from 162 cytosolic phosphoproteins were characterized. Analysis of protein domains and amino acid sequence composition of this data set of cytosolic phosphoproteins revealed that it is significantly enriched in intrinsic sequence disorder, and this enrichment is associated with both cellular location and phosphorylation status. The majority of phosphorylation sites found by MS were located outside of structural protein domains (97%) but were mostly located in regions of intrinsic sequence disorder (86%). 368 phosphorylation sites were located in long regions of disorder (over 40 amino acids long), and 94% of proteins contained at least one such long region of disorder. In addition, we found that 58 phosphorylation sites in this data set occur in 14-3-3 binding consensus motifs, linear motifs that are associated with unstructured regions in proteins. These results demonstrate that in this data set protein phosphorylation is significantly depleted in protein domains and significantly enriched in disordered protein sequences and that enrichment of intrinsic sequence disorder may be a common feature of phosphoproteomes. These encompass conventional fragmentation techniques such as CID (in combination with precursor ion scanning or neutral loss-triggered fragmentation) as well as alternative fragmentation strategies such as electron capture dissociation (11) and electron transfer dissociation (12, 13).

Protein phosphorylation in the brain has been intensively studied from the early 1970s, and since then it has become apparent that nerve cells (like other cell types) are highly regulated by this post-translational modification. The application of phosphoproteomic strategies has identified hundreds of phosphorylation sites in synaptic proteins allowing insights into the complex signaling network that exists at the synaptic membrane (3, 4). However, to describe brain phosphoproteomes comprehensively, it is necessary to analyze other cellular subfractions to reduce protein complexity and sufficiently enrich proteins with lower expression levels or with lower stoichiometry of phosphorylation. In addition, analysis and comparison of phosphorylation in subcellular proteomes may provide functional insights that are not apparent in phosphorylation data sets obtained from whole cell lysates.

The wealth of phosphorylation site information generated in the last few years by mass spectrometry is changing the way we view this post-translational modification. Classically it has been thought that this modification is reserved for regulation of classical signaling cascades, direct modulation of the activity of enzymes and receptors/channels, and providing bind-
ing sites for phosphorylation-dependent interactions. However, the widespread nature of this post-translational modification that is being revealed by proteomics must prompt investigation of other global functions of protein phosphorylation. The most basic effect of phosphorylation is to change the physicochemical characteristics of the polypeptide to which it is added. Phosphorylation can affect protein conformation (14), occurs on accessible regions in the three-dimensional structure (15), and is thought to occur in flexible regions in a protein structure (16); thus investigation of the structural topology of phosphorylation may reveal novel functional aspects of this modification. It has been suggested that intrinsic sequence disorder, which is specified by primary sequence composition and manifests itself as flexible or unstructured regions in proteins, is associated with protein phosphorylation (17). However, the relationship between phosphorylation and disorder in the context of large in vivo phosphoproteomics data sets obtained by MS has thus far not been investigated.

We analyzed protein phosphorylation in the mouse forebrain cytosol by using a sequential immobilized gallium affinity (IMAC) strategy in which phosphoproteins are specifically purified from the cytosol fraction and tryptically digested, and the resultant peptide mixture is applied to a second IMAC column to specifically enrich for phosphopeptides (3, 18). A number of complementary MS-based strategies allowed the unambiguous characterization of over 500 phosphorylation sites from a collection of cytosolic proteins. A striking bias in the location of these phosphorylation sites outside of domains in proteins prompted investigation of structural features of phosphorylated sequences. Intrinsic sequence disorder is a dominant feature of phosphorylated sequences in the cytosolic phosphoproteome and provides a functional explanation for the clustering of phosphorylation sites to defined regions in proteins. We suggest that protein phosphorylation is utilized in intrinsically disordered regions to regulate the increased number of protein interactions (especially phosphorylation-dependent protein interactions) mediated by intrinsically disordered proteins (19). In addition, the use of clustered hyperphosphorylation as a means of creating local negative charge to regulate, for example protein-RNA/DNA backbone interactions, may explain the high stoichiometry of phosphorylation observed for some proteins. Finally we characterized a number of highly disordered phosphoproteins with related roles in the pathogenesis of neurodegenerative disease where phosphorylation may regulate protein folding and function.

**EXPERIMENTAL PROCEDURES**

**Preparation of Cytosolic Extract**

Mouse forebrains were rapidly dissected, frozen immediately in liquid nitrogen, and stored at −80 °C. Forebrains were homogenized with 25 stokes in a Dounce homogenizer in homogenization buffer (50 mM Tris, pH 8, 50 mM sodium fluoride, 20 μM zinc chloride, 1 mM sodium ortho-vanadate, 0.5 mM PMSF, 2 μg/ml aprotinin, 2 μg/ml leupeptin) at a ratio of 0.379 g of tissue/7 ml of buffer. The extract was subjected to centrifugation at 50,000 × g for 30 min, and the supernatant was concentrated in a Vivaspin 6 polyethersulfone membrane 10-kDa molecular mass cutoff spin column (Vivascience) and stored at −80 °C.

**Protein IMAC of Forebrain Cytosolic Fraction**

Fast flow chelating Sepharose with iminodiacetic acid (Amersham Biosciences) chelating groups was charged with GaCl₃. Cytosolic protein (50 mg) was brought to 6 M urea and incubated with 2 ml of the metal-charged resin with mixing for 1 h at room temperature. The unbound protein was washed with buffer A (6 M urea, 50 mM Tris acetate) to base line, and the phosphoproteins were specifically eluted with buffer B (6 M urea, 50 mM Tris acetate, 100 mM EDTA, 100 mM EGTA). Two of these purifications were carried out, and eluted phosphoproteins were pooled, concentrated, and washed with buffer B in a Vivaspin 6 polyethersulfone membrane 10-kDa molecular mass cutoff spin column, resulting in a final yield of 2.5 mg of purified phosphoproteins.

**Peptide IMAC of Protein IMAC**

2.5 mg of the protein IMAC-purified sample was digested with sequencing grade, modified trypsin (Promega) in a ratio of 1:20 in digestion buffer (pH 8) (1 M urea, 25 mM NH₄HCO₃) at 37 °C for 4 h. The resultant digest was desalted and dried, and methyl esterification was performed with 2 M methanolic HCl (10). Self Pack POROS® 20 MC medium (Applied Biosystems) for phosphopeptide purification was charged with GaCl₃ as described above for the iminodiacetic acid resin. Peptide digests were reconstituted in loading buffer (equal volumes of acetonitrile, methanol, and water, pH 2.5–3). 1 ml of this peptide mixture was incubated with 200 μl of POROS-gallium slurry for 1 h at room temperature. The resin was then loaded into a spin column and washed with 10 volumes of loading buffer. Phosphopeptides were eluted with 2 × 100 μl of 200 mM Na₂HPO₄. In addition, another double IMAC enriched sample was generated using 1 mg of phosphoprotein from which phosphopeptides were enriched using peptide IMAC as above except without a methyl esterification step.

**Enzymatic Dephosphorylation of Double IMAC Enriched Phosphopeptides**

One-eighth of the phosphopeptide sample eluted from the peptide IMAC (double IMAC) was desalted on a 300-μM × 5-mm PepMap C₁₈ column. After the peptides were dried down, they were resuspended in 50 μl of NE buffer (New England Biolabs) and were incubated with 2 μl of calf intestinal alkaline phosphatase (New England Biolabs) for 3 h at 37 °C.

**Profiling of the Protein IMAC Enrichment**

1.5 μg of a solution digest of protein IMAC enriched phosphoproteins was analyzed on an LTQ-FT (Thermo Electron), a hybrid linear ion trap, and a 7-tesla Fourier transform ion cyclotron resonance mass spectrometer coupled with an Ultimate 3000 Nano/Capillary LC System ( Dionex). Samples were first loaded and desalted on a trap column (0.3-mm inner diameter × 5 mm) at 25 μl/min with 0.1% formic acid for 5 min and then separated on an analytical column (75-μM inner diameter × 15 cm) (both PepMap C₁₈, LC Packings) over a 60-min linear gradient of 4–92% CH₃CN, 0.1% formic acid. The flow rate through the column was 300 nl/min. The LTQ-FT mass spectrometer was operated in standard data-dependent acquisition mode controlled by Xcalibur 1.4 software. The survey scans (m/z 400–2000) were acquired on the FT-ICR instrument at a resolution of
100,000 at m/z 400, and one microscan was acquired per spectrum. The three most abundant multiply charged ions with a minimal intensity at 500 counts were subjected to MS/MS in the linear ion trap at an isolation width of 3 Thomson. Precursor activation was performed with an activation time of 30 ms, and the activation Q was set at 0.25. The normalized collision energy was set at 35%. The dynamic exclusion width was set at ±5 ppm with one repeat and a duration of 30 s. To achieve high mass accuracy, the automatic gain control target value was regulated at 5 × 10⁶ for FT and 5000 for the ion trap with maximum injection time at 1000 and 200 ms for FT and ion trap, respectively. The instrument was externally calibrated using the standard calibration mixture of caffeine, MRFA, and Ulramark 1600. In the first experiment, the top three most abundant ions in a given chromatographic time window were selected for MS/MS fragmentation, and in the second, the fifth to seventh most abundant ions were selected for MS/MS fragmentation.

Multiple Data-directed Analysis of the Double IMAC Enrichment

A nanoflow HPLC system, Ultimate™ (LC Packings), was coupled to a Q-Tof 1 (Micromass) mass spectrometer. Phosphopeptides (1/6e the elution) from the peptide IMAC purification were loaded in 0.1% aqueous formic acid and desalted on PepMap C₁₈ trapping cartridge (180-μm inner diameter x 30 mm; LC Packings or a BetaMax Neutral (Thermo Hypersil-Keystone)). Peptides on the trap were back-flushed to and separated on the analytical column (PepMap C₁₈, 75-μm inner diameter x 15 cm; LC Packings). The Q-Tof 1 was operated in automated data-dependent acquisition mode. Each cycle had a 1-s MS survey (m/z 400–1500), and up to four of the highest intensity multiply charged ions (+2, +3, and +4) were selected for MS/MS (m/z 50–2000) every 5 s (4 × 1.15 s). The collision energy in MS/MS was set at 35%. The normalized collision energy was set at 5 ppm with one repeat and a duration of 30 s. The three most abundant multiply charged ions (+2, +3, and +4) were selected for MS/MS (m/z range 50–2990 with a solution of sodium/cesium iodide. All data were acquired with lock spray using m/z 785.8426 from [Glu]₁ fibrinopeptide as reference. Data-directed analysis (DDA) was performed where the multiply charged precursors were selected and fragmented automatically with the collision energy in MS/MS varied according to the m/z and the charge state of the precursor ion. The instrument was calibrated over the m/z range 50–2990 with a solution of sodium/cesium iodide. All data were acquired with lock spray using m/z 785.8426 from [Glu]₁ fibrinopeptide as reference. Data-directed analysis (DDA) was performed where the multiply charged precursors were selected and fragmented automatically with the collision energy in MS/MS varied according to the m/z and the charge state of the precursor ion, and the top five precursor ions were selected for MS/MS analysis. The MS/MS switch list was used as an exclusion list for the subsequent DDA experiment. This was carried out four times (2-h acquisition for each). This set of experiments is referred to as iterative DDA with exclusion list (iDDA). In addition, Protein Expression (MS²) (Waters) analysis was performed on the sample with 1-h acquisition time. Alternate low and elevated collision energy scans were performed in alternating MS scans. The data were processed with ProteinLynx Global SERVER using the Protein Expression System software. An exact mass retention time (EMRT) list was generated to use as an inclusion list for a subsequent 1-h acquisition DDA experiment (iDDA) in which the top seven precursor ions were selected for MS/MS analysis.

Database Searching

Q-Tof 1-generated raw data were processed using MassLynx 3.4 (Waters), Q-Tof Premier data were processed using ProteinLynx Global SERVER 2.2.5 with expression analysis (Waters), and LTQ-FT data were processed using BioWorks 3.2 (Thermo Electron) to give peak list files. Processed data were submitted to a local MASCOT V2.0/V2.1 (Matrix Science) server for iterative searching on a non-identical, non-redundant, combined human and mouse International Protein Index database (European Bioinformatics Institute) (113,646 sequences, 53,539,666 residues, September 2004) or a non-identical mouse database generated in house (75,777 sequences, 37,283,622 residues, Ensembl build 43/UniProt/varsplcier/trEMBL release 9 (downloaded January 2007)/Refseq (downloaded January 2007, release 21). Variable modifications used included acetylation (protein N terminus), oxidation (Met), and phosphorylation (STY) and methylation (C terminus and DE) when peptide methyl esterification was performed. A maximum of three missed cleavages by trypsin was allowed for database searching, and the following precursor and fragment ion tolerances were used: 20 ppm and 0.5 Da for the LTQ-FT and 0.4 Da and 0.4 Da for the Q-Tof, respectively.

Protein IMAC Profiling on LTQ-FT—False discovery rates determined by reverse database searches and empirical analyses of the distributions of mass deviation and MASCOT ion scores were used to establish score and mass accuracy filters (two classes of protein identifications were approved with the following minimum requirements: Class A, two or more peptides, one with a MASCOT ion score over the MASCOT identity threshold with a length of >8 residues and the other with a MASCOT ion score over the MASCOT homology threshold with a Δppm of <7; Class B, only one peptide with a MASCOT ion score over the MASCOT identity threshold with a length of >8 residues or two peptides, one with a MASCOT ion score over the MASCOT identity threshold with a length of >8 residues and the other with a Δppm of 5. Using these filters, protein identifications in the protein IMAC enrichment were approved, and random sequence database searching (the random version of the mouse database was generated using a Perl script downloaded from Matrix Science) produced an estimated false discovery rate (FDR) of 0.7%. This was reduced to 0% FDR by excluding proteins identified by one peptide in only one of the two protein IMAC profiling experiments. Phosphopeptides in the protein IMAC profiling experiment were manually approved using the MASCOT homology threshold as an initial filter.

Q-Tof Phosphopeptide Analysis—All phosphopeptides were reported manually inspected (no score cutoff), and assignment of phos-
Phosphorylation sites was verified manually (using neutral loss of phosphoric acid) with the aid of PEAK Studio V4.1 (Bioinformatics Solutions) software (supplemental Fig. 1, A-D). Peptide identifications in the enzymatically dephosphorylated phosphopeptide sample (from a double IMAC purification) were approved using the MASCOT homology threshold (MHT) as a cutoff and corresponded to a 0.3% FDR (as assessed by MASCOT decoy database analysis). Peptides specific to the dephosphorylated analysis (i.e., not found in any other analysis) were manually inspected. Peptides and phosphopeptides were assigned to the longest matching protein sequence in the database used for searching, and identification of isoforms was only possible when isoform-specific peptides were identified. The number of unique phosphorylation sites reported is the number of non-redundant phosphorylation sites that we identified in total in this study. The number of unambiguously assigned phosphorylation sites is the number of sites that we could assign to the precise location in a peptide sequence, and the difference between these numbers is the number of phosphorylation sites that we detected but could not localize precisely to a given Ser, Thr, or Tyr residue by manual inspection of the spectra.

Sequence-based Analysis

All phosphoproteins detected in this study were classified according to Swiss-Prot keywords. Scansite (20) was used for predicting the most likely kinases responsible for the phosphorylation at sites characterized in this study. In addition, for ambiguously defined phosphorylation sites Scansite was used to predict the most likely site of phosphorylation when a number of possibilities were present on a phosphopeptide. Scansite was also used to predict whether phosphorylation sites were localized in phospho-dependent interaction domains. Pfam-A domain information was extracted from the Pfam database (21). Composition Profiler was used to assess significantly enriched amino acid compositional differences between data sets (22). PONDR (Predictors of Natural Disordered Regions) VL-XT Predictor (access to PONDR® was provided by Molecular Kinetics) (23), which predicts order-disorder classification for every residue in a protein, was used to predict phosphoprotein sequence disorder. The significance of enrichment of phosphorylation sites in regions of intrinsic disorder and depletion of phosphorylation sites in protein domains was assessed using a two-tailed Fisher’s exact test. Three-dimensional protein structure data were visualized with the Deep View Swiss-PdbViewer 3.7 (24). Motif-x (25) was used to discover phosphorylation site motifs that were significantly enriched compared with the mouse proteome. WebGestalt was used to determine significantly enriched (using Fisher’s exact test) gene ontology categories in the cytosolic phosphoproteome compared with the mouse proteome.

RESULTS

Strategies for Analysis of the Cytosolic Brain Phosphoproteome

Protein IMAC Profiling

A number of different MS strategies were used to characterize phosphoproteins and phosphopeptides isolated from mouse forebrain cytosol extracts. We used immobilized gallium affinity chromatography in a fast protein LC format to purify phosphoproteins from cytosol extract (Fig. 1). A 6 M urea buffer system was used to prevent protein aggregation and proteolytic and other enzymatic activity. After extensive column washing, specific elution of phosphoproteins in a discrete chromatographic peak was achieved using a chelating buffer. The protein composition of the protein IMAC purification was interrogated in two LC-MS/MS experiments with a hybrid ion trap–FT mass spectrometer. In total, 192 proteins were identified with a 0.7% FDR as assessed by reverse database searching (supplemental Table 1 and supplemental Fig. 1A). Exclusion of proteins identified by one peptide in only one of these experiments resulted in a final list of 152 putative phosphoproteins with an FRD of 0%. In total, 58 phosphopeptide spectra were observed in the protein IMAC puri-
A sequential protein and peptide IMAC enrichment was analyzed by four different LC-MS/MS strategies. Three DDAs with varying LC gradient times were performed (experiments A, B, and C) termed “Multiple DDA.” Enzymatic dephosphorylation (dephos) of the double IMAC was performed and analyzed using the same LC-MS/MS conditions as those used in experiment C (experiment D). Two iterative DDA strategies were also used: in experiment E, an exclusion (excl) list was generated and used as a filter for subsequent DDA experiments, and in experiment F, and EMRT inclusion (incl) list was generated using Protein Expression (MSE) analysis and was used for subsequent DDA analyses. *, eight of 69 proteins were still phosphorylated, and 18 of 145 peptides were still phosphorylated; the remaining 127 previously phosphorylated peptides are described in supplemental Table 5. NR, non-redundant; seq, sequences.

**Analysis of Double IMAC Purifications**

**Multiple DDAs**—To extend the characterization of phosphopeptides, protein IMAC enriched cytosolic phosphoprotein was digested in solution, and phosphopeptides were specifically enriched in a second immobilized gallium affinity chromatography step. LC-MS/MS analyses (experiments A, B, and C) were performed on a Q-ToF to characterize phosphopeptides present in the double IMAC purification. Collectively these analyses resulted in the identification of 238 manually approved non-redundant phosphopeptides from 81 proteins and a total of 559 phosphorylation events (Table I, supplemental Tables 3 and 4, and supplemental Fig. 1C).

**Enzymatic Dephosphorylation**—We assessed the effect of phosphorylation on the rate of phosphopeptide identification (Table I and Fig. 2) by enzymatically dephosphorylating a portion of the peptide IMAC sample and analyzing the resultant (previously phosphorylated) peptides by LC-MS/MS (experiment D). Experiments C and D are directly analytically comparable as the same amount of starting material, LC and MS conditions, and database search parameters were used.

156 peptides were identified above MHT in experiment C with an FDR of 1.29% as assessed by MASCOT decoy analysis, whereas 270 peptides were identified above MHT in the dephosphorylated sample in experiment D with a 0.74% FDR. FDRs were also calculated using the same random database that was used for the protein IMAC profiling experiment, and this gave values of 1.9 and 0%, respectively, for experiments A and B.

**Extended coverage of phosphopeptides using enzymatic dephosphorylation**. A, overlap of non-redundant peptide/phosphopeptide sequences identified in experiments C and D. The majority of phosphopeptides identified in experiment C were also found in other analyses of protein IMAC and double IMAC purifications. 54 additional previously phosphorylated peptides were identified as a result of enzymatic dephosphorylation in experiment D. B, overlap of non-redundant peptide/phosphopeptide phosphoforms (monophosphorylated, diphosphorylated, etc.) identified in experiments C and D. 136 phosphopeptide forms were identified in both experiments C and D consisting of dephosphorylated peptides in experiment D matching and thus validating 80 distinct phosphopeptides from experiment C.

**Table I**

<table>
<thead>
<tr>
<th>Starting material</th>
<th>Forebrain cytosol extract</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purification</td>
<td>Sequential Protein and Peptide IMAC</td>
</tr>
<tr>
<td>Instrument</td>
<td>Q-ToF</td>
</tr>
<tr>
<td>Analysis</td>
<td>Multiple DDA (A, B, C)</td>
</tr>
<tr>
<td>Approval</td>
<td>Manual approval</td>
</tr>
<tr>
<td>Phosphoproteins</td>
<td>81</td>
</tr>
<tr>
<td>Phosphopeptides</td>
<td>238</td>
</tr>
<tr>
<td>Phosphopeptides (NR base seq)</td>
<td>199</td>
</tr>
<tr>
<td>Phosphorylation events</td>
<td>559</td>
</tr>
</tbody>
</table>

The Mouse Brain Cytosolic Phosphoproteome
C and D indicating that the use of the MHT as a cutoff produces an acceptable FDR for these data.

Because 98% of peptides with MASCOT ion scores above MHT identified in experiment C corresponded to phosphorylated peptides, it can be assumed that the vast majority of peptides identified in the dephosphorylation experiment must have been previously phosphorylated. The identification of both phosphorylated and dephosphorylated forms of a peptide increases the confidence of that peptide identification. This was exploited to increase the sensitivity of phosphopeptide identifications by lowering the score cutoff for phosphopeptides in experiment C from the MHT to a MASCOT ion score of 15 (empirically chosen based on the poorer quality of CID spectra below this cutoff) while keeping the MHT as the score cutoff for peptides in experiment D. 62 non-redundant (base sequences) phospho/dephosphopeptide pairs in experiments C (score > 15) and D (score > MHT) were identified, validating 80 distinct phosphopeptides in experiment C (supplemental Tables 5 and 6 and supplemental Fig. 1D). The combined FDR (as assessed by MASCOT decoy or standard random database searching) for this set of 62 phosphopeptides above a MASCOT ion score of 15 matching to dephosphorylated peptides (above MHT) was 0%. Comparison of spectra identifying the same peptides in both experiments (peptides above MHT in experiment D versus peptides above MASCOT ion score of 15 in experiment C) revealed that an additional 36 spectra below MHT in the phosphorylated sample were above MHT in the dephosphorylated sample (supplemental Tables 5 and 6). The average MASCOT ion score of these pairs of spectra increased from 38 in experiment C to 58 in experiment D.

In addition to validating phosphopeptides in experiment C, dephosphorylation of the sample analyzed in experiment D allowed the identification of a further 102 peptides (above MHT), corresponding to 88 non-redundant peptide sequences. This set of peptides can be considered as “previously phosphorylated” with 98% confidence as that was the purity of the sample as assessed in experiment C. 24 of 88 of these peptides were also present in other DDAs of the phosphorylated version of the double IMAC purification (supplemental Table 7), leaving 64 novel previously phosphorylated non-redundant peptide sequences. Manual inspection of these spectra resulted in confident identification of 54 of these peptides from 41 proteins (supplemental Table 8). 35 of these peptides map to 23 proteins characterized in the protein IMAC or double IMAC purifications indicating that not only was increased coverage of individual phosphoproteins achieved but that additional phosphoproteins (18 phosphoproteins) were amenable to detection by LC-MS/MS upon dephosphorylation.

Iterative DDAs—Two iterative DDA strategies were performed. The first approach used an exclusion list based on the first DDA (eDDA). Manual inspection of these data allowed approval of mass spectra corresponding to 143 non-redundant phosphopeptides (127 base sequences) (Table I and supplemental Tables 3 and 4). The second approach used Protein Expression (MSE) analysis followed by the generation of an EMRT list, which was used as an inclusion list for a subsequent DDA experiment (iDDA). Manual inspection of the data from this DDA with an MSE-derived inclusion list resulted in the allowed approval of mass spectra corresponding to 118 non-redundant phosphopeptides (108 base sequences) (Table I, supplemental Fig. 1C, and supplemental Table 3). Collectively these analyses allowed the identification of 185 manually approved phosphopeptides (164 base sequences) containing 267 phosphorylation events from 81 phosphoproteins (Table I). 57 phosphopeptide base sequences were found in both the iDDA and eDDA experiments. An additional 52 phosphopeptide base sequences were specifically detected in the eDDA, and 37 phosphopeptide base sequences were specifically detected in the iDDA. 34 phosphopeptides identified in these iterative DDAs were also found using the multiple DDA approach.

The Cytosolic Phosphoproteome

The combined phosphoproteomic strategies used here to study cytosolic phosphoproteins in the mouse forebrain resulted in the identification of 512 unique phosphorylation sites on 540 phosphopeptides and previously phosphorylated peptides from 162 phosphoproteins (Table I and supplemental Table 3). 92% of these phosphorylation sites (473 sites) were unambiguously assigned exact sites in peptide sequences with the remaining phosphorylation events mapped to a few possible serine, threonine, or tyrosine residues in peptide sequences. The distribution of phosphorylation sites was 87.1, 12.5, and 0.4% for phosphoserine, phosphothreonine, and phosphotyrosine, respectively (Fig. 3). Overall good coverage of singly to highly phosphorylated peptides was obtained with singly and doubly phosphorylated peptides accounting for 80% of the data set. Furthermore the remaining 20% encompassed highly phosphorylated peptides, many of which had four or more phosphorylation sites clustered in a short peptide sequence.

The data set of 162 phosphoproteins was classified according to annotated information in UniProt as well as from literature mining (Fig. 4). The largest class of phosphoproteins has no known function, and the fact that they represent 20% of the observable cytosolic phosphoproteome in our experiments highlights the lack of comprehensive functional annotation of the mouse proteome. The next most significant classes of proteins were involved in splicing and RNA binding and represented 12%, cytoskeletal proteins (12%), DNA-binding proteins/transcription (9%), kinases (9%), G-protein/ modulators (7%), and adaptor proteins (7%) (Fig. 4). Gene ontology analysis of these 162 proteins revealed that the molecular function “binding” was significantly enriched (p value <0.01) compared with the mouse proteome.
cytosolic phosphoproteins were associated with binding, 32 of which were significantly enriched in nucleotide binding ($p$ value <0.01) and 53 of which were significantly enriched in protein binding ($p$ value <0.01).

Protein Phosphorylation and Protein Structure

Depletion of Phosphorylation Sites in Protein Domains—To further characterize this data set of phosphoproteins identified in the double IMAC purification, a number of sequence-based analyses were performed specifically on the set of 141 phosphoproteins for which phosphopeptides were directly characterized and phosphorylation site information was available (as opposed to previously phosphorylated peptides). Initially the distribution of phosphorylation sites with respect to Pfam protein domains was investigated (supplemental Table 9). Pfam-A domains (curated part of Pfam database) were found for the majority of phosphoproteins (79%, 111 proteins) with the most common being the protein kinase domain (nine proteins) (Table II). Intriguingly, however, 94% (482) of phosphorylation sites were located outside of 201 Pfam-A domains (123 distinct domain types) that mapped to 141 proteins. 16 of the remaining 32 phosphorylation sites were located in MAP2 projection domains, and as these domains are unstructured (27, 28), at least 498 phosphorylation (97%) sites lie outside of structural domains in the cytosolic phosphoproteome. Comparison of the distribution of observed (by MS) serine/threonine phosphorylation sites versus all possible phosphorylation sites (all serine/threonine residues in 141 proteins) revealed that structural protein domains were significantly depleted ($p$ value <0.01) in Ser/Thr phosphorylation sites compared with sequences outside of protein domains. Phosphotyrosine sites were excluded from this analysis as the number of observed sites was just 2, but interestingly both of these sites are located in protein kinase domains and in regions of sequence order consistent with order-promoting characteristics of tyrosine residues. The depletion of Ser/Thr phosphorylation sites

![Fig. 3. Distribution of phosphorylation sites identified in the cytosolic phosphoproteome. A, numbers of approved phosphopeptides, previously phosphorylated peptides, phosphorylation events, and a breakdown of the types of phosphorylation sites found are shown. B, the distribution of the number of phosphorylation events per phosphopeptide shows that the majority of phosphopeptides were singly and doubly phosphorylated peptides, but good coverage of highly phosphorylated peptides was also achieved.]

![Fig. 4. Classification of cytosolic phosphoproteins. Phosphoproteins for which phosphopeptides were characterized and manually assigned were functionally classified according to Swiss-Prot. The major class of proteins in this data set have not been functionally characterized to date, but the next most prominent types of proteins are splicing/RNA-binding proteins, cytoskeletal proteins, and protein kinases.]

Molecular & Cellular Proteomics 7.7 1337
Enrichment of Intrinsic Sequence Disorder in Cytosolic Phosphoproteins—The amino acid composition of phosphoproteins characterized in this data set (141 phosphoproteins listed in supplemental Table 3) was analyzed using Composition Profiler (22). When compared with proteins contained in the entire Swiss-Prot database, the cytosolic phosphoproteome was significantly enriched (p < 0.01) in most of the amino acid residues (Glu, Lys, Arg, Gly, Gln, Ser, and Pro) that confer intrinsic sequence disorder (natively unfolded proteins) and significantly depleted (p < 0.01) in most of the amino acid residues that confer ordered structure (Ile, Leu, Val, Trp, Phe, Tyr, Cys, and Asn) (Fig. 5A). To make more specific comparisons, this analysis was repeated using 825 UniProt Mus musculus proteins (annotated as being expressed in the brain and located in the cytoplasm (UP-mbc)) as a background data set (Fig. 5B). This comparison showed the same significant enrichment in disorder-promoting and depletion in order-promoting residues, indicating specifically that phosphorylated proteins are significantly more likely to be disordered (Fig. 5).

Furthermore this UP-mbc data set was also enriched in disorder when compared with the entire Swiss-Prot database indicating that cytosolic proteins are enriched in disorder compared with proteins from other cell compartments (Fig. 5C). Therefore, phosphorylation and subcellular localization appear to be associated with the extent of intrinsic sequence disorder observed in the cytosolic phosphoproteome.

To ensure that these results are not due to a bias of IMAC enrichment for intrinsically disordered proteins, we compared data generated by three different enrichment methods (phosphoramidate chemistry, IMAC, and TiO2) (6) using Composition Profiler analysis. Comparison of phosphoproteins enriched by each of these methods with proteins in the Swiss-Prot database produced very similar profiles of enrichment of disorder-promoting and depletion of order-promoting residues (supplemental Fig. 2, A–C). Also to show that our results are not biased by the analytical platform used and that intrinsic sequence disorder is a common feature of phosphoproteomes, we compared our mouse brain cytosolic phosphoprotein data set to a mouse brain whole tissue lysate phosphoproteome data set (29) (enrichment using strong cation exchange and analysis on an ion trap) as well as a mouse brain synaptosome phosphoproteome data set (30) (enrichment using strong cation exchange and IMAC, analysis using MALDI-TOF, ion trap, and hybrid ion trap-FT) and a mouse brain postsynaptic density phosphoproteome data set (31) (enrichment using TiO2 and analysis with a Q-Tof) (Fig. 5 and supplemental Fig. 2). All of these additional phosphorylation data sets generated using different enrichment and analysis platforms show enrichment of intrinsic sequence disorder because they contain phosphoproteins. However, enrichment in intrinsic sequence disorder in the cytosolic phosphoproteome is more than that observed for these other data sets compared with Swiss-Prot because of the additive effects of the association of intrinsic sequence disorder with both phosphorylation and subcellular localization.

Enrichment of Phosphorylation in Regions of Intrinsic Sequence Disorder—More detailed analysis of the cytosolic phosphoproteome data set (141 phosphoproteins listed in supplemental Table 3) was performed using PONDR (which has a prediction accuracy of 98.3% on a per residue basis for predicting disordered regions of more than 40 residues).

Three general categories of disordered protein were revealed: those that can be considered highly disordered (>70% sequence disorder, 39 proteins), those that are mostly disordered (50–70% sequence disorder, 49 proteins), and those that are partly disordered (<50% sequence disorder, 53 proteins) (Fig. 6, Table III, and supplemental Table 9). The highly and mostly disordered categories are likely to represent completely disordered proteins especially if no obvious protein domains are present, whereas the partly disordered category may represent flexible linkers between regions of order or proteins with disordered N or C termini. Long disordered regions are usually defined as disordered regions of more than 40 residues (32), and as such, 94% of proteins (133) in this data set contain at least one long disordered region. 368 phosphorylation sites were located in regions of disorder over 40 amino acids long, and 205 phosphorylation sites were located in regions of disorder that were over 100 amino acids long. The higher proportion of phosphorylation sites observed in disordered regions of over 600 residues long (Fig. 6B) corresponds to highly phosphorylated and disordered RNA-binding and splicing proteins with the longest continuous region of disorder of 837 residues in the protein Srm2. Additionally we observed a trend for regions of disorder to elongate in highly disordered...
proteins rather than an increase in the number of disordered regions (Fig. 6).

In the cytosolic phosphoproteome data set, 86% (438 of 512) of phosphorylation sites were predicted to lie in intrinsically disordered regions using PONDR analysis (Table II). Comparison of the distribution of observed Ser/Thr phosphorylation sites versus all potential phosphorylation sites (all Ser/Thr residues) revealed that the serine/threonine phosphorylation sites are significantly enriched (p value < 0.01) in regions of intrinsic sequence disorder. In addition, of the remaining 74 phosphorylation sites that were predicted to occur in ordered regions, only 12 (2.3%) could be mapped to Pfam protein domains, three of which are in unstructured MAP2 projection domains. 95% of phosphorylation sites in disordered regions were also located outside Pfam domains (419 of 439), and collectively these two analyses highlight the significant propensity of phosphorylation to occur in unstructured regions in proteins.

Microtubule-associated protein Tau is an intrinsically disordered protein that is highly phosphorylated with at least 35 known phosphorylation sites (19 were identified in this study). 31 of these phosphorylation sites are clustered on either side of its four tandem microtubule-binding domains in regions that are clearly predicted to be disordered by PONDR analysis and have been experimentally (33) confirmed as being disordered (Fig. 7). The remaining four phosphorylation sites in Tau are located in tubulin-binding domains, two of which are in a predicted disordered sequence. CaMKII is a highly structured protein with only 21% disordered sequence and contains a large protein kinase domain that constitutes more than half of the protein sequence. A phosphopeptide in CaMKII and another in CaMKII (homologous sequence) map to the main disordered region in between the protein kinase domain and the association domain (supplemental Fig. 3). The occurrence of these phosphopeptides in a small window of disordered sequence in CaMKII and the large number of phosphorylation sites in disordered regions on either side of microtubule-binding domains in Tau further support the idea that sequence disorder and protein phosphorylation are intimately linked.

**Linear Binding Motifs, Phosphorylation, and Intrinsic Sequence Disorder**—One function of phosphorylation is to provide regulated binding sites on proteins. There are many different types of such phosphorylation-dependent protein interactions,

---

**Fig. 5. Enrichment of sequence disorder in the cytosolic phosphoproteome.** The sequence composition of the data set of cytosolic phosphoproteins was compared with proteins in the entire Swiss-Prot database (A) and 825 UniProt M. musculus proteins (annotated as being expressed in the brain and located in the cytoplasm (UP-mbc)) (B) using Composition Profiler software. In addition, this set of UP-mbc proteins was compared with the entire Swiss-Prot database to determine the enrichment of disorder of these proteins due to their subcellular localization. The ratio of significant enrichment (p < 0.01) is color-coded according to the propensity of amino acids to promote protein disorder (red) or promote protein order (blue). It can be clearly seen that the cytosolic phosphoproteome is enriched in most disorder-promoting and depleted in most order-promoting amino acids when compared with the Swiss-Prot database as well as the UP-mbc data set, which is representative of the starting material (mouse brain cytosolic extract) used for the phosphopurifications. This UP-mbc data set also shows disorder enrichment compared with the Swiss-Prot database indicating that cytosolic proteins tend to be more disordered than proteins in other cellular compartments, and the enrichment of disorder between the cytosolic phosphoproteome and the UP-mbc data set is likely because they are phosphoproteins. Error bars represent fractional differences of the standard deviations of observed relative frequencies of the bootstrap samples.
of the 331 phosphorylation sites (5.4%) were predicted by Scansite to be sites for 14-3-3 binding, there is an apparent enrichment for this motif in this data set.

The majority of 14-3-3 binding sites predicted by Scansite fall into three consensus motifs: RXX(pS/T)XP, RXX(pS/T)(G/D/E/S/T), and KXX(pS/T)(H/D/E/S/T) where pS/T represents phosphoserine/phosphothreonine (Table IV). 14 sites conform to the minimal motif RXX(pS/T)XP, whereas 28 sites conform to the RXX(pS/T)(G/D/E/S/T) motif. 17 of these putative 14-3-3 binding sites are predicted to be substrates of Akt, and 10 are predicted to be substrates for PKA, consistent with data that basophilic kinases such as Akt and PKA have consensus motifs for phosphorylation similar to those 14-3-3 proteins have for phosphorylation-dependent binding. To investigate other phosphorylation motifs in the cytosolic phosphoproteome data set, 13-residue-long sequences surrounding each unambiguously assigned phosphorylation site were analyzed using Motif-x. This analysis uses an iterative strategy that builds successive motifs through comparison with a dynamic statistical background. Comparison of sequences from the cytosolic phosphoproteome with the International Protein Index mouse proteome highlighted nine broad motifs showing significant enrichment (supplemental Fig. 4). Five of these motifs correspond to proline-directed kinases in which a proline residue is in the +1 position in relation to the phosphorylation site. The most significantly enriched motif was TPP, which was enriched by 47-fold compared with the control data set and has recently been reported as a novel and highly enriched motif in the nuclear phosphoproteome from HeLa cells (25). In addition, we also observed significantly enriched RXXX (X is any amino acid) motifs that are characteristic of arginine- and serine-rich (RS) domains in serine-arginine (SR) splicing factors. Finally two motifs (SXXXP and SX) that were also enriched likely reflect the high degree of clustering of phosphorylation sites observed in multiply phosphorylated peptides.

Intrinsic Sequence Disorder and Kinase Specificity—Predictions for kinases that are likely to phosphorylate unambiguously assigned phosphorylation sites (in disordered sequence) in the cytosolic phosphoproteome were performed using Scansite (supplemental Table 9). The top four kinases (of 18) that phosphorylated the most substrates were Cdk5 (71 sites), Cdc2 (50 sites), ERK1 (46 sites), and GSK3 (39 sites); all CMGC group kinases. These four kinases could account for 49% of phosphorylation motifs in the cytosolic phosphoproteome data set, and their optimal binding motifs have been characterized. We investigated the potential of the phosphorylation sites identified in this study to be sites for 14-3-3 binding by screening each site using Scansite, a short sequence motif prediction server. 58 of the 473 (12%) unambiguously assigned phosphorylation sites (11% of the total 512 sites) we characterized by mass spectrometry were predicted by Scansite to be sites for 14-3-3 binding (Table IV). When compared with our previous analysis of the mouse forebrain synapse phosphoproteome (3) in which 18

but the motifs for prediction of most of these interactions are not complete. However, 14-3-3 proteins that bind to phosphorylated serine or threonine residues have been intensively studied, and their optimal binding motifs have been characterized. We investigated the potential of the phosphorylation sites identified in this study to be sites for 14-3-3 binding by screening each site using Scansite, a short sequence motif prediction server. 58 of the 473 (12%) unambiguously assigned phosphorylation sites (11% of the total 512 sites) we characterized by mass spectrometry were predicted by Scansite to be sites for 14-3-3 binding (Table IV). When compared with our previous analysis of the mouse forebrain synapse phosphoproteome (3) in which 18
39 phosphoproteins are predicted to be over 70% disordered as assessed by PONDR analysis. 45% of these highly disordered proteins have no known function, whereas 26% are involved in splicing/RNA binding. This list contains three highly disordered proteins (Atrophin-1, Ataxin-1, and Ataxin-2 in bold) that contain trinucleotide repeats that when expanded cause spinocerebellar ataxias.

<table>
<thead>
<tr>
<th>Functional class</th>
<th>Gene name</th>
<th>Protein name</th>
<th>Longest disordered region (residues)</th>
<th>Disorder %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Splicing/RNA binding</td>
<td>Srm2</td>
<td>Serine/arginine repetitive matrix protein 2</td>
<td>837</td>
<td>95.89</td>
</tr>
<tr>
<td>DNA binding</td>
<td>Nucks1</td>
<td>Nuclear ubiquitous casein and cyclin-dependent kinase substrate</td>
<td>184</td>
<td>93.59</td>
</tr>
<tr>
<td>Signaling</td>
<td>Marcks</td>
<td>Myristoylated alanine-rich protein kinase C substrate</td>
<td>152</td>
<td>92.53</td>
</tr>
<tr>
<td>Unknown</td>
<td>Mtap7d1</td>
<td>Microtubule-associated protein 7 domain-containing 1</td>
<td>479</td>
<td>91.13</td>
</tr>
<tr>
<td>Unknown</td>
<td>B130050I23Rik</td>
<td>B130050I23Rik</td>
<td>209</td>
<td>89.32</td>
</tr>
<tr>
<td>Unknown</td>
<td>MGC42367</td>
<td>MGC42367</td>
<td>389</td>
<td>88.63</td>
</tr>
<tr>
<td>Unknown</td>
<td>8030462N17Rik</td>
<td>8030462N17Rik protein&lt;sup&gt;a&lt;/sup&gt;</td>
<td>145</td>
<td>86.72</td>
</tr>
<tr>
<td>Splicing/RNA binding</td>
<td>Acin1</td>
<td>Apoptotic chromatin condensation inducer in the nucleus</td>
<td>399</td>
<td>85.72</td>
</tr>
<tr>
<td>Splicing/RNA binding</td>
<td>1500011J06Rik</td>
<td>pre-mRNA splicing factor SRP55</td>
<td>161</td>
<td>85.71</td>
</tr>
<tr>
<td>Splicing/RNA binding</td>
<td>Srm1</td>
<td>Serine/arginine repetitive matrix protein 1</td>
<td>687</td>
<td>85.41</td>
</tr>
<tr>
<td>Unknown</td>
<td>Zc3h13</td>
<td>Zinc finger CCCH domain-containing protein 13</td>
<td>146</td>
<td>85.38</td>
</tr>
<tr>
<td>Unknown</td>
<td>Gm1568</td>
<td>Gene model 1568</td>
<td>298</td>
<td>84.28</td>
</tr>
<tr>
<td>Transcriptional regulation</td>
<td>Atn1</td>
<td>Atrophin-1</td>
<td>618</td>
<td>84.00</td>
</tr>
<tr>
<td>Unknown</td>
<td>Dact3</td>
<td>Dapper homolog 3&lt;sup&gt;a&lt;/sup&gt;</td>
<td>274</td>
<td>82.46</td>
</tr>
<tr>
<td>Unknown</td>
<td>Hdgfrp2</td>
<td>Hepatoma-derived growth factor-related protein 2</td>
<td>322</td>
<td>82.45</td>
</tr>
<tr>
<td>Cytoskeletal</td>
<td>Synpo</td>
<td>Synaptopodin</td>
<td>149</td>
<td>82.35</td>
</tr>
<tr>
<td>Cytoskeletal</td>
<td>Mapt</td>
<td>Microtubule-associated protein Tau</td>
<td>235</td>
<td>81.01</td>
</tr>
<tr>
<td>Cytoskeletal</td>
<td>MAP4</td>
<td>Microtubule-associated protein 4</td>
<td>357</td>
<td>80.71</td>
</tr>
<tr>
<td>Unknown</td>
<td>2010300C02Rik</td>
<td>2010300C02Rik</td>
<td>344</td>
<td>80.36</td>
</tr>
<tr>
<td>Unknown</td>
<td>Otd4</td>
<td>Otd4</td>
<td>60</td>
<td>79.78</td>
</tr>
<tr>
<td>Transcriptional regulation</td>
<td>Ncor2</td>
<td>Nuclear receptor corepressor 2&lt;sup&gt;a&lt;/sup&gt;</td>
<td>234</td>
<td>79.17</td>
</tr>
<tr>
<td>Unknown</td>
<td>Als2cr13</td>
<td>Amyotrophic lateral sclerosis 2 chromosome region, candidate 13&lt;sup&gt;a&lt;/sup&gt;</td>
<td>170</td>
<td>79.11</td>
</tr>
<tr>
<td>Splicing/RNA binding</td>
<td>Sf1</td>
<td>Splicing factor 1</td>
<td>291</td>
<td>78.41</td>
</tr>
<tr>
<td>Splicing/RNA binding</td>
<td>PPIG</td>
<td>Peptidyl-prolyl cis-trans isomerase G</td>
<td>516</td>
<td>77.19</td>
</tr>
<tr>
<td>Unknown</td>
<td>Gcc1</td>
<td>Glucocorticoid-induced transcript 1 protein</td>
<td>168</td>
<td>77.09</td>
</tr>
<tr>
<td>Unknown</td>
<td>AB30010M20Rik</td>
<td>Novel</td>
<td>166</td>
<td>76.57</td>
</tr>
<tr>
<td>Unknown</td>
<td>AB30010M20Rik</td>
<td>Hypothetical ATP/GTP-binding site motif A-containing protein</td>
<td>166</td>
<td>76.57</td>
</tr>
<tr>
<td>Translational regulation</td>
<td>Atxn2</td>
<td>Ataxin-2</td>
<td>123</td>
<td>76.26</td>
</tr>
<tr>
<td>Splicing/RNA binding</td>
<td>Sfrs4</td>
<td>Splicing factor, arginine/serine-rich 4</td>
<td>334</td>
<td>76.07</td>
</tr>
<tr>
<td>Unknown</td>
<td>Speccl</td>
<td>Sperm antigen with calponin homology and coiled-coil domains 1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>194</td>
<td>75.35</td>
</tr>
<tr>
<td>Splicing/RNA binding</td>
<td>Sfrs11</td>
<td>Splicing factor, arginine/serine-rich 11 homolog</td>
<td>165</td>
<td>75.35</td>
</tr>
<tr>
<td>Cytoskeletal</td>
<td>Map1a</td>
<td>Microtubule-associated protein 1A</td>
<td>200</td>
<td>75.29</td>
</tr>
<tr>
<td>Unknown</td>
<td>Phkb1</td>
<td>Peckstrin homology-like domain family B member 1</td>
<td>358</td>
<td>73.52</td>
</tr>
<tr>
<td>Transcriptional regulation</td>
<td>Gatad2b</td>
<td>Transcriptional repressor p66&lt;sup&gt;b&lt;/sup&gt;</td>
<td>151</td>
<td>73.40</td>
</tr>
<tr>
<td>G-protein modulator</td>
<td>Rabp1</td>
<td>RabA-binding protein 1</td>
<td>229</td>
<td>73.26</td>
</tr>
<tr>
<td>Splicing/RNA binding</td>
<td>YTHDC1</td>
<td>YTH domain-containing protein 1</td>
<td>226</td>
<td>73.18</td>
</tr>
<tr>
<td>Splicing/RNA binding</td>
<td>Zntrt2</td>
<td>Zinc finger protein 265</td>
<td>155</td>
<td>73.13</td>
</tr>
<tr>
<td>Translational regulation</td>
<td>Atn1</td>
<td>Ataxin-1</td>
<td>156</td>
<td>73.11</td>
</tr>
<tr>
<td>Unknown</td>
<td>Kiaa0284</td>
<td>Kiaa0284 protein&lt;sup&gt;a&lt;/sup&gt;</td>
<td>132</td>
<td>71.92</td>
</tr>
</tbody>
</table>

<sup>a</sup> Six phosphoproteins identified in the dephosphorylation experiment only.
number of LC-MS/MS strategies to characterize phosphorylation in these purifications. 152 proteins were identified in a solution digest of an aliquot of a protein IMAC sample, and 58 phosphopeptide spectra were observed, corresponding to 25 non-redundant phosphopeptide sequences. The value of profiling the protein IMAC enrichment lies in the fact that enriched phosphopeptides from many phosphoproteins may not be amenable to MS analysis due to charge, size, and hydrophobicity and would therefore not be detected and identified. In addition, the most abundant phosphopeptides in the protein IMAC enrichment were characterized.

Multiple DDA of a double IMAC purification resulted in the identification of 238 phosphopeptides (199 non-redundant base sequences) from 81 phosphoproteins. DDA of an enzymatically dephosphorylated double IMAC sample was performed under analytical conditions identical to those in experiment C. This approach generated dephosphorylated peptides that served as reference peptide identifications for the validation of 80 phosphopeptides that reinforced confidence in these phosphopeptide identifications. Furthermore the exceptional purity of this double IMAC enriched sample (98%) allowed dephosphorylated peptides to be considered.

**Table IV**

<table>
<thead>
<tr>
<th>14-3-3 binding sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>UniProt</td>
</tr>
<tr>
<td>NP_067468</td>
</tr>
<tr>
<td>Q52K18</td>
</tr>
<tr>
<td>Q6G205</td>
</tr>
<tr>
<td>Q6G205</td>
</tr>
<tr>
<td>Q6G205</td>
</tr>
<tr>
<td>Q6G205</td>
</tr>
<tr>
<td>Q6G205</td>
</tr>
<tr>
<td>Q6G205</td>
</tr>
<tr>
<td>Q6G205</td>
</tr>
<tr>
<td>Q6G205</td>
</tr>
</tbody>
</table>

Fig. 7. Phosphorylation and sequence disorder in microtubule-associated protein Tau. A, schematic illustration of the domain profile of Tau with all known phosphorylation sites and phosphorylation sites characterized in this study (red). The majority of phosphorylation sites cluster on either side of the four tubulin-binding domains in Tau protein. B, disorder plot of Tau that shows that the regions of disorder map to the regions on either side of the tubulin-binding domains where the majority of phosphorylation sites lie.
to be previously phosphorylated with high confidence. Other studies have reported the use of enzymatic dephosphorylation to validate phosphopeptide identifications; however, between 33 and 63% of peptides found after phosphopeptide enrichment were unphosphorylated, and therefore only phosphopeptides characterized in the dephosphorylated sample that matched phosphopeptides in the phosphorylated sample were useful (34, 35). It is clear that a subset of phosphopeptides were not amenable to detection under our LC-MS/MS conditions in their phosphorylated form. This is highlighted by the fact that the majority of phosphopeptides discovered in experiment C were also found in other DDAs but that 54 additional previously phosphorylated peptides were only observable after enzymatic dephosphorylation. Iterative DDAs (with an exclusion list, eDDA) of the double IMAC sample allowed the identification of 143 non-redundant phosphopeptides from 63 phosphoproteins. Additionally MS² analysis was performed on the sample to generate an inclusion list for a subsequent DDA experiment (iDDA), resulting in the identification of 118 non-redundant phosphopeptides from 58 phosphoproteins.

We probed phosphorylation in the brain cytosol at the phosphoprotein as well as phosphopeptide level with and without enzymatic dephosphorylation as well as using different analytical platforms to increase our coverage of this complex phosphoproteome. Overall complementary data and different segments of a phosphoproteome were observed when different phosphopeptide enrichment techniques, LC conditions, and mass spectrometers were used.

Protein Phosphorylation Occurs in Disordered Regions

Analysis of the distribution of phosphorylation sites with respect to protein domains revealed that in this data set the vast majority of phosphorylation sites were located outside of known Pfam-A domains. This indicates that in most cases protein phosphorylation and ordered protein structure are mutually exclusive. We postulated that if protein phosphorylation does not usually occur and cannot regulate proteins from within structural domains it must do so in other regions of proteins. The presence of a number of known disordered proteins led us to analyze the amino acid composition of the entire data set. When compared with proteins in Swiss-Prot (Fig. 5), Protein Data Bank (supplemental Fig. 2), and the postsynaptic proteome (supplemental Fig. 2), this data set of cytosolic phosphoproteins was significantly enriched for residues that confer protein sequence disorder and depleted for order-promoting residues. This apparent enrichment for disorder was reinforced when we compared the data set with a more specific and similar data set, 825 cytosolic mouse proteins expressed in the brain (UP-mbc) (Fig. 5). Collectively the enrichment in disorder between the cytosolic phosphoproteome and the UP-mbc data set and between the UP-mbc data set and the Swiss-Prot database is equivalent to the enrichment observed in the cytosolic phosphoproteome compared with the Swiss-Prot database (Fig. 5). This highlights that, in the cytosolic phosphoproteome, enrichment of sequence disorder is due to both the subcellular localization of these proteins and to the fact that they are phosphorylated with approximately equal contributions. Other phosphoproteomic data set (whole cell lysate, postsynaptic density, and synaptosomes) are also enriched in sequence disorder because of the relationship between phosphorylation and disorder, but the extent of enrichment was less than that observed for the cytosolic phosphoproteome (supplemental Fig. 2).

A general classification of cytosolic phosphoproteins according to the extent of sequence disorder highlighted that the majority of proteins (88 proteins, 62%) were over 50% disordered. More detailed analysis of the location of phosphorylation sites revealed that 86% of sites are predicted to lie in regions of sequence disorder, and of the remaining 14%, only 2.3% could be mapped to Pfam-A domains. Ser/Thr phosphorylation sites are significantly depleted in structural domains and significantly enriched in disordered regions of proteins. The complimentary nature of these two data types reinforces the specific distribution of phosphorylation sites in flexible unstructured protein sequence. Inspection of the relative topology of phosphorylation sites, protein domains, and sequence disorder for several proteins clearly highlights that most phosphorylation sites cluster to disordered regions, and this occurs even in relatively ordered proteins such as CaMKII (supplemental Fig. 3). In this example only a small portion of its sequence is not in a structural domain and is predicted to be disordered, and we characterized phosphopeptides clustered to this disordered linker region.

Kinase Accessibility and Specificity

A primary requirement of kinases and phosphatases for acting on a substrate protein is accessibility. This is of course not a novel concept but is interesting in the context of proteins that are natively unfolded in a state where much more of its backbone is accessible for phosphorylation or dephosphorylation. Such natively disordered proteins can become phosphorylated to a high degree when they exist as monomers, and upon binding to DNA or other proteins the presence or absence of this phosphorylation can regulate their function (36, 37). Therefore, if accessible disordered regions of proteins are the major requirements for protein phosphorylation then the specific residues in these sequences confer whatever differential specificity exists between different kinases.

Analysis of kinases that potentially phosphorylate sites identified in the cytosolic phosphoproteome showed that the CMGC group of kinases could phosphorylate the most sites. The specificity of these top four kinases includes a preference for proline and arginine/lysine residues in sequences surrounding target phosphorylation sites, consistent with disor-
order-promoting residues that are strongly enriched in this data set. Indeed intrinsic disorder propensity and position-specific amino acid frequencies have been combined to create an algorithm for predicting phosphorylation sites (17). It is not yet clear whether all serine/threonine kinases require completely disordered sequences as substrates (17), but clearly kinases that require such disorder-promoting residues in their consensus sequences will tend to phosphorylate completely disordered regions in proteins. Furthermore kinases such as ERK require docking sites in substrate proteins to which they bind prior to phosphorylation of the substrate protein (38). Such docking interactions are believed to increase specificity of phosphorylation as well as increasing the rate of phosphorylation (39). Furthermore docking site sequences for ERK contain arginine and glutamine residues (e.g. LAQRRX₄L where X is any residue (40)), and this would suggest that these docking sites are also enriched in disorder-promoting residues.

**Intrinsic Sequence Disorder and Phosphoprotein Function**—Intrinsically disordered proteins or regions of intrinsic disorder in proteins are associated with many cellular functions, including regulation of transcription and translation, signal transduction and protein phosphorylation, and assembly of multiprotein complexes such as the ribosome (41). Intrinsic disorder also appears to be a common feature of hub proteins (highly connected proteins in protein interaction networks) (42), and it has been suggested that this flexibility of protein conformation allows interaction with a greater number of proteins than conventional protein domain-domain interactions (43). In fact, gene ontology analysis of the cytosolic phosphoproteome revealed that terms “nucleotide binding” and “protein binding” were significantly enriched compared with the mouse proteome. This may reflect that enrichment of intrinsic sequence disorder permits increased binding capability in these proteins.

Intrinsic protein disorder has evolved from relatively low levels in bacteria (2%) and viruses (7%) to between 18 and 32% in *Caenorhabditis elegans* and *Drosophila melanogaster*, respectively (17). It has been suggested that intrinsically unstructured proteins evolve by repeat expansion, and two of the putatively unstructured phosphoproteins identified in this study (MAP2 and Tau) have statistically significant satellite sequences (44). The reason for positive evolutionary selection appears to be that unstructured proteins are capable of more versatile molecular functions compared with structured proteins. Their unstructured or flexible conformation may permit interaction with multiple binding partners at once and may have many more accessible sites for post-translational modifications (45). Unstructured proteins have little or no hydrophobic core, and consequently more of the protein can form binding interfaces compared with structured proteins, which are limited by their reduced surface area. The functional relationship between protein phosphorylation and sequence disorder is discussed for the two main groups of disordered phosphorylated proteins in this data set, namely RNA-binding/spliceosomal and cytoskeletal proteins.

**Phosphorylation and Intrinsic Sequence Disorder in the Spliceosome**—A quarter of the highly disordered proteins found in this study (Table III) are involved in RNA binding and splicing. The majority of these belong to the SR family of non-small nuclear ribonucleoprotein splicing factors, and we identified 76 phosphorylation sites on these proteins. This family is characterized by the presence of an RS domain and an RNA recognition motif. A characteristic feature of SR proteins is extensive serine phosphorylation, which is essential to their function in early stages of spliceosome assembly (46, 47). In addition, RS domain phosphorylation can regulate the activity of SR proteins (48, 49) by the introduction of negatively charged phosphogroups that influence both RNA and protein binding.

SRRM1 (Srm160) and SRRM2 (Srm300) are SR proteins that form the splicing coactivator, which functions in splicing by promoting critical interactions between splicing factors bound to pre-mRNA (50). We characterized 19 and 33 phosphorylation sites on these proteins, respectively, and in addition, we found that they are predicted to be highly disordered (85.4 and 95.9% disordered sequence, respectively). We also characterized phosphorylation sites on an SR protein kinase, PRP4 (62.7% disorder). PRP4 has been shown to bind to and is a substrate of Cik1, another SR protein kinase, and its activity has been mapped to the N terminus at possibly one of the sites we identified (Ser-143 and Ser-145) (51). There is evidence to suggest that Cik1 is not a direct regulator of SR proteins and that PRP4 is more active than Cik1 in phosphorylating SR proteins (51). SR-cyp (77.2% sequence disorder) interacts with Cik1, which hyperphosphorylates SR proteins causing their localization to change from nuclear speckles (zones of accumulation of transcriptional and mRNA splicing factors) to a diffuse nucleoplasmic localization (52). Similarly SR-cyp regulates the localization of SR proteins, including SRRM2, redistributing them from nuclear speckles to a diffuse nucleoplasmic localization (53).

SR proteins and other splicing-associated phosphoproteins characterized in this study were isolated from the cytosolic fraction, supporting a function in RNA export where they would shuttle from the nucleus to the cytoplasm. In addition, these proteins were found in their phosphorylated state, their presence in RNA export complexes or their cytoplasmic shuttling is likely to be regulated by phosphorylation. The requirement for protein and RNA binding, in conjunction with extensive phosphorylation and interaction with other disordered chaperones such as SR-cyp, is rudimentary to participation in highly regulated processes such as RNA splicing and export. The demands placed on SR proteins in terms of molecular binding capabilities and such extensive regulation by phosphorylation appear to be satisfied by their highly disordered nature.
Phosphorylation and Intrinsic Sequence Disorder in the Cytoskeleton—Three phosphorylated components of the cytoskeleton that are known to be disordered or contain experimentally determined disordered regions were found in the cytosolic phosphoproteome; tubulin polymerization-promoting protein (TPPP) (46.8% disorder), Tau (microtubule-associated protein Tau) (81.0% disorder), and β-adducin (48.3% disorder). β-Adducin binds to spectrin-actin complexes and promotes the association of actin with spectrin (54). The C-terminal tail of β-adducin is intrinsically disordered and was identified as the site for binding to spectrin-actin complexes (54). We identified six phosphorylation sites on two consecutive phosphopeptides spanning 53 amino acids of this unstructured C-terminal tail. This multisite phosphorylation cluster is well placed to regulate interaction with spectrin-actin complexes, lying N-terminal to protein kinase C phosphorylation sites that inhibit the activity of β-adducin in promoting spectrin-actin complexes (55).

TPPP was originally identified as a natively unfolded protein (56) but has subsequently been found to be partially ordered with an extended structure (57). TPPP stimulates aberrant tubulin polymerization that gives rise to microtubule assemblies in inclusion bodies of human pathological brain tissues such as in Alzheimer and Parkinson diseases. We characterized four phosphorylation sites in TPPP, two of which are located in the unstructured N terminus, which is missing in shorter forms of TPPP encoded by two separate genes in mammals (58). A phosphorylation motif (TPPKSP) within this region is also present in Tau and is attributed to Cdk5 phosphorylation (59). We found both phosphorylation sites in the motif (pTPPKpS) in Tau and also in TPPP, and this shared motif may have a function in TPPP similar to that in Tau.

Tau is a highly phosphorylated protein with 17 of 35 known phosphorylation sites characterized in this study (Fig. 7). Phosphorylated Tau has been implicated by many studies in the pathology of Alzheimer disease (AD), a neurodegenerative disease characterized by deposits of amyloid A-β peptides in plaques and by Tau deposits in the form of paired helical filaments (PHFs). Most neurodegenerative disorders are disorders of protein folding and are therefore classified as foldopathies. Tau promotes microtubule assembly and stability, and it is known to interact with α- and β-tubulins as well as other microtubule-associated proteins (60). Mutations in exon 10 of the Tau gene are associated with FTDP-17, an autosomal dominant hereditary neurodegenerative disorder (61). These mutations lie in an enhancer region, which SFRS10 (an intrinsically disordered SR protein characterized in this study) binds and regulates (61). Exon 10 encodes one of four microtubule-binding motifs, and aberrant splicing of this exon has implicated SFRS10 (five phosphorylation sites identified) and the increased affinity of Tau for microtubules in the pathogenesis of tauopathies (62).

The region surrounding the microtubule-binding repeats is highly phosphorylated by multiple kinases and is thought to regulate microtubule binding (63). Hyperphosphorylated Tau shows defective microtubule binding and fails to promote microtubule assembly (64). Tau is natively disordered (33), especially on either side of the microtubule-binding repeats where most phosphorylation is concentrated (Fig. 7). Hyperphosphorylated Tau is thought to cause aggregation and the formation of PHFs, and it has been shown that Tau displays an increased level of β-structure in PHFs (33). It therefore seems likely that the combination of disorder and phosphorylation surrounding the microtubule-binding repeats is important in PHF formation.

Relationship between Phosphorylation and Intrinsic Sequence Disorder in Neurodegenerative Diseases—We identified many phosphoproteins that are involved in neurological diseases, and here we discuss two main groups of proteins involved in AD/HD and spinocerebellar ataxias. Dysregulation of cellular signaling pathways is fundamental to many pathologies and especially in neurodegenerative diseases where aberrant phosphorylation of key proteins such as Tau has been implicated in disease formation and progression. The identification of sites of phosphorylation or mapping of the normal phosphorylation state of a phosphoproteome is necessary to understand how changes in the phosphorylation state of proteins can lead to disease. In addition to mapping many phosphorylation sites on Tau and TPPP, which are implicated in AD, we also identified phosphorylation sites on proteins that have been reported to have reduced expression in AD brain (Cy7B (65), Drebrin (66), and MADD (67)).

GIT1 and its interacting protein, collapsin response mediator protein 1 (CRMP1) (68), which in turn interacts with CRMP2, were all characterized in this cytosolic phosphoproteome data set. GIT1, a G-protein-coupled receptor kinase-interacting protein, has numerous cellular functions including acting as a scaffold for the kinases ERK1/2 and MEK1 in focal adhesions (69). Interestingly GIT1 was discovered through a yeast two-hybrid screen to directly interact with huntingtin protein (HTT) (68). Its localization to neuronal inclusions and selective cleavage in HD brains further endorsed its role in HD pathogenesis (68). Increased GSK3β activity is associated with AD, and CRMP2 has been found to be a physiological substrate for GSK3β (70) (we observed the autophosphorylation site of GSK3β). A hyperphosphorylated region of CRMP2 is an Alzheimer disease epitope and is physically associated with neurofilibrillary tangles (70, 71). We characterized this phosphorylated AD epitope and found that it contained seven phosphorylated residues in a stretch of 16 amino acids, three of which are novel. It appears that phosphorylation of this epitope regulates CRMP2 binding to tubulin and that GSK3β is at least partly responsible for this regulation (72). Also GSK3β phosphorylation of CRMP2 regulates axon elongation in primary neurons possibly by promoting microtubule assembly (70).

Spinocerebellar ataxias are neurodegenerative diseases that are caused by expanded CAG trinucleotide repeats en-
coding polyglutamine tracts in different genes. We characterized three such proteins in this study (Table III). We identified a phosphorylation site on serine 752 on Ataxin-1, a protein that when accumulated causes spinocerebellar ataxia type 1 (73). Analysis of the sequence surrounding the site (using Scansite) led to the prediction that Akt could phosphorylate the site and that it was also a consensus 14-3-3 binding site. Interestingly retrospective literature searching revealed that it had been experimentally demonstrated that this site was indeed phosphorylated by Akt, which created a binding site for 14-3-3 proteins (74). The binding of 14-3-3 to Ataxin-1 mediates the neurotoxicity of Ataxin-1 by stabilizing the protein, which slows down its normal degradation resulting in a striking buildup of the protein in nuclear inclusions (74). Mutation of the phosphorylation site to an alanine residue abolished the ability of Ataxin-1 to cause neurodegeneration in flies (75), further supporting the pathogenicity of the single phosphorylation site in combination with repeat expansion. This example reinforces the potential usefulness of detailed sequence-based analyses of phosphorylation sites from phosphoproteomics studies.

The Ataxin-2 gene contains a trinucleotide repeat that encodes a polyglutamine stretch that when expanded causes spinocerebellar ataxia type 2 (76). Another trinucleotide repeat-containing phosphoprotein, Atrophin-1, that was identified in this study also causes a types of spinocerebellar ataxia (dentatorubral-pallidoluysian atrophy) when repeat expansion occurs (77). As well as characterization of phosphorylation sites implicated in neurodegenerative disorders we found phosphorylation sites on 82-FIP, a fragile X mental retardation protein-interacting protein. We characterized a phosphorylation site on 82-FIP at serine 652 that is predicted to be a site for 14-3-3 binding and may be involved in the observed regulated localization of this protein in a manner similar to that described for Ataxin-1. In addition, we characterized phosphorylation sites on NF1 (mutations in which cause neurofibromatosis (78) and Kit1b (Charcot-Marie-Tooth disease type 2A (79)) and dephosphorylated peptides from ATRX (mutations in which cause X-linked α-thalassemia with mental retardation syndrome) (80).

Concluding Remarks

Protein phosphorylation requires that kinases and phosphatases that attach and remove phosphate groups are able to access the target sequence of residues in a protein. In addition, other proteins require accessible and disordered regions of proteins for phosphorylation-dependent binding, such as 14-3-3 proteins. These features of disordered proteins and the local structural requirements for protein phosphorylation point to the fact that sequence disorder and phosphorylation are closely associated, and enrichment of intrinsic sequence disorder is a common feature of phosphoproteomes.

Acknowledgments—We are grateful to Sajani Swamy for performing statistical analyses for this manuscript and to Holger Husi for advice on sample preparation.

* 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.

† Supported by the Wellcome Trust Sanger Institute.

‡ Supported by the Wellcome Trust Sanger Institute and by the Wellcome Trust Genes to Cognition program.

§§ To whom correspondence should be addressed. Tel.: 44-1223-834244; Fax: 44-1223-494919; E-mail: jc4@sanger.ac.uk.

REFERENCES


31. Trinidad, J. C., Thalhammer, A., Specht, C. G., Lynn, A. J., Baker, P. R.,
21. Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V.,
18. Collins, M. O., Yu, L., Husi, H., Blackstock, W. P., Choudhary, J. S., and
30. Munton, R. P., Tweedie-Cullen, R., Livingstone-Zatchej, M., Weinandy, F.,
35. Marcantonio, M., Trost, M., Courcelles, M., Desjardins, M., and Thibault, P.
18. Collins, M. O., Yu, L., Husi, H., Blackstock, W. P., Choudhary, J. S., and
re-assessing the protein structure-function paradigm. J. Mol. Biol. 293, 
60. Takahashi, M., Tomizawa, K., Ishiguro, K., Sato, K., Omori, A., Sato, S.,


