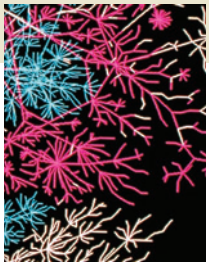

An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets

Daniel Schwartz & Steven P Gygi

With the recent exponential increase in protein phosphorylation sites identified by mass spectrometry, a unique opportunity has arisen to understand the motifs surrounding such sites. Here we present an algorithm designed to extract motifs from large data sets of naturally occurring phosphorylation sites. The methodology relies on the intrinsic alignment of phospho-residues and the extraction of motifs through iterative comparison to a dynamic statistical background. Results show the identification of dozens of novel and known phosphorylation motifs from recently published serine, threonine and tyrosine phosphorylation studies. When applied to a linguistic data set to test the versatility of the approach, the algorithm successfully extracted hundreds of language motifs. This method, in addition to shedding light on the consensus sequences of identified and as yet unidentified kinases and modular protein domains, may also eventually be used as a tool to determine potential phosphorylation sites in proteins of interest.

As research in molecular biology moves forward it has become increasingly clear that few cellular processes are unaffected by protein phosphorylation. Protein degradation, localization and conformation as well as protein/protein interactions are only some of the functions in which protein phosphorylation has been implicated[1,2]. Furthermore, protein phosphorylation levels are central to our current understanding of cell division and signal transduction pathways in both normal and diseased cell states[3]. Yet, relatively little is known about the majority of protein kinases in the human proteome. Only approximately one-tenth of the estimated 500–600 human protein serine, threonine and tyrosine kinases have known consensus sequences for their sites of phosphorylation[4]. Even when consensus sequences are known, *in vivo* protein substrates are often lacking.

To date, the task of understanding kinase recognition sequences has progressed mainly by a 'kinase-driven' approach whereby a kinase of interest is incubated with a combinatorial peptide library and ATP. Edman degradation of the phosphorylated peptides, which have been enriched using a ferric column, leads to the creation of a position-weight matrix of the data and hence the consensus sequence[5]. Though the kinase-driven approach has had much success in identifying optimal kinase consensus sequences and substrates, it has suffered from the fact that optimal *in vitro* binding is often kinetically unfavorable in the cellular environment, thus leading to motifs that are rarely found in the proteome.

Here we present an attempt to start with known biologically phosphorylated substrates from unknown kinases and discover motifs through a 'substrate-driven' approach. In the past, the low number of localized phosphorylation sites cited in the literature made substrate-driven approaches to determining kinase consensus motifs difficult. However, refinements of several affinity-based strategies such as immunoaffinity[6], immobilized metal affinity chromatography (IMAC)[7] and strong cation exchange (SCX) chromatography[8], coupled with the enabling technology of tandem mass spectrometry have more than doubled the number of phosphorylation sites identified in the past year alone, with several studies reporting from several hundred to several thousand sites[6,8–13].

Two of these recently published large-scale mass spectrometry studies were chosen as test sets for our motif-building algorithm. The first study used SCX for the enrichment of phosphopeptides from HeLa cell nuclei, resulting in the elucidation of 1,594 unique phosphoserine and 195 unique phosphothreonine sites[8]. The second study used an antiphosphotyrosine antibody to enrich for phosphorylated tyrosine residues in pervanadate-treated Jurkat cells (151 sites), cells expressing constitutively active NPM-ALK fusion kinase (237 sites) and cells expressing constitutively active Src kinase (185 sites)[6].

Overview of the method

A schematic of the motif extraction algorithm is shown in **Figure 1**. The method commences with the establishment of two parallel sequence data sets: the phosphorylated peptide data set from which motifs will be built, and a peptide data set used for background probability calculations. Next, the two data sets are converted into position-weight matrices of equal dimensions whereby each matrix contains information on the frequency of all residues at the six positions upstream and downstream of the phosphorylation site. Using the information encoded in these two matrices, a third matrix, the binomial probability matrix, is created. Specifically, this matrix contains the probability of observing s or more occurrences of residue x at position j (taken from the phosphorylation

Department of Cell Biology, 240 Longwood Ave., Harvard Medical School, Boston, Massachusetts 02115, USA. Correspondence should be addressed to D.S. (dschwartz@hms.harvard.edu).
Published online 4 November 2005; doi:10.1038/nbt1146

matrix), given a background probability P for residue x at position j (taken from the background matrix).

The motif-building step of the algorithm is a greedy recursive search of the sequence space to identify highly correlated residue/position pairs with the lowest P values. Each recursive iteration identifies the most statistically significant residue/position pair meeting a user-defined binomial probability threshold (in this study taken as $P < 10^{-6}$) and occurrence threshold (which represents the minimal number of sequences in the phosphorylation data set needed to match the residue/position pair). When such a pair is found, the sequence spaces of the phosphorylation and background matrices are reduced by retaining only those sequences containing the selected residue/position pair, and a new binomial probability matrix is calculated (see **Fig. 1**). This recursive pruning procedure is repeated until no more statistically significant residue/position pairs that meet the occurrence threshold are detected. At this point the motif is identified by the tally of residue/position pairs selected during this step.

The next major step of the algorithm involves set reduction of the phosphorylation and background data sets by removing all of those sequences that match the motif identified in the motif-building step. The purpose of this step is to remove the effects of those peptides with identified motifs from confounding the search for other significant motifs. Thus, performing the sequential loop of motif building followed by set reduction results in a decomposition of the phosphorylation sequence database into a list of significant motifs. The algorithm is complete (that is, the loop exits) when the motif-building step fails to identify any significant residue/position pairs.

Algorithm validation

To test the effectiveness of our algorithm, we applied it to a linguistic data set to determine its ability to extract English language motifs, that is, English words or common word fragments. Using a framework previously conceptualized by Bussemaker *et al.*[14] to test their algorithm for the detection of regulatory DNA motifs, we ran our motif-building strategy on the first ten chapters of the classic English novel *Moby Dick*[15] with random characters (at frequencies identical to those found in the original text) inserted between words[14]. Using the criteria $P < 10^{-6}$ and *occurrences* ≥ 10, we extracted 384 unique motifs of which 371 mapped back to English words in the original text, indicating a false positive rate of 3.4% (**Supplementary Tables 1–4** online). Additionally, the motifs

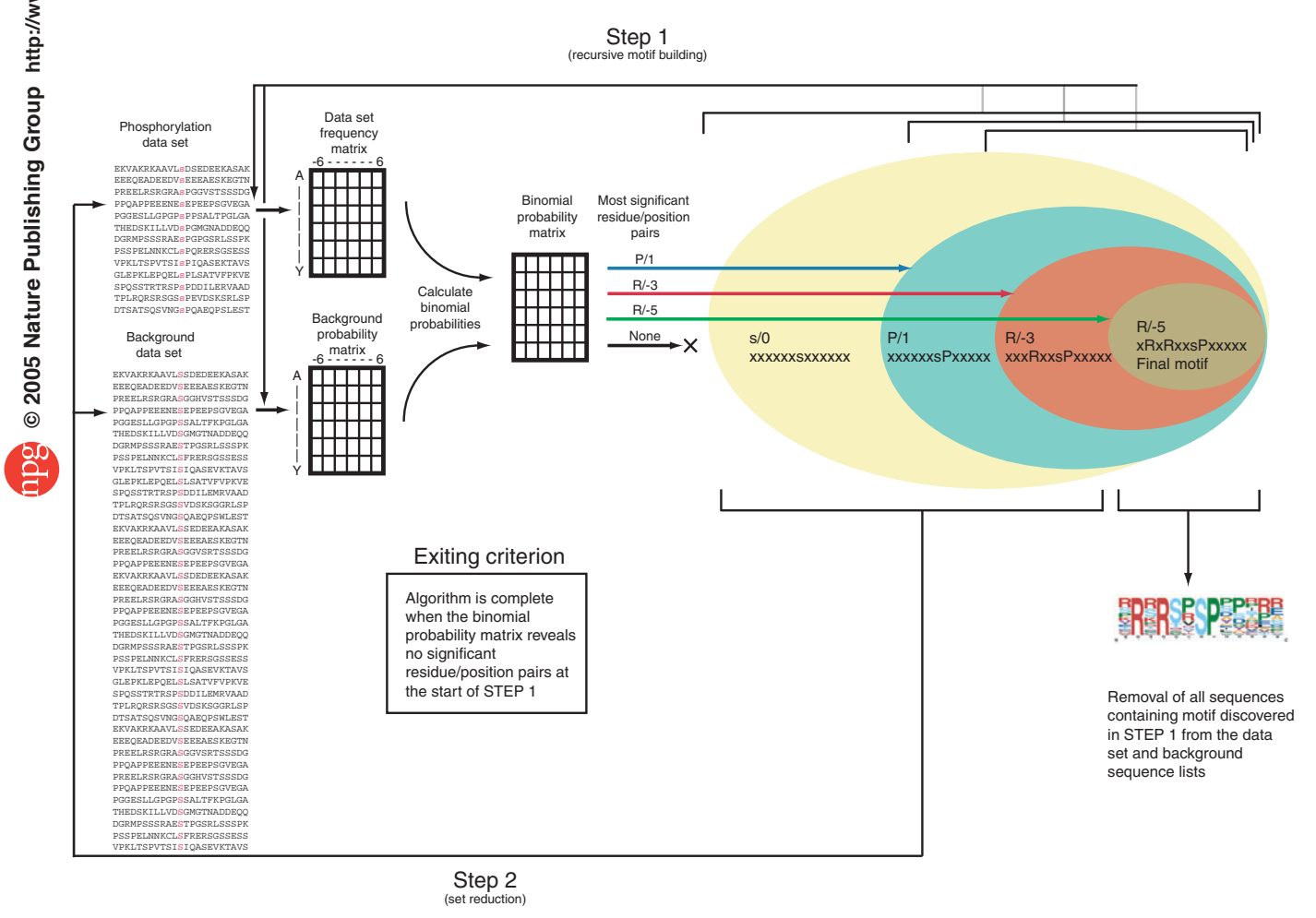

Figure 1 Overview of motif-building strategy. The algorithm can be divided into two major steps. In step 1, the motif is built through a recursive search of the sequence space with each iteration yielding the most significant residue/position pair until no more significant pairs could be detected. This is shown here by P/1, R/-3, and R/-5, thus making the first motif RxRxxsP (where 'x' denotes any residue, and lowercase letters denote phosphorylated residues). Set reduction occurs in step 2 whereby all the sequences containing the motif are removed from both the phosphorylation and background data sets. The motif logo is created from just the sequences removed from the phosphorylation data set following step 1. Though this figure depicts only one iteration of the algorithm, it is the sequential iteration of steps 1 and 2 that ultimately leads to the final list of motifs. See text for greater detail.

extracted covered 93.4% of the English words in the original text (3,886 out of 4,160). If we required the motifs to have at least two fixed letter/position pairs (aside from the central letter), a scenario more suitable for a large linguistic data set, then 317 unique motifs were extracted with only one false positive (FP) (FP rate = 0.32%) and a coverage of 67.1%. It is important to note that because approximately half of the data in this analysis were composed of random characters, the false-positive rate was substantially higher than would be expected in the phosphorylation data set analysis where all of the data were centered on true phosphorylation sites. Nevertheless, to remain conservative, we retained these same stringent *P* value thresholds in our biological analyses.

Table 1 S-centered motifs extracted from an *in silico* generated protein data set containing the motifs RxSxxL, RxSxxP, TVxSxE, DxxSQxN and KSxxxI

Motif*	Score**	"S"–centered data set (*Matches*/Size)		Background data set (*Matches*/Size)	
....R.S..L...	32.00	*199*	9,774	*758*	111,506
....R.S..P...	28.82	*192*	9,575	*547*	110,748
...TV.S.E....	40.85	*137*	9,383	*154*	110,201
...D..SQ.N...	40.46	*128*	9,246	*135*	110,047
.....KS...I..	24.15	*158*	9,118	*413*	109,912

* $P < 10^{-6}$, *occurrences* ≥ 20. ** Score = $\sum$-log(*P*).

In order to more closely mimic the biological situation, we further validated our approach using two additional data sets. The first of these consisted of 300 *in silico*–generated artificial proteins (**Supplementary Table 5** online). These synthetic proteins were created using human proteome residue frequencies and varied between 50 and 700 residues in length. The proteins were then studded at random positions with the following five motifs, DxxSQxN, RxSxxL, TVxSxE, RxSxxP, and KSxxxI ('x' residues retained background residue frequencies). To ensure that the artificial data set was sufficiently challenging, we inserted each motif at most once in only ~50% of the proteins. To deal with the difficulty of an unaligned data set, we created a 'pseudo-alignment' by taking a sliding window of all 13-mers in the data set and dividing this into 20 subsets based on each of the central residues. The motif-extraction algorithm was then run independently on each of these subsets (with the set of all 13-mers as a background data set). Using the same parameters established in the linguistic analysis, and with a run time of under 5 min, the method was able to build and extract all five motifs in various alignments with no false positives. Because each of the motifs contained an 'S,' the S-centered analysis extracted all five motifs at once (**Table 1**). However, the D-, E-, Q-, N-, R-, L-, T-, P-, K-, V- and I-centered analyses also extracted the correct motifs containing those particular residues (data not shown). For example, the Q-centered analysis extracted only the motif DxxSQxN while the R-centered analysis extracted the motifs RxSxxL and RxSxxP. Thus the algorithm does not depend on a priori knowledge of any specific residues contained in the motif which may

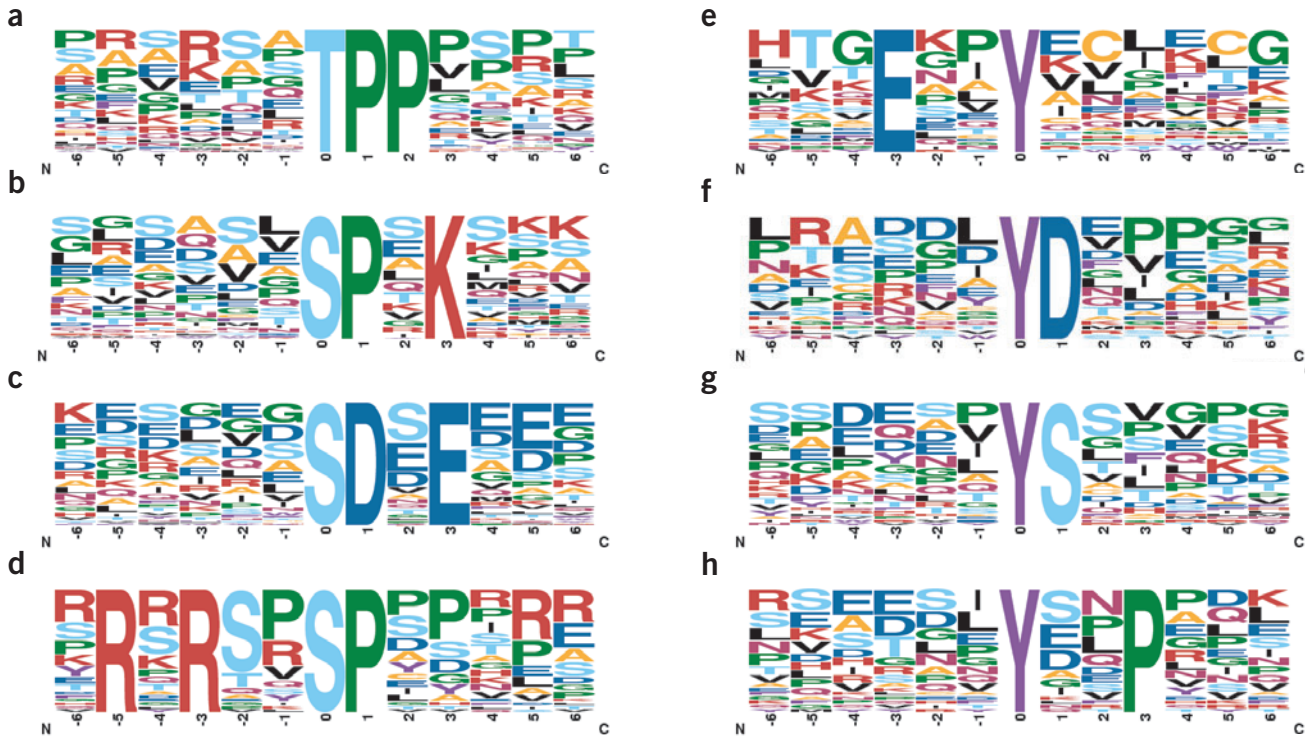

Figure 2 Sequence logo representations of various extracted motifs. (**a**) Most significant motif from the threonine phosphorylation data set. (**b**,**c**) Examples of significant motifs from the serine phosphorylation data set. (**b**) Extracted cyclin-dependent kinase motif. (**c**) Extracted casein II kinase motif. (**d**) Most significant motif from the serine phosphorylation data set. (**e**) One of the NPM-ALK fusion kinase motifs showing a phosphorylated tyrosine residue in C_2H_2 zinc finger domain. (**f**) Candidate c-Src motif similar to that found by combinatorial peptide library screening approaches. (**g**) Unique candidate c-Src motif. (**h**) Motif similar to (**f**) extracted from pervanadate-treated Jurkat cell data set consistent with known Src activation within these cells.

allow for the discovery of biologically important residues. As in the linguistic analysis, it should be noted that this was a much more complex data set than would be seen in a phosphorylation study because only ~1–2% of the 13 mers contained a given motif.

Though the next data set used to validate our algorithm was significantly less complex than the aforementioned ones, it closely resembled the intended application of the algorithm (**Supplementary Table 6** online). To generate this data set we used the Phospho.ELM database[16] to extract serine-phosphorylated peptides experimentally determined to be substrates of the following four kinases: Ataxia Telangiectasia Mutated (ATM) (43 sites), Casein II (184 sites), Calcium/Calmodulin-dependent protein Kinase II (CaMK II) (41 sites) and Mitogen-Activated Protein Kinase (MAPK) (30 sites). Application of the motif-building algorithm to this combined data set resulted in the extraction of six motifs corresponding to the precise consensus sequences for ATM, Casein II and CaM II kinases (**Table 2**). In the case of MAP kinase, the small size of the initial MAPK data set yielded an sP motif instead of the canonical PxsP motif.

Comparison to other algorithms

Despite the wealth of motif-discovery algorithms designed to predict transcription factor binding sites, tools for the extraction of protein motifs have not kept pace. Though no algorithms exist with the specific intention of extracting protein phosphorylation motifs, we have chosen four of the most popular protein motif discovery programs against which to benchmark our algorithm. We applied the TEIRESIAS[17], Pratt[18], Gibbs motif sampler[19] and eMOTIF[20] algorithms through their online servers to the aforementioned *in silico* and Phospho.ELM-derived data sets used to test our approach.

The Gibbs motif sampler is an iterative Monte Carlo procedure, which results in a position-weight matrix representation of a motif. When applied to the two test data sets the Gibbs sampler extracted only DxxSQxN from the artificial-protein data set and HS[IS][PY][SPHE] (a false positive) from the Phospho.ELM-data set.

Pratt operates through pruned depth-first search of the sequence space and returns an almost unlimited number of highly generalized patterns, which vary drastically in accordance with the large number of parameters. Of the top 1,000 motifs returned by the Pratt algorithm on our artificial-protein data set, several appeared to be related to our inputted motifs, but none were exact matches and the overwhelming majority did not resemble any of the five inserted motifs. When applied to our Phspho.ELM-derived test set, Pratt returned 1,000 motifs which were almost entirely acidic in nature. The first ATM-like motif appeared as motif number 492 (SQxxxS).

The TEIRESIAS algorithm is based on an exhaustive search of small patterns followed by a convolution phase in which the small patterns are joined to form longer ones. Using our artificial protein data set TEIRESIAS extracted the motifs TVxSxE and DxxSQxN as the 4th and 6th hits, respectively. These were then followed by a long list of motifs containing serine and leucine presumably because of their increased frequency in the human proteome and the lack of background filtering in the algorithm. None of the other three motifs were found in the top 500 patterns returned by TEIRESIAS. When tested against our Phospho.ELM-derived data set, the program returned all four kinase motifs: casein II kinase motif sxxE (hit no. 1), ATM motif sQ (hit no. 17), CaM II kinase motif Rxxs (hit no. 30) and MAP kinase motif sP (hit no. 40). While TEIRESIAS was the only algorithm tested that returned all four kinase motifs, the high number of false-positive motifs between these true positives unfortunately limits the applicability of the algorithm to most biological data sets of this nature.

The final protein motif discovery tool we tested was eMOTIF. By using a prealigned data set this program uses a pruned exhaustive search to find motifs with high specificity and coverage. Since eMOTIF requires a prealigned data set, we used the same set of 13 mers centered on 'S' used for our motif-building analysis as input for the artificial-protein data set. Probably because this data set was not a true alignment, eMOTIF was unable to form any motifs. However, when applied to the Phospho.ELM-derived data set, eMOTIF returned 497 motifs. Inspection of these motifs revealed a majority of highly generalized acidic motifs. The specific motifs for two of the expected kinases, namely ATM and CaMK II, were found in the middle of the list.

Analysis of mass spectrometry phosphorylation data sets

Table 3 shows the results of the motif extraction algorithm applied to the set of threonine phosphorylated peptides from the Beausoleil, Jedrychowski *et al.* [8] data set. The most significant motif identified, tPP (where lowercase letters indicate phosphorylated residues) appeared in 62 unique sequences from the data set, representing ~32% of the phosphorylation data. The same motif was identified in only 0.7% of the background data set, thereby highlighting the statistical significance of this motif. To further visualize the identified motifs, we used sequences from the subset of the phosphorylation data set containing the motifs to construct sequence logos in which the height of each residue in the logo was proportional to its frequency in the subset[21,22]. It is evident from the sequence logo for the tPP motif that a preference for basic residues exists at the –3 position of the motif (**Fig. 2a**).

To address the issue of conservative amino acid substitutions, often found in kinase motifs, we created degenerate phosphorylation and background data sets whereby the 20 amino acids were condensed into an 11-amino-acid code based on residue characteristics (see Methods and shaded portions of **Table 3**). Not surprisingly, the degenerate analysis yielded a more specific analog of the initial motif, [RK]xxtPP (where 'x' denotes any residue, 't' represents the phosphorylated threonine and [RK] denotes R or K at that position), which was 168-fold overrepresented in the phosphorylation data set as compared to the background. Sequences from the data set that contained this phosphorylated motif indicated a significant number of transcription-related proteins including eIF4γ2, eIF3, HOMEZ, PPRB, TCF20, RUNX1 and DRPLA. Despite their overwhelming statistical significance in this data set, it is interesting that the biological significance of these motifs has yet to be reported in the literature.

The motif analysis of serine phosphorylated peptides from this data set indicated successful decomposition into 12 previously identified kinase motifs and 6 novel motifs (**Table 3**,

Table 2 Motifs extracted from an experimentally validated data set of ATM, Casein II, CaMK II, and MAPK kinase substrates

Motif*	Kinase	Score**	Phospho data set (*Matches*/Size)		Background data set (*Matches*/Size)	
.....sD.E...	Casein II	29.83	*33*	298	*5,574*	1,279,892
.....s..E...	Casein II	16.00	*70*	265	*77,819*	1,274,318
.....s..D...	Casein II	16.00	*51*	195	*61,044*	1,196,499
.....sQ.....	ATM	16.00	*38*	144	*51,451*	1,135,455
.....sP.....	MAPK	9.33	*29*	106	*80,073*	1,084,005
...R..s.....	CaMK II	8.78	*21*	77	*57,969*	1,003,932

* $P < 10^{-6}$, *occurrences* ≥ 20. ** Score = Σ-log(P).

Table 3 Normal and degenerate analyses of phosphorylated threonine and serine motifs from the Beausoleil, Jedrychowski *et al.*[8] data set

Motif	Literature	Score*	Phospho data set (*Matches*/Size)		Background data set (*Matches*/Size)	
Threonine motifs**						
.....tPP....	Novel	32.00	*62*	195	*142*	20,208
.....tP.....	Proline-directed	16.00	*79*	133	*1,424*	20,066
...[KR]..tPP....	Novel	41.12	*26*	195	*16*	20,208
.....tPP....	Novel	28.11	*36*	169	*126*	20,192
...[KR]..tP....[KR]	Novel	35.18	*10*	133	*2*	20,066
.....tP.....	Proline-directed	16.00	*69*	123	*1,422*	20,064
Serine motifs***						
.R.R..sP.....	Novel	48.00	*34*	1,594	*2*	27,980
...R..sP.P...	Novel	38.04	*36*	1,560	*12*	27,978
.....sP...RR	Novel	39.24	*25*	1,524	*7*	27,966
...R..sP.....	Novel	28.07	*70*	1,499	*69*	27,959
....P.sP.....	MAPK	28.23	*169*	1,429	*265*	27,890
.....sP.R...	CDK	29.48	*69*	1,260	*75*	27,625
.....sPP....	Novel	24.87	*79*	1,191	*127*	27,550
.....sP.K...	CDK	24.54	*57*	1,112	*87*	27,423
..R...sP.....	Novel	23.21	*40*	1,055	*60*	27,336
.....sP.....	Pro-directed	16.00	*359*	1,015	*1,490*	27,276
.....sD.E...	CK II	32.00	*95*	656	*208*	25,786
.....sE.E...	CK II	32.00	*68*	561	*273*	25,578
.R.RS.s.....	AKT	41.31	*22*	493	*10*	25,305
...RS.s.....	AKT-like	25.21	*29*	471	*93*	25,295
...R..s.....	CaMK II	16.00	*63*	442	*1,043*	25,202
.....DsD.....	CK II-like	24.58	*25*	379	*125*	24,159
.....s.D....	CaMK II	9.69	*55*	354	*1,459*	24,034
.....s.E....	G-CK	10.44	*60*	299	*1,802*	22,575
.[KR].[KR]..sP....[KR]	Novel	64.00	*20*	1,594	*0*	27,980
...[KR][ST].sP...[KR].	Novel	56.68	*26*	1,574	*1*	27,980
...[KR]..sP.P...	Novel	41.84	*47*	1,548	*20*	27,979
.....sP.[KR]...	CDK	31.95	*157*	1,501	*203*	27,959
....P.sP.....	MAPK	30.52	*158*	1,344	*240*	27,756
.....sP...[KR][KR]	Novel	39.61	*38*	1,186	*20*	27,516
...[KR]..sP.....	Novel	26.93	*81*	1,148	*126*	27,496
.....sPP....	Novel	23.48	*62*	1,067	*114*	27,370
.....sP.....	Proline-directed	16.00	*349*	1,005	*1,470*	27,256
.....s[DE].[DE].[DE].	CK II	44.15	*107*	656	*218*	25,786
.....[DE]s[DE].[DE]...	CK II	39.00	*44*	549	*115*	25,568
.....s[DE][DE][DE]...	CK II	33.83	*29*	505	*106*	25,453
.[KR].[KR][ST].s.....	AKT	38.33	*29*	476	*57*	25,347
.....s.[DE].[DE]..	CK II-like	22.35	*44*	447	*548*	25,290
...[KR][ST].s.....	AKT-like	24.43	*35*	403	*281*	24,742
.....s..[DE]....	CK II	9.54	*97*	368	*3,409*	24,461
.....s.[DE]....	CaMK II/G-CK	6.63	*60*	271	*2,365*	21,052

* Score = Σ-log(*P*). ** $P < 10^{-6}$, *occurrences* ≥ 10. *** $P < 10^{-6}$, *occurrences* ≥ 20.

second nonshaded region) which together were able to account for 85% of the starting phosphorylated data set. A computational analysis of the sequence space of the 1,594 phosphorylated peptides revealed an upper bound of 246,383 potential motifs (containing 1, 2 or 3 nonwildcard positions). If we make the highly conservative assumptions that ~100 serine phosphorylation motifs exist in the literature and all are represented in the data set, then the probability of extracting a single known motif by chance is 0.0004 (100/246,383) and the odds of extracting 12 known motifs is vanishingly small. Thus, the identification of 12 known motifs served both as a validation of the methodology, and a positive control for the data. Among these were motifs for MAPK, Cyclin-Dependent Kinase (CDK), Casein II kinase, AKT, CaMK II and Golgi Casein Kinase (G-CK) (see **Figs. 2b**,**c** and **Table 3**). Surprisingly however, the novel motifs RxRxxsP, RxxxsPxP and sPxxxRR were among the most significant motifs identified. Although they share similarities with both basophilic and proline-directed kinases, these motifs have not been previously characterized (example in **Fig. 2d**). Inspection of the proteins identified from the data set containing these novel motifs revealed a

disproportionate number of the so-called RS domain–containing proteins involved in RNA binding and splicing[23]. It should also be noted that the serine motifs were robustly identified even when differing background data sets were chosen (see **Box 1**, **Table 4** and **Supplementary Tables 7–10** online).

Although we demonstrated the ability of the motif-building algorithm to decompose a large data set into constitutive motifs, we also wished to test the performance of the algorithm on small data sets containing fewer motifs. To this end we used the data sets from the Rush *et al.*[6] tyrosine phosphorylation immunoaffinity- tandem mass spectrometry (MS/MS) study. The first of these data sets contained tyrosine phosphorylated peptides from two cell lines, Karpas 299 and SUD-DHL-1, both of which are known to express constitutively active Nucleophosmin-Anaplastic Lymphoma Kinase (NPM-ALK) [24], an oncogenic fusion tyrosine kinase. The degenerate motif-building analysis was able to extract four novel motif classes (**Table 5**, first shaded region). These motifs represent candidate consensus sequences for NPM-ALK, a kinase which currently has no known consensus.

Figure 2e shows the sequence logo for the Exxy NPM-ALK motif. Inspection of this sequence logo indicated a clear C_2H_2 zinc finger domain consensus sequence, with the phosphorylation site falling between the second histidine (position –6) and first cysteine (position 2) of the domain[25]. Interestingly, there are 14 unique phosphorylated C_2H_2 zinc fingers identified by the data set, all of which contain an invariable glutamic acid at the –3 position despite the fact that this residue is not well conserved among all members of the C_2H_2 zinc finger family of proteins. Though not mentioned in the Rush *et al.*[6] study, this represents the first example of tyrosine phosphorylation in a zinc finger domain. The possibility exists that this phosphorylation event interferes with zinc coordination and may shed light on the poorly understood process by which zinc fingers associate and dissociate from their cognate DNA sequences.

The next data set we analyzed from this study contained tyrosine phosphorylated sequences from cells expressing constitutively active c-Src kinase. The optimal substrate sequence for c-Src has been determined by combinatorial library screening methods to be DEEIy[GE]EFF[26]. Comparison of the consensus to the motifs identified in this study yielded some striking similarities and differences (**Table 5**). The sequence logo for the c-Src motif yD (**Fig. 2f**), despite containing only 26 unique sequences, shares similar residue characteristics at nearly all positions with the *in vitro*–determined consensus, whereas the most significant motif identified, yS, bears only slight resemblance to the library-based c-Src motif (**Fig. 2g**).

The normal and degenerate motif analysis from the pervanadate-treated Jurkat cells in the third Rush *et al.*[6] data set also revealed several motifs, all of which are indicative of the known Src activation in these cells[27]. One such significant motif (**Fig. 2h**) contained a proline at the +3 position (accounting for approximately one-fifth of the data set), consistent with recent work that has indicated the ability of Src kinase to also phosphorylate YxxP motifs[28].

Box 1 Sensitivity to background data sets

One of the defining features of the presented algorithm relates to the use of a large background data set upon which dynamic statistical calculations could be performed to ensure motif relevance. Therefore, we sought to test the algorithm's sensitivity to various background data sets. To carry out this analysis we used four differing backgrounds on the serine phosphorylation data from the Beausoleil, Jedrychowski, *et al.* [8] study. Surprisingly, the algorithm was very robust, yielding very similar sets of motifs even when a randomized peptide data set was used as a background (see **Table 4**). In most cases, the differing motifs were simply analogs of motifs from the MS/MS-based analysis. For example, the more specific casein kinase motifs sDxExE and DsExE, were found in three out of four of the tested background data sets (**Supplementary Tables 7–10** online). The similarity of extracted motifs was likely due to our choice of using high significance and occurrence thresholds in the analyses. The background data set primarily acts as a filter for nonsignificant motifs, however, owing to the set reduction strategy that limits the total number of motifs that can be extracted, coupled with the fact that most motifs become more significant in the face of a random background, the overall character of the extracted motifs does not change dramatically. It should be noted that the background cannot act as a useful filter without a fairly large number of sequences because of the dynamic pruning of the background data set in the motif-building procedure (we suggest an order of magnitude more sequences in the background than in the data set for reasonable probability estimates).

Table 4 Similarity of extracted motifs using various background data sets

Background data set used	Number of motifs extracted	Exact motif overlap with full MS/MS-based serine centered data set
Half-sized MS/MS-based data set (13,990 sequences)	17	17 (100%)
Quarter-sized MS/MS-based data set (6,995 sequences)	21	12 (57%)
All human S centered peptides (1,279,892 sequences)	18	12 (67%)
Randomized MS/MS-based data set (28,008 sequences)	20	13 (65%)

Outlook

As proteomic sequence data sets grow ever larger, tools for the extraction of biologically relevant motifs will become even more useful. The algorithm presented here represents an attempt to extract biologically relevant motifs based on sequence information from large-scale mass spectrometry–based data sets, and is meant to serve as a starting point for future research. Using a statistical framework that does not assume independence of positions in the motif owing to a dynamic statistical background, a two-step methodology of motif building and set reduction is used to decompose a given data set into its constitutive motifs. The strategy taken is substantially different from previous work in the realm of protein motif discovery, and its validity has been demonstrated through direct comparison to other motif discovery algorithms. Furthermore, the approach's usefulness has been exemplified through its ability to extract known and novel motifs from several large-scale MS/MS-based phosphorylation data sets. Since the motifs and their cognate position weight matrices are generated from actual *in vivo* phosphorylation sites as opposed to synthetic peptide libraries, the method may lead to an improvement in phospho-site prediction. It should be noted however, that the derived position weight matrices, and not consensus sequences alone, should be used in the search for potential phosphorylation sites, as they contain information on residue frequencies surrounding the 'locked' sites and will help reduce false-positive rates.

We also envision the novel motifs to be used as peptide baits to identify corresponding uncharacterized kinases.

Given the success of the algorithm on a linguistic data set, it is apparent that the method has the versatility to extract motifs from a wide range of data sets including, but not limited to, other post-translational modifications and genomic data. Additionally, the described algorithm may lead to the discovery of novel biologically relevant protein motifs directly from a raw proteome.

METHODS

Phosphorylation data formatting. The Beausoleil, Jedrychowski, *et al.* [8] phosphoserine and phosphothreonine data set and the Rush *et al.*[6] phosphotyrosine data sets were used as starting points for the analysis. Only those tryptic peptides where the exact site of phosphorylation was known were selected. Peptides were mapped back to their prospective proteins and six residues upstream and downstream of the phosphorylation sites were reextracted from the proteome sequence. This step removed the nonuniformity of tryptic peptide fragments to produce a data set of known phosphorylation sites in a uniform 13-residue context. In cases where the phosphorylation site was within six residues of a protein terminus, the sequence was discarded. Thus, only those sites which had sequence information for 12 residues surrounding the phosphorylation site were used. The sequences were then filtered for redundancy, so that only unique sequences remained. This formatting procedure gave rise to 1,594 phosphoserine-centered sequences, 195 phosphothreonine-centered sequences and 573 phosphotyrosine-centered sequences.

Background data set creation. This analysis was dependent upon the ability to compare the distribution of residue frequencies in the phosphorylation data set with a background model. Background data sets were created in three ways. For the phosphoserine and phosphothreonine analyses, background data sets were created using fully-tryptic peptides generated from mass spectrometry data in the Gygi lab and searched using the human protein database with high Sequest[29] thresholds (Xcorr > 2.5, dCn > 0.1). Data were centered on nonphosphorylated serine or threonine residues and formatted according to the same procedure described in the previous section, thus yielding two similarly aligned background data sets containing 27,980 sequences centered on serine and 20,208 sequences centered on threonine. Owing to the lower abundance of tyrosine residues, a second type of background data set for the phosphotyrosine analysis was created by taking six residues upstream and downstream of every tyrosine residue in the human proteome. This resulted in a background data set centered on tyrosine, containing 441,343 sequences. It should be noted, however, that despite our intention of using a mass spectrometry–based background to avoid mass spectrometry–specific residue biases, performing the serine and threonine analyses with a proteomic background as opposed to a mass spectrometry–based background did not significantly alter the motifs extracted (**Supplementary Table 9** online). Finally, a third type of background data set with randomized residue positions for each of the serine-centered mass spectrometry peptides was created (see **Box 1**).

Degenerate residue positions. To allow for conservative amino acid substitutions at various positions, we condensed our phosphorylated peptide and background lists from a 20-amino-acid code to a degenerate 11-amino-acid code based on chemical properties as follows: A = AG, D = DE, F = FY, K = KR, I = ILMV, Q = QN, S = ST, C = C, H = H, P = P, W = W. The analysis was then carried out as described for the nondegenerate analysis (see shaded regions of **Tables 3** and **5**).

Significance analysis. The motif-building strategy was carried out by finding successive significant residue/position pairs. Though the significance threshold is a parameter of the algorithm, for all analyses in this paper, residue/position pairs were deemed significant if they had random probabilities less than 10^{-6} according to a binomially distributed model (equation (1) below),

$$P(m,c_{xj},p_{xj}) = \sum_{i=c_{xj}}^{m} \binom{m}{i} p_{xj}^{\,i}(1-p_{xj})^{m-i} \qquad (1)$$

where m equaled the number of sequences in the data set matrix, c_{xj} equaled the count of residue x at position j in the data set matrix, and p_{xj} equaled the

Table 5 Normal and degenerate analyses of NPM-ALK, c-Src and pervanadate-treated Jurkat cell phosphorylated tyrosine motifs from the Rush *et al.*[6] data set

Motif	Literature	Score*	Phospho data set (*Matches*/Size)		Background data set (*Matches*/Size)	
NPM-ALK motifs**						
......y..V...	Novel	*16.00*	*47*	237	*24,781*	441,343
...E..y......	Novel	*8.90*	*34*	180	*24,251*	416,562
......y[DE].[ILVM]...	Novel	*24.63*	*37*	237	*10,533*	441,343
......y..[ILVM]...	Novel	*10.69*	*80*	200	*84,036*	430,810
...[DE]..y......	Novel	*7.35*	*34*	120	*36,258*	346,774
......y....[FY].	Novel	*6.22*	*22*	86	*24,426*	310,516
c-Src motifs**						
......yS.....	Novel	*8.58*	*40*	185	*34,153*	441,343
......yD.....	Src consensus	*7.75*	*26*	145	*20,509*	407,190
...E..y......	Src consensus	*6.40*	*23*	119	*22,643*	386,681
...[DE]..y......	Src consensus	*12.76*	*56*	185	*46,522*	441,343
......y[DE].....	Src consensus	*7.37*	*34*	129	*38,047*	394,821
......y[ST].....	Novel	*6.00*	*33*	95	*52,558*	356,774
......y[AG].....	Src consensus	*6.27*	*26*	62	*46,997*	304,216
Jurkat cell line motifs**						
......y..P...	Src consensus	*8.75*	*29*	151	*23,459*	441,343
...D..y......	Src consensus	*6.61*	*21*	122	*19,539*	417,884
...E..y......	Src consensus	*6.75*	*22*	101	*24,466*	398,345
...[DE]..y......	Src consensus	*15.65*	*54*	151	*46,522*	441,343

* Score = Σ-log(P). ** $P < 10^{-6}$, *occurrences* ≥ 10.

fractional percentage of residue x at position j in the current background matrix. The result was calculated using the pbinom function in the Math::CDF PERL module. The function could not calculate probabilities below 10^{-16}. Since each recursive iteration of the algorithm chose the residue/position pair with the lowest binomial probability, if more than one pair had probabilities of 10^{-16}, then the pair with the greater frequency in the data set matrix was selected.

Motif scores. Despite the statistical significance of every motif extracted, heuristic scores for the motifs were calculated as the sum of the negative log of the binomial probabilities used to generate the motifs (equation (2) below),

$$Score\ (motif) = \sum -\log\ (P_{binomial}) \qquad (2)$$

Linguistic analysis. Using the analytical framework previously created by Bussemaker *et al.*[14], text from the first ten chapters of *Moby Dick* by Herman Melville[15] with random characters inserted between words was retrieved from http://www.physics.rockefeller.edu/siggia/projects/mobydick/. By taking a sliding 13-character window, the text was then transformed into a matrix of all 13-character strings, thus constituting the background data set. From this background data set, 26 subsets were created, each being centered on a different letter of the alphabet. Using the background data set and each of the subsets, the motif-building methodology (with $P < 10^{-6}$, and *occurrences* ≥ 10) was carried out 26 times, thus yielding motifs centered on every letter of the alphabet (**Supplementary Tables 1–4** online).

Comparison to other algorithms. To compare our algorithm to other motif discovery tools, we input our synthetically generated list of 300 proteins and the manually curated phosphorylation list containing 298 13-mers to four websites: Pratt at http://www.ebi.ac.uk/pratt/ with parameters C% = 2%, PL = 13, PN = 50, PX = 5, FN = 2, and FL = 1; TEIRESIAS at http://cbcsrv.watson.ibm.com/Tspd.html with option 'exact discovery' and parameters L = 2 or 3, W = 5 and K = 2; eMOTIF at http://motif.stanford.edu/emotif/emotif-maker.html with a 10% match threshold; Gibbs motif sampler at http://bayesweb.wadsworth.org/cgi-bin/gibbs.9.pl?data_type=protein with number of patterns = 5, max sites per sequences = 1, motif width = 5, estimated total sites = 40.

Public access to algorithm. Access to the algorithm will be available through a website currently under construction at, http://motif-X.med.harvard.edu/ which will allow users to input their sequence data and adjust the various algorithm parameters to retrieve motif results.

Programming and sequence logos. All programming and analysis was done using the PERL programming language on a Linux workstation (2.2 GHz microprocessor with 1.5 GB RAM). Sequence logos were generated online using Weblogo[21] at http://weblogo.berkeley.edu/

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The authors thank John Rush and Cell Signaling Technology for providing access to the tyrosine phosphorylation data sets prior to their publication. Additionally, D.S. wishes to thank Michael Chou for assistance with the *Moby Dick* analysis as well as numerous stimulating conversations regarding the algorithm and critical reading of the manuscript. This work was supported in part by National Institutes of Health grant HG03456 (S.P.G.).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Schlessinger, J. & Lemmon, M.A. SH2 and PTB domains in tyrosine kinase signaling. *Sci. STKE* **2003**, RE12 (2003).
2. Ang, X.L. & Wade Harper, J. SCF-mediated protein degradation and cell cycle control. *Oncogene* **24**, 2860–2870 (2005).
3. Pawson, T. & Scott, J.D. Protein phosphorylation in signaling—50 years and counting. *Trends Biochem. Sci.* **30**, 286–290 (2005).
4. Obenauer, J.C., Cantley, L.C. & Yaffe, M.B. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.* **31**, 3635–3641 (2003).
5. Manning, B.D. & Cantley, L.C. Hitting the target: emerging technologies in the search for kinase substrates. *Sci. STKE* **2002**, PE49 (2002).
6. Rush, J. *et al.* Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat. Biotechnol.* **23**, 94–101 (2005).
7. Ficarro, S.B. *et al.* Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **20**, 301–305 (2002).
8. Beausoleil, S.A. *et al.* Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl. Acad. Sci. USA* **101**, 12130–12135 (2004).
9. Collins, M.O. *et al.* Proteomic analysis of *in vivo* phosphorylated synaptic proteins. *J. Biol. Chem.* **280**, 5972–5982 (2005).
10. Ballif, B.A., Villen, J., Beausoleil, S.A., Schwartz, D. & Gygi, S.P. Phosphoproteomic analysis of the developing mouse brain. *Mol. Cell. Proteomics* **3**, 1093–1101 (2004).
11. Gruhler, A. *et al.* Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol. Cell. Proteomics* **4**, 310–327 (2005).
12. Nuhse, T.S., Stensballe, A., Jensen, O.N. & Peck, S.C. Phosphoproteomics of the *Arabidopsis* plasma membrane and a new phosphorylation site database. *Plant Cell* **16**, 2394–2405 (2004).
13. Loyet, K.M., Stults, J.T. & Arnott, D. Mass spectrometric contributions to the practice of phosphorylation site mapping through 2003: a literature review. *Mol. Cell. Proteomics* **4**, 235–245 (2005).
14. Bussemaker, H.J., Li, H. & Siggia, E.D. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. USA* **97**, 10096–10100 (2000).
15. Melville, H. *Moby-Dick, or, The whale* (Signet Classic, New York, 1998).
16. Diella, F. *et al.* Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* **5**, 79 (2004).
17. Rigoutsos, I. & Floratos, A. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics* **14**, 55–67 (1998).
18. Jonassen, I., Collins, J.F. & Higgins, D.G. Finding flexible patterns in unaligned protein sequences. *Protein Sci.* **4**, 1587–1595 (1995).
19. Thompson, W., Rouchka, E.C. & Lawrence, C.E. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.* **31**, 3580–3585 (2003).
20. Nevill-Manning, C.G., Wu, T.D. & Brutlag, D.L. Highly specific protein sequence motifs for genome analysis. *Proc. Natl. Acad. Sci. USA* **95**, 5865–5871 (1998).
21. Schneider, T.D. & Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
22. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
23. Boucher, L., Ouzounis, C.A., Enright, A.J. & Blencowe, B.J. A genome-wide survey of RS domain proteins. *RNA* **7**, 1693–1701 (2001).
24. Fujimoto, J. *et al.* Characterization of the transforming activity of p80, a hyperphosphorylated protein in a Ki-1 lymphoma cell line with chromosomal translocation t(2;5). *Proc. Natl. Acad. Sci. USA* **93**, 4181–4186 (1996).
25. Iuchi, S. Three classes of C2H2 zinc finger proteins. *Cell. Mol. Life Sci.* **58**, 625–635 (2001).
26. Songyang, Z. & Cantley, L.C. Recognition and specificity in protein tyrosine kinase-mediated signalling. *Trends Biochem. Sci.* **20**, 470–475 (1995).
27. Branch, D.R. & Mills, G.B. pp60c-src expression is induced by activation of normal human T lymphocytes. *J. Immunol.* **154**, 3678–3685 (1995).
28. Shin, N.Y. *et al.* Subsets of the major tyrosine phosphorylation sites in Crk-associated substrate (CAS) are sufficient to promote cell migration. *J. Biol. Chem.* **279**, 38331–38337 (2004).
29. Yates, J.R. III, Eng, J.K. & McCormack, A.L. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal. Chem.* **67**, 3202–3210 (1995).