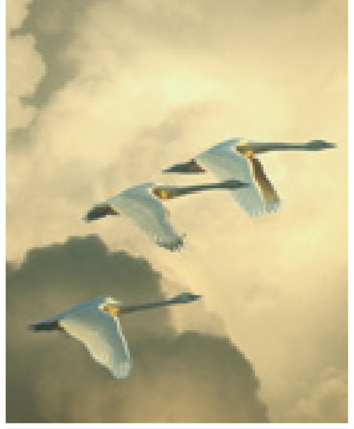


Utility of evolutionary conservation in ranking phosphorylation sites discovered by high-throughput methods.

Julie Nardone, Yi Wang, Bin Zhang, Jon Kornhauser, Elzbieta Skrzypek, Vicky Yang, David Merberg, Roberto Polakiewicz, Peter Hornbeck • Cell Signaling Technology, Danvers, MA

Abstract

Mass spectrometry allows the observation of thousands of phosphorylation sites on hundreds of proteins in a single experiment. Because of the large numbers, not all phosphorylation events can be evaluated for their biological importance; even the task of prioritizing sites for evaluation is prohibitively time-consuming and may be subject to bias. In the current work, we use the published phosphorylation sites collected in PhosphoSite® to assess the utility of using evolutionary conservation of the sequence surrounding the phosphorylation site as one measure of the biological relevance of a site.

Introduction

A common problem in bioinformatics is prioritizing results of a high-throughput (HTP) study for follow-up by wet-lab experiments. Desirable properties of such prioritizing systems include objectivity, basis in complete and reliable sources of information, the ability to be automated and correlation of an object's rank with its likelihood of passing additional screens.

Our interest is in ranking phosphorylation sites from mass spectrometry (MS) experiments for their likely biological importance. For this study, we evaluate the predictive value of evolutionary conservation. We assume that, in order to interact with a specific kinase and downstream regulatory molecules, phosphorylation sites are subjected to evolutionary constraints.

For our test sets, we use the phosphorylation sites catalogued in PhosphoSite® because it is a comprehensive, manually curated source and contains experimental method annotations in a controlled vocabulary. We divide the sites into two overlapping sets, those described in journal articles using HTP methods and those described in journal articles using low-throughput (LTP) methods. We reasoned that articles focussing on one or a few sites indicate that those sites are likely to be biologically important, justifying research funding, scientific labor and peer-reviewed publication. HTP articles, in contrast, are likely to contain sites of all levels of biological relevance, including spurious and redundant ones. By using the LTP sites as a model, we attempt to identify properties that can be used to rank the HTP sites.

Methods

We downloaded PubMed IDs, number of associated phosphorylation sites and curated experimental methods from PhosphoSite®. Articles were divided into HTP (> 25 associated sites) and LTP (≤ 25 associated sites). Manual inspection of the experimental methods showed that all of the papers with more than 25 sites used HTP methods. We collected a non-redundant set of sequences surrounding the sites (seq15, i.e. +/- 7 amino acids) for each group of papers, using human as the reference organism.

NCBI Entrez Gene IDs were determined for each protein and used to retrieve NCBI Homologene data. From this, the orthologous mouse gene ID and the gene ID of the most distant species from human in the Homologene set were used to retrieve all of the associated proteins in the two species.

Clustalw (Nucleic Acids Research 31: 3497) was used to align human, mouse and most distant species proteins. Seq15 was mapped onto the alignment, and the score for that segment was determined for human paired with each of the other proteins. The BLOSUM62 matrix was used for scoring human/vertebrate pairs, BLOSUM45 for human /invertebrate pairs, and BLOSUM40 for human/other pairs. The highest-scoring alignment within a species was reported.

In the protein with the best alignment, we also scored segments of equal length, 3 on each side of seq15, offset by half the length of seq15. These were used as a measure of the overall conservation level of the protein sequence around the phosphorylation site but not close enough to include an entire kinase binding site.

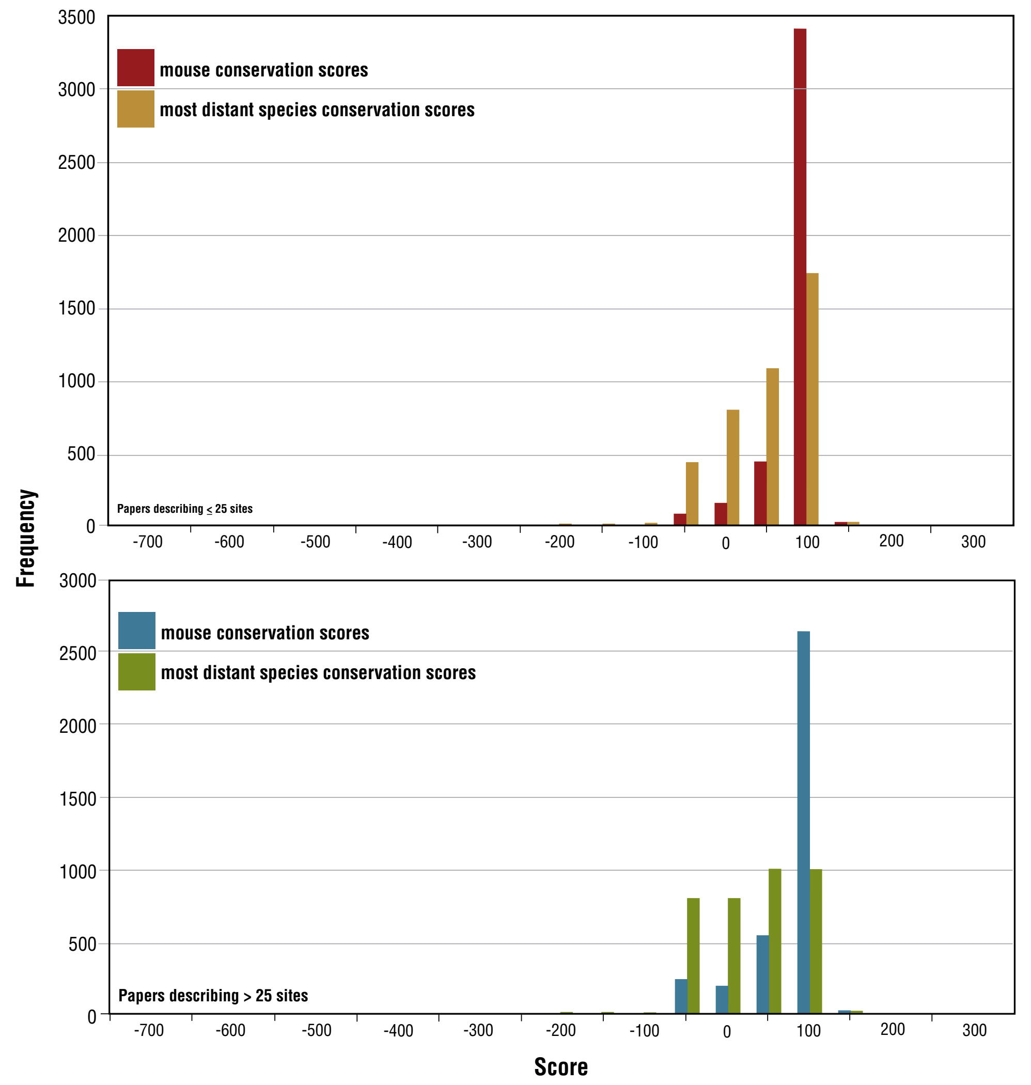
Sites were eliminated from the test set if they belonged to a protein reported to be conserved in a species more distant than mouse but not in mouse itself. These sites would have required a prohibitive investment of time to verify that the proteins were, in fact, absent in mouse.

Proteins were analyzed for enriched Gene Ontology (GO) categories using the Babelomics online software tools (Nucleic Acids Research 33: W460). Data retrieval, alignment with Clustalw, and alignment scoring were automated with custom Perl scripts. All other analyses were implemented with either Microsoft Excel or custom Perl scripts.

www.cellsignal.com

Frequency of Evolutionary Conservation Scores for Phosphorylated Peptides.

Conservation was measured for the alignment of the human seq15 with either mouse or the most distant species represented in a Homologene set. Higher scores imply greater conservation. Although the distribution of scores is slightly higher in the LTP set, the difference is not sufficient to support a ranking system solely based on these scores.



-800	0	0
-750	0	0
-700	0	2
-650	0	0
-600	0	0
-550	0	1
-500	0	0
-450	0	0
-400	0	0
-350	0	2
-300	0	2
-250	0	1
-200	0	5
-150	0	8
-100	0	14
-50	84	459
0	155	785
50	460	1117
100	3443	1750
150	30	31
200	2	0
250	0	0
300	3	0

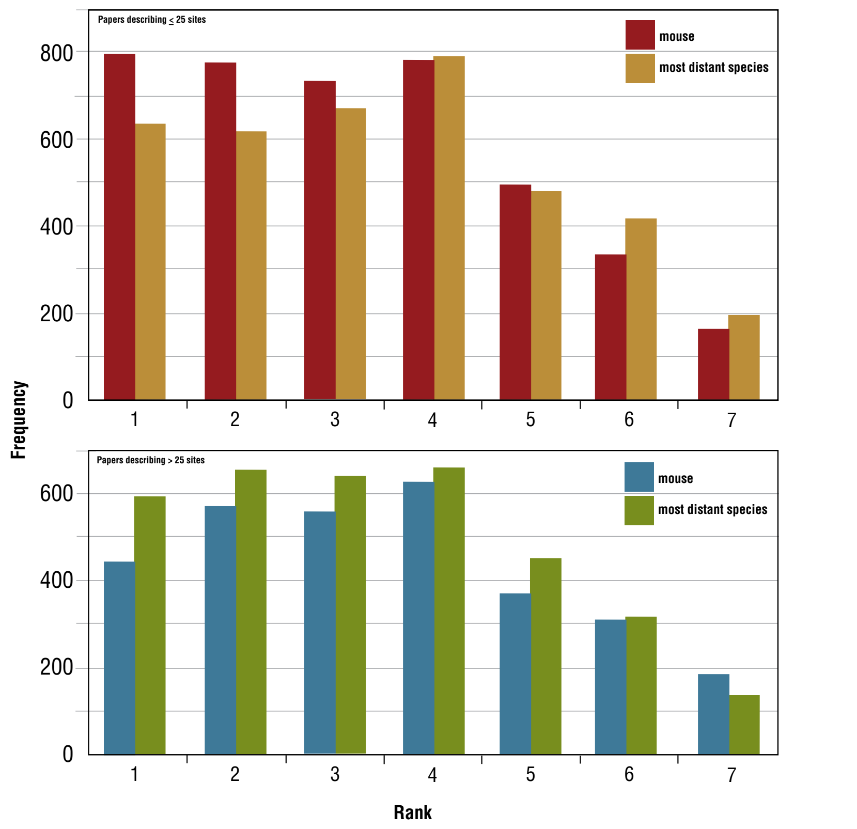
-700	0	0
-650	0	0
-600	0	0
-550	0	0
-500	0	0
-450	0	0
-400	0	0
-350	0	0
-300	0	0
-250	0	1
-200	0	6
-150	0	5
-100	1	7
-50	240	805
0	192	799
50	539	1006
100	2656	1005
150	22	19
200	3	0
250	0	0
300	0	0
More	0	0

	Papers describing ≤ 25 sites	Papers describing > 25 sites	Common
# papers	5053	22	
# phosphorylation sites	4443	4026	390
# proteins	1435	1687	366
mean # sites per paper	2.50	215	
median # sites per paper	2.00	66.5	
range of # sites per paper	1 - 25	26 - 1959	
% sites containing pS	58	69	
% sites containing pT	18	12	
% sites containing pY	22	17	

	Sites in papers describing ≤25 sites	Sites in papers describing >25 sites
arabidopsis	222	326
rice	367	605
S. pombe	15	10
S. cerevisiae	85	71
N. crassa	10	34
malaria parasite P. falciparum	17	34
Eremothecium gossypii	2	0
rice blast fungus	1	1
C. elegans	560	518
mosquito	382	270
drosophila	34	54
chicken	1641	1102
dog	711	486
rat	99	73
primates	30	59

For each protein, we determined the most distant species with an orthologue by using NCBI's Homologene. The range of species represented as well as the distribution of proteins among them was remarkably similar in LTP and HTP sets.

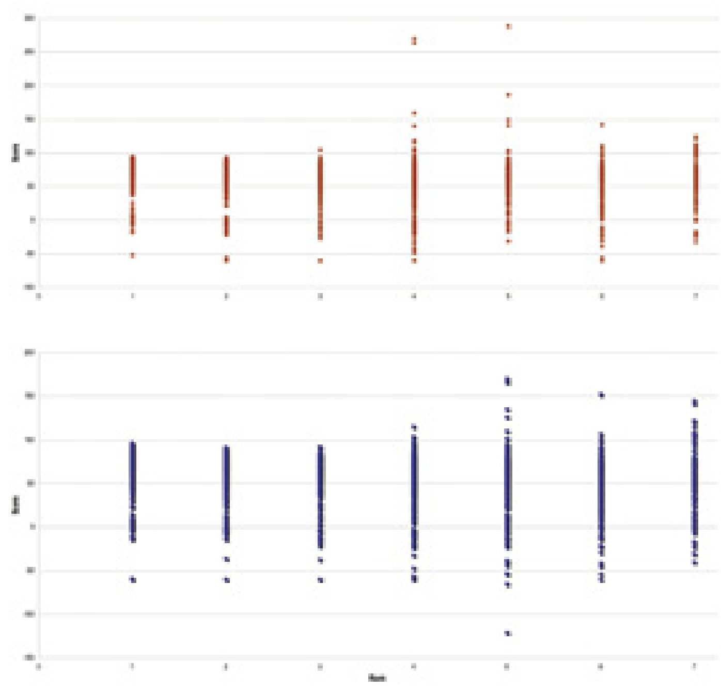
Rank of Phosphorylated Peptide's Conservation Score Compared to 6 Neighboring Peptides.



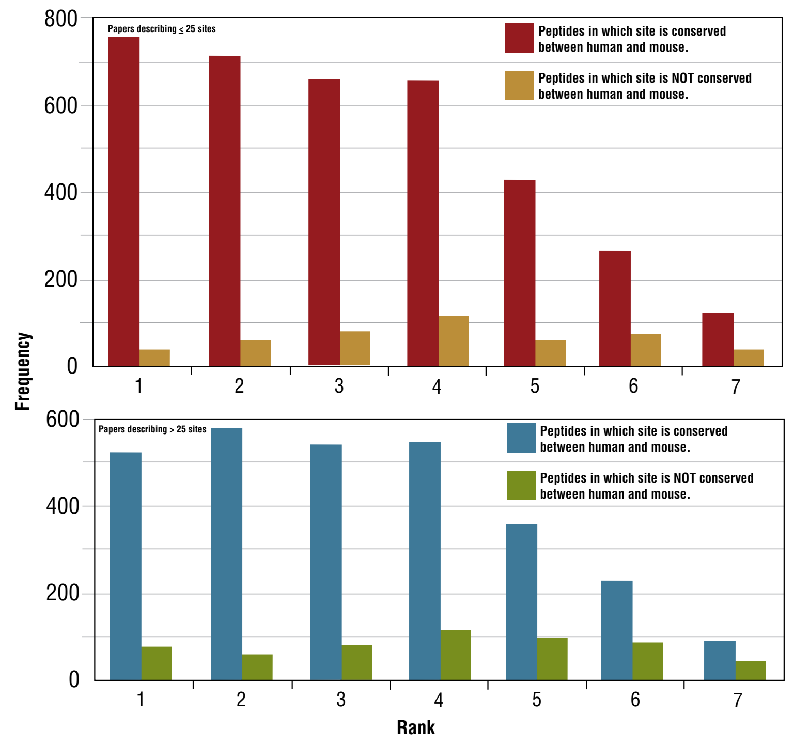
To assess the conservation of the sequence immediately surrounding the phosphorylation site relative to neighboring sequences, we determined the rank of the seq15's score compared to 6 other sequences nearby: *. A rank of 1 means that the seq15 had the highest conservation score of the 7 sequences. For both human/mouse and human/most distant species comparisons, most seq15s were at least as well conserved as were adjacent sequences.*

Score vs. Rank

We examined the relationship between conservation score and rank in order to put very high and low scoring seq15s in context. Low scoring seq15s occurred in every rank, suggesting that they carry no special significance. High scoring seq15s tended to occur in the middle or lower ranks, although the numbers are small, suggesting that it would be difficult to detect particular conservation of a phosphorylation site sequence in the context of a highly conserved domain.



Rank of Phosphorylated Peptide's Conservation Score Compared to 6 Neighboring Peptides.



Both LTP and HTP contain phosphorylation sites that are not conserved between human and mouse, although orthologous proteins exist in both species. We examined the effect on conservation of treating this class of sites separately. As we expected, the fraction of seq15s that were more conserved than their neighbors was higher in the set of proteins that had human/mouse site conservation than in the full set; however, the increase was not substantial.

Significantly enriched molecular function GO categories (level 4) from Fatigo.

Papers describing ≤ 25 sites	p-value	Papers describing > 25 sites	p-value
transmembrane receptor activity	<1e-05	RNA binding	<1e-05
transferase activity, transferring phosphorus-containing groups	<1e-05	ATP-dependent helicase activity	0.00012
α-type channel activity	<1e-05	DNA binding	0.00043
ion channel activity	<1e-05	hydrolase activity, acting on acid anhydrides	0.00105
cation transporter activity	0.00006	RNA helicase activity	0.01895
peptide receptor activity	0.00281	transferase activity, transferring one-carbon groups	0.03665
receptor signaling protein serine/threonine kinase activity	0.03154		
calmodulin binding	0.04809		

LTP and HTP sets do differ in the types of proteins that are represented. LTP papers are significantly enriched in transmembrane proteins and kinases. This may be because typical protein isolation protocols for MS do not efficiently extract membrane proteins. These classes of protein may also be in low abundance in the cell and, therefore, less likely to appear in MS analyses.

HTP experiments, in contrast, are enriched in RNA- and DNA-binding proteins. More than half the sites in the 22 papers were found in cancer cell lines, and the type of protein detected may reflect the abundance of proteins involved in cell division.

Conclusion:

Although phosphorylation sites are embedded in sequences that tend to be at least as conserved as neighboring segments, evolutionary conservation is unlikely to be useful as a practical predictor of biologically relevant phosphorylation sites. There is no clear correlation of the extent of conservation with biological relevance.

It is possible that the lack of correlation that we observe is the result of mistaken assumptions. In particular, if spurious or redundant sites are present in low abundance in the cell, then HTP articles may detect too few of these to distinguish their set from the LTP group, and, therefore, too few to make worthwhile a ranking system to eliminate them.

Another possibility is that the density of highly conserved residues is too low to make a robust scoring system based on a large sequence window. If this is the case, then motif detection methods may prove more successful than conservation scoring.