JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

pubs.acs.org/JACS

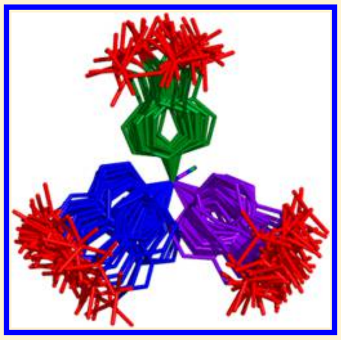
Conformational Clusters of Phosphorylated Tyrosine

Maha Abdelrasoul,[†] Komala Ponniah,[‡] Alice Mao,[§] Meghan S. Warden,[‡] Wessam Elhefnawy,[†] Yaohang Li,*[,†] and Steven M. Pascal*[,‡]

[†]Department of Computer Science, Old Dominion University, Norfolk, Virginia 23529, United States
[‡]Department of Chemistry and Biochemistry, Old Dominion University, Norfolk, Virginia 23529, United States
[§]Ocean Lake High School, Virginia Beach, Virginia 23454, United States

Supporting Information

ABSTRACT: Tyrosine phosphorylation plays an important role in many cellular and intercellular processes including signal transduction, subcellular localization, and regulation of enzymatic activity. In 1999, Blom et al., using the limited number of protein data bank (PDB) structures available at that time, reported that the side chain structures of phosphorylated tyrosine (pY) are partitioned into two conserved conformational clusters (Blom, N.; Gammeltoft, S.; Brunak, S. *J. Mol. Biol.* **1999**, *294*, 1351−1362). We have used the spectral clustering algorithm to cluster the increasingly growing number of protein structures with pY sites, and have found that the pY residues cluster into three distinct side chain conformations. Two of these pY conformational clusters associate strongly with a narrow range of tyrosine backbone conformation. The novel cluster also highly correlates with the identity of the *n* + 1 residue, and is strongly associated with a sequential pYpY conformation which places two adjacent pY side chains in a specific relative orientation. Further analysis shows that the three pY clusters are associated with distinct distributions of cognate protein kinases.



INTRODUCTION

Protein production is energetically expensive. For example, translation of a protein with 101 amino acids requires an expenditure of 400 ATP molecules to form 100 peptide bonds. To toggle protein activity solely by production on demand and degradation when no longer needed is not only inefficient, but also time-consuming. Nature has developed more efficient and reversible strategies for protein regulation, namely, post-translational modification (PTM). Common PTMs include acetylation, glycosylation, methylation, and ubiquitination.[1] But by far the most common PTM is phosphorylation.

While prokaryotes commonly phosphorylate a wide variety of residue types including serine, threonine, tyrosine, histidine, glutamic acid, and aspartic acid, eukaryotes principally phosphorylate the hydroxyl-containing residues serine, threonine, and tyrosine.[2−4] It is estimated that one-third of all eukaryotic proteins are targets for phosphorylation at some point.[5,6] Many of these are targeted at more than one site and by more than one kinase.[7,8] They can also be targeted by phosphatases to remove the phosphate when necessary.[5,9] Each phosphorylation event requires only one ATP molecule, an obvious energetic advantage over the translation/degradation cycle. Phosphorylation/dephosphorylation can also be achieved far more rapidly than protein translation/degradation.[10] For these reasons, phosphorylation/dephosphorylation is used by cells to efficiently regulate protein activity, interactions, localization, and stability.[11]

Tyrosine phosphorylation was first identified in studies of polyomavirus.[12] The same studies provided the first identification of a protein kinase, the now well-known Src protein that is also the first identified oncogene.[13−15] Apart from its well-known role in signal transduction due to these seminal studies, tyrosine phosphorylation is also involved in control of localization, enzymatic activity and other processes.[16]

Nearly 20 years ago, Blom et al.[17] identified two conserved conformational clusters of phosphorylated tyrosine (pY) side chains. These clusters were identified via analysis of the limited number of pY-containing structures available from the Protein Data Bank (PDB). In the past 20 years, the number of PDB entries has increased 20-fold. This explosion in structural biology creates many bioinformatics opportunities, including the present work: revisiting the structural analysis of pY residues. In doing so, we have identified a novel pY structural cluster. The novel cluster is highly correlated with the identity of the residue following the tyrosine, often forming a distinct two residue conformational motif. A comprehensive analysis of kinase recognition sites of all three pY clusters also produces a strong correlation between cognate kinase and cluster.

RESULTS

Identification and Structural Characterization of Three pY Clusters. The χ_1 and χ_2 angle values of pY residues in the PDB were extracted. In tyrosine (and pY) residues, χ_1 is the N−C$_\alpha$−C$_\beta$−C$_\gamma$ dihedral angle, while χ_2 is the C$_\alpha$−C$_\beta$−C$_\gamma$−C$_{\delta 1}$ dihedral angle. In unmodified tyrosine residues, χ_1 typically falls near one of the three rotamer values of 60°, 180°, and 300° (or −60°), while χ_2 is less restricted, but tends to cluster around

Received: September 28, 2017
Published: November 9, 2017

DOI: 10.1021/jacs.7b10367

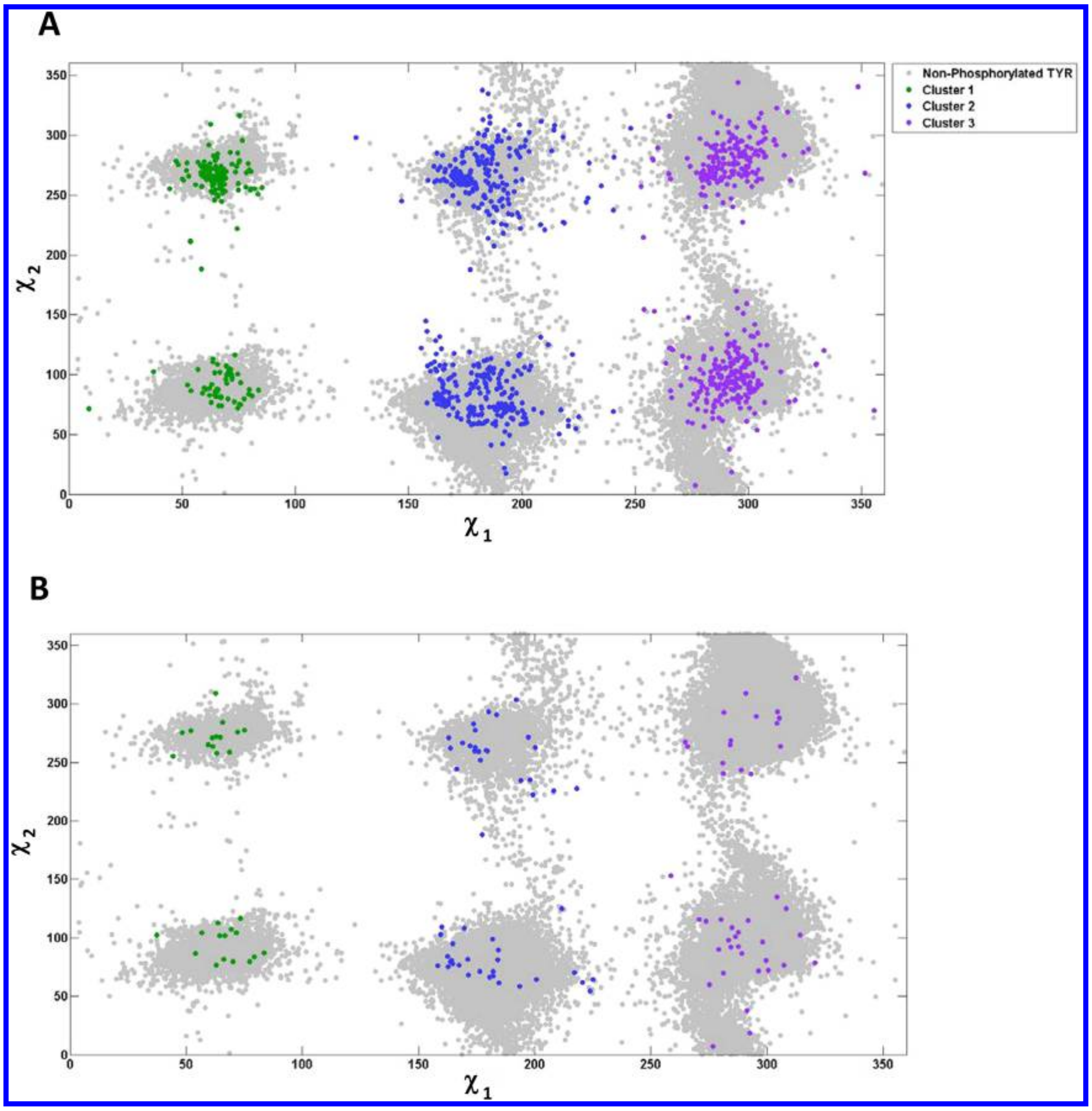


Figure 1. χ_1/χ_2 plots of Y and pY residues from the PDB. (A) All instances from the PDB. (B) Reduced set eliminating pY residues from redundant and short sequences as described in the text. Colors: nonphosphorylated Y (gray), pY cluster 1 (green), pY cluster 2 (blue), pY cluster 3 (violet).

the values of 90° and 270° (gray dots in Figure 1A). Blom et al. previously analyzed 210 pY sites in the PDB and identified two conformational clusters, characterized by χ_1 near 180° and near 300°, respectively. Here, we have analyzed 998 pY sites in 467 PDB structures. Results confirm the existence of the two previously identified clusters (blue and violet dots in Figure 1A). However, we have found a third structural cluster, characterized by χ_1 near 60° (green dots in Figure 1A). Ordering the three clusters by increasing χ_1 values, we will refer to the novel cluster as CLUSTER 1. The numbers of instances of CLUSTER 1, 2, and 3 found were 189, 411, and 398, respectively.

In each pY cluster, χ_2 values gather near 90° and 270° (Figure 1A), though the novel CLUSTER 1 appears to have less spread in χ_2 values than do clusters 2 and 3. This reflects the relative degree of spread seen in unmodified Tyrosine χ_2 values (gray dots in Figure 1A). While the appearance of Figure 1 may suggest the presence of six clusters, note that a change of χ_2 from 90° to 270° amounts to a 180° "flip" of the symmetric aromatic ring, merely exchanging $C_{\delta 1}$ with $C_{\delta 2}$: it does not produce a distinct conformation. Thus, Figure 1 appears nearly symmetric about the χ_2 = 180° line. Any nonsymmetry is due to the statistics of a small sample size.

We next reduced the number of pY sites from 998 down to 124 by eliminating repeats within a single PDB entry, PDB entries with sequences surrounding the pY that are identical to another entry, and PDB entries with short sequences surrounding the pY (see Methods). From this reduced set of 124 pY sites, we identified 29, 52, and 43 distinct sites with pY conformations falling into CLUSTER 1, 2, and 3, respectively. The χ_1 and χ_2 angles of the reduced set are plotted in Figure 1B. The most notable difference in comparing Figure 1A and 1B is that most CLUSTER 1 outliers, green dots with χ_2 near 200 degrees, are removed in the reduced set, increasing the correlation of the novel CLUSTER 1 with a narrow range of χ_1 and χ_2 angles. The correspondence of the three clusters to three χ_1 rotamer conformations can be seen clearly in Figure 2, where the novel cluster is colored green.

Figure 3A shows ϕ/ψ plots (Ramachandran plots) of all unmodified tyrosine residues (gray dots) and pY residues

DOI: 10.1021/jacs.7b10367

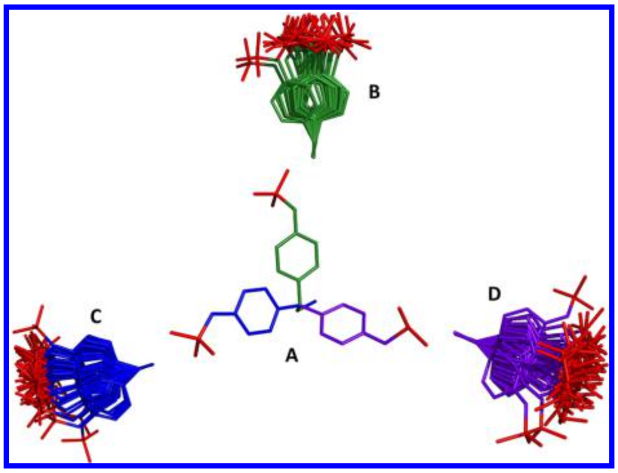


Figure 2. Superimposition of reduced set of pY residues from the PDB. Backbone N, C_α, and C_β atoms were superimposed, with the $C_\alpha-C_\beta$ bond pointing out of the paper. (A) Superimposition of a representative member of each of the three clusters, showing the correspondence between the three clusters and the three χ_1 rotamers; (B) side chains of cluster 1, $\chi_1 \approx 60°$; (C) side chains of cluster 2, $\chi_1 \approx 180°$; (D) side chains of cluster 3, $\chi_1 \approx 300°$. The phosphate groups are colored red, and clusters 1, 2, and 3 are colored as in Figure 1.

(colored dots) from the PDB. While the previously identified clusters 2 and 3 are distributed both in the canonical β sheet and α helix range, CLUSTER 1 pY residues (green dots) fall nearly exclusively at the top left-hand corner of the β sheet range, an area correlated with antiparallel β sheet formation. The distribution of CLUSTER 1 ϕ/ψ angles is even more striking when using the reduced set of 124 pY sites (Figure 3B), which again removes most CLUSTER 1 outliers, strengthening the correlation between CLUSTER 1 and the antiparallel β sheet region. Note that the CLUSTER 2 sites (blue dots) also have a much more restricted range in the reduced set (Figure 3B), suggesting that pY CLUSTER 2 is correlated with parallel β sheet conformations, while CLUSTER 3 appears to be present in both β sheet and helix-associated regions.

Analysis of the Novel Cluster: The p_1Yp_2Y Motif. We next sought to understand how CLUSTER 1 had evaded previous detection. As mentioned previously, the number of proteins extracted from the PDB with pY in CLUSTER 1 is smaller than the other two clusters. More importantly, all PDB entries in the reduced pY CLUSTER 1 were deposited into the PDB post 1999: maturation of the PDB was required for recognition of CLUSTER 1.

It also became apparent that CLUSTER 1 pY residues often occur in the first of two consecutive tyrosine residues: for CLUSTER 1 pY sites, the $n + 1$ residue is also a tyrosine in 86% of the cases (25 of 29 CLUSTER 1 pY sites). We will call this the tandem YY sequence.

We observed two types of modified tandem YY sequence, pYY and pYpY, i.e. only the first, or both Y are phosphorylated, respectively. We did not find any YpY sites in our search. Of the 25 modified tandem YY sites that we found, 23 were pYpY sequences and two were pYY. For pYpY sequences, the first pY and the second pY were in CLUSTER 1 and CLUSTER 2, respectively, in 22 of the 23 instances. This p_1Yp_2Y motif, where the numbers denotes the cluster, is therefore dominantly representative of the novel cluster.

Interestingly, although the backbone torsion angles (Figure 1) suggest that the p_1Yp_2Y motif may occur in a beta strand, closer inspection (see Figure 4) reveals the positions of the two pY side chains to be orthogonal to one another. This relative orientation of sequential pY side chains is therefore a nearly ubiquitous feature of the pYpY motif. This raises the possibility that orthogonality may facilitate simultaneous recognition of the two pY residues by separate binding partners, or perhaps enable a switch between recognition by two serial partners.

Further inspection revealed that each of the p_1Yp_2Y motifs in our reduced set was found in either a protein kinase or in the kinase domain of a growth factor receptor (see Table 1). Several instances each of Jak1, Jak2, and Jak3 kinase are present in Table 1. These PDB entries have survived the reduction process because each has a different degree/position of apparent disorder, indicated by "x" notation in the sequence (see Table 1 legend). Disorder was common C-terminal to the p_1Yp_2Y motif (found in Jak2, Jak3, Syk and HGFR), but was found N-terminal to the motif only in Jak1 kinase. Positionally, p_1Yp_2Y the motif is located in a similar site in each PDB entry: in a long loop that is close in space to a short loop connecting two alpha helices (see Figure 5). In most cases, the alignment of the two loops is consistent with the formation of a very short two-stranded antiparallel β sheet as seen in Figure 5. In other cases, the two loops are in contact, but do not orient properly for β sheet formation. However, in every case, regardless of the presence or absence of a short antiparallel β sheet, and regardless of the degree/position of disorder, the orthogonal conformation of the p_1Yp_2Y motif is conserved.

Perhaps not surprisingly, the conformation of the p_1Yp_2Y motif appears to be stabilized through interactions with basic residues. Taking Jak2 as an example (see Figure 5), the hydrophobic part of the p_1Y residue is surrounded by the hydrophobic regions of three lysine residues, while the negatively charged phosphate group is surrounded by the three positively charged lysine amino groups. The p_2Y position also appears to be stabilized by a nearby lysine residue. This level of coordination certainly plays a role in stabilizing a single conformation of the p_1Yp_2Y motif, although the motif resides in a loop region.

The last line of Table 1 shows the consensus features of the sequence surrounding the p_1Yp_2Y motif. The most highly conserved regions are not immediately adjacent to the motif: a highly hydrophobic 10-amino acid region ending five residues N-terminal to the p_1Yp_2Y motif, and a PΦXWXAP sequence beginning nine residues C-terminal to the motif (see Table 1 legend for symbol definitions). The first of these conserved regions contributes the lysine that appears to stabilize the p_2Y position in Figure 5. The three basic residues stabilizing the p_1Y position are two residues away from the p_1Y in each direction (lysines at the −2 and +1 positions) and a lysine at the +22 position.

Other p_1Y and Other pYpY Sites. Excluding the 22 p_1Yp_2Y motifs discussed above, there were six additional instances of CLUSTER 1 pY residues in the reduced set. Four were found in phosphopeptides that are recognized by SH2 domains, where the SH stands for Src Homology. Additional short SH2-interacting phosphopeptides, that may fall into the p_1Y class, would have been removed by our search criteria that required the presence of at least seven amino acids on either side of the pY residue.

The two remaining CLUSTER 1 occurrences were in the two monophosphorylated pYY sites mentioned above. These will henceforth be referred to as p_1YY sites. Thus, whether singly or doubly phosphorylated, the first Y in the tandem YY

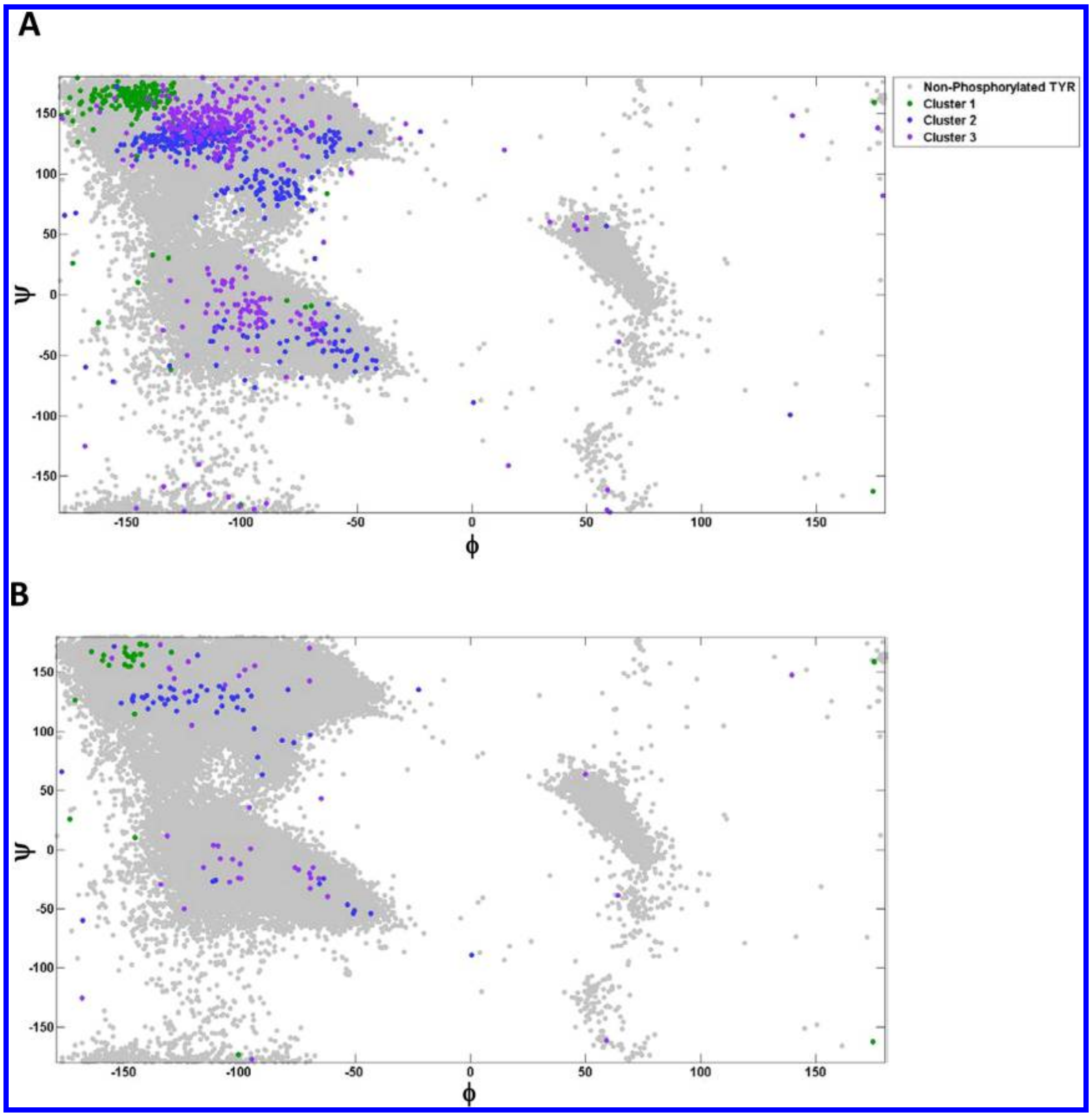


Figure 3. φ/ψ plots of Y and pY residues from the PDB. (A) All instances from the PDB. (B) Reduced set eliminating pY residues from redundant and short sequences as described in the text. Coloring is as in Figure 1.

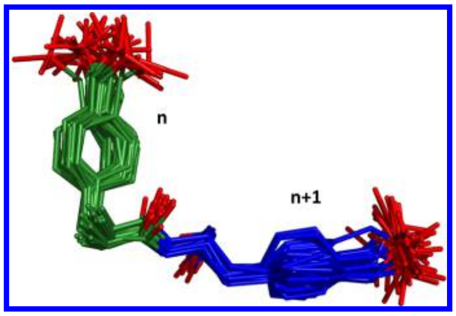


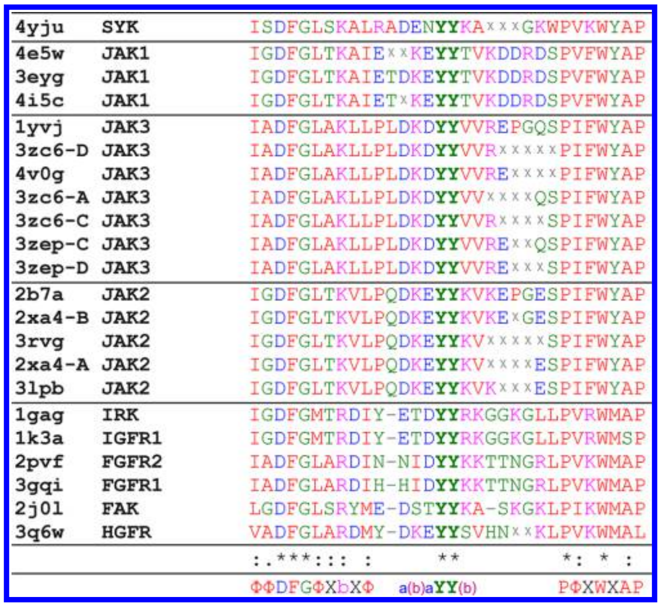
Figure 4. All-atom superimposition of p_1Yp_2Y motifs. p_1Yp_2Y motifs, wherein two sequential Y (labeled n and n + 1 in the Figure) are each phosphorylated. Colors: pY cluster 1 (green), pY cluster 2 (blue), phosphorus and oxygen atoms (red).

sequence was always found in CLUSTER 1. Interestingly, one of the two p_1YY sites was in SYK, which was also found as a p_1Yp_2Y site. The other p_1YY site was in TYK2. The TYK2 sequence differs from the other sequences in Table 1 in that the six residues immediately N-terminal to the YY motif are VPEGHE, where the underlined residues represent divergence from all other sequences in Table 1. TYK2 also diverges in its cognate kinases: GPS 3.0 predicts that the second Y in the TYK2 YY motif can be phosphorylated by Axl, which had been predicted only to phosphorylate CLUSTER 2 and CLUSTER 3 sites (see Figure 6A). Care must be taken in drawing conclusions from analysis of a single PDB entry, but it seems possible that the six residues N-terminal to YY may contain a determinant influencing single vs double phosphorylation of the YY motif.

The one pYpY sequence that did not form a p_1Yp_2Y motif instead forms a p_3Yp_3Y motif, in which each pY was in CLUSTER 3. This p_3Yp_3Y motif occurs in the insulin receptor kinase (IRK, PDB entry 2Z8C), and in a similar location as are the p_1Yp_2Y motifs. In fact, one of the two p_1Yp_2Y motifs in the reduced set that did not form a short antiparallel β sheet was also a crystal structure of IRK (PDB entry 1GAG) confirming that the IRK pYpY site is structurally divergent from the other kinase domains that contain tandem pYpY motifs.

Analysis of Cognate Kinases. Next, we sought a correlation between each cluster and their cognate kinases.

Table 1. Alignment of Sequences Surrounding the p_1Yp_2Y Motif[a]



4yju	SYK	ISDFGLSKALRADEN**YY**KAxxxGKWPVKWYAP
4e5w	JAK1	IGDFGLTKAIExxKE**YY**TVKDDRDSPVFWYAP
3eyg	JAK1	IGDFGLTKAIETDKE**YY**TVKDDRDSPVFWYAP
4i5c	JAK1	IGDFGLTKAIETxKE**YY**TVKDDRDSPVFWYAP
1yvj	JAK3	IADFGLAKLLPLDKD**YY**VVREPGQSPIFWYAP
3zc6-D	JAK3	IADFGLAKLLPLDKD**YY**VVRxxxxxPIFWYAP
4v0g	JAK3	IADFGLAKLLPLDKD**YY**VVRExxxxPIFWYAP
3zc6-A	JAK3	IADFGLAKLLPLDKD**YY**VVxxxxQSPIFWYAP
3zc6-C	JAK3	IADFGLAKLLPLDKD**YY**VVRxxxxSPIFWYAP
3zep-C	JAK3	IADFGLAKLLPLDKD**YY**VVRExxQSPIFWYAP
3zep-D	JAK3	IADFGLAKLLPLDKD**YY**VVRExxxSPIFWYAP
2b7a	JAK2	IGDFGLTKVLPQDKE**YY**KVKEPGESPIFWYAP
2xa4-B	JAK2	IGDFGLTKVLPQDKE**YY**KVKExGESPIFWYAP
3rvg	JAK2	IGDFGLTKVLPQDKE**YY**KVxxxxxSPIFWYAP
2xa4-A	JAK2	IGDFGLTKVLPQDKE**YY**KVxxxxESPIFWYAP
3lpb	JAK2	IGDFGLTKVLPQDKE**YY**KVKxxxESPIFWYAP
1gag	IRK	IGDFGMTRDIY-ETD**YY**RKGGKGLLPVRWMAP
1k3a	IGFR1	IGDFGMTRDIY-ETD**YY**RKGGKGLLPVRWMSP
2pvf	FGFR2	IADFGLARDIN-NID**YY**KKTTNGRLPVKWMAP
3gqi	FGFR1	IADFGLARDIH-HID**YY**KKTTNGRLPVKWMAP
2j01	FAK	LGDFGLSRYME-DST**YY**KA-SKGKLPIKWMAP
3q6w	HGFR	VADFGLARDMY-DKE**YY**SVHNxxKLPVKWMAL
		:.***::: : ** *: * :
		ΦΦDFGΦXbXΦ a(b)aYY(b) PΦXWXAP

[a] The p_1Yp_2Y motif is shown as bold YY in the center of the sequences. A dash in the sequence indicates a gap introduced for alignment purposes. An χ in the sequence indicates the residue is present in the protein, but absent in the crystal structure. Amino acids are colored by type: hydrophobic (red), acidic (blue), basic (pink), polar (green). A summary of the consensus regions of the sequence is at the bottom, with the symbols Φ (hydrophobic), a (acidic), b (basic), X (not conserved). A b in parentheses means "often basic". The symbols in the second to last line have the following meaning: * (conserved), : (highly similar), . (somewhat similar). Alignment initially performed with Clustal Omega and adjusted manually to properly align the χ regions.

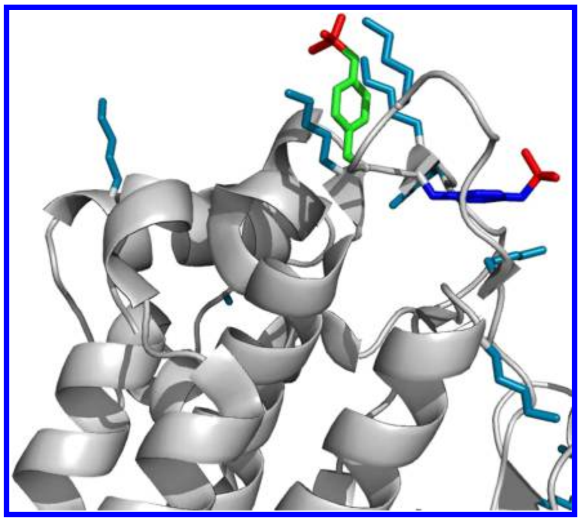


Figure 5. Positioning of p_1Yp_2Y motifs. Cartoon depiction of a representative structure (Jak2 kinase from PDB entry 2B7A[20]) containing the p_1Yp_2Y motif. Colors: p_1Y side chain (green), p_2Y side chain (blue), phosphate groups (red), lysine side chains (teal).

We used a 31 amino acid sequence fragment centered on the pY residue (32 amino acids in the case of sequential pYpY) to predict the identities of kinases that may phosphorylate the site(s). Figure 6A presents the normalized results (see Methods), with cognate kinases grouped into families. The figure is organized so as to present kinases most commonly predicted to phosphorylate CLUSTER 1, 2, and 3 sites on the left-hand side, center and right-hand side of the chart, respectively.

The tendencies charted in Figure 6A can be useful to scientists studying phosphorylated proteins. First, if the phosphorylating kinase is known, the information in Figure 6A can be used to predict the pY conformation. For instance, if a substrate is known to be phosphorylated by the Ack or Ret kinase, a CLUSTER 2 pY conformation will likely result. Clearly, some kinases, such as Met, are less discriminating in the type of pY conformation that they create. Second, if the cluster conformation is known, Figure 6A can be used to predict the kinases most likely involved. For instance, if a CLUSTER 1 site is detected, then the kinases on the left side of the chart are the best candidates for creating that site.

The kinase distributions were further analyzed via linear discriminant analysis (LDA) which provides a precise means to distinguish between clusters. Figure 6B shows, for each of the kinases predicted to phosphorylate each cluster member, a vector representing the potential for that kinase to distinguish the cluster from the other two clusters. For instance, relatively large green bar above Syk indicates that Syk is useful to distinguish CLUSTER 1, while the large blue bar above Ack shows that Ack helps to distinguish CLUSTER 2. Other kinases, such as Src, are not useful for distinguishing between clusters, and therefore have short bars in Figure 6B. This information was used to create the two-dimensional LDA plot in Figure 7, that shows a clear discrimination between clusters, with only a few outliers. Note the number of dots in Figure 7 appears to be less than 24. However, this is due to overlap of identically colored dots (two PDBs from the same cluster) that also have identical kinase distributions. Such an overlap is a clear indication of the compactness of the clusters.

CLUSTER 1 Feedback Loop. The fact that each of the p_1Yp_2Y motifs in the reduced set was found in a kinase prompted us to investigate whether kinases containing p_1Yp_2Y motifs are capable of phosphorylating YY sequences in other proteins, creating additional p_1Yp_2Y motifs. Analysis shows that all but one of the kinases containing a p_1Yp_2Y motif is found on the left-hand side of Figure 6A, the area reserved for kinases that create mostly CLUSTER 1 sites (green bars). The relationship is therefore nepotic: a given group of kinases both contains and creates p_1Y sites, most of these being p_1Yp_2Y motifs.

■ CONCLUSION

In conclusion, we have identified a novel pY conformational cluster, characterized by χ_1 values near 60°. This conformation most commonly occurs in the tandem p_1Yp_2Y motif, in which both Y are phosphorylated and fall into CLUSTER 1 and CLUSTER 2, respectively. It also occurs in peptides recognized by SH2 domains and in singly phosphorylated YY sequences. While the backbone dihedral angles of CLUSTER 1 and CLUSTER 2 residues in the p_1Yp_2Y motif are characteristic of antiparallel and parallel beta sheets, respectively, the p_1Yp_2Y motif was found exclusively in loop regions that in most cases participate in formation of a short two-stranded antiparallel β sheet. The conformation of the p_1Yp_2Y motif places the two pY side chains orthogonal to one another, which is consequential for interactions with binding partners. We have also correlated pY clusters with cognate kinases. Furthermore, examination of Figure 1, which includes all Y and pY residues from the PDB,

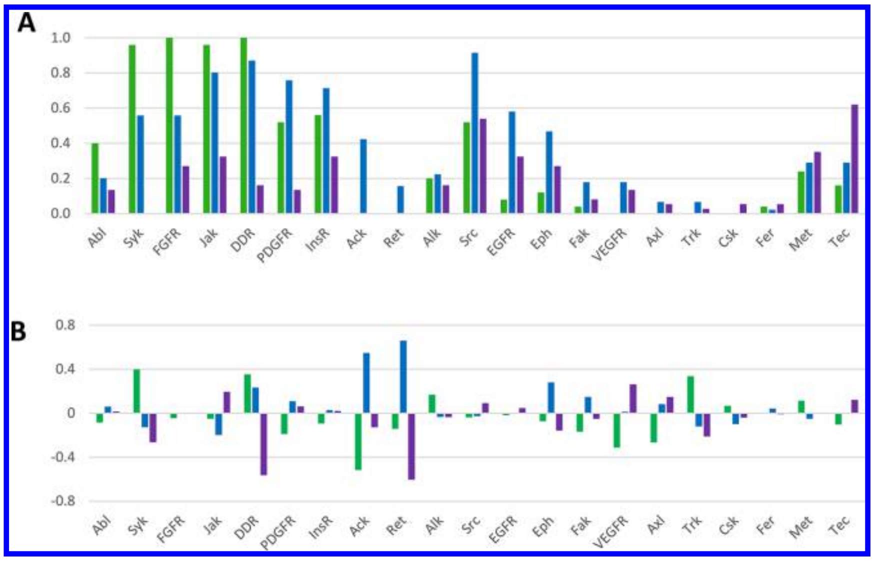


Figure 6. Cognate kinase histograms. The amino acid sequence surrounding the pY residues in the reduced set of PDB entries was used to predict the kinases most closely associated with the phosphorylation event. (A) Number of times that each kinase was identified for each cluster, normalized as described in Methods. (B) Linear discriminant analysis (LDA) of the kinase distributions. The magnitude of the bars represent the potential for that kinase to distinguish the cluster from the other two clusters. Positive bars represent positive correlation, while negative bars represent anticorrelation (kinase not likely to associate with that cluster). Cluster coloring is as in Figure 1.

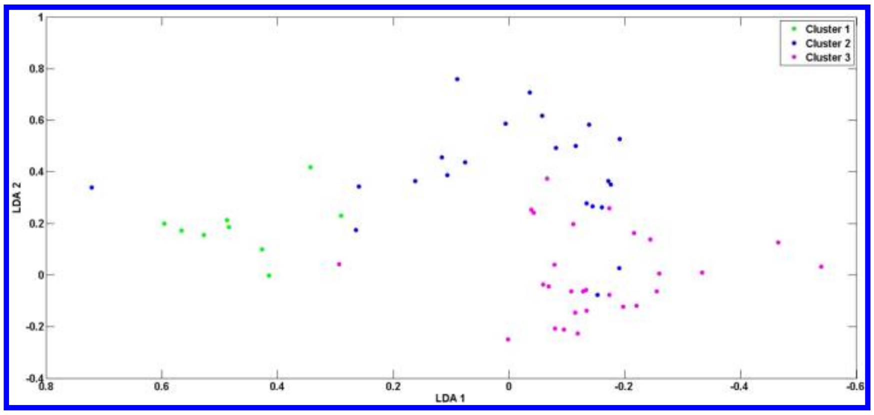


Figure 7. Two dimensional reduction of the kinase-based LDA analysis in Figure 6B. Cluster coloring is as in Figure 1. The tendency for members of each cluster to group into distinct areas of the graph indicates that the three clusters can be separated via LDA analysis of cognate kinases.

suggests that all significantly populated clusters of pY side chain conformations have now been identified.

METHODS

Identification of pY Sites in the PDB. The Protein Data Bank was screened for all experimentally determined phosphorylated tyrosine (pY) sites (PDB annotations of PTR). A total of 467 PDB entries with 998 pY sites were extracted.

Conformational Clustering Algorithm. We applied the normalized cuts algorithm,[18] a kind of spectral clustering algorithm popularly used for image segmentation,[19] to identify structural clusters of the pY conformations. Representing the data points as a weighted undirected complete graph $G = (V, E, w)$, the fundamental idea of the normalized cuts algorithm is to find a cut $(S,\overline{S})$ in G, where partitions $S,\overline{S} \subseteq V, S + \overline{S} = V$, and $S \cap \overline{S} = \emptyset$, in order to minimize

$$ncut(S, \overline{S}) = \frac{w(S, \overline{S})}{w(S, V)} + \frac{w(S, \overline{S})}{w(\overline{S}, V)}$$

where $w(S,\overline{S})$ is the weight function summing weights between two partitions and $ncut(S,\overline{S})$ measures the similarity between S and $\overline{S}$. Compared to commonly used clustering algorithms such as k-means, the graph-cut based clustering algorithms such as normalized cuts have the advantages of generating stable clusters with nonconvex boundaries, achieving a theoretical optimum, and providing an efficient way to estimate the number of clusters.

First, the pY conformations are superimposed by their backbone N$-$C$_\alpha$$-$C atoms. Then, we calculate the root-mean-square (RMS) distance of the side chain atoms (C_β, C_γ, $C_{\delta1}$, $C_{\delta2}$, $C_{\varepsilon1}$, $C_{\varepsilon2}$, C_ζ, O) and the phosphorus atom in each conformation pair. The pairwise RMS distance d_{ij} is used to construct the entry of the weight matrix W by defining

$$W_{ij} = \begin{cases} e^{-d_{ij}/\sigma^2} & i \neq j \\ 0 & i = j \end{cases}$$

where σ is a customizable parameter and here we use the standard deviation of the RMS distances to calculate σ. Afterward, a diagonal matrix D is defined where

$$D_{ii} = \sum_{j=1}^{n} W_{ij}$$

Theoretical analysis shows that minimizing $ncut(S,\overline{S})$ is equivalent to

$$\min_x \frac{x^T(D - W)x}{x^T Dx}$$

subject to the constraints $x_i \in \{1, -b\}$ for some constant b and $x^T D\mathbf{1} = 0$. By relaxing the constraints on x, the normalized cuts problem can be solved by finding the eigenvector corresponding to the second smallest eigenvalue of the normalized Laplacian matrix L where

$$L = D^{-1/2}(D - W)D^{-1/2}$$

The second smallest eigenvector of L is used to bipartition the graph G according to the eigenvector sign. The above bipartitioning process is repeated until all pairwise RMS distances in all subgraphs are within a 1.0 Å clustering cutoff.

Formation of a Reduced Set of pY Sites. The amino acid sequences surrounding the extracted 998 pY sites in 467 PDB entries were analyzed. If the pY site was not sequential with a second Y residue, the identity of the prior and subsequent 15 amino acids were extracted to create a series of 31-mer amino acid sequences with pY at position 16. If the pY site was part of a tandem YY sequence, a 32-mer amino acid sequence was extracted with the two tyrosines at position 16 and 17. Redundant sequences, in which all 31 or 32 amino acids matched another entry, were removed, as were short sequences which did not have at least seven residues preceding and seven residues following the pY residue. The remaining reduced set contained 124 distinct pY sites.

Cognate Kinase Identification. The reduced set of 124 sequences was modified by replacing all pY with Y and was then analyzed via GPS 3.0, a kinase-specific phosphorylation site prediction software.[20,21] The 31 and 32 residue sequences were entered into GPS 3.0 in batches in FASTA format, and the threshold was set at high. The output was a series of potential kinases for each sequence with a score and cutoff for each kinase. The score was then divided by cutoff to create the parameter "ratio". For each sequence, the kinase with the top ratio was retained, along with all other kinases with a ratio ≥ 1.5 (personal communication, Yongbo Wang, Yu Xue). All other kinases were discarded. The kinase family corresponding to each remaining kinase was then taken as the result and used to generate the kinase association histogram and for the LDA analysis discussed below.

Normalization of Kinase Association. To emphasize the probability of association between a kinase and a pY conformational cluster, the number of times that a kinase was predicted in the above analysis was divided by a factor of the number of entries in the associated reduced cluster (29 or 53 or 42). This adjustment prevents the under-emphasis of kinases associated with the least represented CLUSTER 1. All values were then normalized to a maximum value of one.

LDA Analysis. Linear discriminant analysis (LDA)[22,23] was used to examine the likelihood of discriminating between each cluster using cognate kinase information. First, the kinases predicted to phosphorylate each pY site (Figure 6A) were used to build a feature vector where each entry specifies a certain kinase, 1 for predicted and 0 for not predicted. Then, for each cluster, we applied LDA to find the LDA vector that maximizes the separation of this cluster from the other two. The resulting LDA vectors (Figure 6B) form the feature subspace revealing the kinases which are statistically different in the three clusters. The magnitude of a component in an LDA vector is commensurate with the contribution of the corresponding kinase in separating the cluster. Therefore, the LDA vectors shown in Figure 6B identify the most discriminating kinases that can be used to distinguish the clusters.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/jacs.7b10367.

Three dimensional plots indicating the number of PDB entries with pY dihedral angles in each 5° by 5° region of Figures 1 and 3 (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*yaohang@cs.odu.edu
*spascal@odu.edu

ORCID

Steven M. Pascal: 0000-0002-9492-6167

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation under Grant No. 1066471 and the Natural Science Foundation of China under Grant No. 61728211. We wish to thank Adam Boudion and Erich O'Saben.

■ REFERENCES

(1) Prabakaran, S.; Lippens, G.; Steen, H.; Gunawardena, J. *Wiley Interdis. Rev.-Sys. Bio. Med.* **2012**, *4*, 565−583.
(2) Hunter, T. *Philos. Trans. R. Soc., B* **1998**, *353*, 583−605.
(3) Krebs, E. G. *Philos. Trans. R. Soc., B* **1983**, *302*, 3−11.
(4) Yan, J. X.; Packer, N. H.; Gooley, A. A.; Williams, K. L. *J. Chromatogr. A* **1998**, *808*, 23−41.
(5) Zolnierowicz, S.; Bollen, M. *EMBO J.* **2000**, *19*, 483−488.
(6) Cohen, P. *Trends Biochem. Sci.* **2000**, *25*, 596−601.
(7) Cohen, P. *Trends Biochem. Sci.* **2000**, *25*, 596−601.
(8) Salazar, C.; Hofer, T. *FEBS J.* **2009**, *276*, 3177−98.
(9) Alonso, A.; Sasin, J.; Bottini, N.; Friedberg, I.; Friedberg, I.; Osterman, A.; Godzik, A.; Hunter, T.; Dixon, J.; Mustelin, T. *Cell* **2004**, *117*, 699−711.
(10) Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. *Cell* **2006**, *127*, 635−648.
(11) Hunter, T. *Cell* **2000**, *100*, 113−27.
(12) Eckhart, W.; Hutchinson, M. A.; Hunter, T. *Cell* **1979**, *18*, 925−33.
(13) Stehelin, D.; Varmus, H. E.; Bishop, J. M.; Vogt, P. K. *Nature* **1976**, *260*, 170−3.
(14) Martin, G. S. *Nature* **1970**, *227*, 1021−3.
(15) Duesberg, P. H.; Vogt, P. K. *Proc. Natl. Acad. Sci. U. S. A.* **1970**, *67*, 1673−80.
(16) Hunter, T. *Curr. Opin. Cell Biol.* **2009**, *21*, 140−146.
(17) Blom, N.; Gammeltoft, S.; Brunak, S. *J. Mol. Biol.* **1999**, *294*, 1351−1362.
(18) Shi, J. B.; Malik, J. *IEEE T. Pattern Anal.* **2000**, *22*, 888−905.
(19) Cai, W. C.; Wu, J.; Chung, A. C. S. *IEEE Image Proc.* **2006**, 1101−1104.
(20) Xue, Y.; Liu, Z. X.; Cao, J.; Ma, Q. A.; Gao, X. J.; Wang, Q. Q.; Jin, C. J.; Zhou, Y. H.; Wen, L. P.; Ren, J. A. *Protein Eng., Des. Sel.* **2011**, *24*, 255−260.
(21) Xue, Y.; Ren, J.; Gao, X. J.; Jin, C. J.; Wen, L. P.; Yao, X. B. *Mol. Cell. Proteomics* **2008**, *7*, 1598−1608.
(22) Fisher, R. A. *Ann. Eugenics* **1936**, *7*, 179−188.
(23) Guyon, I.; Elisseeff, A. *J. Machine Learning Res.* **2003**, *3*, 1157−1182.