A helix scaffold for the assembly of active protein kinases

Alexandr P. Kornev*, Susan S. Taylor*†‡, and Lynn F. Ten Eyck†§

*Howard Hughes Medical Institute, †Department of Chemistry and Biochemistry, and §San Diego Supercomputer Center, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093

Contributed by Susan S. Taylor, August 12, 2008 (sent for review July 18, 2008)

Structures of set of serine-threonine and tyrosine kinases were investigated by the recently developed bioinformatics tool Local Spatial Patterns (LSP) alignment. We report a set of conserved motifs comprised mostly of hydrophobic residues. These residues are scattered throughout the protein sequence and thus were not previously detected by traditional methods. These motifs traverse the conserved protein kinase core and play integrating and regulatory roles. They are anchored to the F-helix, which acts as an organizing "hub" providing precise positioning of the key catalytic and regulatory elements. Consideration of these discovered structures helps to explain previously inexplicable results.

graph theory | hydrophobic motifs | structure comparison

Protein kinases represent a large protein superfamily that regulates numerous processes in living cells (1). Malfunction of this regulation typically leads to various diseases, including immunodeficiencies, cancers, and endocrine disorders (2). Multiple sequence alignment identified the most conserved motifs and defined universal subdomains in protein kinases (3). Solving crystal structures of different protein kinases demonstrated not only a conserved core but also the exceptional flexibility of protein kinases. This indicated an important role of dynamics and plasticity for this family (4, 5). Substantial progress has been made in understanding the regulatory mechanisms, although many questions still remain unanswered (6). Recently, we reported a new bioinformatics method that is capable of detecting conserved patterns formed by residues in space without any relation to protein sequence or main chain geometry. Originally, it was created for comparison of protein surfaces (7, 8), but later the method was expanded for analysis of the whole molecule and was termed "Local Spatial Patterns (LSP) alignment" (9). Application of the method to a set of serine/threonine and tyrosine kinases led to the discovery of an unusual structure, which we termed a "spine" (8). The most remarkable feature of the spine is that it is assembled during the protein kinase activation process and provides coordinated movement of the two kinase lobes. In deactivated kinases, the spine is usually broken because of the rearrangement of the C-helix and/or activation loop. Disassembly of the spine leads to general destabilization of the kinase molecule, which was previously observed in hydrogen–deuterium exchange studies (10, 11) and MD simulations (12). It was demonstrated subsequently that mutation of the spine residues leads to increased flexibility of the activation loop in MAP kinase ERK2 (13) and to a total inactivation of p38α MAP kinase (14).

Despite the fact that the spine is a conserved feature, present in all active eukaryotic protein kinases, it was not detected earlier as a conserved spatial motif. This is due, in part, to the highly unusual nature of its formation. It is comprised of four single residues coming from four different subdomains of the protein kinase molecule (III, IV, VIb and VII)[¶], which do not form a sequence "motif" in a traditional sense. 3D alignment of different kinases was also incapable of detecting this ensemble, because it does not form any contiguous main-chain pattern. This discovery drew our attention to the internal structure of the kinase catalytic core. Quite often, it is considered as an amorphous clustering of hydrophobic residues, a result of hydrophobic collapse in the protein folding process. However, it was shown that large ensembles of residues can be formed inside proteins, thereby creating allosteric signaling pathways (15, 16). Residues in these formations are precisely positioned, and their mutation abolishes the allosteric signal propagation. Existence of the hydrophobic spine demonstrated that such unconventional structures not only may be a part of allosteric signaling systems but also can perform structural and/or regulatory functions. Detection of such ensembles is not a trivial task, because they can be formed by residues that come from different parts of the polypeptide chain and do not form any motifs in terms of sequence or secondary/tertiary structure. Usually it requires a complicated multiple sequence alignment of hundreds or even thousands of sequences to achieve statistical equilibrium (15). In contrast, the LSP alignment does not need any sequence alignment, although it does require knowledge of the 3D structures. Its advantage, however, is that comparison of only two structures may provide meaningful insight into protein structure and function. The reason is that the LSP alignment not only detects the conserved patterns of residues but also ranks these residues according to their involvement in the patterns.

In our previous work, we analyzed only water-accessible residues (8). In this study, we used the LSP alignment to compare whole molecules of serine/threonine and tyrosine kinases. Inside the hydrophobic core of the kinases, we detected conserved unconventional motifs, similar to the spine reported earlier. We show that the F-helix, which is positioned in the middle of the large lobe, plays an integrating role by anchoring many hydrophobic motifs. These motifs traverse the entire molecule and orchestrate the catalytic process. They also provide diverse mechanisms for regulation. We define a second "spine," which we term a catalytic spine, because it traverses both lobes but is completed by the adenine ring of ATP. This is distinct from the previous spine that we now refer to as the regulatory spine. The R and C spines are anchored to the N and C termini of the F-helix, respectively. We furthermore demonstrate that the APE motif, which is bound to the F-helix, nucleates substrate binding and allostery. A discussion of several published works shows that newly defined structures can be helpful for understanding the intramolecular machinery of protein kinases and their interactions with other proteins.

Author contributions: A.P.K., S.S.T., and L.F.T.E. designed research; A.P.K. performed research; A.P.K. contributed new reagents/analytical tools; A.P.K., S.S.T., and L.F.T.E. analyzed data; and A.P.K. and S.S.T. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

‡To whom correspondence should be addressed. E-mail: staylor@ucsd.edu.

¶Subdomains are numbered according to Hanks and Hunter (3).

This article contains supporting information online at www.pnas.org/cgi/content/full/0807988105/DCSupplemental.

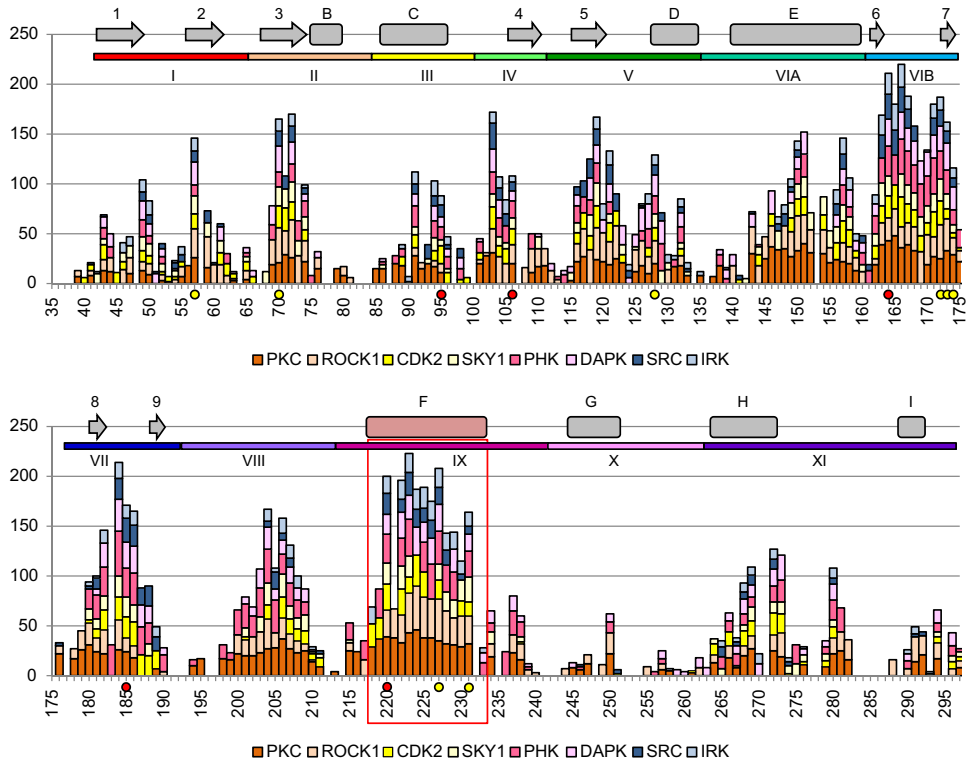

Fig. 1. LSP alignment of active conformation of PKA to active conformations of other kinase. Four different families were used in the comparison: AGC kinases (PKC and ROCK1), CMGC kinases (CDK2 and SKY1), calcium/calmodulin kinases (PHK and DAPK), and tyrosine kinases (SRC and IRK). Residues that constitute the two conserved spines are spread along the kinase sequence: R spine residues are marked as red dots; yellow dots mark the C spine residues. Highly scored αF-helix is marked by red square.

Results

LSP Alignment of Different Serine/Threonine and Tyrosine Kinases. The most important information produced by the LSP alignment is expressed by the so-called involvement score (IS). It is defined for every residue of the compared proteins. If the score equals zero, this residue is not a part of any conserved spatial pattern. The higher the IS value, the more this residue is involved in formation of the conserved patterns, and the higher the probability that it is important for protein functionality [see supporting information (SI) *Text* for detailed explanation of the concept]. Although in the previous work (8), we described the major changes related to protein kinase activation, in this work, we have analyzed only active kinases and considered all residues, not just solvent accessible residues, as was done previously. We compared protein kinase A (PKA) from the AGC group to eight kinases from four groups: AGC (PKC and ROCK1), CMGC (CDK2 and SKY1), CaMK group (PHK and DAPK), and protein tyrosine kinases (SRC and IRK). All IS values obtained in eight comparisons are listed in the Table S1. Fig. 1 shows accumulated IS for each PKA residue.

Newly detected residues with high IS were localized mostly in the hydrophobic core of the large lobe, predominantly in helices E and F, around the P + 1-loop and the APE motif (residues 201–210; here and subsequently, we will use mammalian PKA sequence numbering) (Fig. 1, Table S1). Several residues scored in every kinase were found in the H-helix. In addition, residues with exceptionally high IS, which were not appreciated previously, were detected in the N lobe and catalytic and activation loops: L^{103}, M^{128}, I^{150}, L^{172}, and V^{182}.

Defining Conserved Structural Motifs. Traditionally conserved motifs are related to an amino acid sequence pattern (like DFG or APE motifs in protein kinases) or a combination of secondary structures (e.g., helix–turn–helix motif in DNA-binding proteins). However, as we showed earlier (8), proteins can contain unconventional structural motifs, which are not related to any sequence or secondary structure motifs. This raises important questions: what constitutes a conserved structural motif? Where is a border between two motifs? In the case of traditional motifs, the answer to the second question is intuitive; two different motifs have to be separated "far enough" from each other. However, in our case, when motifs are formed by amino acid residues coming from different parts of the primary sequence, this approach is irrelevant. In the current work, we have attempted to separate the large set of highly conserved residues, identified by the LSP alignment, into several different structural motifs based on their catalytic or regulatory roles.

For several reasons, we considered the F-helix as the central hub for these motifs. First, the IS level obtained by the F-helix residues was the highest through the entire molecule. Only the catalytic loop region together with the β_6 and β_7 strands was comparable. Second, it is positioned in the middle of the rigid hydrophobic core of the large C lobe, which means it is one of the most immobilized secondary structures in the molecule. Third, it is known that the F-helix serves as a signal-integrating element, which connects several key areas such as the substrate-binding residues and the catalytic loop (5, 17). These observations define the F-helix as a robust scaffold for the whole protein kinase molecule, where all of the motifs can be precisely positioned in space with respect to each other.

R Spine. Earlier, we defined the regulatory spine as a motif of four hydrophobic residues, connecting the two lobes of the kinase (8). In the current work, we demonstrate that all structural motifs known to be important for catalysis are connected to the F-helix. In the case of the hydrophobic regulatory spine, this role is

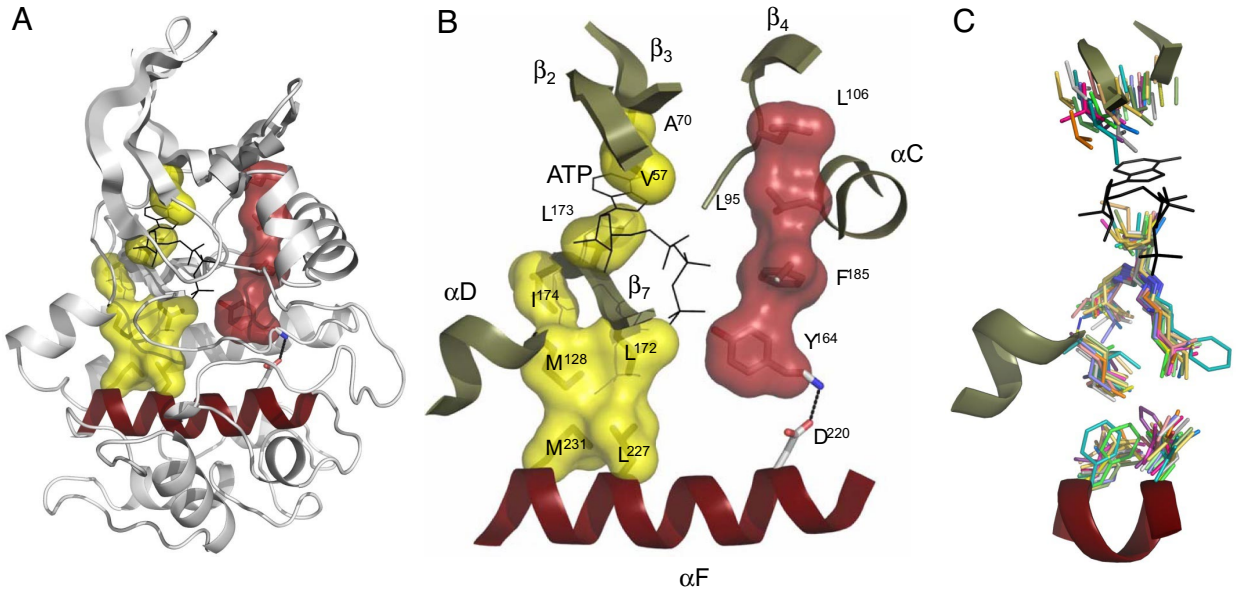

Fig. 2. R and C spines flank the F-helix and span the protein kinase core. The PKA structure is shown as a prototype. (*A* and *B*) The hydrophobic part of the R spine is shown as a red molecular surface. The C spine is colored yellow. The adenine ring of ATP completes the C spine. (*C*) Eight hydrophobic residues that constitute the C spine in 22 active kinases from six different groups (see Table S2 for the list of kinases).

played by the aspartic acid D^{220} (Fig. 2), which is universally conserved in all eukaryotic and eukaryotic-like kinases (5, 17) and is the first F-helix residue with high IS. Further, we will refer to this five-residue motif (L^{95}, L^{106}, Y^{164}, F^{185}, and D^{220}) as regulatory spine, or R spine.

C Spine. The last highly scored residue at the C terminus of the F-helix was M^{231} (Fig. 1). To the best of our knowledge, this residue was never considered as an important part of the kinase structure. However, our analysis shows that it starts a highly scored hydrophobic structure, which spans the large lobe and contacts the adenine moiety of ATP (Fig. 2). This structure includes two residues from the F-helix: $L^{227}$and M^{231}; one residue from the D-helix, M^{128}; and the three hydrophobic residues, which comprise the β_7-strand: L^{172}, L^{173} and I^{174}. There are two highly ranked residues that contact the other side of the adenine ring: V^{57} and A^{70}. Thus, in the presence of ATP, this ensemble connects the two lobes together, in a way similar to the R spine. By analogy, we termed this motif catalytic spine, or C spine. The major difference between these two spines is that the R spine formation depends on activation loop configuration, whereas the C spine depends on the presence of ATP. The C spine thus defines the adenine ring as a primary driving force for integration of the small and large lobes. As shown on the Fig. 2*C*, the C spine is conserved in both serine/threonine and tyrosine kinases. However, in more distant kinases, such as 3′,5′′-aminoglycoside phosphotransferase (APH) (18), this part of the molecule has a substantially different organization. It should be noted that the R spine, unlike the C spine, is present in APH (8). This suggests that the C spine is more sensitive to the nature of the substrate to be phosphorylated, whereas the R spine remains conserved in different kinases as soon as they have the conserved HRD and DFG motifs and the regulatory C-helix.

H-Helix Anchor. There are three hydrophobic residues (W^{222}, I^{228}, and Y^{229}) and glycine (G^{225}) from the middle part of the F-helix, which scored very high. These residues form a tight hydrophobic cluster with the four conserved leucines in the H-helix (L^{268}, L^{269}, L^{272}, and L^{273}). As Fig. 3 shows, other residues from previously described structures are involved in the cluster: W^{222}, V^{226}, and I^{228}. An apparent function of the structure is to protect and stabilize the hydrophobic core around the F-helix. A characteristic feature of the structure is the position of L^{273}, which is inserted into the void space between W^{222} and V^{229}, where a strictly conserved G^{225} is located. Such insertion is a characteristic of tetratricopeptide repeats (19), which provide a robust connection between two antiparallel α-helices. Most likely, the strong conservation of G^{225} is dictated by this feature.

Catalytic and Activation Loops Anchoring. The maximal involvement score in our computation was received by another residue from the F-helix: A^{223}. It was slightly higher than IS for such catalytically crucial aspartates as D^{166} and D^{184} (Fig. 1, Table S1). As shown in Fig. 4 *A* and *B*, A^{223} and neighboring L^{224} perform a very important function of anchoring L^{167} from the catalytic loop. This leucine interacts both with R and C spines and provides exact positioning of D^{166} and K^{168} main chain (Fig. 4*A*). These residues are directly involved in phosphoryl transfer (20) and are universally conserved.

L^{224} also serves as a basis for another highly scored residue from the E-helix: I^{150} that defines the position of N terminus of the activation loop with the D^{184} (Fig. 4 *B* and *C*). Stabilization of this aspartate is critical, because it binds a magnesium atom that bridges the β- and γ-phosphates of ATP. Such positioning is secured by a contact between I^{150} and two hydrophobic residues from the β_8 strand: I^{180} and V^{182}. They interact with both spines and are a part of the β_7-β_8 sheet, a structural feature that is conserved in all protein kinases.

Substrate-Binding Structure. A common feature for all protein kinases is that they phosphorylate polypeptide chains. This separates them from all other phosphotransferases and makes their substrate-binding structure substantially different (21). This structure is known to be formed by two major structural elements: the G-helix and the C-terminal part of the activation segment (5), referred to as the *P* + 1 loop (residues 198–205), where the substrate is bound to provide positioning of the phosphorylation site next to the γ-phosphate of ATP. Our analysis identified a set of highly scored residues in the subdomain VIII, with the conserved APE motif at the end of the

BIOCHEMISTRY

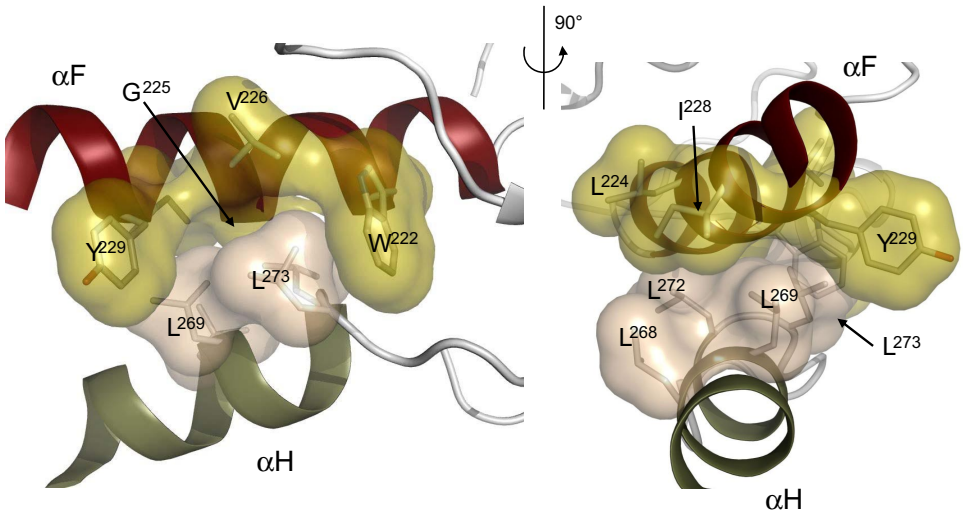

Fig. 3. αH-helix anchoring structure in PKA. Molecular surfaces of the highly scored residues are shown: olive, from the αF-helix, tan, from the αH-helix. L^{273} fits as a ''knob'' into a ''hole'' formed by conserved W^{222}, G^{225}, and Y^{229}.

activation segment. Residues from Y^{204} through E^{208} were detected in both serine/threonine and tyrosine kinases (Fig. 1, Table S1). Several residues received high scores only in serine/threonine kinases: T^{201}, E^{203}, and I^{209}. The major difference between serine/threonine and tyrosine kinases is that the size of tyrosine side chain is significantly larger. This explains the variation in geometry and sequence of the $P + 1$ Loop, where a hydrophobic residue following the phosphorylation site is accommodated (Fig. 5*B*).

The highest levels of IS were obtained by Y^{204} and A^{206}, which are in direct contact with the highly scored residues from the F-helix: W^{222} and V^{226} (Fig. 5*B*). This is consistent with our suggestion that all of the important structural motifs are connected to the F-helix. As we showed earlier, these two residues from the F-helix constitute a part of the rigid H-helix anchoring structure. Fig. 5 shows that they also play an important role as a foundation of the substrate-binding structure. V^{226} provides the exact positioning of the conserved tyrosine (Y^{204}) in serine/threonine kinases (Fig. 5*A*) or tryptophan in tyrosine kinases. This tyrosine was reported to be an important allosteric ''hot spot'' in PKA (22–24). It is involved in hydrophobic interaction with the aliphatic part of K^{168} from the catalytic loop and, together with another residue from the $P + 1$ loop (T^{201}), stabilizes the side chain of this catalytically important residue. W^{222} tightly binds to the Ala-Pro pair from the APE-motif (Fig. 5 *B* and *C*) and nucleates the hydrophobic structure that positions four hydrophobic residues comprising the $P + 1$ loop (Fig. 5*B*). The APE-glutamate (E^{208}) anchors the side chain of the highly conserved arginine from the loop between the H and I helices: R^{280} (Fig. 5*C*).

Discussion

During last two decades, protein kinases were studied intensively using sequence and structure alignments. Our recent work demonstrated that comparison of protein kinase surfaces can discover unconventional structural motifs that play an important regulatory role (8). Further development of the surface comparison method allowed us to include both water-exposed and buried residues in the analysis (9). The major objective of this work is to analyze conserved structural features in the hydrophobic core of different protein kinases. As we expected, all residues, which were known to be important for catalysis or for its regulation were highly scored, thus confirming the credibility of our method. Along with the well known residues, we discov-

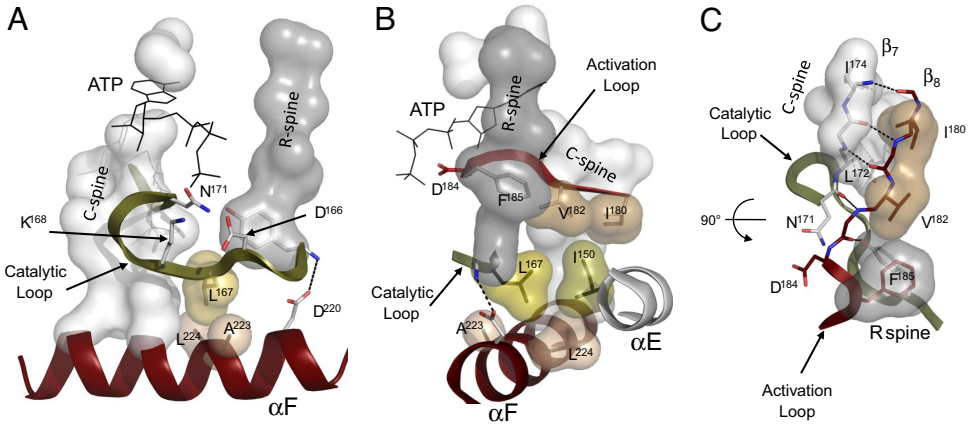

Fig. 4. Catalytic and activation loop anchoring in PKA. (*A*) N- and C-terminal parts of the catalytic loop are secured by the spines. The middle part is docked to A^{223} and L^{224} from the αF-helix. Three catalytically active residues are shown: D^{166}, K^{168}, and N^{171}. (*B* and *C*) Activation loop anchoring. Its N terminus binds to the β_8-strand that makes both polar and hydrophobic bonds to the β_7-strand from the C spine. The β_8-strand is also bound to the αF-helix via hydrophobic interaction with the conserved I^{150}. This positions the N terminus of catalytically active D184, whereas its C terminus is secured by the R spine.

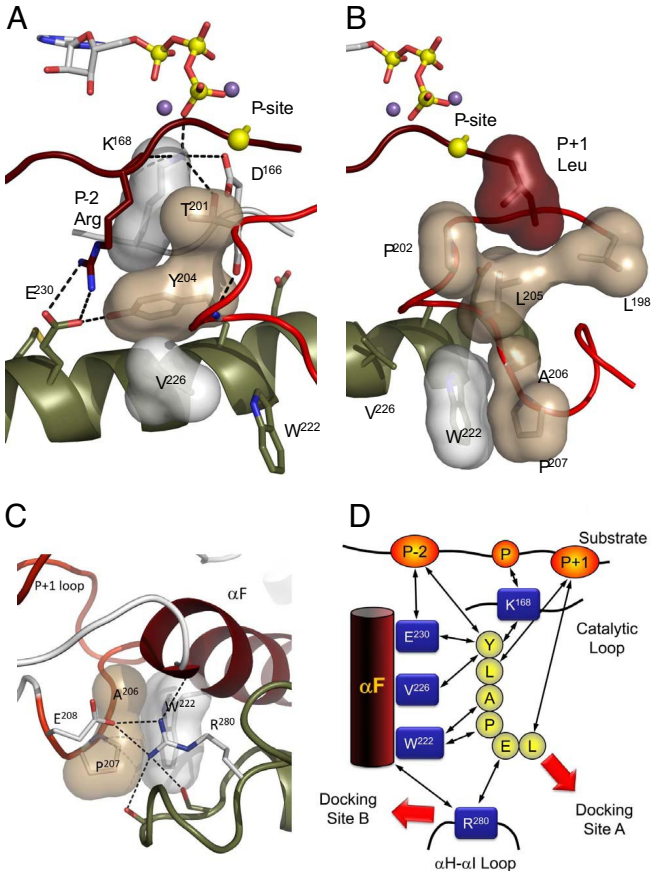

Fig. 5. Organization of the substrate-binding network in PKA. (*A*) V^{226} stabilizes Y^{204}, which anchors the activation segment (colored bright red) and orients side chain of K^{168}. E^{230} interacts with the substrate arginine. (*B*) W^{222} serves as a docking platform for the APE motif, which serves as a foundation for the *P* + 1 loop. (*C*) Closeup of the APE-motif interactions. R^{280} forms multiple hydrogen bonds to the αF-helix, the APE motif, and the αH-αI loop. (*D*) General chart of the YLAPEL motif that mediates interaction between the αF-helix, substrate binding, and docking sites.

ered a set of residues that were not considered previously to be important elements of the protein kinase structure or function (Table S3). The major problem, however, was to separate the pool of the detected residues into meaningful structural elements. In the analysis, we followed a simple principle of functionality: there has to be a structure that coordinates the binding of ATP and a structure that coordinates the binding of the substrate. These structures have to be positioned with respect to each other in a precise and robust way. In our analysis, we found that the highest levels of the IS were obtained by the F-helix. Almost every residue of the helix was highly conserved and precisely positioned with respect to the other key residues responsible for catalysis, substrate binding, or regulation. Earlier, the helix was described as a signal integrating motif for the protein kinase molecule (5). In this work, we demonstrate that it serves as a central scaffold for active protein kinase assembly. Two conserved structures that flank the F-helix and span the protein kinase molecule are the most significant feature of the intramolecular architecture (Fig. 2). We termed these structures R and C spines. They represent unconventional structural motifs formed by residues coming from different parts of the protein sequence and do not form any sequential motifs in a traditional sense (Fig. 1). An important property of these structures is that they can be assembled and disassembled depending on the presence of ATP, the phosphorylation state of the activation loop, or the C-helix movement. This provides a significant conformational plasticity, which is very important for protein kinase functionality (4, 5). Furthermore, because the spines are comprised of hydrophobic residues, this makes the connection between different parts of the molecule firm but flexible. When the two spines are assembled, all catalytically important residues are securely positioned and ready for the catalysis (Fig. 4).

Substrate binding is also defined by residues from the F-helix: W^{222}, V^{226} and E^{230} (Fig. 5). The latter directly binds to the substrate and, for this reason, is relatively substrate specific. It was shown that presence of glutamate in this position is a characteristic of kinases with substrate preference for P-2 or P-5 arginines (25). These three residues position the C-terminal portion of the activation segment, which is usually associated with the APE-motif. Fig. 5*D* shows the general organization of the substrate binding elements. E^{230}, together with V^{226}, anchors Y^{204}, which is highly conserved in serine/threonine kinases but substituted by tryptophan in tyrosine kinases. However, both tyrosine and tryptophan play a common role by providing a rigid platform for the N-terminal part of the substrate and by orienting the side chain of the catalytically active K^{168} (arginine in tyrosine kinases) (Fig. 5*A*). W^{222} serves as a docking site for binding alanine and proline in the conserved APE-motif (Fig. 5 *B* and *D*). The hydrophobic residue right before the alanine (L^{205} in PKA) defines the position of the *P* + 1 loop, which stabilizes the C-terminal part of the substrate (Fig. 5*B*). The main chain geometry of the six C-terminal residues of the activation segment is universally conserved through all serine/threonine and tyrosine kinases, even in casein kinase I (PDBID 2CHL), which does not have the APE-motif, W^{222} or R^{280}. This geometry is also precisely mimicked in the transphosphorylating protein kinases, when two interacting kinases exchange their activation segments (26). It emphasizes the crucial role of the YLAPEL-motif. Fig. 5*D* shows that it is related not only to substrate binding but also to docking sites on the protein kinase. Docking site A is located between the activation segment and the G-helix. It is a major binding site between the catalytic subunit of PKA and the A-domain of its regulatory subunits (27, 28). It was shown that docking of the CDK-interacting protein phosphatase KAP to CDK2 also occurs on docking site A (29). The second site is located in the loop between the H and I helices. This was shown to be an important interface in the PKA type RII holoenzyme, where the B-domain of the regulatory subunit binds to the catalytic subunit (28). One can suggest that, because of the proximity of this area to the substrate binding site and the universally conserved interaction between the APE-glutamate and R^{280}, docking site B is important for other protein kinases as well.

Thus, we defined several conserved spatial structures in the hydrophobic core of protein kinases, which position the following three major elements: the peptide substrate, ATP, and the set of residues directly involved in the catalysis. Definition of the structures and their roles helps to interpret previously inexplicable results. For example, it was not clear why the $E^{230}Q$ mutant of PKA crystallized in an open apo-form, despite the fact that ATP was present in the solution (30). E^{230} is usually considered a part of substrate-binding complex, and it was not clear how it could influence ATP binding. However, a close look at the $E^{230}Q$ mutant structure shows that such a glutamate–glutamine substitute significantly destabilizes R^{133} from the D-helix, which is bonded to the E^{230} in wild-type PKA. The arginine side chain in the $E^{230}Q$ mutant was not resolved, and the whole D-helix had elevated temperature factors. According to our classification, the D-helix constitutes the middle part of the C spine, which is responsible, in particular, for ATP binding (Fig. 2). Destabilization of the C spine foundation would explain the observed instability of ATP binding. This would lead to inability of the mutant to complete the C spine formation, promoting the apo-configuration.

A similar effect of the C spine destabilization was observed in the $F^{314}A$ mutant of PKA (31). This phenylalanine is located in the C-terminal tail and contacts H^{131} and I^{135} from the D-helix. Clearly, the phenylalanine-to-alanine mutation had to destabilize both the C-terminal tail and the C spine. Indeed, this mutation led to a significant decrease in thermal stability, a moderate decrease in affinity for ATP, and a nearly 20-fold decrease in the catalytic activity. Mutation of the neighboring residue I^{315}, which is not in contact with the D-helix, and thus the C spine, decreased thermal stability of the mutant and affinity for ATP and Kemptide but did not decrease activity. This indicates that stability of the C spine is important for optimization of the catalytic process.

Interconnectivity of the conserved spatial structures sheds light on long-range communication within kinase molecules, which has been observed by numerous authors (13, 14, 22, 24). The unphosphorylated apo-structures are typically the most disorganized: both the R and C spines are broken, and movements of the lobes are not coordinated. R spine assembly induced by phosphorylation or interaction with other activating entities such as cyclin for cyclin-dependent kinases causes significant ordering of the kinase core structure (11, 12). ATP binding completes the C spine formation and makes the molecule even more compact and primed for catalysis. Finally, substrate binding connects all parts of the molecule. The phosphorylation site is in the central cross-point for all conserved structures, and residues positioned close to it can substantially influence the intramolecular connectivity and thus, communication. A well examined case of such influence is Y^{204} from the YLAPEL-motif. Multiple studies demonstrated that mutation of this tyrosine to alanine decreased substrate binding and enzyme activity (22), caused destabilization in distal parts of the C lobe (23), and disrupted general intermolecular communication (24).

Methods

Modification of the Computational Method. LSP alignment is a graph-theory-based method that compares two protein structures and detects similar spatial patterns made by residues in the proteins. The patterns are described by a pair of isomorphic graphs, where vertices correspond to the detected residues with edges connecting pairs of similar residues whose mutual orientation in space is conserved. Our previous work showed that functionally important residues are usually positioned in the middle of the graph with numerous connections to their neighbors. Alternatively, residues that play a supportive role are on a periphery of the graph with a fewer number of connections. The number of connections on the similarity graph was termed IS (see *SI Text* for detailed explanation of the involvement score concept). IS calculation in the current work was made according to the previously published algorithm (8) with a certain modification of similarity specification for residues. Earlier, we considered residues to be similar if their BLOSUM62 coefficient (32) was ≥ 2. However, this approach sometimes leads to suggestions that are not adequately justified. For example, valine and methionine are considered similar to isoleucine but not to leucine. A decrease of the threshold to 1 resolves this problem but, on the other hand, provides a rather doubtful suggestion that threonine can be substituted by proline, glycine, or aspartic acid. In the current work, we based our substitution matrix on one of the optimized substitution matrices (33) with consideration of the strong cysteine hydrophobicity (34) (see Table S4).

Protein Structures. The following structures of protein kinases were used in the analysis: AGC kinases: PKA–PDBID: 2CPK; PKC–PDBID: 2JED; ROCK1–PDBID: 2ESM. CMGC kinases: CDK2–PDBID: 1FIN; SKY1–PDBID:1HOW. Calcium/calmodulin kinases: PHK–PDBID: 2PHK; DAPK–PDBID: 1JKK. Tyrosine kinases: IRK–PDBID: 1IR3; SRC–PDBID 1Y57.

Molecular graphics were prepared by using PyMOL (DeLano Scientific). Molecular surface was rendered with a probe radius of 1.4 Å.

ACKNOWLEDGMENTS. This work was supported by National Institute of General Medical Sciences Grant GM70996 (to L.F.T.E.) and National Institutes of Health Grant GM19301 (to S.S.T.).

1. Manning G, Plowman GD, Hunter T, Sudarsanam S (2002) Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci* 27:514–520.
2. Ortutay C, Valiaho J, Stenberg K, Vihinen M (2005) KinMutBase: A registry of disease-causing mutations in protein kinase domains. *Hum Mutat* 25:435–442.
3. Hanks SK, Hunter T (1995) Protein kinases 6. The eukaryotic protein kinase superfamily: Kinase (catalytic) domain structure and classification. *FASEB J* 9:576–596.
4. Huse M, Kuriyan J (2002) The conformational plasticity of protein kinases. *Cell* 109:275–282.
5. Johnson DA, Akamine P, Radzio-Andzelm E, Madhusudan M, Taylor SS (2001) Dynamics of cAMP-dependent protein kinase. *Chem Rev* 101:2243–2270.
6. Taylor SS, Ghosh G (2006) Protein kinases: Catalysis and regulation. *Curr Opin Struct Biol* 16:665–667.
7. Berman HM, *et al.* (2005) The cAMP binding domain: An ancient signaling module. *Proc Natl Acad Sci USA* 102:45–50.
8. Kornev AP, Haste NM, Taylor SS, Ten Eyck LF (2006) Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proc Natl Acad Sci USA* 103:17783–17788.
9. Kornev AP, Taylor SS, Ten Eyck LF (2008) A generalized allosteric mechanism for cis-regulated cyclic nucleotide binding domains. *PLoS Comput Biol* 4:e1000056.
10. Iyer GH, Moore MJ, Taylor SS (2005) Consequences of lysine 72 mutation on the phosphorylation and activation state of cAMP-dependent kinase. *J Biol Chem* 280:8800–8807.
11. Lee T, Hoofnagle AN, Resing KA, Ahn NG (2005) Hydrogen exchange solvent protection by an ATP analogue reveals conformational changes in ERK2 upon activation. *J Mol Biol* 353:600–612.
12. Barrett CP, Noble ME (2005) Molecular motions of human cyclin-dependent kinase 2. *J Biol Chem* 280:13993–14005.
13. Emrick MA, *et al.* (2006) The gatekeeper residue controls autoactivation of ERK2 via a pathway of intramolecular connectivity. *Proc Natl Acad Sci USA* 103:18101–18106.
14. Bukhtiyarova M, Karpusas M, Northrop K, Namboodiri HV, Springman EB (2007) Mutagenesis of p38alpha MAP kinase establishes key roles of Phe169 in function and structural dynamics and reveals a novel DFG-OUT state. *Biochemistry* 46:5687–5696.
15. Suel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* 10:59–69.
16. Swain JF, Gierasch LM (2006) The changing landscape of protein allostery. *Curr Opin Struct Biol* 16:102–108.
17. Kannan N, Neuwald AF (2005) Did protein kinase regulatory mechanisms evolve through elaboration of a simple structural component? *J Mol Biol* 351:956–972.
18. Burk DL, Hon WC, Leung AK, Berghuis AM (2001) Structural analyses of nucleotide binding to an aminoglycoside phosphotransferase. *Biochemistry* 40:8756–8764.
19. Blatch GL, Lassle M (1999) The tetratricopeptide repeat: A structural motif mediating protein–protein interactions. *BioEssays* 21:932–939.
20. Valiev M, Yang J, Adams JA, Taylor SS, Weare JH (2007) Phosphorylation reaction in cAPK protein kinase-free energy quantum mechanical/molecular mechanics simulations. *J Phys Chem B* 111:13455–13464.
21. Scheeff ED, Bourne PE (2005) Structural evolution of the protein kinase-like superfamily. *PLoS Comput Biol* 1:e49.
22. Moore MJ, Adams JA, Taylor SS (2003) Structural basis for peptide binding in protein kinase A. Role of glutamic acid 203 and tyrosine 204 in the peptide-positioning loop. *J Biol Chem* 278:10613–10618.
23. Yang J, *et al.* (2005) Allosteric network of cAMP-dependent protein kinase revealed by mutation of Tyr204 in the P+1 loop. *J Mol Biol* 346:191–201.
24. Masterson LR, Mascioni A, Traaseth NJ, Taylor SS, Veglia G (2008) Allosteric cooperativity in protein kinase A. *Proc Natl Acad Sci USA* 105:506–511.
25. Zhu G, *et al.* (2005) A single pair of acidic residues in the kinase major groove mediates strong substrate preference for P-2 or P-5 arginine in the AGC, CAMK, and STE kinase families. *J Biol Chem* 280:36372–36379.
26. Pike AC, *et al.* (2008) Activation segment dimerization: A mechanism for kinase autophosphorylation of non-consensus sites. *EMBO J* 27:704–714.
27. Kim C, Cheng CY, Saldanha SA, Taylor SS (2007) PKA-I holoenzyme structure reveals a mechanism for cAMP-dependent activation. *Cell* 130:1032–1043.
28. Wu J, Brown SH, von Daake S, Taylor SS (2007) PKA type II-alpha holoenzyme reveals a combinatorial strategy for isoform diversity. *Science* 318:274–279.
29. Song H, *et al.* (2001) Phosphoprotein-protein interactions revealed by the crystal structure of kinase-associated phosphatase in complex with phosphoCDK2. *Mol Cell* 7:615–626.
30. Wu J, *et al.* (2005) Crystal structure of the E230Q mutant of cAMP-dependent protein kinase reveals an unexpected apoenzyme conformation and an extended N-terminal A helix. *Protein Sci* 14:2871–2879.
31. Batkin M, Schvartz I, Shaltiel S (2000) Snapping of the carboxyl terminal tail of the catalytic subunit of PKA onto its core: Characterization of the sites by mutagenesis. *Biochemistry* 39:5366–5373.
32. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89:10915–10919.
33. Hourai Y, Akutsu T, Akiyama Y (2004) Optimizing substitution matrices by separating score distributions. *Bioinformatics* 20:863–873.
34. Nagano N, Ota M, Nishikawa K (1999) Strong hydrophobic nature of cysteine residues in proteins. *FEBS Lett* 458:69–71.

Supporting Information

Kornev *et al*. 10.1073/pnas.0807988105

SI Text

Involvement Score Rationale. Involvement Score is a novel concept that is related to the nature of the LSP-alignment as a graph-theory based method. *Graph* is a way to represent information about a set of objects and relations between them. The objects are usually depicted as *vertices*, and relations between them as *edges* (Fig. S1). For example, vertices can represent airports and edges direct flights between them. Two identical graphs are called "isomorphic," that means that they consist of identical objects and are connected in the same way. In other words there is a "one-to-one" correspondence between all elements of these graphs. Fig. S1 shows a pair of similar graphs that have such correspondence only between parts of the graphs, colored red. These identical parts are called "isomorphic subgraphs."

In our method, protein molecule is presented as a graph. Residues are considered as vertices of the graph while edges describe spatial relations between them. By "spatial relation" we mean distance between these two residues and mutual orientation of their side chains. Orientation of a side chain in space can be defined by multiple ways. In our case, we define it as orientation of Cα-Cβ vector. The LSP alignment procedure is capable to compare two graphs that represent two different proteins and to discover isomorphic subgraphs between them. Thus we identify "one-to-one" correspondence between residues of these proteins, i.e., define an alignment between them. The unique feature of such alignment is that it is presented as a graph that consists not only from vertices but also from edges. Each vertex can have different number of edges. Our experience shows that simple counting of the edges can provide important information about the residue. This number we define as the "Involvement Score." As we mentioned earlier, edges describe spatial relations between residues. That is if two residues in the isomorphic subgraphs are connected, they are positioned in space in both proteins in the same way. If a residue has numerous connections on the isomorphic subgraph it means that the position of the residue in space and the orientation of its side chain is conserved with respect to multiple residues. This suggests that the residue is involved in formation of a precisely organized cluster of similar residues. From this we derive a conclusion that residues with high Involvement Scores can play an important functional and/or structural role.

The LSP-alignment method is still under development and the way of the Involvement Score calculation is the most simple: a sum of edges on isomorphic subgraph. Such approach does not consider precision of the geometrical fit between residues. As soon as they are positioned in space within a predefined tolerance, the edge is created. Another simplification is level of similarity between two residues. Instead of using elaborated similarity scores we considered them either similar or dissimilar according to the similarity matrix (Table S4). Improvement of the Involvement Score calculation is a matter of future research. In this work we tried to compensate these weak points by multiple comparisons of PKA to different protein kinases and combining all Involvement Scores.

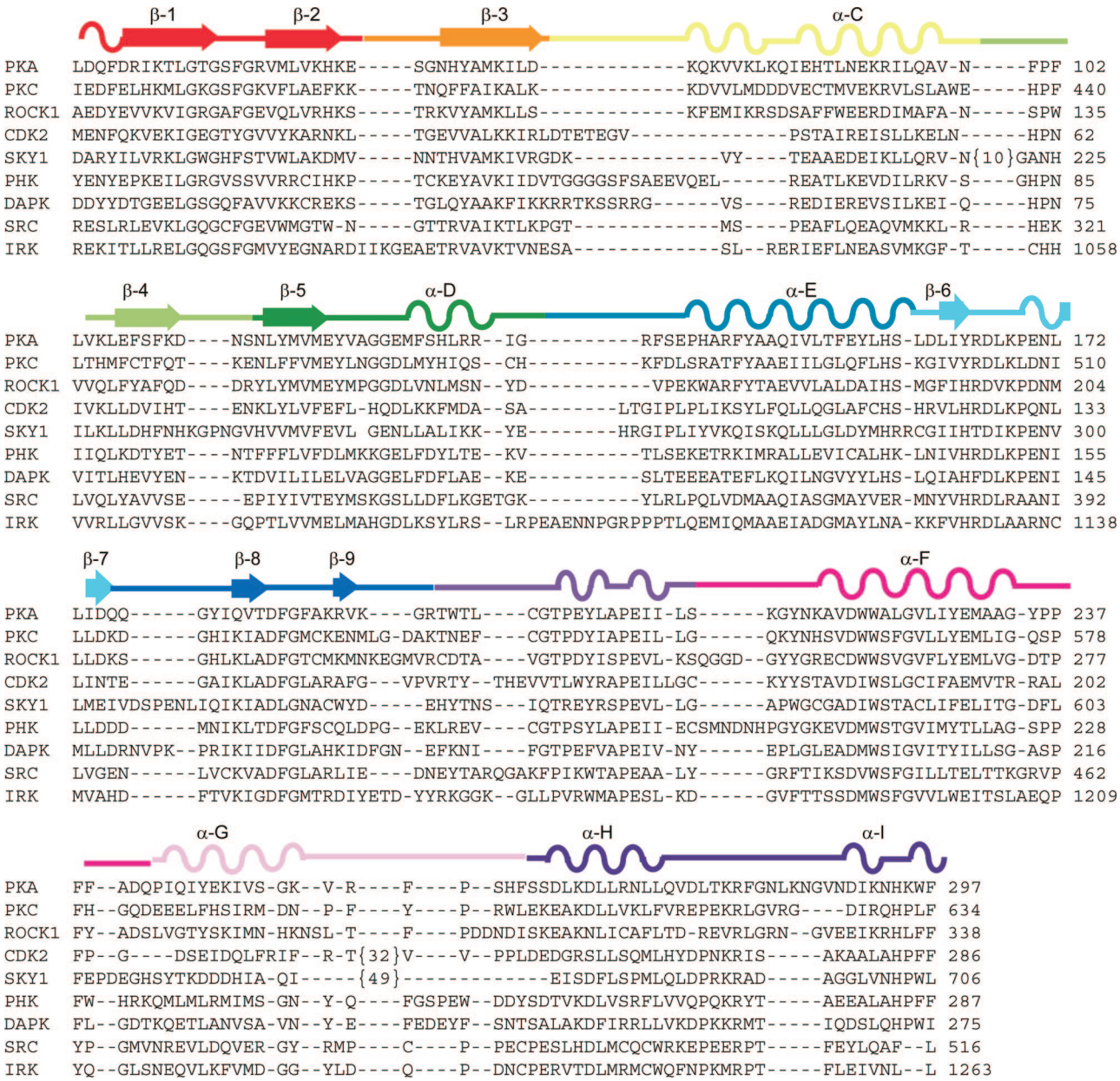

Fig. S1. Illustration of similarity between graphs A and B. Isomorphic subgraphs are colored red. Red vertices contain the same letters, red edges connect identical vertices.

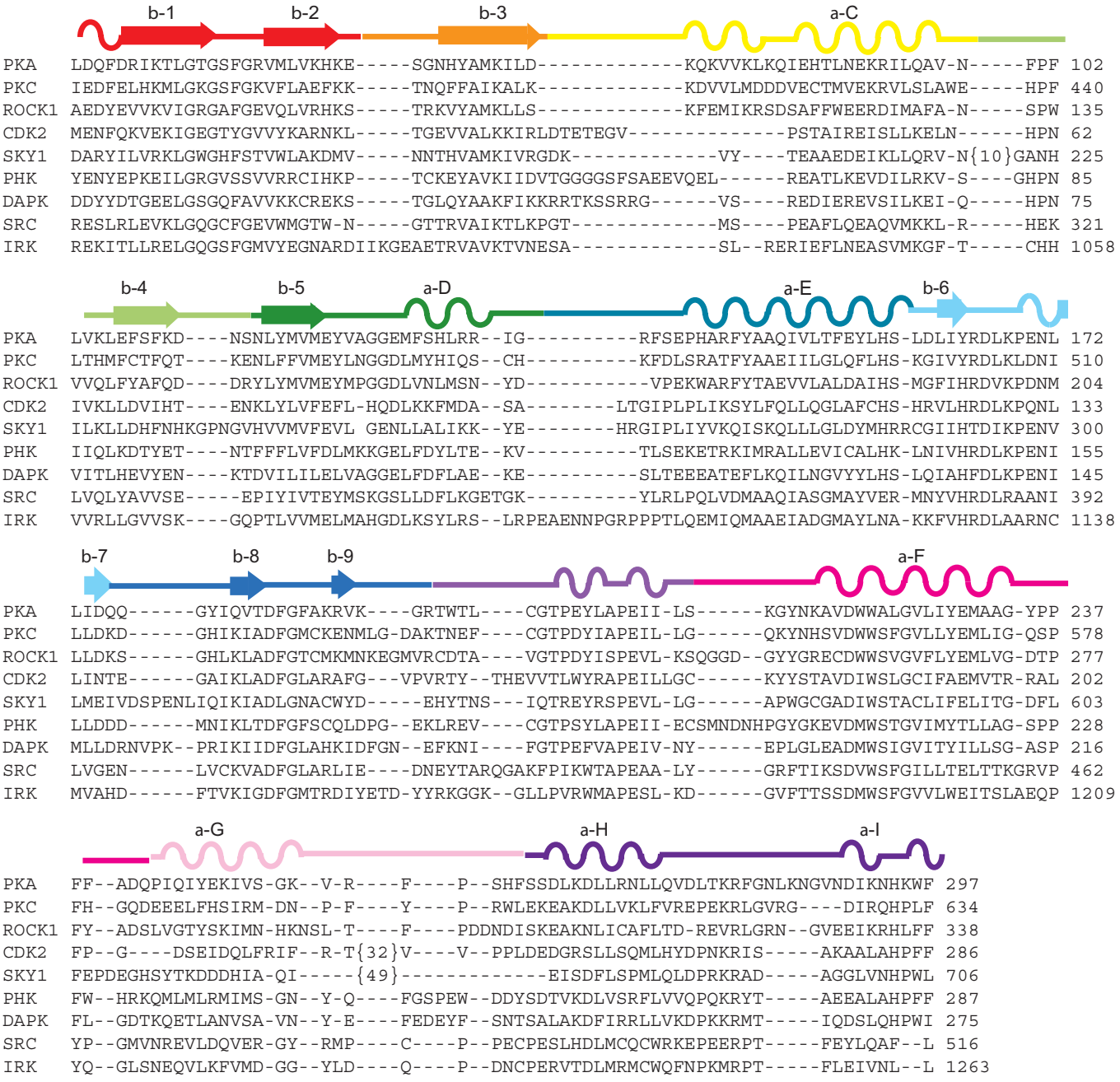

Fig. S2. Structure-based multiple sequence alignment of kinases used in the LSP-alignment. All protein kinases were aligned to PKA in pairwise way by Combinatorial Extension method [Shindyalov IN, Bourne PE (1998), *Protein Eng* 11:739–747]. The alignments were combined and manually curated to provide maximum consistency. Secondary structures of PKA are shown as colored cartoons. Different colors represent conserved subdomains of the kinase core.

Table S1. Involvement scores obtained by different residues of PKA in the LSP-alignment to eight different protein kinases

	PKC	ROCK1	CDK2	SKY1	PHK	DAPK	SRC	IRK	Total
L-40			6						6
D-41	8	5	6				2		21
Q-42				10				2	12
F-43	13	11	15	6	11	11	2		69
D-44	12	14			11	13			50
R-45			11						11
I-46		14	8	9				10	41
K-47	10	16		12				9	47
T-48									0
L-49	13	9	11	14	17	19	9	12	104
G-50	9	11	8	12	8	12	9	14	83
T-51						10	2		12
G-52	3	6	3	8	7	8	5		40
S-53	2		5						7
F-54		4	3	6		2	6	2	23
G-55	7	10	5				7	8	37
R-56	18								18
V-57	26	29	15	18	11	23	11	13	146
M-58									0
L-59	16	31		14			12		73
V-60		19					2		21
K-61		19	12	12		14	3		60
H-62		8	12		10				30
K-63	3	2	5			2			12
E-64									0
S-65	4	9	9	5	4	5			36
G-66			7			6			13
N-67									0
H-68				12					12
Y-69	19	25			15	19			78
A-70	21	35	22	19	15	26	15	12	165
M-71	29	24	23	19			13		108
K-72	26	38	18	20	18	22	16	12	170
I-73		28		15	20				63
L-74	22	21	11	13	16	7	6	3	99
D-75					8				8
K-76	15	11				6			32
Q-77									0
K-78									0
V-79	15								15
V-80	8	9							17
K-81		6							6
L-82									0
K-83									0
Q-84									0
I-85	15								15
E-86	15			4	3	3			25
H-87									0
T-88	20				8				28
L-89	18		9		8			4	39
N-90				2				5	7
E-91	28	8	13	8	21	12	10	12	112
K-92	15	6							21
R-93	20			5			14		39
I-94	17	6	10	13	17	15	10	15	103
L-95	11	9	18	6	13	10	12	9	88
Q-96			11	4	12	8	4	8	47
A-97									0
V-98			4		11	11	9		35
N-99			6						6
F-100									0
P-101	20	4	7		11	3			45
F-102	28	3							31
L-103	31	24	22	13	22	23	26	11	172
V-104		27	11	9	19	18	12	11	107

	PKC	ROCK1	CDK2	SKY1	PHK	DAPK	SRC	IRK	Total
K-105		20	15	6	17		12	14	84
L-106	20	19	16	11	17	10	9	6	108
E-107									0
F-108		16							16
S-109	11	24			15				50
F-110	17	18		12		3			50
K-111	18	17							35
D-112		13				7			20
N-113		4			3				7
S-114					9	4			13
N-115	3			8		6			17
L-116	22	18	11	10	17	11	8		97
Y-117	27	28	16		16		16		103
M-118		34	14	21	15	22	19		125
V-119	24	32	22	23	20	18	16	12	167
M-120	22	29				13		16	80
E-121	19	23	19	9	12	18	19	14	133
Y-122	25	33	15				17		90
V-123	21		8	10		19			58
A-124						6	7	6	19
G-125	12	22		3		12			49
G-126	18	25			10	23	2	2	80
E-127	10	17	9		19	25		10	90
M-128	20	23	13	14	18	21	10	10	129
F-129				13	27	23	8		71
S-130		14		7				8	29
H-131	17				7			4	28
L-132	18	10	7	6	16	14	6	8	85
R-133	8			4			3	6	21
R-134									0
I-135	8							4	12
G-136									0
R-137	7								7
F-138	18				11	5			34
S-139					5	9		2	16
E-140		18				11			29
P-141			4				4		8
H-142				5					5
A-143	30	27				13	2		72
R-144		18	11	8	2				39
F-145	25	22							47
Y-146	37	21	12			23			93
A-147	34			19			7	7	67
A-148	35	23					12	9	79
Q-149	27	25	14	15		11	5	8	105
I-150	33	35	17	15	15	12	7	9	143
V-151	40	29	18	20	23	22			152
L-152	31	23		17					71
T-153									0
F-154	30	24	15	18					87
E-155	21	15		17					53
Y-156	24		17	16		12	11	14	94
L-157	22	19	22	16	20	16	15	16	146
H-158	20	17	17	13	10	17		12	106
S-159	13	13	7			10		7	50
L-160		12		7	12	10	7		48
D-161					13		6		19
L-162	19	6	13	12	18	12		9	89
I-163	39	22	17	23	25		23	20	169
Y-164	43	23	22	20	30	27	25	21	211
R-165	48	27	24		30		25	26	180
D-166	36	21	30	23	35	27	25	23	220
L-167	39	22	19	19	34	23	19	13	188
K-168	33	24	17	17	28	24	15		158
P-169		32	16	19	33	23			123
E-170	19	30	12	17	30	24		2	134

	PKC	ROCK1	CDK2	SKY1	PHK	DAPK	SRC	IRK	Total
N-171	27	20	25	18	36	23	17	14	180
L-172	25	34	24	22	28	25	16	13	187
L-173	33	32	10	15	24	27	12	9	162
I-174	29	17	5	8	15	23	9	10	116
D-175	22	11	3		18				54
Q-176	22	8					3		33
Q-177									0
G-178	17	10							27
Y-179	26	19							45
I-180	31	22	3	11	20		3	4	94
Q-181	24	14	9	10	24	6	11	2	100
V-182	22	24	20	14	29	24		13	146
T-183					31				31
D-184	26	30	23	21	45	32	21	16	214
F-185	24	14	21	20	29	26	24	13	171
G-186	18	12	24	17	32	27	21	14	165
F-187			21		28	22	17		88
A-188			20	12	21	17	20		90
K-189	7		18				14	10	49
R-190		4			17	7			28
V-191									0
K-192									0
G-193									0
R-194	10				6				16
T-195	17								17
W-196									0
T-197									0
L-198	17				14				31
C-199	16				7				23
G-200	22	19			25				66
T-201	20	16	3	9	24	7			79
P-202	20	19			21	9			69
E-203	23	21		15	29	19			107
Y-204	27	30	14	22	34	22	6	12	167
L-205	28	15			33	11	17	4	108
A-206	37	21	21	23	20	14	7	15	158
P-207	27	15	15	13	23	16	9	13	131
E-208	23	12	12	11	18	12		12	100
I-209	25	7	13	16	13	13			87
I-210	14	2	6	2		2		3	29
L-211	9		9	4			3		25
S-212									0
K-213	4								4
G-214									0
Y-215	25	8		3	17				53
N-216	24								24
K-217		16			19				35
A-218	29		23					17	69
V-219	37		21		29				87
D-220	39	27	26	21	29	20	21	17	200
W-221	38	29							67
W-222	33	28	26	23	28	26	13	19	196
A-223	43	40	16	21	37	24	23	19	223
L-224	46	44	31			28	18	20	187
G-225	38	41	25		30	20	14	21	189
V-226	38	39			35	26	18	19	175
L-227	35	42	19	16	33	27	17	19	208
I-228	32	33		19	27		15	17	143
Y-229	31	27		18	28	23		17	144
E-230	29	31	15	21			6	13	115
M-231	32	28	14	25	26	17	8	14	164
A-232									0
A-233					13	10		5	28
G-234	19	13		12	9	12			65
Y-235									0
P-236					24				24

	PKC	ROCK1	CDK2	SKY1	PHK	DAPK	SRC	IRK	Total
P-237	23	18			24	15			80
F-238	16	16	2	5	15	6			60
F-239		6			3	3			12
A-240		3							3
D-241									0
Q-242									0
P-243									0
I-244		7							7
Q-245	8					5			13
I-246	6				3		2		11
Y-247	12	9							21
E-248									0
K-249		11							11
I-250	22	21			11	8			62
V-251			2				4		6
S-252									0
G-253									0
K-254									0
V-255		9							9
R-256					4				4
F-257	6	6			5	8			25
P-258	5							2	7
S-259						6			6
H-260						4			4
F-261	3		2						5
S-262		8				10			18
S-263						8			8
D-264	13	8	11	5					37
L-265				7	12		10	6	35
K-266	18	12	14		11	8			63
D-267	8			2	12	8	4	4	38
L-268	20	10	15	6	11	15	7	9	93
L-269	27	13	15	14	10	12	10	8	109
R-270						12		10	22
N-271									0
L-272	25	16	22	11	16	17	10	10	127
L-273	19	24	19	12	22	25			121
Q-274			2	8			4	4	18
V-275				12	19				31
D-276	11		3	7		6		2	29
L-277									0
T-278									0
K-279	9		6	7	7	6			35
R-280	22	16	11	7	20	16	6	10	108
F-281	25	19			24				68
G-282	16	20							36
N-283									0
L-284									0
K-285									0
N-286									0
G-287									0
V-288		16							16
N-289									0
D-290		7			8	7		4	26
I-291	14	25					3	7	49
K-292	21	19					4		44
N-293				2				2	4
H-294	17	16	6	5	9	13			66
K-295									0
W-296		7	7	6	10	13			43
F-297	8	7	4	5	3				27

Table S2. Description of the 22 active protein kinases used for alignment of the C-spines in Fig. 2*C*

Name	Group	Require phosphorylation	RD kinase	PDB ID	Chain
Aurora A	AGC	Yes	Yes	1OL5	A
PKA	AGC	Yes	Yes	2CPK	E
PKB	AGC	Yes	Yes	1O6K	A
PDK1	AGC	Yes	Yes	1H1W	A
CDK2	CMGC	Yes	Yes	1FIN	A
ERK2	CMGC	Yes	Yes	2ERK	-
CDK5	CMGC	No	Yes	1H4L	A
CDK6	CMGC	No	Yes	1JOW	B
GSK-3β	CMGC	No	Yes	1O9U	A
CK2	CMGC	No	Yes	1DAW	A
SKY1	CMGC	No	No	1HOW	A
IRK	TK	Yes	Yes	1IR3	A
IGF1RK	TK	Yes	Yes	1K3A	A
LCK	TK	Yes	Yes	3LCK	-
EGFR	TK	No	Yes	1M14	A
C-KIT	TK	No	Yes	1PKG	A
CSK	TK	No	Yes	1K9A	A
PHK	CAMK	No	Yes	2PHK	A
CHK1	CAMK	No	Yes	1IA8	A
DAPK	CAMK	No	No	1JKK	A
CK1	CK1	No	Yes	1CKJ	A
PKNB	Prokaryotic	Yes	Yes	1MRU	A

Table S3. Residues with high level of the Involvement Score (≥ 100) and their proposed functions (PKA sequence)

Residue	Total Involvement Score	Secondary structure	Function
L-49	104	β1	ATP binding
V-75	146	β2	C-spine
A-70	165	β3	C-spine
M-71	108	β3	Conserved hydrophobic feature of the N-lobe
K-72	170	β3	ATP binding, αC-helix anchoring
E-91	112	αC	Anchors the αC-helix to K-72
I-94	103	αC	Anchors the αC-helix to the R-spine and β6
L-103	172	αC-β4 loop	Anchors the αC-β4 loop to the R-spine
L-106	108	β4	R-spine
M-118	125	β5	Conserved hydrophobic feature of the N-lobe
V-119	167	β5	Conserved hydrophobic feature of the N-lobe
E-121	133	β5-αD loop	Conserved feature of the linker region
M-128	129	αD	C-spine
Q-149	105	αE	Anchoring β8 to αE
I-150	143	αE	Anchoring β8 to the F-helix
V-151	152	αE	Anchoring αE to αF and αI
L-157	146	αE	Anchors the E-helix to the R-spine
H-158	106	αE	Anchors the E-helix to the F-helix
I-163	169	β6	Supports side chain of R-165
Y-164	211	Catalytic Loop	R-spine
R-165	180	Catalytic Loop	Activation Segment anchoring
D-166	220	Catalytic Loop	Catalysis
L-167	188	Catalytic Loop	Anchors the Catalytic Loop to the F-helix
K-168	158	Catalytic Loop	Catalysis
E-170	134	Catalytic Loop	Substrate binding
N-171	180	Catalytic Loop	Catalysis
L-172	187	β7	C-spine
L-173	162	β7	C-spine
I-174	116	β7	C-spine
Q-181	100	β8	? Possible anchoring to β5-αD loop via E-121
V-182	146	β8	Activation Segment anchoring
D-184	214	Activation Segment	Catalysis
F-185	171	Activation Segment	R-spine
G-186	165	Activation Segment	Positioning of D-184
Y-204	167	Activation Segment	Substrate binding, Activation Segment anchoring
L-205	108	Activation Segment	Substrate binding
A-206	158	Activation Segment	Activation Segment anchoring
P-207	131	Activation Segment	Activation Segment anchoring
E-208	100	Activation Segment	Activation Segment anchoring
D-220	200	αF	R-spine anchoring
W-222	196	αF	Activation Segment and αH-helix anchoring
A-223	223	αF	Catalytic Loop anchoring
L-224	187	αF	Catalytic Loop anchoring
G-225	189	αF	αH-helix anchoring
V-226	175	αF	Activation Segment and αH-helix anchoring
L-227	208	αF	C-spine
I-228	143	αF	αH-helix anchoring
Y-229	144	αF	αH-helix anchoring
E-230	115	αF	Substrate binding
M-231	164	αF	C-spine
L-269	109	αH	αH-helix anchoring
L-272	127	αH	αH-helix anchoring
L-273	121	αH	αH-helix anchoring
R-280	108	αH-αI loop	αH-αI loop anchoring

Residues with IS > 140 are shown in bold font.

Table S4. Amino acid substitution matrix used in the LSP alignment

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
R	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
N	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
D	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0
Q	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0
E	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
I	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	1
L	0	0	0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	0	0	1
K	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
M	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	1	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
S	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0
Y	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0
V	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1

Residues were considered similar if they have score ≥ 12 in the optimized matrix presented by Hourai *et al*. [Hourai Y, Akutsu T, Akiyama Y (2004) Optimizing substitution matrices by separating score distributions. *Bioinformatics* 20:863–873]. We also took into account hydrophobicity of cysteine [Yang J, *et al.* (2005) Allosteric network of cAMP-dependent protein kinase revealed by mutation of Tyr204 in the P+1 loop. *J Mol Biol* 346:191–201.26] and considered it to be similar to methionine, leucine, and isoleucine.