Comparative surface geometry of the protein kinase family

Elaine E. Thompson,¹ Alexandr P. Kornev,² Natarajan Kannan,^{3,4} Choel Kim,² Lynn F. Ten Eyck,^{1,5*} and Susan S. Taylor^{1,2,6}

¹Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, CA 92093

²Department of Pharmacology, Baylor College of Medicine, Houston, TX 77030

³Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602-7229

⁴Institute of Bioinformatics, University of Georgia, Athens, GA 30602-7229

⁵San Diego Supercomputer Center, University of California at San Diego, La Jolla, CA 92093

⁶Department of Pharmacology, University of California at San Diego, La Jolla, CA 92093

Received 23 March 2009; Revised 15 June 2009; Accepted 24 June 2009

DOI: 10.1002/pro.209

Published online 16 July 2009 proteinscience.org

Abstract: Identifying conserved pockets on the surfaces of a family of proteins can provide insight into conserved geometric features and sites of protein–protein interaction. Here we describe mapping and comparison of the surfaces of aligned crystallographic structures, using the protein kinase family as a model. Pockets are rapidly computed using two computer programs, FADE and Crevasse. FADE uses gradients of atomic density to locate grooves and pockets on the molecular surface. Crevasse, a new piece of software, splits the FADE output into distinct pockets. The computation was run on 10 kinase catalytic cores aligned on the α F-helix, and the resulting pockets spatially clustered. The active site cleft appears as a large, contiguous site that can be subdivided into nucleotide and substrate docking sites. Substrate specificity determinants in the active site cleft between serine/threonine and tyrosine kinases are visible and distinct. The active site clefts cluster tightly, showing a conserved spatial relationship between the active site and α F-helix in the C-lobe. When the α C-helix is examined, there are multiple mechanisms for anchoring the helix using spatially conserved docking sites. A novel site at the top of the N-lobe is present in all the kinases, and there is a large conserved pocket over the hinge and the α C- β 4 loop. Other pockets on the kinase core are strongly conserved but have not yet been mapped to a protein–protein interaction. Sites identified by this algorithm have revealed structural and spatially conserved features of the kinase family and potential conserved intermolecular and intramolecular binding sites.

Keywords: kinase; protein kinase A; pockets; protein surface

The clustering analysis for this article was generated using SAS software, Version 9.1 of the SAS System for Windows. Copyright © 2002–2003 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Grant sponsor: National Institute of Health; Grant number: GM19301; Grant sponsor: National Institute of General Medical Sciences; Grant number: R01GM070996; Grant sponsor: UCSD Bioinformatics Graduate Program.

*Correspondence to: Lynn F. Ten Eyck, San Diego Supercomputer Center, University of California at San Diego, La Jolla, CA 92093. E-mail: teneyckl@sdsc.edu

Introduction

The surface of a protein is crucial for intermolecular function because interactions with other proteins, ligands, and nucleic acids all happen at the surface. The specificity at the site is determined by local surface properties including charge, the availability of hydrogen bonding, hydrophobicity, and shape. Families of proteins often have conserved binding sites, such as the ligand binding site in G-protein coupled receptors, or the ATP binding site in the kinase family. Similarly shaped pockets on structurally related surfaces suggest similar function, while differences in the

shape and location of pockets can suggest diversity within the family. To detect similarities in the spatial organization and location of pockets on the protein surface, we have developed a method for spatially clustering pockets computed on aligned protein crystallographic structures. Here, we present results that show conserved interactions at the surface of the protein kinase catalytic core, including regulatory elements and substrate recognition.

Protein kinases are of interest because they are essential to cellular function and survival, and central to multiple signal transduction pathways. The importance of the kinase family in cancer was discovered in the 1970s when the oncogenic protein in Rous Sarcoma Virus, v-Src, was determined to be a constitutively active protein tyrosine kinase.¹ Kinases have since been extensively studied and have become targets for oncology drug development. Fourteen tyrosine kinase inhibitors have either been FDA approved or are in clinical development as new chemotherapeutic agents.² Comparative studies of the kinase family have been undertaken both within and across species using sequence similarity. The human kinome has at least 518 putative kinase genes, with 244 mapped to known disease loci.³ Kinases share a conserved catalytic core where ATP and substrate bind and phosphoryl transfer occurs. The catalytic core may be the majority of the molecule, as with cyclin dependent protein kinase (CDK2), or it may be only a small part of a much larger molecule, as in the receptor tyrosine kinases. Catalysis is regulated by inter- and intramolecular interactions with the conserved catalytic core. Structural comparisons of the kinase family have been undertaken, but they are largely limited to main chain structural alignments. Here we compare and contrast pockets on the conserved catalytic core of the kinase family to provide insight into substrate recognition, activation mechanisms, and possible conserved sites of protein–protein interaction.

Pockets are initially identified using FADE (Fast Atomic Density Evaluator)⁴ and a new computer program we have named Crevasse. FADE maps the surface of a protein by computing the local atomic density gradient, which is closely related to the local surface curvature. Crevasse segments the FADE output into individual pockets. FADE does not seek to reproduce the well-established methods for finding small, “drugable” pockets such as Q-SiteFinder,⁵ AutoLigand,⁶ or DrugSite⁷ but rather identifies all the concavities on a protein surface. Along with ligand binding sites, we have tuned FADE to identify larger pockets and grooves on the protein surface that could be sites of protein–protein interactions. Crevasse is new software written as companion software for FADE. It improves the FADE signal-to-noise ratio by discarding isolated outlying points and clustering the remaining points into distinct binding sites and grooves. A further level of clustering applied to all of the pockets on the set of

studied kinases shows surface features that are spatially conserved across the family. The method is purely geometric, so sites with very different amino acids, and therefore different surface properties, can still be identified as conserved.

First, the computation is shown on a new, high-resolution structure of Protein Kinase A (PKA). In addition to the catalytic core, PKA has well-studied C-terminal and N-terminal tails that are essential for kinase activity.^{8,9} We show that these cis-regulatory elements fit into well-defined pockets on the conserved catalytic core. Second, a set of 10 active kinase crystal structures was chosen for comparisons. Active kinases were used because the active catalytic core has high spatial conservation.^{10,11} The consistency of the active kinase structure makes surface geometry comparisons possible without needing to align different parts of the kinase substructure to compare pockets. All of these kinases have well-described interactions between the kinase cores and protein regulatory agents such as subdomains and subunits that allow us to test the predictions made by the software and identify areas of interest on other kinases.

The pocket clustering shows similarities between regulatory mechanisms in the kinase family. Recent work has shown that the kinase catalytic core is organized around the α F-helix, with two conserved spines of four amino acids each connecting the α F-helix and the active site cleft.¹¹ By aligning the proteins on the α F-helix and running the computation, we can test the spatial location of the active site cleft relative to the α F-helix. Differences between the substrate binding sites of serine/threonine and tyrosine kinases are apparent. This is the first algorithm to discriminate between serine/threonine and tyrosine kinase catalytic cores based solely on the shape of the peptide binding site.

There are mechanisms for stabilizing the α C-helix that have been described on the EGF receptor kinase and the cyclin/cyclin dependent kinase (CDK2) complex,¹² and in the AGC kinase family.⁹ Using FADE and Crevasse, we are able to compute a cluster of pockets at the α C-helix that shows the similarities between the mechanism of α C-helix stabilization in CDK2, PKA, mitogen activated protein (MAP) kinase, and death associated protein kinase (DAPK). In addition to highlighting known features of the family, the software can identify orphan pockets without known inter- or intramolecular interactions. Protein kinases have unmapped interactions with regulatory proteins, so these sites could be considered for sites of interaction with other proteins.

Results

Figure 1 shows a visual representation of the computational results on the PKA catalytic core. On the left, the individual grid points are shown as small spheres. The active site cleft and substrate binding sites are

Figure 1. Representations of FADE and Crevasse computations. Left, computation showing individual points returned by FADE. Right, pockets represented as surfaces over the points. The surfaces were rendered in PyMOL.

lined with points. The right-hand side shows a different representation of the data. The points are rendered as a surface in PyMOL, with a Van der Waals radius setting of 0.6 Å. The surface representation is useful to show amino acids or ligands in the pockets.

Protein Kinase A

On the PKA catalytic core, the algorithm identifies two distinct pockets adjacent to the α C-helix, one at the middle of the N-lobe surface of the helix and a second

near the C-terminal end of the C-lobe surface of the helix. Figure 2 shows the computationally identified pockets on the kinase core and the how they are occupied in the complete structure. The pocket on the N-lobe surface of the α C-helix is filled by the FxxF motif formed by residues 347–350 from the C-terminal tail. Residues F26 and W30 from the N-terminal tail fit into the pocket on the C-lobe surface, and stabilize the α C-helix from the C-lobe surface.

Purified mammalian PKA has been shown to have a myristic acid covalent modification¹³ on the N-terminal tail. Zheng identified an acyl binding pocket adjacent to the n-terminal tail, and when nonmyristylated recombinant PKA crystals are grown in the presence of octanoyl-*N*-methylglucamide (MEGA-8) detergent, a detergent molecule occupies the site.¹⁴ When recombinant PKA is crystallized in the absence of detergent, the acyl pocket is still present, and is occupied by solvent. Crevasse successfully identifies the empty acyl myristylation site on the recombinant enzyme as a pocket. When the algorithm is run, co-crystallized molecules such as detergents, water molecules, and glycerol are removed from the protein. Examination of the PKA structure with MEGA-8 shows a computationally identified pocket in the same spot that is neatly filled by the detergent molecule. The pocket is completely absent in the myristylated mammalian protein, where the computation was run with the myristic acid included. The myristic acid

Figure 2. Cis-regulatory elements on protein kinase A. A: N-terminal tail, showing an empty pocket at the myristylation site. B: Myristylation site occupied by Mega-8 (green). Mega-8 was removed from the PKA molecule for the computation. C: Myristylated PKA. The myristic acid is shown in yellow and fills the pocket. D: Pockets at the α C-Helix. Pockets are identified on the catalytic core where the C-terminal FxxF motif binds, and where the N-terminal F26 and W30 bind. The pockets highlight the importance of geometric fit in the cis-regulation of PKA.

Figure 3. Kinases relationships and active site cleft. A: The kinome tree,³ with the 10 kinases marked. The colors are indicated in part B. B: Clustering results on the active site cleft. The second column is the distance in angstroms of the pocket center from the *k*-means cluster centroid. The cluster centroid is the computed center of the group of pockets. C: All 10 kinase cores aligned on the α F-helix. D: Pocket computation on all 10 kinases, showing the active site cleft and a conserved pocket at the top of the N-lobe. Points from each kinase are colored as indicated in part B.

packs well enough to smooth the surface of the protein and fill the acyl pocket. This demonstrates the ability of Crevasse to identify potential small molecule binding sites, and shows the stability of the acyl site in nonmyristylated PKA. The empty pocket under the N-terminal tail present in all three structures is also of interest. Upon RII subunit binding, the N-terminal tail has been shown to have increased backbone flexibility, with the myristyl group flipping out to associate with membranes.¹⁵ Movement of the N-terminal tail would expose the pocket for a potential protein-protein or protein-membrane interaction.

Kinase Clustering

Ten active kinase crystal structures in closed conformations were chosen to test the pocket clustering methodology. The relationship of the kinases to the human kinome is illustrated in Figure 3(A) on a map of the human kinome.³ The AGC, CMAK, CMGC, and TK families are all represented. The kinases were aligned on the conserved α F-helix backbone, shown in Figure 3(B) and the computation run in all the kinases. The α F-helix was chosen because it has been

shown to be a highly conserved element in the kinase catalytic core.¹⁰ There are two sets of spatially conserved hydrophobic amino acids, termed spines, which originate at the α F-helix and end in the active site cleft. Clustering on a small, spatially conserved element rather than by backbone RMS alignment preserves the relationships of the surface pockets to the catalytic machinery.

The kinase active site cleft was identified by the software as the largest crevice in all 10 proteins and the centers of the active sites clustered tightly. Computational results for the active site cleft are depicted in Figure 3(D), with results from each kinase shown in a different color. Figure 3(B) shows the clustering results for the active site cleft. The active site is 30 Å long and irregularly shaped, yet the site centers in all 10 kinase cores fall no more than 2.57 Å from the cluster centroid. In all 10 kinases, the nucleotide and peptide binding sites are identified as a single large, geometrically contiguous active site cleft with similar geometry. This demonstrates a universal, precise relationship of the active site cleft to the α F-helix.

N-lobe Cap

At the top of the N-lobe, there is a pocket present on all 10 kinases. Figure 4(A) shows the overlaid points from the computation, with each kinase in a different color. Differences in the position of the pocket are partly reflective of subtle differences in the position of the glycine loop, since the kinases were aligned on the α F-helix rather than the N-lobe. The pocket tends to be above the beta sheet, located above a residue

Figure 4. N-lobe Cap. N-lobe Cap on all 10 kinases. A: All of the points returned by the computation are shown on PKA ribbons, colored according to Figure 3(B). Three examples of N-lobe caps are shown on PKA, ERK2, and PHK. The pocket is shown as a gold surface.

Table I. *N-lobe Cap Sequences*

Protein	Residue	Subdomain I
CDK2	1	MENFQKVEKIGEGTYGVVYKARNKL
ERK2	17	EDVGPRTNLSYIGEGAYGMVCSAYDNL
DAPK	8	NVDYDYDTGEELGSGQFAVVKCREKS
PHK	15	FYENYEPKEILGRGVSSVVRRCIHKP
SRC	262	IPRESLRLEVKLGGCGFGEVWMTWNG
LCK	240	VPRRTLKLVERLGAGQFGEVWMTWNG
CSK	189	ALNMKELKLLQTIGKGEFGDVMLGDYRG
IRK	991	VSRKELITLRELGGQSGFGMVYEGNARD
PKA	40	LDQFDRIKTLGTGSFGRVMLVKHKE
Sky1P	153	EPYARYILVRKLGWGHFSTVWLAKDMV

The 10 kinase sequences are shown aligned on Hanks and Hunter¹⁶ subdomain I. The residues in white on black have side chain or backbone atoms that fall into the computationally identified pocket.

homologous to PKA M71. This location is of interest, because it is adjacent to K72 and A70. K72 is a highly conserved residue that hydrogen bonds to ATP, and A70 is the top of the C-spine. Because of the highly conserved location of A70 and K72, the position of M71 is strictly defined. The structural motif of residues filling the top of the N-lobe over the spines is a new feature of the kinase core. When the 10 kinases are examined, all of them have between one and four N-terminal residues that occupy the computationally defined pocket. Figure 4 shows specific examples of the motif on PKA, ERK2, and PHK. Table I shows the kinase sequences, aligned by homology on Subdomain I to sequences in Hanks and Hunter.¹⁶ The residues on a black background in the table had atoms in the pocket. There is no clear sequence homology between the residues outside of closely related sequences like IRK and LCK; the N-lobe Cap motif is purely geometric. Our results suggest that amino acids N-terminal to the β_1 strand of the N-lobe should be studied further for a possible structural role in the kinase core.

Substrate Specificity

The computation showed a difference between tyrosine and serine/threonine kinases outside of the active site cleft where phosphorylation substrates bind. Figure 5(B) shows the results of the clustering computation. Two different clusters of pockets were identified, one cluster with only serine/threonine kinases and the other with only tyrosine kinases. PHK has two sites in the cluster because Crevasse separated the points into two small, adjacent sites rather than one larger site. The two sets of clustered pockets are depicted in Figure 5(C,D), with the different colored points representing different kinases. Three of the kinases in the set of 10 have peptides co-crystallized at the active site cleft. The PKA crystal structure used in the clustered experiment was chosen because it has a very good resolution, but lacks ATP. To examine the PKA active site, FADE and Crevasse were run on residues 41–297 of a new crystal structure with ATP and residues 5–24 of

the protein kinase inhibitor peptide PKI. The active site cleft, pictured in Figure 5(A), shows the extremely important contacts between PKI and PKA that occur in the pocket. The specificity determinants for PKA phosphorylation are arginine residues two, three, and six residues before the substrate serine or threonine. The residues are termed P-2, P-3, and P-6. All three of these arginine residues occupy pockets on the PKA surface. The pocket unique to the serine/threonine kinase family is occupied by the P-2 and P-6 arginines, and the P-3 arginine falls into the large active site cleft pocket. The rest of the active site cleft is filled by F327 from the C-terminal tail.

PHK is also co-crystallized with a short inhibitor peptide. Similar to PKA, specificity determinants are located in the pocket. The P-2 asparagine residue occupies the pocket, and the P-3 arginine falls into the active site cleft. The peptide ends at P-3, so the binding of the rest of the specificity determinants cannot be inferred from the structure but the pocket has room for additional amino acids. In contrast, when the full IRK structure is overlaid with the core used in the computation, the backbone of the substrate tyrosine at the phosphorylation site falls into the computationally identified pocket. The pocket serves to extend the active site cleft so that the substrate peptide backbone is positioned for the bulkier tyrosine residue to fit into the phosphorylation site. The geometric fit of the peptide backbone and tyrosine residue may be part of the specificity determinant between serine/threonine and tyrosine kinases.

Figure 5. Substrate specificity determinants. Results of the Crevasse computation. A: The computation shows a large, contiguous active site on the new, high-resolution structure of PKA (PDB ID 3FJQ). The pocket is shown on PKA ribbons as a translucent gold surface. B: Computational results showing two separate clusters of pocket centers. As in Figure 3(B), the distance column is the distance in angstroms of each pocket center from its respective *k*-means cluster centroid. C: Cluster identified on serine/threonine kinases, pictured on PKA ribbons. D: Cluster identified on tyrosine kinases, pictured on IRK ribbons. The points are colored by kinase, using the colors in Figure 3(B).

Figure 6. α C-helix pockets and amino acids filling them. The four proteins shown each have different regulatory elements that fill pockets at the α C-helix. Computed pockets are shown as gold surfaces, the N-lobe in grey, and amino acids outside the kinase core are shown in red. Cyclin is in teal. ERK2 and CDK2 are in a different orientation to show the phenylalanine rings clearly. CDK2 has pockets filled by cyclin, ERK2 by its C-tail, DAPK by its N-tail, and PKA by both C- and N-tails.

α C-Helix Stabilization

Pockets were identified on multiple kinases at the N-lobe and C-lobe surfaces of the α C-helix. There are pockets on both surfaces in PKA, ERK2, IRK, and DAPK. There are also single pockets on c-Src, CSK, and PHK. As the PKA FxxF motif falls into the pocket on the PKA core,⁹ we examined the clusters of pockets near the α C-helix in the aligned structures. Amino acids occupy the pockets in four of the serine/threonine kinases, PKA, DAPK, CDK2, and ERK2. The α C-helices and the amino acids that occupy the pockets are pictured in Figure 6. As shown before, the pockets on PKA are occupied by the C-terminal and N-terminal tails. In ERK2, a long extension of the C-terminal tail occupies both the N-lobe surface and C-lobe surface pockets. Residues F327, M331, and D334 occupy the N-lobe surface, and an α -helix formed by residues 341–350 lies along the C-lobe surface of the α C-helix. Residues in the pocket are L341, I345, T349, and F352. T349 forms a hydrogen bond to R89; the rest of the interactions within the computationally defined pocket are nonpolar. In DAPK, the short N-terminal tail folds over the N-lobe and fills the pocket. Amino acids in the pocket are T1 and F3. In CDK2, there is a similar interaction, only the α C-helix pockets are occupied by amino acids from cyclin. The N-lobe surface pocket is occupied by cyclin residues L299 and F304. Residue H296 is not in the computational pocket, but

it is next to the α C-helix. Cyclin residues H296 and F304 are structurally homologous PKA residues F346 and F350. The C-lobe surface pocket is occupied by cyclin residues F267, E268, and I270. PHK has nothing occupying the pocket, but PHK has very few amino acids in the crystal structure beyond the kinase core. Our analysis suggests that another protein might interact at that site on PHK. There is no crystal packing near the α C-helix, and water molecules are in the pockets.

The pocket was not detected in the computation on Sky1P. Sky1P is the only constitutively active serine/threonine kinase of the six and it has a unique, extended α C-helix.¹⁷ Rather than a direct interaction with the α C-helix, the Sky1P C-terminal tail binds in a different site on the C-lobe of the kinase to stabilize the α C-helix. Similar to other kinases with cis-regulatory elements, deletion of the tail results in a loss of the constitutive activity. In IRK, the pocket on the N-lobe surface of the α C-helix is filled in the crystal by the N-terminus of an adjacent molecule. A common feature in the amino acids packed against the N-lobe surface of the α C-helix in these structures is the presence of at least one phenylalanine residue.

A small computationally distinct pocket was present at the C-terminal end of the α C-helix on two of the four tyrosine kinases, c-Src and IRK. The pocket is filled by a highly conserved tryptophan residue, W260 in c-Src. Because Crevasse was tuned to select larger

Figure 7. Orphan pockets. Pockets were found on multiple kinases that were not filled in the crystal structures. Points are colored by kinase, as shown in Figure 3(B). There is a pocket in all 10 kinases over the hinge region and α C- β 4 loop. A second pocket is in all 10 kinases at the N-terminal end of the α F-helix. The third pocket is over the DE loop and the C-terminal end of the α F-helix. It is only absent in PKA.

pockets that might be sites of protein–protein interaction, a second computation with a shorter, 4 Å length cutoff was run to look for smaller pockets at the α C-helix of the other two kinases. In that computation, a third pocket was found on LCK. In c-Src, W260 is positioned at the C-terminal end of the α C-helix with two hydrogen bonds to D258, and E97 from the SH3 domain. In inactive c-Src, the α C-helix is moved away from the kinase core, and W260 is positioned between the α C-helix and the kinase core.¹⁸ There is no hydrogen bond to the tryptophan nitrogen atom. When the four tyrosine kinases are aligned on the α C-helix rather than the α F-helix, the conserved tryptophan at the C-terminal end of the helix is tightly structurally conserved in c-Src, IRK, and LCK. All three kinases also have a nearby glutamic acid residue within hydrogen bonding distance of the tryptophan backbone. CSK shares a homologous tryptophan, but no computational pocket. The position of the CSK tryptophan is slightly different, and the glutamic acid residue is not present in the structure. Instead, F183 is packed next to the C-lobe surface of the α C-helix similar to the serine/threonine kinases. The results suggest that c-Src and other tyrosine kinases have a different mechanism of α C-helix stabilization. In Hck, a Src family kinase, mutation of the homologous tryptophan to alanine increases activity and disrupts the activation by ligand binding to the SH2 and SH3 domains. All that is required for full activity is autophosphorylation.¹⁸ In contrast, mutation of W362 in c-Raf inactivates the kinase.¹⁹

Orphan Pockets

There are conserved pockets on the kinase core that cannot be related to a known function. Some are present on all the kinase cores in the survey, and others show some selectivity. Three sites are depicted in Figure 7. There is a cluster of pockets over the α C- β 4 loop and the N-lobe and C-lobe linker region that is present on all 10 kinases. None of the 10 crystal structures have anything occupying the pocket. There are two large pockets common to most of the kinases on the C-lobe as well. There is a large orphan pocket located over the loop between the D and E helices, and at the C-terminal end of the α F-helix. This pocket is present in 9 of the 10 kinases, and is missing only in PKA. There is nothing crystallized at that particular site except in c-Src. In c-Src, the last amino acid on the C-terminal tail, L533, folds into the identified site. The function of the site and the reason it is missing on PKA are difficult to determine from these data.

The second C-lobe orphan pocket is on the other side of the α F-helix. It is located at the N-terminal end of the α F-helix and the C-terminal end of the α E-helix. The pocket is near where the PKA regulatory subunit binds in the holoenzyme, but is not occupied by regulatory subunit amino acids in any structure crystallized thus far. This site in Sky1P is filled by the Sky1P C-ter-

minal tail. The pocket is occupied by I729, W732, and part of F733. As mentioned earlier, the Sky1P C-terminal tail is important in positioning the α C-helix¹⁷ and required for the constitutive activity.

Methods

FADE and Crevasse

The protein surfaces are first mapped using FADE. FADE uses the gradients of radial counting functions to characterize the shape of the molecular surface without actually computing the surface. FADE counts the number of atoms within spheres of varying radii to estimate the gradient. Atomic centers from a protein crystal structure are mapped onto a grid and the atomic density gradient is computed at each grid point. The atomic density gradient contains information about the protein surface and can map changes in surface curvature. By restricting the output to points where the FADE score is high and removing interior points, FADE can output a set of points that lines crevices and grooves on the molecular surface. FADE is fast, full source code is available, it can run successfully on any structure that fits into memory, and it is potentially extensible to any molecular property that can be mapped onto a grid.

In the original publication, FADE mapped crystallographic structures onto a grid with 1.0 Å spacing. The input set included only those atoms that were resolved in the crystallographic structures. Hydrogen atoms, which are not usually visible in X-ray crystallographic structures, were not included in the computation. The software was able to map grooves on the protein surface, but it did not fill ligand pockets well, or flood deeper pockets differently from shallower ones. The FADE results are improved dramatically by using finer grids and computationally adding hydrogen atoms to the structure. Because each atom is counted irrespective of type, adding hydrogen atoms improves the FADE computation by filling the interior of the protein more completely and representing the number of atoms at the molecular surface more accurately. The finer grid provides a better representation of atomic density and more detail in the output.

Crevasse was written as companion software for FADE. Raw FADE output is a set of grid points that line the surfaces of pockets, and it can be noisy. Crevasse is a small piece of software that segments the FADE output into individual pockets. It reads the FADE grid, finds sets of connected points, calculates the longest axis through the points and filters the pockets to a user-defined minimum number of points and/or axis length. The number of connections and the way it defines connections is also user-configurable. As part of the computation Crevasse finds the mean of the points, which can be output as an estimate of the center of the pocket. Crevasse is available for free download at <http://www.sdsc.edu/CCMS/>.

Table II. List of Kinases Used in the Comparison

Type	Abbrev.	Name	PDB ID	Residues	α F-Helix
Ser/ Thr	CDK2	Cyclin-dependent protein kinase	1FIN (A)	2–286	183–195
	DAPK	Death-associated protein kinase	1JKK	11–275	197–209
	PHK	Phosphorylase kinase	2PHK	17–287	209–221
	PKA	Camp-dependent protein kinase	2CPK	41–297	218–230
	Sky1P	Sky1p SR protein kinase	1Q97 (A)	156–706	584–596
Tyr	ERK2	MAP kinase	2ERK	21–311	206–218
	IRK	Insulin receptor kinase	1IR3	994–1263	1189–1201
	CSK	Carboxyl terminal Src kinase	1K9A (A)	193–443	366–378
	LCK	Lymphocyte-specific protein kinase	3LCK	243–494	420–432
	c-Src	c-Src Proto-oncogene	1Y57	265–516	442–454

The kinases used in the comparison are listed in Table I, with their PDB ID, chain, and the residues defined as the core. The α F-helix residues listed were used for the structural alignments.

Kinase Processing

Ten kinase core X-ray crystal structures were selected for the comparison, six serine/threonine and four tyrosine kinases. Only active structures in a closed conformation with the R-spine in a catalytically competent position were used.¹¹ The structures were truncated to include only the catalytic core. The alignment and subdomain definition provided in Hanks and Hunter¹⁶ was used to determine where to truncate the proteins. The beginning of subdomain I was defined as a residue homologous to PKA D41, and the end of subdomain XI was defined as homologous to PKA F297. Water molecules, ligands, and metals were removed, leaving only the polypeptide. Alternate atom locations were also removed. No attempt was made to rebuild missing side chains or side chain atoms because of the difficulty of determining a location for the rebuilt atoms. Hydrogen atoms were added using Reduce software on default settings.²⁰ All 10 kinase cores were pairwise structurally aligned on the backbone atoms of the α F-helix against PKA residues 218–230 using PyMOL.²¹ The PDB IDs, names, and residues of the 10 kinases are listed in Table II.

After the crystal structures were aligned and processed, FADE and Crevasse were run on each structure. FADE was set to use 0.5 Å grid spacing and return points 1.5 Å from atomic centers with scores above 4.6. These FADE settings return points that are very close to the Van der Waals surface of the protein, and only in regions with strong convex surface curvature. Crevasse was set to discard groups of points with fewer than 80 points or with the longest axis through the points shorter than 5 Å. The settings for FADE and Crevasse were determined by examining the distribution of FADE scores and the sizes of the resulting pockets. As part of the Crevasse computation, the center of each crevice is computed as the mean of the points. To identify groups of potentially conserved binding sites, the crevice centers were spatially clustered using a *k*-means procedure in SAS® software.^{22,23} New clusters were initiated when the Euclidean distance between a point in the cluster and the

seed exceeded 10 Å. The pocket centers clustered into 30 distinct spatial clusters. Many of the clusters high-light regions of known conserved function.

Crystallography

The catalytic subunit was expressed and purified in *E. coli* as described previously.²⁴ The catalytic subunit (Cat) was dialyzed against 10 mM MOPS buffer (pH 7.0), 50 mM NaCl, 3 mM MnCl₂, and 0.2 mM ATP. Approximately 10-fold molar excess amounts of IP20 (PKI peptide 5–24) was added at the time of crystallization to the protein. The Cat:IP20 complex was crystallized at 4°C in hanging drops using the vapor diffusion method in 0.5M (NH₄)₂SO₄, 0.1M HEPES (pH 7.5), and 10% MPD. The crystals were transferred to a cryoprotectant solution (mother liquor containing 15% glycerol) and flash cooled in liquid nitrogen. X-ray diffraction data were collected at beamline 8.2.1 (Advanced Light Source, Berkeley, CA). Diffraction data were processed and scaled using HKL2000.²⁵ The final data were integrated and scaled in P2₁2₁2₁ (*a* = 57.5 Å, *b* = 80.5, *c* = 96.6 Å) with satisfactory statistics. Initial phases of the Cat:IP20 complex were generated by molecular replacement using the same structure solved reported in 1993 (Protein Data Bank ID code 1ATP²⁶) as a search model after the protein kinase inhibitor peptide removed using Phaser.²⁷ The C-subunit was clearly located in the asymmetric unit showing a rotational *Z*-score of 30.5 and a translational *Z*-score of 40.9. Electron densities for the protein kinase inhibitor were clearly visible except for the residues 25. A combination of simulated-annealing composite omit and sigmaA-weighted 3Fobs-2Fcalc, 2Fobs-Fcalc and Fobs-Fcalc maps were used for model building in COOT.²⁸ The final refinement implementing TLS refinement²⁹ followed by manual model building resulted in a model with *R*_{work} and *R*_{free} of 18.01 and 20.4 respectively. The coordinates were deposited in the Protein Data Bank under the ID code 3FJQ. The crystallographic data and refinement factors are listed in Table III.

Table III. Data and Refinement Statistics

Data Set	SSRL Beamline 8.2.1
Space group	P2 ₁ 2 ₁ 2 ₁
Cell constants (Å)	<i>a</i> = 57.517; <i>b</i> = 80.548; <i>c</i> = 97.622
Number of crystals	1
Wavelength (Å)	1.00
<i>D</i> _{min} (Å)	1.6
Mosaicity (°)	0.23
Unique reflections	57,176
Average redundancy	12.3
<i>R</i> _{sym} (%) ^a	9.2 (48.3) ^b
Completeness (%)	99.5 (87.3) ^b
<i>I</i> / <i>σI</i>	15.3 (1.5)
Resolution range for refinement (Å)	50–1.6
Total reflections used	57,176
Number of protein atoms	3252
Number of water molecules	413
Rmsd bond lengths (Å)	0.007
Rmsd bond angles (°)	1.128
<i>R</i> _{work} (%)	18.01
<i>R</i> _{free} (%)	20.4
Average <i>B</i> factor (Å ²)	15.453

Crystallographic data and refinement statistics for the high-resolution crystal structure of the Protein Kinase A catalytic subunit with ATP and PKI 5-24.

^a $R_{\text{sym}} = \frac{\sum h \sum i |I(h) - I(h)i|}{\sum h \sum i I(h)i}$, where $I(h)$ is the mean intensity after rejections.

^b Numbers in parentheses correspond to the highest resolution shell of data, which were 1.72–1.6 Å.

Conclusions

Clustering surface pockets within a protein family can be used to make comparisons and find new features within a related group of structurally homologous proteins. In the kinase family, clustering FADE and Crevasse results shows both similarities and differences between the serine/threonine and tyrosine kinases. The tight active site cleft cluster both validates the pocket clustering methodology and demonstrates the strong degree of conservation of active kinase catalytic cores. The spatial relationship between the active site cleft and the α F-helix has been tightly conserved in these 10 kinases.

The cluster of pockets at the kinase α C-helix shows how this method can pick out similar mechanisms that do not necessarily share sequence homology. Four of the six serine/threonine kinases in this survey have something packed tightly against the α C-helix, presumably to maintain the active conformation. The pocket is conserved, but the mechanisms to fill it are quite diverse. PKA has both C- and N-terminal tails packed against the helix. DAPK has “solved” the problem with an N-terminal tail, while ERK has an extended C-terminal tail. CDK2 has a fourth mechanism, with cyclin packed in against the helix. There is specificity from kinase to kinase, as cyclin does not activate kinases other than CDK2; kinase tails do not promiscuously activate other kinases.

The pocket at the top of the N-lobe is a new finding for the kinase family. The pocket is occupied in all

10 of the active structures studied. It is tempting to speculate that the N-lobe cap plays a role in N-lobe stabilization because of its proximity to the conserved lysine (K72 in PKA) and the residue at the top of the R-spine. Packing at the top of the N-lobe could help maintain the orientation of these residues in the active conformation. More kinases should be studied to see whether the top of the N-lobe is consistently occupied in kinase structures. If so, the classical definition of the kinase catalytic core should be revised to include the amino acids that fold onto the top of the N-lobe.

Another new finding is the difference in shape at the substrate recognition site and peptide docking site. This algorithm shows an extra surface at the peptide docking site that may function to position the bulkier tyrosine residue at the phosphorylation site. It also shows a conserved pocket at the substrate recognition site in serine/threonine kinases. Recently a new structure of c-Fes with bound substrate was solved.³⁰ FADE and Crevasse were run on the core of c-Fes and a pocket matching the other tyrosine kinases in this study was present at the active site. Figure 8(A) shows phosphorylated c-Fes (PDB ID 3CD3), with the co-crystallized substrate in the computed pocket. This is the first method that has shown potential for differentiating between serine/threonine and tyrosine kinases by purely geometric methods. More structures with bound peptide substrates and inhibitors would have to be studied to see whether the difference is consistent,

Figure 8. FADE and Crevasse computation on c-Fes and dual specificity kinases. A: FADE and Crevasse were run the catalytic core, residues 559–815, of phosphorylated c-Fes. The pocket at the substrate binding site is shown as a gold surface. B: Computation on c-Fes with the SH2 domain organized around a sulfate atom. A large pocket is visible next to the sulfate ion. There is a series of crevasses around the α C-helix and leading to the active site that are large enough for kinase-substrate interactions. C: Computation on CLK1 (blue), DYRK1A (gold), and PKA (dark red) pictured on CLK1 ribbons with points returned by Crevasse drawn as spheres. The substrate recognition site has a similar shape to that of the serine/threonine kinases in Figure 5.

and whether it is maintained in open conformations and inactive kinases.

The algorithm was run on two dual specificity kinases, CLK1 (PDB ID [2VAG](#)) and DYRK1A, the tyrosine-phosphorylation-regulated kinase 1A (PDB ID [2VX3](#)). The dual specificity kinases autophosphorylate on tyrosine residues, but the target substrates for both CLK1 and DYRK1A are serine or threonine residues.^{31–33} The results, shown in Figure 8(C), show that these two kinases have a surface like the serine/threonine kinases, with a substrate recognition site similar to that of PKA and no extended area for tyrosine at the phosphorylation site. No peptide substrate is co-crystallized, so the question remains as to how these proteins accommodate the bulky tyrosine for autophosphorylation.

The large, conserved pockets with unknown functions (orphan pockets) are also potential areas of study. The hinge and α C- β 4 loop was identified as a possible regulatory region of the kinase core.³⁴ The observed orphan pocket supports the hypothesis of protein-protein interaction at that region of the kinase. The conserved C-lobe site where the Sky1p tail binds may also have functions in other kinases. It is close to where the PKA regulatory subunit binds. There are not full structures of all of the PKA holoenzymes available yet, so it remains to be seen whether the site is occupied in PKA. The third orphan pocket at the DE loop shows a difference between PKA and the rest of the kinases studied. It is a rather large pocket in the nine other kinases, but the surface on PKA is quite flat at that site.

Another interesting use of the algorithm is to form hypotheses about known protein-protein interactions with unmapped locations. In the publication on c-Fes³⁰ it is shown that SH2 domain binding to phosphotyrosine on a substrate primes the active site for phosphorylation. Figure 8(B) shows the computation c-Fes (PD ID [3BKB](#)) where the SH2 domain is organized around a sulfate atom. There is a large, unoccupied pocket next to the sulfate atom, and sets of deep grooves leading from the SH2 domain, around the α C-helix, and down to the active site cleft. The computation provides a possible map of the path of substrate interaction with c-Fes. Overall, this method is an efficient way of comparing protein surfaces independent of amino acid sequence. The method has identified some of the important motifs in the kinase family, shown a possible way of discriminating between serine/threonine and tyrosine kinases, and highlighted potential regions of interest for further study.

References

- Brickell PM (1992) The p60c-src family of protein-tyrosine kinases: structure, regulation, and function. *Crit Rev Oncog* 3:401–446.
- Pytel D, Sliwinski T, Poplawski T, Ferriola D, Majsterek I (2009) Tyrosine kinase blockers: new hope for successful cancer therapy. *Anticancer Agents Med Chem* 9:66–76.
- Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. *Science* 298:1912–1934.
- Mitchell JC, Kerr R, Ten Eyck LF (2001) Rapid atomic density methods for molecular shape characterization. *J Mol Graph Model* 19:325–330, 388–390.
- Laurie AT, Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* 21:1908–1916.
- Harris R, Olson AJ, Goodsell DS (2008) Automated prediction of ligand-binding sites in proteins. *Proteins* 70:1506–1517.
- An J, Totrov M, Abagyan R (2004) Comprehensive identification of “druggable” protein ligand binding sites. *Genome Inf* 15:31–41.
- Johnson DA, Akamine P, Radzio-Andzelm E, Madhusudan M, Taylor SS (2001) Dynamics of cAMP-dependent protein kinase. *Chem Rev* 101:2243–2270.
- Kannan N, Haste N, Taylor SS, Neuwald AF (2007) The hallmark of AGC kinase functional divergence is its C-terminal tail, a cis-acting regulatory module. *Proc Natl Acad Sci USA* 104:1272–1277.
- Kornev AP, Haste NM, Taylor SS, Eyck LF (2006) Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proc Natl Acad Sci USA* 103:17783–17788.
- Kornev AP, Taylor SS, Ten Eyck LF (2008) A helix scaffold for the assembly of active protein kinases. *Proc Natl Acad Sci USA* 105:14377–14382.
- Zhang X, Gureasko J, Shen K, Cole PA, Kuriyan J (2006) An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell* 125:1137–1149.
- Zheng J, Knighton DR, Xuong NH, Taylor SS, Sowadski JM, Ten Eyck LF (1993) Crystal structures of the myristylated catalytic subunit of cAMP-dependent protein kinase reveal open and closed conformations. *Protein Sci* 2:1559–1573.
- Narayana N, Cox S, Nguyen-huu X, Ten Eyck LF, Taylor SS (1997) A binary complex of the catalytic subunit of cAMP-dependent protein kinase and adenosine further defines conformational flexibility. *Structure* 5:921–935.
- Gangal M, Clifford T, Deich J, Cheng X, Taylor SS, Johnson DA (1999) Mobilization of the A-kinase N-myristate through an isoform-specific intermolecular switch. *Proc Natl Acad Sci USA* 96:12394–12399.
- Hanks SK, Hunter T (1995) Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J* 9:576–596.
- Nolen B, Yun CY, Wong CF, McCammon JA, Fu XD, Ghosh G (2001) The structure of Sky1p reveals a novel mechanism for constitutive activity. *Nat Struct Biol* 8:176–183.
- LaFevre-Bernt M, Sicheri F, Pico A, Porter M, Kuriyan J, Miller WT (1998) Intramolecular regulatory interactions in the src family kinase hck probed by mutagenesis of a conserved tryptophan residue. *J Biol Chem* 273:32129–32134.
- McPherson RA, Taylor MM, Hershey ED, Sturgill TW (2000) A different function for a critical tryptophan in c-raf and hck. *Oncogene* 19:3616–3622.
- Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol* 285:1735–1747.

21. DeLano WL (2002) The PyMOL Molecular Graphics System [computer program]. DeLano Scientific, Palo Alto, CA, USA. Available at: <http://www.PyMOL.org>.
22. Hartigan J (1975) Clustering algorithms. John Wiley & Sons, Inc., New York, NY, 351 p.
23. Milligan GW, Sokol LM (1980) A two-stage clustering algorithm with robust recovery characteristics. *Educ Psychol Meas* 40:755–759.
24. Gangal M, Cox S, Lew J, Clifford T, Garrod SM, Aschbaher M, Taylor SS, Johnson DA (1998) Backbone flexibility of five sites on the catalytic subunit of cAMP-dependent protein kinase in the open and closed conformations. *Biochemistry* 37:13728–13735.
25. Otwinowski Z, Minor W, Processing of X-ray diffraction data collected in oscillation mode. In: Carter CW, Jr, Ed. (1997) *Methods in enzymology*. NY: Academic Press, pp 307–326.
26. Zheng J, Trafny EA, Knighton DR, Xuong NH, Taylor SS, Ten Eyck LF, Sowadski JM (1993) 2.2 Å refined crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MnATP and a peptide inhibitor. *Acta Cryst D* 49:362–365.
27. McCoy AJ (2007) Solving structures of protein complexes by molecular replacement with phaser. *Acta Cryst D* 63: 32–41.
28. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Cryst D* 60:2126–2132.
29. Winn MD, Isupov MN, Murshudov GN (2001) Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Cryst D* 57:122–133.
30. Filippakopoulos P, Kofler M, Hantschel O, Gish GD, Grebien F, Salah E, Neudecker P, Kay LE, Turk BE, Superti-Furga G, Pawson T, Knapp S (2008) Structural coupling of SH2-kinase domains links fes and abl substrate recognition and kinase activation. *Cell* 134:793–803.
31. Bullock AN, Das S, Debreczeni JE, Rellos P, Fedorov O, Niesen FH, Guo K, Papagrigoriou E, Amos AL, Cho S, Turk BE, Ghosh G, Knapp S (2009) Kinase domain insertions define distinct roles of CLK kinases in SR protein phosphorylation. *Structure* 17:352–362.
32. Seifert A, Allan LA, Clarke PR (2008) DYRK1A phosphorylates caspase 9 at an inhibitory site and is potently inhibited in human cells by harmine. *FEBS J* 275:6268–6280.
33. Nayler O, Stamm S, Ullrich A (1997) Characterization and comparison of four serine- and arginine-rich (SR) protein kinases. *Biochem J* 326:693–700.
34. Kannan N, Neuwald AF, Taylor SS (2008) Analogous regulatory sites within the alphaC-beta4 loop regions of ZAP-70 tyrosine kinase and AGC kinases. *Biochim Biophys Acta* 1784:27–32.