Kincore: a web resource for structural classification of protein kinases and their inhibitors

Vivek Modi

Roland Dunbrack Jr.

Institute for Cancer Research

Fox Chase Cancer Center,

Philadelphia PA 19111

USA

Abstract

Protein kinases exhibit significant structural diversity, primarily in the conformation of the activation loop and other components of the active site. We previously performed a clustering of the conformation of the activation loop of all protein kinase structures in the Protein Data Bank (Modi and Dunbrack, PNAS, **116**:6818-6827, 2019) into 8 classes based on the location of the Phe side chain of the DFG motif at the N- terminus of the activation loop. This is determined with a distance metric that measures the difference in the dihedral angles that determine the placement of the Phe side chains (the ϕ, ψ of X, D, and F of the X-DFG motif and the χ_1 of the Phe side chain). The nomenclature is based on the regions of the Ramachandran map occupied by the XDF residues and the $\chi 1$ rotamer of the Phe residue. All active structures are "BLAminus", while common *inactive* DFGin conformations are "BLBplus" and "ABAminus". Type II inhibitors bind almost exclusively to the DFGout "BBAminus" conformation. In this paper, we present Kincore (http://dunbrack.fccc.edu/kincore), a web resource providing access to the conformational assignments based on our clustering along with labels for ligand types (Type I, Type II, etc.) bound to each kinase chain in the PDB. The data are annotated with several properties including PDBid, Uniprotid, gene, protein name, phylogenetic group, spatial and dihedral labels for orientation of DFGmotif residues, C-helix disposition, ligand name and type. The user can browse and query the database using these attributes individually or perform advanced search using a combination of them like a phylogenetic group with specific conformational label and ligand type. The user can also determine the spatial and dihedral labels for a structure with unknown conformation using the web server and standalone program. The entire database can be downloaded as text files and structure files in PyMOL sessions and mmCIF format. We believe that Kincore will help in understanding conformational dynamics of these proteins and guide development of inhibitors targeting specific states.

Introduction

Protein kinases are catalytic molecular switches that regulate signaling pathways in cells by phosphorylating protein substrates [1]. Their catalytic activity is achieved by a remarkably flexible active site which is observed in multiple different conformations when the enzyme is in inactive state but adopts a unique conformation in the catalytically active state. The dysregulation of this mechanism due to a mutation or upregulation of expression can lead to a variety of diseases including cancer [2, 3]. Protein kinases are widely studied as drug targets with molecules targeted to inhibit the active state or stabilize a specific inactive state [4, 5]. Thus, the understanding of conformational dynamics in protein kinases is critical for development of better drugs and novel biological insights.

There are 484 typical protein kinase genes with 497 kinase domains in the human genome [6, 7]. This number includes several pseudokinases but excludes atypical protein kinase genes, some of which are distantly related to the typical protein kinase fold [7]. Among the 497 domains, currently the structures of 283 have been experimentally determined either in apo form or in complex with ligands. The protein kinase fold consists of an N-terminal lobe, which is formed by five beta sheets and one alpha helix called the C-helix, and a C-terminal lobe which consists of five or six alpha helices. The two lobes form a deep cleft in the middle region of the protein creating the ATP-binding active site. This site is surrounded by several structural elements critical for catalysis which occupy a unique conformation in the active state and exhibit flexibility across different inactive states of the enzyme. One of the most critical elements is the activation loop which adopts a unique extended orientation in the active state of the kinase and multiple types of folded conformations in inactive states. It begins with a conserved motif called the DFGmotif (Asp-Phe-Gly) whose orientation is tightly coupled with active/inactive status of the protein. In addition, the C-helix displays inwards disposition in the active state while exhibiting a range of positions and orientations in other states.

The DFGmotif conformations were previously addressed by using a simple convention of DFGin and DFGout. The DFGin group consists of all the conformations in which DFG-Asp points in ATP pocket and DFG-Phe is adjacent to the C-helix. The structures solved in the active state conformation of the enzyme form a subset of this category. In DFGout conformations, the DFG-Asp and DFG-Phe residues swap their positions so that DFG-Asp is removed from the ATP binding site and replaced with DFG-Phe. All the Type II inhibitors bind to DFGout conformations [8].

The DFGin and DFGout groups, however, provide only a broad description of a more complex conformational landscape [9, 10]. In our previous work, we developed a scheme for clustering and labeling different conformations of protein kinase structures [11]. Our clustering scheme is based on the spatial location and backbone and side-chain dihedrals of the conserved DFGmotif in the activation loop. We clustered all the conformations into three spatial groups (DFGin, DFGinter, DFGout) based on the proximity of the DFG-Phe side chain to two different residues in the N-terminal domain. Within these groups, we further clustered the structures by the dihedral angles that determine the location of the DFG-Phe side chain: the backbone dihedrals of the X, D and F residues (where X is the residue before the DFGmotif) and the χ_1 dihedral angle of the Phe side chain. The kinase states are therefore named after the region of the Ramachandran map occupied by the X, D, and F residues (A for alpha, B for beta, L for

left-handed) and the Phe χ_1 rotamer (plus, minus, or trans for the +60°, -60°, or 180° conformations). As a result, among the DFGin structures, we distinguished between the catalytically active kinase conformation (labeled BLAminus) and five inactive conformations (BLBplus, BLBminus, BLBtrans, ABAminus, BLAplus). Among DFGout structures, we identified one dominant conformation labeled BBAminus, which is strongly correlated with Type II kinase inhibitors, such as imatinib. Finally, among the small set of DFGinter structures, where the Phe side chain is intermediate between the DFGin and DFGout positions, we distinguished one cluster based on clustering the dihedral angles (BABtrans). Our nomenclature strongly correlates with other structural features associated with active and inactive kinases, such as the positions of the C-helix and the activation loop and the presence or absence of the N-terminal domain salt bridge. Since our clustering and nomenclature is based on backbone dihedrals, it is intuitive to structural biologists and easy to apply in a wide variety of experimental and computational studies, as demonstrated recently in identifying the conformation in crystal structure of IRAK3 [12], molecular dynamics simulations of Abl kinase [13] and structural analyses of pseudokinases [14].

Developing small molecule inhibitors is one of the most common therapeutic strategies against protein kinases. These inhibitors occupy the ATP binding pocket and allosteric sites on the surface of the protein. There have been two approaches used to classify inhibitors – a) based on the region of the protein to which the inhibitor binds; b) based on the conformation of the protein to which it binds. The first approach was used by Dar and Shokat [15] who defined three types of inhibitors: Type I – inhibitors which bind to the adenosine pocket but do not require a specific conformation of structural elements including the C-helix and DFGmotif; Type II – inhibitors that occupy the adenosine pocket and induce DFGout conformations because they extend into the pocket adjacent to the C-helix occupied by DFG-Phe in DFGin structures; Type III – inhibitors that block kinase activity but without displacing ATP. This classification was extended by Zuccotto and coworkers who introduced Type I½ inhibitors as molecules which bind to the ATP region like Type I compounds but extend into the back cavity making additional contacts with the residues involved in Type II binding [16]. Rauh et. al. defined Type IV as the allosteric inhibitors which bind to a site distant to the ATP binding region inducing an inactive conformation in the active site [17, 18]. van Linden et al. defined the ligand types by identifying three regions in the active site - a front cleft, the gate area, and the back cleft, which are further divided into subpockets [19] without the use of labels like Type I, II etc.

Roskoski used the second approach and redefined all the inhibitors based on the conformation of the protein [20]. According to this scheme, Type I inhibitors bind only to the active conformation; Type I½ are the inhibitors which bind to DFGin inactive conformations and Type II inhibitors bind to DFGout conformation. Each of these categories were divided into two subtypes A and B. However, this scheme is inadequate because, as we have shown, some inhibitors such as Bosutinib and Sunitinib can bind to different conformations across proteins [11]. For example, according to Roskoski's classification Sunitinib will be labeled Type I in 6NFZ_A (DFGin-BLAminus) and Type IIB in 3G0F_A (DFGout-BBAminus), even though they bind to the kinase domain in an identical manner.

In this paper, we present the Kinase Conformation Resource, Kincore – a web resource which automatically collects and curates all protein kinase structures from the Protein Data Bank (PDB) and assigns conformational and inhibitor type labels. The website is designed so that the information for all

the structures can be accessed at once using one database table and instances of it through individual pages for kinase phylogenetic groups, genes, conformational labels, PDBids, ligands and ligand types. The database can be searched using unique identifiers such as PDBid or gene, and queried using a combination of attributes such as phylogenetic group, conformational label and ligand type. We also provide several options to download data – database tables as a tab separated files; the kinase structures as PyMOL sessions and coordinate files in mmCIF format. The structures have been renumbered by Uniprot and our common numbering scheme, which is derived from our structure-based alignment of all 497 human protein kinase domains [7]. We have also developed a webserver and standalone program which can be used to determine the spatial and dihedral labels for a structure with unknown conformation.

We automatically label ligand types based on the pockets to which an inhibitor binds defined by specific residues in the kinase domain. Thus, we use five labels for different ligand types: Type I – bind to ATP binding region only (both active and inactive DFGin states); Type I½ – ATP binding region and extending into the back pocket (both active and inactive DFGin states); Type II – ATP binding region and extending to back pocket regions exposed only in DFGout structures; Type III – back pocket only without displacing ATP; and Allosteric – outside the active site cleft.

Results

Kincore provides conformational assignments and ligand type labels to protein kinase structures from PDB. The current update contains structures from 283 kinase genes from humans (7129 chains) and from 55 genes (707 chains) from seven model organisms. The PK structures were identified from the PDB [21] using PSI-BLAST [22] using a kinase PSSM matrix as a query (Methods). The PDB files are split by chain, renumbered by Uniprot numbering [23, 24] and our common residue numbering scheme, and annotated by conformational and ligand type labels as described below.

The conformational labels are assigned using the structural features and clusters described in our previous work [11]. The scheme assigns two types of labels to each chain – 1) A spatial label (DFGin, DFGinter, DFGout) by computing the distance of the DFG-Phe-CZ atom from the Cα atoms of two conserved residues – the strand β3-Lys involved in the N-terminal domain salt bridge formed in active kinase structures (and some inactive structures) and the residue four amino acids past the C-helix-Glu involved in the same salt bridge and assigning a label using distance cutoff criteria (Methods); 2) A dihedral label –the dihedral angles (ϕ,ψ of X-DFG, Asp, Phe and χ_1 for Phe) for each chain in a spatial group are used to calculate the distance of the structure from the precomputed cluster centroids and assigned a label if its distance satisfies defined cutoff criteria (Methods). All the kinase conformations are represented by a set of eight labels: DFGin-BLAminus, DFGin-BLAplus, DFGin-ABAminus, DFGin-BLBminus, DFGin-BLBplus, DFGin-BLBtrans; DFGout-BBAminus; DFGinter-BABtrans. The chains that do not satisfy the dihedral distance cutoff criteria for any cluster or are missing some of the relevant coordinates are labeled as 'Unassigned'.

Additionally, we have also labeled the C-helix disposition by computing the distance between the C-helix-Glu-Cβ atom from the B3-Lys-Cβ-atom (as a proxy for the conserved salt bridge interaction) and labeled it as C-helix-in and C-helix-out (Methods).

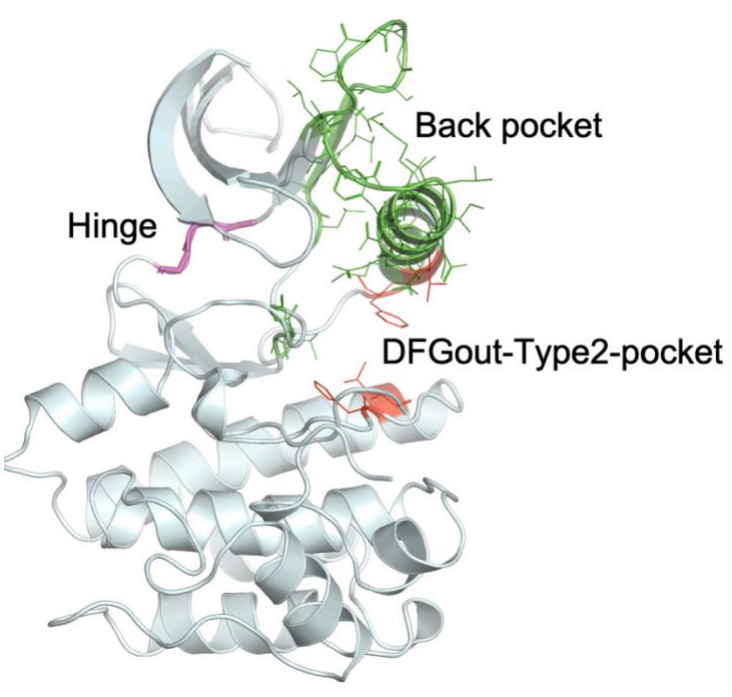

Figure 1: Representative protein kinase structure (3ETA_A) displaying the residues used to define inhibitor binding regions.

To assign labels to ligands, we have used specific residue positions to identify regions of the binding pocket – the ATP binding pocket (including the hinge residues), back pocket and Type II-only region (Figure 1). The structures are first renumbered by our common numbering scheme so that all the aligned residues have the same residue number across all the kinases. A ligand is then assigned a label based on its contacts with different binding regions. We have used the following five ligand type labels to annotate all the ligand-bound structures of protein kinases (Figure 1):

1. Type I – bind to ATP binding region only
2. Type I½ – bind to ATP binding region and extend into the back pocket (subdivided as Type I½-front and Type I½-back depending on contact with N-terminal or C-terminal residues of the C-helix, respectively)
3. Type II – bind to the ATP binding region and extend into the back pocket and Type II-only region
4. Type III – bind only in the back pocket without displacing ATP
5. Allosteric - any pocket outside the ATP-binding region

The distribution of different ligand types across kinase conformations is provided in Table 1. It shows that Type I and Type I½ are the most commonly observed inhibitors. However, except Type II, all the inhibitor types are observed in complex with multiple conformational states.

Table 1: Distribution of ligand types across protein kinase conformations (Number of chains).

Spatial label	Dihedral label	Type I	Type I½ (front+back)	Type II	Type III	Allosteric	Total (%)
DFGin	BLAminus (active)	2926	196	-	12	199	3333 (55.0)
	BLBplus	443	76	-	59	15	593 (9.8)
	ABAminus	479	36	-	1	19	535 (8.8)
	BLBminus	162	11	-	5	10	188 (3.1)
	BLBtrans	175	6	-	-	5	186 (3.1)
	BLAplus	91	86	-	-	1	178 (2.9)
	Noise	282	38	-	1	18	339 (5.6)
DFGout	BBAminus	20	9	288	69	24	410 (6.8)
	Noise	43	17	79	26	12	177 (2.9)
DFGinter	BABtrans	14	1	-	-	-	15 (0.2)
	Noise	89	16	-	3	3	111 (1.8)
Total (%)		4724 (77.9)	492 (8.1)	367 (6.1)	176 (2.9)	306 (5.0)	

Many inhibitors are observed in multiple crystal structures bound to one or more different kinases. We counted the number of unique inhibitors that occur bound to kinase chains in two (or more) states across entries in the PDB. In Table 2, we show a table that provides the number of unique inhibitors that occur in each pair of states (excluding the unclassified spatial or dihedral labels). The numbers along the diagonal are the counts of unique inhibitors observed in at least one structure of the given state. A total of 259 inhibitors occur in two or more kinase states.

Table 2. Counts of inhibitors that are bound to chains in two or more states.

	DFGin-BLAminus	DFGin-ABAminus	DFGin-BLBplus	DFGin-BLBminus	DFGin-BLBtrans	DFGin-BLAplus	DFGout-BBAminus	DFGinter-BABtrans
DFGin-BLAminus	1686							
DFGin-ABAminus	48	334						
DFGin-BLBplus	39	11	344					
DFGin-BLBminus	26	9	11	210				
DFGin-BLBtrans	29	4	11	4	134			
DFGin-BLAplus	15	6	13	8	2	107		
DFGout-BBAminus	7	3	2	2	2	3	254	
DFGinter-BABtrans	6	2	4	4	1	3	1	8

Numbers along the diagonal provide the number of unique inhibitors in each state. The off-diagonal values are the number of unique inhibitors bound to chains in the two states shown in the row and column headers.

Website

The web pages on Kincore are designed in a common format across the website to organize the information in a consistent and uniform way. Each page retrieved from the database is organized in two parts – the top part provides a summary of the number of structures in the queried groups or conformations, with representative structures from each category listed and displayed. This is followed by a table from the database with each unique PDB chain as a row providing different kinds of information including conformational and ligand type labels and C-helix position, kinase family, gene name, Uniprot ID, ligand PDB ID, and ligand type. The kinase group, gene name, PDB code, conformational labels, ligand name and ligand type are hyperlinked to their specific pages. Each page also contains three tabs on the top to list 'Human', 'Non-human' and 'All' structures. There are buttons provided on each page to

download the database table as a tab separated file, and to download all of the kinase structures on the page as PyMOL sessions, and renumbered coordinate files.

Group ▲▼	Gene ▲▼	UniprotID ▲▼	Protein ▲▼	PDB ▲▼	Spatial Label ▲▼	Dihedral Label ▲▼	C-helix ▼	Ligand ▲▼	Type ▲▼
TYR	KDR	VGFR2_HUMAN	Vascular endothelial growth factor receptor 2	2XIRA	DFGout	BBAminus	in	00J	Type1.5_Back
CAMK	PIM1	PIM1_HUMAN	Serine/threonine-protein kinase pim-1	3MA3A	DFGin	BLAminus	in	01I	Type1
CMGC	CSNK2A1	CSK21_HUMAN	Casein kinase II subunit alpha	3MB6A	DFGin	BLAminus	in	01I	Type1
TYR	IGF1R	IGF1R_HUMAN	Insulin-like growth factor 1 receptor	3QQUC	DFGin	BLAminus	in	01P	Type1
TYR	IGF1R	IGF1R_HUMAN	Insulin-like growth factor 1 receptor	3QQUA	DFGin	BLBminus	in	01P	Type1
TYR	IGF1R	IGF1R_HUMAN	Insulin-like growth factor 1 receptor	3QQUD	DFGin	None	in	01P	Type1
TYR	IGF1R	IGF1R_HUMAN	Insulin-like growth factor 1 receptor	3QQUB	DFGin	None	in	01P	Type1
TYR	BTK	BTK_HUMAN	Tyrosine-protein kinase BTK	3PIXA	DFGin	BLBplus	out	027	Type1
CMGC	CDK2	CDK2_HUMAN	Cyclin-dependent kinase 2	3RZBA	DFGin	BLAminus	out	02Z	Type1
TKL	BRAF	BRAF_HUMAN	Serine/threonine-protein kinase B-raf	4RZVB	DFGin	BLAplus	out	032	Type1.5_Back
TKL	BRAF	BRAF_HUMAN	Serine/threonine-protein kinase B-raf	4RZVA	DFGin	BLAplus	out	032	Type1.5_Back
TKL	BRAF	BRAF_HUMAN	Serine/threonine-protein kinase B-raf	3OG7A	DFGin	BLAplus	out	032	Type1.5_Back
TKL	MAP3K20	M3K20_HUMAN	Mitogen-activated protein kinase kinase kinase kinase MLT	5HESB	DFGin	BLAplus	out	032	Type1
TKL	MAP3K20	M3K20_HUMAN	Mitogen-activated protein kinase kinase kinase kinase MLT	5HESA	DFGin	BLAplus	out	032	Type1.5_Back
TYR	BTK	BTK_HUMAN	Tyrosine-protein kinase BTK	3PIZA	DFGin	BLAplus	in	03C	Type1
CMGC	CDK2	CDK2_HUMAN	Cyclin-dependent kinase 2	4EK5A	DFGin	BLBtrans	out	03K	Type1

Figure 2: Snapshot of database table displaying entries for PDB chains on Browse page.

The information from the database can be accessed using two main pages:

1. **Browse page:** This page provides statistics and labels for all the kinase structures in the database (Figure 2). The 'Summary' table on top of the page displays the distribution of protein kinase chains in the PDB across conformational states and phylogenetic groups. This is followed by 'Database' table which contains annotation for all individual PDB chains retrieved from the database. The entire table with additional information like resolution, Rfactor, activation loop residue etc. can be downloaded as a tab separated file.

2. **Search page:** This page offers two options to query the database:

    - Unique identifier: The database can be queried by PDB entry code (e.g., 2GS6), UniProt identifier (e.g., EGFR_HUMAN), gene name (e.g., EGFR), and ligand identifier (e.g., STI). The result will take the user to the page dedicated to the specific query item. For reference the list of all genes in the database is provided for the user through a 'Help' button above the search box.

- Advanced query: The database can be queried by selecting kinase phylogenetic group, conformational label, and ligand type using a drop-down menu. If 'All' option is selected for all the three categories, then the entire database table can be accessed at once. A subset of chains in the database can be retrieved by selecting a specific group name, conformational label, and ligand type, for example selecting TYR group + DFGout-BBAminus + Type II ligand type will retrieve all the structures which have these three annotations. If all the structures in complex with Type I½ ligand are desired, then the user can select 'All group' + 'All conformations' + 'Type I½ ligand'.

The website contains several webpages which are dynamically generated and retrieve queried instances of the database. These pages can be accessed as a result of individual queries or by clicking on the hyperlinks on the Browse page table. They are,

1. Phylogenetic group page: typical protein kinases are divided into nine phylogenetic groups – AGC, CAMK, CMGC, CK1, NEK, RGC, STE, TKL and TYR [ref]. Each group is assigned a page on Kincore displaying information about the structures in that group. On each page, the Summary table provides the number of kinase chains in the group across different conformations with their representative structures (best resolution and least missing residues). These representative structures are also displayed on the page in 3D using NGL viewer.

2. Gene page: A page for each kinase gene in the PDB can be accessed through the hyperlinks on Browse page or by unique identifier Search feature and contains information for all the structures of a specific gene. The summary table on the page gives the number of structures available and their distribution across different conformations with representative example for each. It also provides hyperlinks to the phylogenetic group page (described above) for the gene and the corresponding protein entry on the Uniprot website. In addition to the data provided on the Browse page, the Database table on this page also contains for each chain information on mutations, phosphorylation with total length of the structure and number of residues resolved in the activation loop.

3. PDB page: The PDB page provides information on individual PDB entries and can be accessed by the hyperlinks on the Browse page or by the unique identifier Search feature (Figure 3). Each PDB entry is annotated with information on gene, protein name, phylogenetic group, UniProt id, organism, domain boundary, resolution, conformation, and ligand type labels for every chain. Additionally, the page also contains a sequence feature displaying the UniProt sequence of the protein in the structure. The residues which are unresolved in the structure are displayed in lower case letters to distinguish them from residues with coordinates in the entry. Further, mutated and phosphorylated residues are shown in red and green color, respectively.

4. Ligand page: The ligand page provides access to all chains in complex with a specific ligand. For example, all the structures in complex with ATP can be retrieved by querying for 'ATP' on the Search page or clicking on the hyperlinks on the Browse page. The Summary table provides the

number of chains in complex with the ligand across different conformations. Like other pages, the Database table provides the list of all the PDB chains with conformational labels and ligand annotations. This page facilitates the comparison of conformations and ligand binding mode across structures from one or multiple kinases in complex with the same ligand. For example, Bosutinib (PDB identifier DB8) which is an FDA-approved drug, is found in complex with structures from 10 kinases in 5 different conformations (Figure 4).

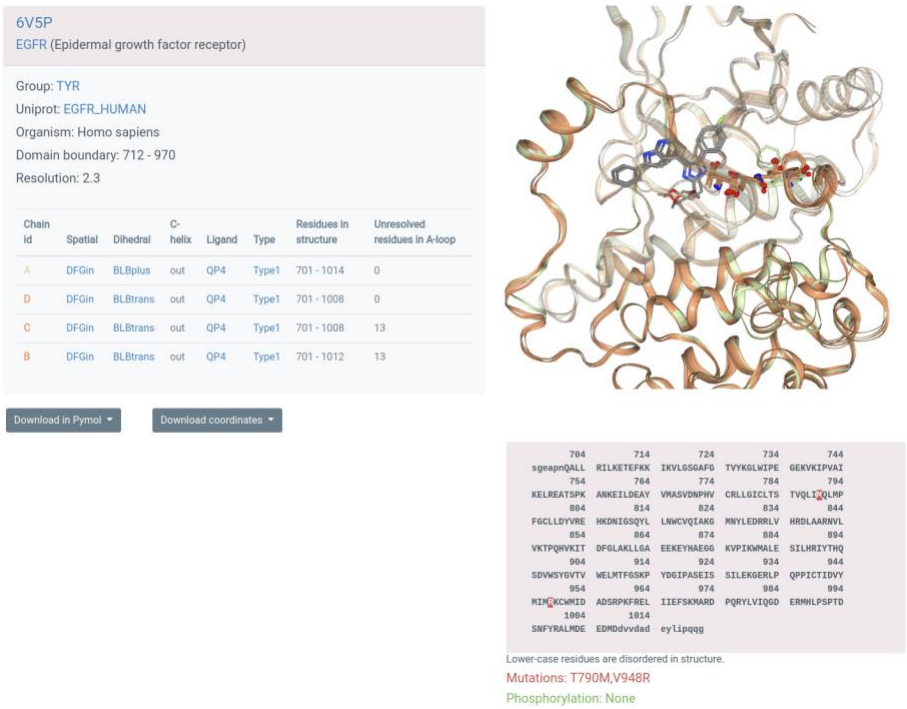

Figure 3: Snapshot of PDB page with the sequence feature.

Alignment Page

In our previous work, we developed a structure-based multiple sequence alignment (MSA) for 497 human protein kinase domains [7]. This alignment contains 17 blocks of aligned regions conserved across human kinases with intermittent regions of low sequence similarity in lower case letters. The alignment is annotated with gene name, UniProt id, and protein residue numbers. On Kincore, we provide access to this MSA through the Alignment page which contains basic information about the alignment with a table of conserved regions across human kinases. The alignment can be visualized inside the browser window through 'Open in browser' button created using Jalview's BioJS feature. This feature provides multiple options for quick analysis including buttons to filter, color, or sort the sequences within the browser window. The alignment is also available to download as a Jalview session as well as Clustal- and FASTA-formatted files.

Phylogeny Page

Using our multiple sequence alignment, we also updated the protein kinase phylogenetic tree [7]. This tree was used to assign a set of ten kinases previously categorized as "OTHER" to the CAMK group, consisting of Aurora kinases, Polo-like kinases, and calcium/calmodulin-dependent kinase kinases. On our resource the tree can be accessed through the Phylogeny page. It provides basic information about the tree, the number of kinase genes and domains in different phylogenetic groups, and links to visualize and download the tree.

Figure 4: Snapshot of ligand page displaying Bosutinib (PDB ligand identifier DB8) in complex with structures from 10 kinase genes and in 5 different conformations.

Download Options

We provide multiple data download options on Kincore to assist the user in different kinds of analysis. These download options are created for all the pages or any instance of database retrieved by a query, e.g. structures of a specific gene, ligand etc. or structures from an advanced query like TYR kinases with DFGout state and Type II ligands. These options are:

1. **Coordinate Files**
   We provide structure files in mmCIF and PDB format with three different numbering systems: the original author residue numbering; renumbered by Uniprot protein sequence; and a

common residue numbering scheme derived from our multiple sequence alignment of kinases [7].

2. **PyMOL Sessions**

   We provide PyMOL [25] sessions for the structures retrieved from any query from the database. Two PyMOL sessions are provided for each query – All chains and Representative chains (best resolution, least missing residues). Across all the PyMOL sessions, the chains are labeled in a consistent format as – PhyloGroup_Gene_SpatialLabel_DihedralLabel_PDBidChainid (e.g., TYR_EGFR_DFGin_BLAminus_2GS6A). Additionally, we also provide PyMOL scripts (.pml format) which the user can download and run on a local machine to create the sessions.

3. **Database Files**

   We provide the information retrieved from the database on every page as tab separated files which can be downloaded using 'Database table as tsv' button. When clicked on the 'Browse' page, this button will download the information in the entire database in one file. On the other pages specific for a gene or conformation, this file will contain only the subset of the information from the database which is queried. The tsv file has the following header,

   "Organism Group Gene UniprotID PDB Method Resolution Rfac FreeRfac SpatialLabel DihedralLabel C-helix Ligand LigandType DFG_Phe Edia_X_O Edia_Asp_O Edia_Phe_O Edia_Gly_O ProteinName"

4. **Bulk download**

   The 'Download' page provides different options to download structure files and PyMOL sessions in bulk. The page is divided into two sections – coordinate files and PyMOL sessions. The user can download coordinate files for all the structures in one zip folder or in subsets of specific phylogenetic group, gene, and conformational label. The tab on the top of the page gives the option to download files with original author residue numbering or renumbered by Uniprot protein sequence and common residue numbering from our alignment. The second part of the 'Download' page provides PyMOL sessions for phylogenetic groups, genes and ligands.

We have developed a webserver which the user can use to upload a kinase structure file in PDB or mmCIF format to determine its conformation. The program extracts the sequence from structures file and identifies residue positions by aligning it with precomputed HMM profiles of kinase groups. It then determines the conformation of the protein by assigning Spatial and Dihedral labels (Methods). On the output page, the server prints the kinase phylogenetic group which is the closest match to the sequence of the input structure, dihedrals of X-DFG, DFG-Asp, DFG-Phe residues, spatial group, dihedral label and C-helix disposition.

We have written a standalone program using Python3 which the user can download to assign conformational labels to an unannotated structure. The program can be run in two ways: a) with flag align=True: alignment with precomputed HMM profiles is done to identify the residue numbers for B3-Lys, C-helix-Glu and DFG-Phe. The program then computes inter-residue distances and dihedral angles to label the conformation in the structure (Methods); b) with flag align=False: alignment with an HMM profile is not done, and the residue numbers are provided by the user. This option is faster and more useful for identifying conformations in a large number of structures generated from a molecular dynamics simulations.

Discussion

Experimentally determined protein kinase structures in apo-form or in complex with a ligand display an extremely flexible active site. However, examining the conformational dynamics of kinases and its role in ligand binding require combining two pieces of information – the conformational state of the protein and the type of ligand in complex. Currently, there are two main resources, Kinametrix and KLIFS, that address protein kinase conformations and inhibitors. However, they provide either conformational assignments or ligand type information, but not both. Kinametrix (http://kinametrix.com/) offers a simple scheme of DFGin and DFGout coupled with C-helix conformation [26]. The resource does not provide information on ligands and lacks any download options for structures. This resource has not been updated with structures since May 2017. KLIFS (https://klifs.vu-compmedchem.nl/index.php) – also offers a simple DFGin and DFGout classification [19, 27] and does not distinguish active and inactive DFGin structures. This resource is more focused on providing information about ligand binding to kinases. It is regularly updated and allows bulk downloads for the results of each search.

Kincore fills a gap by providing a sophisticated scheme for kinase conformations, with ligand type labels. The information can be accessed as individual queries for example, getting a list of all chains in complex with Type II ligand; or a combination of queries like, AGC group kinases + DFGin-BLBplus conformation + Type I½ ligand. A feature that distinguishes Kincore from many structural bioinformatics resources is the ability to download coordinate files for the result of any query in one click. For example, a search for AURKA produces a list of 191 protein chains from 154 PDB entries. These can be downloaded in mmCIF format with one click with residue numbering in original PDB numbering, renumbered according to the UniProt sequences, or in our common residue numbering scheme from the kinase multiple sequence alignment. Each coordinate file is labeled by spatial label and dihedral angle cluster, e.g. CAMK_AURKA_DFGin_BLAminus_1OL6A.cif. A user can also download a PyMOL session file with all of the structures for a given query. In addition, an important part of our resource is the web server and standalone program which can label the unknown conformation of a new structure. The standalone program can run on structure files with multiple chains and models. We believe it will be extremely useful to batch process the structures generated from a molecular modeling protocol or molecular dynamics simulation.

Several experimental and computational studies have reported applying the nomenclature from our previous work in structural analyses of kinases [11]. Lange and colleagues have solved the crystal structure of the pseudokinase IRAK3 (PDBID 6RUU) and identified its conformation as BLAminus, similar to the active state of a typical protein kinase [12]. Paul et.al. have studied the dynamics of ABL kinase by various simulation techniques with Markov state models and analyzed the transition between different metastable states by using our nomenclature [13]. Kirubakaran et. al. have identified the catalytically primed structures (BLAminus) from the PDB to create a comparative modeling pipeline for the ligand bound structures of CDK kinases [28]. Paul and Srinivasan have done structural analyses of pseudokinases in *Arabidopsis thaliana* and compared with typical protein kinases by applying our conformational labels [14]. Therefore, we believe that the development of Kincore database and webserver will greatly benefit a larger research community by making the labeled kinase structures more accessible and facilitating identification of kinase conformations in a wide range of studies.

Methods

Identifying and renumbering protein kinase structures

The database contains protein kinase domains from *Homo sapiens* and seven model organisms consisting *Bos taurus, Danio rerio, Drosophila melanogaster, Mus musculus, Rattus norvegicus, Sus scrofa* and *Xenopus laevis*. To identify structures from these organisms the sequence of human Aurora A kinase (residues 125-391) was used to construct a PSSM matrix from three iterations of NCBI PSI-BLAST on the PDB with default cutoff values [22]. This PSSM matrix was used as query to run command line PSI-BLAST on the *pdbaa* file from the in the PISCES server (http://dunbrack.fccc.edu/pisces) [29]. *pdbaa* contains the sequence of every chain in every asymmetric unit of the PDB in FASTA format with resolution, R-factors, and SwissProt identifiers (e.g. AURKA_HUMAN). A total of 4908 PDB entries with 7277 kinase chains were identified. Some poorly aligned kinases and non-kinase proteins that were homologous to kinases but distantly related were removed.

The structure files were split by individual kinase chains in the asymmetric unit and renumbered by UniProt protein numbering scheme. The mapping between PDB author numbering and UniProt was obtained from Structure Integration with Function, Taxonomy and Sequence (SIFTS) database [24]. The SIFTS files were also used to extract mutation, phosphorylation, and missing residue annotations.

The structure files were also renumbered by a common residue numbering scheme using our protein kinase multiple sequence alignment. Each residue in a kinase domain was renumbered by its column number in the alignment. Therefore, aligned residues across different kinase sequences get the same residue number. For example, in these renumbered structure files the residue number of the DFGmotif across all kinases is 1338 – 1340. The conserved motifs for all the structures were identified from the same alignment.

Assigning conformational labels

Each kinase chain is assigned a spatial group and a dihedral label using our previous clustering scheme as a reference [11]. Our clustering scheme has three spatial groups – DFGin, DFGinter, and DFGout. These are sub-divided into dihedral clusters DFGin -- BLAminus, BLAplus, ABAminus, BLBminus, BLBplus, BLBtrans; DFGinter – BABtrans; and DFGout – BBAminus.

To determine the spatial group for each chain, the location of DFG-Phe in the active site was identified using the following criteria:

1. D1≤11 Åand D2≥11 Å– DFGin
2. D1>11 Å and D2<=14 Å– DFGout
3. D1≤11 Å and D2≤11 Å – DFGinter, where D1= αC-Glu(+4)-Cα to DFG-Phe-Cζ and
   D2 = β3-Lys-Cα to DFG-Phe-Cζ

Any structure not satisfying the above criteria is considered an outlier and assigned the spatial label "None."

To identify the dihedral label the DFG-Phe rotamer type in each chain was first identified (minus, plus, trans). The chains for each rotamer type were then represented with a set of 6 backbone (Φ, Ψ) dihedrals from X-DFG, DFG-Asp, DFG-Phe residues. Using these dihedrals, the distance of each kinase chain was calculated from precomputed cluster centroid points for each cluster with the same rotamer type in the given spatial group. For example, the dihedral distance for all DFGin with Phe-minus structures was computed against BLAminus, ABAminus and BLBminus. The dihedral angle distance is computed using the following formula,

$$D(i,j) = \frac{1}{6}\left(D\left(\emptyset_i^X, \emptyset_j^X\right) + D\left(\psi_i^X, \psi_j^X\right) + D\left(\emptyset_i^D, \emptyset_j^D\right) + D\left(\psi_i^D, \psi_j^D\right) + D\left(\emptyset_i^F, \emptyset_j^F\right) + D\left(\psi_i^F, \psi_j^F\right)\right)$$

where, $D(\theta_1, \theta_2) = 2(1 - \cos(\theta_1 - \theta_2))$

A chain is assigned to a dihedral label if the distance from that cluster centroid is less than < 0.45. The chains which have any motif residue missing or are distant from all the cluster centroids are assigned the dihedral label "None."

The C-helix disposition is determined using the distance between Cβ atoms of B3-Lys and C-helix-Glu(+4). A distance of <10 Å indicates that the salt bridge between the two residues is present suggesting a C-helix-in conformation. A value of >10 Å suggests a C-helix-out conformation.

Ligand classification

The different regions of the ATP binding pocket are identified by specific residues using our common numbering scheme (Supplementary figure 1):

- ATP binding region – hinge residues – residues 426-428

- Back pocket - C-helix and partial regions of B4 and B5 strands, DFGmotif backbone – residues 106-147, 150-152, 184, 187-195, 420-422 and 1337-1339
- Type II-only pocket – exposed only in DFGout conformation – residues 153, 149, 959 and 1011

A contact between ligand atoms and protein residues is defined if the distance between any two atoms is ≤ 4.5 Å (hydrogens not included). Based on these contacts we have labeled the ligand types as follows:

1. Allosteric: Any small molecule in the asymmetric unit whose minimum distance from the hinge region and C-helix-Glu(+4) residue is greater than 6.5 Å.
2. Type I½: subdivided as – Type I½-front – at least three or more contacts in the back pocket and at least one contact with the N-terminal region of the C-helix. Type I½_back - at least three or more contacts in the back pocket but no contact with N-terminal region of C-helix.
3. Type II – at least three or more contacts in the back pocket and at least one contact in the Type2-only pocket.
4. Type III – minimum distance from the hinge greater than 6 Å and at least three contacts in the back pocket.
5. Type I – all the ligands which do not satisfy the above criteria.

Identify conformation using webserver

The program uses the structure file uploaded by the user to extract the sequence of the protein. It aligns the sequence with precomputed HMM profiles of kinase phylogenetic groups (e.g. AGC.hmm, CAMK.hmm). The alignment with the best score is identified and used to determine the positions of the DFGmotif, B3-Lys, and C-helix-Glu(+4) residues. The program then computes the distance between specific atoms and dihedrals to identify spatial and dihedral labels using the assignment method described above.

Standalone program

The standalone program is written in Python3.7. The program is available to download from https://github.com/vivekmodi/Kincore-standalone and can be run in a MacOS or Linux machine terminal window. The user can provide individual .pdb or .cif (also compressed .gz) file or a list of files as an input. It identifies the unknown conformation from a structure file in the same way as described for the webserver.

Software and libraries used

All the scripting and analysis is done using Python3 and depends on Pandas (https://pandas.pydata.org), and Biopython [30] libraries.

Website and Database

Kincore is developed using Flask web framework (https://flask.palletsprojects.com/en/1.1.x/). The webpages are written in HTML5 and style elements created using Bootstrap v4.5.0

(https://getbootstrap.com/). The 3D visualization is done by using NGL Viewer (http://nglviewer.org/ngl/api/). PyMOL (v2.3) is used for creating download sessions [25]. The entire application is deployed on the internet using Apache2 webserver.

Acknowledgements

The authors want to thank Maxim Shapovalov for his help in deploying the server. This work was funded by NIH grant R35 GM122517 to R.L.D.

References

1. Adams, J.A., *Kinetic and catalytic mechanisms of protein kinases.* Chem Rev, 2001. **101**(8): p. 2271-90.
2. Blume-Jensen, P. and T. Hunter, *Oncogenic kinase signalling.* Nature, 2001. **411**(6835): p. 355-365.
3. Lahiry, P., et al., *Kinase mutations in human disease: interpreting genotype-phenotype relationships.* Nat Rev Genet, 2010. **11**(1): p. 60-74.
4. Zhang, J., P.L. Yang, and N.S. Gray, *Targeting cancer with small molecule kinase inhibitors.* Nat Rev Cancer, 2009. **9**(1): p. 28-39.
5. Ferguson, F.M. and N.S. Gray, *Kinase inhibitors: the road ahead.* Nature Reviews Drug Discovery, 2018. **17**(5): p. 353-377.
6. Manning, G., et al., *The protein kinase complement of the human genome.* Science, 2002. **298**(5600): p. 1912-34.
7. Modi, V. and R.L. Dunbrack, Jr., *A Structurally-Validated Multiple Sequence Alignment of 497 Human Protein Kinase Domains.* Sci Rep, 2019. **9**(1): p. 19790.
8. Vijayan, R., et al., *Conformational analysis of the DFG-out kinase motif and biochemical profiling of structurally validated type II inhibitors.* Journal of medicinal chemistry, 2015. **58**(1): p. 466-479.
9. Möbitz, H., *The ABC of protein kinase conformations.* Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 2015. **1854**(10): p. 1555-1566.
10. Ung, P.M.-U., R. Rahman, and A. Schlessinger, *Redefining the protein kinase conformational space with machine learning.* Cell chemical biology, 2018. **25**(7): p. 916-924. e2.
11. Modi, V. and R.L. Dunbrack, *Defining a new nomenclature for the structures of active and inactive kinases.* Proceedings of the National Academy of Sciences, 2019. **116**(14): p. 6818-6827.
12. Lange, S.M., et al., *Dimeric Structure of the Pseudokinase IRAK3 Suggests an Allosteric Mechanism for Negative Regulation.* Structure, 2020.
13. Paul, F., Y. Meng, and B. Roux, *Identification of Druggable Kinase Target Conformations Using Markov Model Metastable States Analysis of apo-Abl.* J Chem Theory Comput, 2020. **16**(3): p. 1896-1912.
14. Paul, A. and N. Srinivasan, *Genome-wide and structural analyses of pseudokinases encoded in the genome of Arabidopsis thaliana provide functional insights.* Proteins, 2020. **88**(12): p. 1620-1638.
15. Dar, A.C. and K.M. Shokat, *The evolution of protein kinase inhibitors from antagonists to agonists of cellular signaling.* Annu Rev Biochem, 2011. **80**: p. 769-95.

16. Zuccotto, F., et al., *Through the "gatekeeper door": exploiting the active kinase conformation.* J Med Chem, 2010. **53**(7): p. 2681-94.
17. Gavrin, L.K. and E. Saiah, *Approaches to discover non-ATP site kinase inhibitors.* MedChemComm, 2013. **4**(1): p. 41-51.
18. Fang, Z., C. Grutter, and D. Rauh, *Strategies for the selective regulation of kinases with allosteric modulators: exploiting exclusive structural features.* ACS Chem Biol, 2013. **8**(1): p. 58-70.
19. van Linden, O.P., et al., *KLIFS: A knowledge-based structural database to navigate kinase-ligand interaction space.* J Med Chem, 2013.
20. Roskoski, R., Jr., *Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes.* Pharmacol Res, 2016. **103**: p. 26-48.
21. consortium, w., *Protein Data Bank: the single global archive for 3D macromolecular structure data.* Nucleic Acids Research, 2018. **47**(D1): p. D520-D528.
22. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of database programs.* Nucleic Acids Research, 1997. **25**: p. 3389-3402.
23. UniProt Consortium, *UniProt: a hub for protein information.* Nucleic Acids Res, 2015. **43**(Database issue): p. D204-12.
24. Velankar, S., et al., *SIFTS: Structure Integration with Function, Taxonomy and Sequences resource.* Nucleic Acids Research, 2013. **41**(D1): p. D483-D489.
25. DeLano, W.L., *The PyMOL molecular graphics system*. 2002, Schrödinger, Inc.: San Carlos, CA.
26. Rahman, R., P.M.-U. Ung, and A. Schlessinger, *KinaMetrix: a web resource to investigate kinase conformations and inhibitor space.* Nucleic acids research, 2018. **47**(D1): p. D361-D366.
27. Kanev, G.K., et al., *KLIFS: an overhaul after the first 5 years of supporting kinase research.* Nucleic Acids Research, 2021. **49**(D1): p. D562-D569.
28. Kirubakaran, P., et al., *Comparative Modeling of CDK9 Inhibitors to Explore Selectivity and Structure-Activity Relationships.* bioRxiv, 2020: p. 2020.06.08.138602.
29. Wang, G. and R.L. Dunbrack, Jr., *PISCES: recent improvements to a PDB sequence culling server.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W94-8.
30. Cock, P.J., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-3.