Systematic identification and quantification of substrate specificity determinants in human protein kinases

(Identificación y cuantificación sistemática de determinantes de la especificidad por sustrato en las proteínas quinasas de humano)

Manuel Alejandro Alonso Tarajano

Tesi Doctoral

Universitat de Barcelona
Facultat de Farmàcia
Departament de Bioquímica i Biologia Molecular
Programa de Doctorat en Biomedicina
2013

Systematic identification and quantification of substrate specificity determinants in human protein kinases

(Identificación y cuantificación sistemática de determinantes de la especificidad por sustrato en las proteínas quinasas de humano)

Memòria presentada per Manuel A. Alonso Tarajano, per optar al títol de doctor per la Universitat de Barcelona

This thesis was realized under the supervision and tutorship of ICREA research professor Dr. Patrick Aloy Calaf and under the co-supervision of Dr. Roberto Mosca, in the Structural Bioinformatics and Network Biology group of the Institute for Research in Biomedicine (IRB) Barcelona.

Manuel A. Alonso Tarajano Dr. Patrick Aloy Calaf Dr. Roberto Mosca

Contents

List of Figures

List of Tables

List of Acronyms

3D Three dimensional

AKAP A-kinase anchoring proteins

aPK Atypical protein kinases

ATP Adenosine triphosphate

AUC-ROC Area under the receiver operating characteristic curve

cAMP Cyclic adenosine monophosphate

CDK Cyclin-dependent kinase

ePK Eukaryotic kinase domain

FPR False positive rate

GO Gene Ontology

IC Information content

kAS Known adaptor/scaffold protein

KSR Kinase suppressor of Ras

MAPK Mitogen-activated protein kinase

MF Molecular Function category, Gene Ontology database

pAS Potential adaptor/scaffold protein

PIKK Phosphatidyl inositol 3' kinase-related kinases

PPI Protein-protein interaction

PTM Post-translational modification

ROC Receiver operating characteristic

SDR Specificity-determinant residue

SH2 Src homology-2 domain

SH3 Src homology-3 domain

SLiM Short linear motif

TPR True positive rate

1 Resumen global

1.1 Introducción

Las modificaciones postraduccional constituyen uno de los mecanismos más frecuentes de regulación de proteínas. La fosforilación es la modificación postraduccional más común en eucariontes, y ha sido estimado que al menos el 30 % de las proteínas de estos organismos son objeto de fosforilación. La fosforilación es la adición de un grupo fosfato (proveniente de una molécula de ATP) a residuos de serina, treonina o tirosina de una proteína diana, o sea, el sustrato. La fosforilación es una reacción rápida y reversible, que puede modificar uno o varios aspectos de la proteína afectada como por ejemplo su función, su localización celular y sus interacciones con otras proteínas.

Las proteínas quinasas son las enzimas encargadas de catalizar la reacción de fosforilación, una acción que es contrarrestada por las fosfatasas. En humano han sido reportadas 518 proteínas quinasas, las cuales abarcan aproximadamente el 2 % de los genes y constituyen una de las familias de proteínas más numerosas. Las quinasas están involucradas en un gran número de procesos y funciones celulares como por ejemplo la señalización, la transcripción, el transporte, la duplicación, el crecimiento y la proliferación entre otros. Debido a la importancia de los procesos en los cuales están involucradas, la desregulación o abolición de las funciones de numerosas quinasas han sido relacionadas con importantes patologías en humanos como son el cáncer y la diabetes. En consecuencia, muchas proteínas quinasas constituyen importantes dianas terapéuticas, para las cuales han sido registradas desde el 2001 — tan sólo en los Estados Unidos — más de 10'000 solicitudes de patentes de inhibidores.

Las proteínas quinasas poseen una gran variedad de secuencias así como de funciones biológicas, muchas de las cuales son aportadas por la presencia de diversos dominios funcionales. No obstante, la mayoría de las quinasas poseen un dominio catalítico de entre 250 y 300 residuos cuya estructura tridimensional ha sido conservada durante la evolución. Este dominio es comúnmente conocido como 'dominio quinasa de eucariontes'. En dicho dominio, varios motivos de secuencia y estructurales han sido particularmente conservados, fundamentalmente los relacionados con la unión de ATP y con la transferencia del grupo fosfato al sustrato. Sin embargo, existe un grupo de quinasas cuyos dominios catalíticos no poseen homología de secuencia con el antes mencionado 'dominio quinasa de eucariontes'. A este otro grupo se le denomina comúnmente 'quinasas atípicas'.

En general, las proteínas quinasas poseen una gran variedad en la especificidad mostrada *in vivo* por los sustratos. Dicha especificidad tiene una relación muy limitada con la secuencia primaria de las quinasas, no obstante, se ha observado que quinasas de una misma familia son más propensas a compartir sustratos. Durante finales de los años ochenta y principios de los noventa, se pensó que la especificidad por sustrato de las quinasas era determinada, fundamentalmente, por el segmento de secuencia del sustrato que contiene el sitio (residuo) de

fosforilación. Varios de estos estudios identificaron secuencias consenso para varias quinasas. Sin embargo, para la mayoría de las quinasas, las secuencias consenso no resultaban suficientes para explicar la especificidad observada *in vivo*. Estudios posteriores han demostrado que la selección *in vivo* de los sustratos es guiada, además, por otros factores como por ejemplo la colocalización celular de la enzima y el sustrato, las interacciones entre dominios y/o sitios de acoplamiento y la asociación de las quinasas con proteínas adaptadoras y/o plataformas (A/P), las cuales facilitan las asociaciones entre varias proteínas mediante la formación de complejos macromoleculares.

1.2 Objetivos

El objetivo general de esta tesis es la cuantificación de la contribución de varios elementos que contribuyen (*in vivo*) a la especificidad por sustratos de las proteínas quinasas de humano. En la práctica, hemos dividido este objetivo en dos partes:

- **Cuantificación de la contribución a la especificidad del sitio de fosforilación y sus residuos vecinos en secuencia.**
- **Cuantificación de la contribución a la especificidad de la asociación de las quinasas con proteínas adaptadoras y/o plataformas.**

1.3 Materiales y Métodos

Para nuestro análisis fue necesario contar con la lista de proteínas quinasas en humano, así como con su clasificación filogenética. Dichos datos fueron descargados desde el sitio web http://www.kinase.com/human/kinome.

Para la colección de datos de fosforilaciones en proteínas humanas, integramos información proveniente de las tres mayores bases de datos públicas al respecto, HPRD, Phospho-ELM y PhosphoSitePlus. En cada caso, sólo incluimos sitios de fosforilación que hayan sido determinados experimentalmente y para los cuales se conoce la quinasa responsable. En adición, a fin de evaluar posteriormente el rendimiento de las matrices de puntuación por posiciones generadas para cada quinasa y familia de quinasas, compilamos un conjunto negativo de fosforilaciones en humano. En dicho conjunto sólo incluimos proteínas para las cuales no han sido reportadas fosforilaciones y que tampoco contienen ninguna de las secuencias presentes en los 5946 sitios de fosforilación de nuestro set. Posteriormente, eliminamos la redundancia del conjunto de proteínas usando un umbral del 100 % de identidad de secuencia. De este modo, obtuvimos un set compuesto por 8876 proteínas.

Los datos de interacciones entre proteínas humanas fueron obtenidos a partir de la base de datos local PPI-DB. Esta base de datos integra información de numerosos recursos públicos (*e.g.*, Intact, MINT, DIP y HPRD) y contiene alrededor de 45'000 interacciones binarias de alta confiabilidad, determinadas experimentalmente.

Los logotipos de secuencias fosforiladas fueron generados usando el programa `WebLogo`. Como entrada al programa, proporcionamos los alineamientos de las secuencias fosforiladas

por cada quinasa o familia de quinasas. Dichas secuencias fueron definidas como segmentos de 9 residuos, donde el residuo fosforilado se encuentra en la posición central.

Para generar, evaluar el rendimiento y estimar la significación estadística de las matrices de puntuación por posiciones (PSSM, por sus siglas en inglés) desarrollamos el programa `genpssm`. Como entrada, `genpssm` requiere varios datos, entre ellos un alineamiento de secuencias fosforiladas, la frecuencia de cada amino ácido en el proteoma humano, un valor umbral de significación estadística (valor p) para la identificación de secuencias que concuerden con el modelo representado por la PSSM y el conjunto de secuencias no fosforiladas. Para calcular la puntuación de cada residuo en cada posición de la secuencia fosforilada, `genpssm` toma en consideración el cociente de probabilidades de cada residuo en dicha posición así como la frecuencia de dicho residuo en el proteoma de referencia. Para estimar la significación estadística de una PSSM, `genpssm` toma como estadísticos el contenido de información así como la sensitividad de dicha matriz. En este caso, la sensitividad es la fracción de secuencias semilla (*i.e.*, aquellas presentes en el alineamiento a partir del cual se generó la PSSM) que concuerdan con la propia PSSM a un nivel de significación estadístico deseado. La significación estadística es estimada mediante el cálculo de valores p empíricos, generados a partir de distribuciones nulas. Para obtener el rendimiento de una PSSM, `genpssm` calcula la curva Característica Operativa del Receptor (ROC, por sus siglas en inglés) correspondiente. Una curva ROC es la representación gráfica de la sensibilidad frente a (1 – especificidad) para un clasificador binario. Mediante una curva ROC podemos evaluar cuán bien una PSSM puede distinguir entre sus secuencias semillas (positivos) y secuencias en el conjunto negativo de fosforilaciones (negativos).

Por otra parte, también hemos empleado las PSSMs para identificar residuos — en las secuencias fosforiladas — que probablemente contribuyan de forma significativa a la especificidad de las quinasas. A dichos residuos los hemos denominado 'residuos determinantes de especificidad' (SDR, por sus siglas en inglés). Para clasificar un residuo como SDR, comparamos su puntuación en la PSSM con la puntuación del residuo fosforilado. Dada una PSSM, si la puntuación de un residuo en una posición determinada es igual o mayor que la mitad de la puntuación del residuo fosforilado, entonces dicho residuo es clasificado como un SDR.

Con el propósito de recopilar — de manera automática — un conjunto de proteínas humanas para las cuales se ha descrito que actúan como A/P, empleamos la base de datos UniProt. Basados en la anotación funcional en UniProt, hemos compilado un primer conjunto de proteínas que contienen al menos uno de los términos 'adaptor' (adaptadora) o 'scaffold' (plataforma). Seguidamente, descartamos aquellas proteínas para las cuales no existen evidencias de una interacción binaria con al menos una quinasa. Finalmente, comprobamos que los términos 'adaptor' o 'scaffold' de la anotación funcional estén asociados directamente a la función molecular de la proteína de interés.

Para la identificación de proteínas con posible función A/P de quinasas, nuestro método se basa en la selección de aquellas proteínas que interaccionan con una cantidad estadísticamente significativa de los sustratos de una quinasa. En este modelo asumimos que dichas proteínas tienen mayores posibilidades de promover interacciones entre las quinasas y sus correspondientes sustratos. La significación estadística es estimada utilizando como estadígrafo la fracción de sustratos (de una misma quinasa) que interacciona con la proteína de interés. Para el estadígrafo calculamos un valor p empírico a partir de distribuciones nulas, las cuales

toman en cuenta la cardinalidad del conjunto de sustratos de cada quinasa. Para la construcción de dichas distribuciones utilizamos una subred del interactoma humano compuesta por las quinasas, sus sustratos y los vecinos de primer grado de ambos. Finalmente, obtenemos 10'000 valores del estadígrafo seleccionando aleatoriamente (de la subred) un conjunto de proteínas en representación de los sustratos y calculando la fracción de los mismos que interacciona con un mismo vecino. Durante el cálculo de estas fracciones, las asociaciones entre proteínas de la subred son intercambiadas de forma aleatoria.

Por otra parte, también evaluamos si quinasas que comparten proteínas A/P, son más propensas a tener más sustratos en común. En este modelo, asumimos que las proteínas A/P contribuyen a la especificidad de las quinasas por los sustratos que las mismas comparten. En este modelo usamos como estadígrafo el número de sustratos compartidos por dos o más quinasas. Para dicho estadígrafo calculamos un valor *p* empírico a partir de distribuciones nulas, las cuales consideran la cardinalidad del grupo de quinasas que comparten una proteína A/P. Las distribuciones nulas son obtenidas a partir del cómputo del estadígrafo para grupos de quinasas con cardinalidad 2 o 3. En total, las distribuciones contienen 1000 valores del estadígrafo.

En adición, hemos evaluado si proteínas A/P, interaccionan con una fracción estadísticamente significativa de los sustratos de las quinasas con las cuales se asocian. Para esto, empleamos un conjunto curado de 51 asociaciones entre 31 quinasas y 36 proteínas A/P. Para cada asociación, verificamos la existencia de al menos una publicación donde se haya demostrado que la proteínas A/P desempeña dicha función sobre la quinasa.

En este análisis hemos usado como estadígrafo el número de sustratos de cada quinasa que interacciona con una proteína A/P dada, y estimamos su significación estadística mediante el cálculo de un valor *p* empírico. En el cálculo del valor *p* (usando distribuciones nulas) tomamos en consideración la cardinalidad de sustratos de las quinasas analizadas. La subred del interactoma humano empleada para generar las distribuciones nulas fue compilada tomando los primeros vecinos de 111 quinasas humanas (con al menos cinco sustratos *in vivo*) y los de sus respectivos sustratos. Dicha subred contiene 15'046 interacciones entre 5939 proteínas. Para generar distribuciones del estadígrafo, primeramente seleccionamos de forma aleatoria un nodo Q (quinasa hipotética) de la subred. Posteriormente, tomamos el conjunto de primeros vecinos de Q y los dividimos de forma aleatoria en dos subgrupos, sustratos hipotéticos (S, de cardinalidad igual al número de sustratos de Q) y A/P hipotéticos (AP). Finalmente, obtenemos valores del estadígrafo calculando cuántos elementos en S interaccionan con cada elemento en AP. Para cada valor de cardinalidad de sustratos, el algoritmo realiza iteraciones hasta alcanzar 10'000 valores del estadígrafo. En cada iteración, las asociaciones entre proteínas de la subred son intercambiadas de forma aleatoria.

Para analizar el enriquecimiento de dominios funcionales en grupos de proteínas, hemos empleado la base de datos de familias de proteínas Pfam. A fin de estimar dicho enriquecimiento implementamos un test hipergeométrico, realizamos correcciones de test múltiples usando los métodos de Bonferroni y Benjamini-Hochberg y definimos un umbral de significación estadística α <0.05. Como conjunto control utilizamos el proteoma humano, en el cual incluimos únicamente proteínas presentes en la base de datos Swiss-Prot y que hayan sido detectadas a nivel de proteína, de ARN o que hayan sido inferidas por homología de secuencia.

Para detectar el enriquecimiento funcional de grupos de proteínas, hemos empleado las anotaciones de la base de datos 'Gene Ontology' (GO). GO emplea un vocabulario controlado de términos funcionales para describir las características de genes y sus productos, en base a tres categorías: 'proceso biológico', 'componente celular' y 'función molecular'. El análisis de enriquecimiento lo hemos realizado empleando el programa `GOstats` (test hipergeométrico), con corrección para test múltiples mediante el método de Bonferroni y definiendo un umbral de significación estadística α <0.05. Como grupo de control empleamos el conjunto de sustratos y vecinos de primer nivel de las quinasas humanas.

1.4 Resultados y discusión

En nuestros datos de fosforilación en humano están representadas 325 proteínas quinasas humanas (62.7 %) y 1856 sustratos. En total, hemos compilado 5946 sitios de fosforilación distintos, de los cuales 3583 (60 %) han sido determinados *in vivo*. Mediante la integración de distintas bases de datos, conseguimos incrementos del 18 %, 58 % y 59 % en las cantidades de quinasas, sustratos y sitios de fosforilación (respectivamente) con respecto a la media contenida en las bases de datos de referencia.

La comparación visual de varios de nuestros logotipos de secuencia con los previamente reportados en la literatura, muestra una concordancia en cuanto a los motivos de secuencia reconocidos por varias quinasas y familias de quinasas. Mediante estas comparaciones, hemos comprobado que de manera general nuestros datos de fosforilación reflejan correctamente los patrones de secuencias reconocidos por distintas clases de quinasas como son los casos de las basofílicas, acidofílicas, las guiadas por prolina y las guiadas por glutamina.

El análisis de SDRs para quinasas guiadas por prolina, nos permitió identificar para tres familias (CDK, MAPK y GSK) varios residuos relevantes en el reconocimiento de sustrato. Entre las secuencias fosforiladas por las tres familias mencionadas, los SDRs de mayor relevancia resultaron ser residuos de prolina (P) en las posición +1 (P+1) y -2 (P-2) con respecto del residuo de fosforilación. Dichos SDRs están presentes, como promedio, en el 74.85 % y 25.12 % — respectivamente — de los eventos de fosforilación de las familias mencionadas. En comparación, las frecuencias promedio de P+1 y P-2 entre eventos de fosforilación de quinasas que no son guiadas por prolina son del 5.95 % y 5.83 % respectivamente. En el caso particular de la familia CDK también identificamos como SDR un residuo de lisina en la posición +3 (K+3), con una frecuencia promedio del 21.25 %. Por otra parte, para la familia GSK identificamos otros tres SDRs, serina -4 (S-4), prolina +2 (P+2) y serina +4 (S+4) con frecuencias promedio del 38.49 %, 27.70 % y 48.56 % — respectivamente — entre eventos de fosforilación de quinasas GSK. Por el contrario, para la familia MAPK no encontramos ningún SDR en adición a los ya mencionados P-2 y P+1.

En el caso de las quinasas guiadas por glutamina analizamos la familia PIKK, perteneciente al grupo de las quinasas atípicas. Para esta familia sólo detectamos como SDR el propio residuo de glutamina en posición +1 (Q+1). Dicho residuo está presente en el 80.83 % de los eventos de fosforilación de quinasas PIKK y, como promedio, sólo en en el 3.98 % de los eventos de fosforilación de las otras 21 familias de quinasas incluidas en nuestro análisis.

Las quinasas basofílicas muestran preferencias por sitios de fosforilación rodeados de re-

siduos básicos (*e.g.*, arginina y lisina). Para 5/8 de las familias basofílicas analizadas, identificamos como SDRs a residuos de arginina en las posiciones -2 y/o -3 (R-2, R-3). Dichos SDRs (R-2, R-3) tienen frecuencias del 35.67 % y 45.43 % — respectivamente — entre los eventos de fosforilación de las familias basofílicas y los mismos son identificados como SDRs únicamente entre este tipo de quinasas. Por otra parte, las frecuencias de residuos R-2 y R-3 entre los eventos de fosforilación de otras clases de quinasas son, como promedio, del 5.09 % y 5.14 % respectivamente. Cuatro de las ocho familias de quinasas basofílicas analizadas pertenecen al grupo de quinasas AGC. Para dos de estas cuatro familias (AKT y PKC), en adición a R-2 y R-3, encontramos los SDRs triptófano +1 (W+1, en la familia AKT) y arginina +2 y lisina +2 (R+2, K+2, en la familia PKC). La frecuencia de W+1 entre los eventos de fosforilación de la familia AKT es considerablemente baja (3.85 %), y en contraste, las frecuencias de R+2 y K+2 para la familia PKC son comparables con las frecuencias de los SDRs que caracterizan a las quinasas basofílicas (*i.e.*, R-2, R-3). En la familia CAMK2, también de quinasas basofílicas, pero perteneciente al grupo CAMK (quinasas reguladas por calcio/calmodulina), identificamos SDRs compuestos por residuos de naturaleza ácida (*e.g.*, aspártico +2 (D+2) y glutámico +2 (E+3)). Dichos SDRs ocurren en el 23.08 % y el 23.72 % de los eventos de fosforilación de al familia CAMK2 y sólo en un 6.85 % y 6.17 % entre los eventos de fosforilación de quinasas que no tienen D+2 o E+2 como SDRs.

Otra clase de quinasas, las acidofílicas, muestran preferencia por sitios de fosforilación enriquecidos en residuos ácidos (*e.g.*, aspártico y glutámico). Una de las familias estudiadas es la de caseína quinasas 1 (CK1, perteneciente al grupo homónimo), para la cual los SDRs identificados fueron residuos de serina en las posición -3 y +3 (S-3 y S+3 respectivamente). Resulta evidente que S-3 y S+3 no constituyen ejemplos clásicos de SDRs preferidos por quinasas acidofílicas. Sin embargo, las secuencias reconocidas por miembros de la familia CK1 suelen ser previamente fosforiladas por otras quinasas en la posición S-3. De este modo, la fosforilación inicial en S-3 confiere características ácidas a la vecindad del sitio reconocido por miembros de la familia CK1. Otra familia analizada entre las quinasas acidofílicas es la también caseína quinasa 2 (familia CK2, perteneciente al grupo CMGC). En esta familia identificamos un total de ocho SDRs, todos compuestos por residuos ácidos localizados mayoritariamente las posiciones C-terminal con respecto al residuo fosforilado. Un residuo de glutámico en posición +3 (E+3), es el SDR con mayor frecuencia promedio (45.83 %) entre los eventos de fosforilación de la familia. Además, este E+3 está presente sólo en el 6.15 % de los eventos de fosforilación de quinasas que no presentan E+3 como SDR. Otro SDR identificado para la familia CK2 es aspártico en posición +1 (D+1), el cuál está presente en el 27.4 % de los eventos de fosforilación de la familia y sólo en un 4.6 % de los eventos de fosforilación de otras familias de quinasas.

A partir del análisis de las 325 PSSMs obtenidas para las quinasas en nuestros datos, hemos encontrado una correlación negativa entre el número de secuencias semilla empleadas para generar cada PSSM y la sensitividad de dicha PSSM (R = -0.59, valor $p = 2{,}38e^{-31}$); así como también entre el número de secuencias semillas y el contenido de información (CI) de la PSSM (R = -0.4, valor $p = 9{,}8e^{-14}$). Estas correlaciones sugieren que el incremento del número de secuencias semilla provoca un degeneración de la señal contenida en la PSSM, la cual afecta negativamente la sensitividad y el CI de las PSSMs. Al evaluar la significación

estadística de las PSSMs usando el CI como estadígrafo, nuestros resultados sugieren que las PSSMs estadísticamente significativas (163), fueron generadas a partir de conjuntos de secuencias semillas 10.2 veces más numerosos que las PSSMs no significativas (162). Al comparar ambos conjuntos de PSSM con respecto a su sensitividad, nuestros resultados muestran que las PSSM no significativas tienen una sensitividad media 1.4 veces mayor que la correspondiente a las PSSM significativas. Este último resultado es causado por la degeneración provocada en la señal de la PSSM al aumentar el número de secuencias semilla. De modo similar, ambos sets de PSSMs difieren significativamente en cuanto sus valores del área bajo la curva ROC (AUC-ROC) y además, hemos encontrado una correlación negativa (R = -0.63, valor $p = 6{,}2e^{-37}$) entre el AUC-ROC y el número de secuencias semilla. Al analizar las PSSMs correspondientes a las 93 familias en nuestros datos, encontramos resultados en la misma dirección que los descritos anteriormente.

En esta tesis hemos creado una estrategia para la identificación automatizada de proteínas con función conocida como adaptadoras o plataformas (cA/P) de quinasas humanas. Mediante esta estrategia, compilamos un conjunto de 191 proteínas cA/P asociadas a 287/518 (55.4 %) quinasas, las cuales representan el 72.3 % de las familias de quinasas humanas. Estos datos sugieren que, las asociaciones con proteínas con función A/P es un evento común entre las quinasas. El conjunto de las 191 cA/Ps está enriquecido en varios dominios funcionales, de los cuales se conoce promueven interacciones entre proteínas (*e.g.*, SH2, SH3 y PDZ entre otros). El análisis de términos de la ontología de genes, sugiere que el conjunto también está enriquecido en funciones moleculares relacionadas con interacciones entre proteínas (*e.g.*, 'protein binding, bridging'; 'SH3/SH2 adaptor activity' y 'protein complex scaffold' entre otros). El conjunto de 191 proteínas cA/P será empleado posteriormente, como set de prueba, para estimar la eficacia de estrategias enfocadas en la identificación de potenciales A/Ps de quinasas humanas. En adición, a partir del conjunto cA/P — identificado de manera automatizada — depuramos manualmente un conjunto de asociaciones quinasa-cA/P de elevada confiabilidad ('gold standard set', GSS). Para cada asociación quinasa-cA/P, verificamos la existencia de al menos una referencia en la literatura donde haya sido demostrada, experimentalmente, una función adaptadora o de plataforma de la proteína en cuestión sobre la quinasa a la cual está asociada. En total, el set está compuesto por 75 asociaciones entre 47 quinasas y 46 cA/Ps.

En este trabajo desarrollamos una estrategia para la identificación de proteínas con posible función A/P (pA/P) de quinasas humanas. Dicha estrategia consiste en la identificación de proteínas del interactoma humano que se asocian con un número significativo de sustratos de quinasas. Como prueba de concepto, demostramos que las 191 proteínas A/P de quinasas antes descritas, son cinco veces más propensas a interaccionar con un número significativamente mayor de los sustratos de quinasas que proteínas del interactoma humano seleccionadas al azar (razón de probabilidades = 5.04, valor $p = 1{,}08e^{-15}$). Con la presente estrategia identificamos 279 proteínas pA/P asociadas a 78 (49.7 %) de las quinasas evaluadas, para un total de 706 asociaciones quinasa-pA/P. Como conjunto, las 279 proteínas están enriquecidas en dominios funcionales (*e.g.*, 14-3-3, SH2, SH3) y términos de ontología (*e.g.*, 'SH2 domain binding' y 'SH3 domain binding', 'protein kinase binding') relacionados con interacciones entre proteínas. Sin embargo, no encontramos enriquecimientos en términos de ontología directamente relacionados con funciones como adaptadoras ('adaptor') o pla-

taformas ('scaffolds'), de modo que las proteínas identificadas no han sido relacionadas, de manera sistemática, con estas funciones. No obstante, varios términos de función molecular enriquecidos entre las 279 pA/Ps, sugieren que las mismas están vinculadas a procesos de señalización celular (*e.g.*, 'protein phosphatase binding', 'NF-kappaB binding', 'transcription factor binding'). Posteriormente, las 706 relaciones quinasa-pA/P antes mencionadas fueron filtradas sobre la base de evidencia de colocalización celular de la pA/P con los sustratos de la quinasa correspondiente. Como resultado, obtuvimos un conjunto de 527 relaciones quinasa-pA/P que involucra 41 quinasas, 156 pA/Ps y 35 compartimentos celulares distintos. En nuestra opinión, estos resultados sugieren que para la mayoría de las relaciones quinasa-pA/P identificadas — 527/706, 74.6 % — , la proteína pA/S podría desempeñar un papel importante facilitando la colocalización de la quinasa con sus sustratos.

En adición, hemos intentado aportar más evidencia con relación al rol de las proteínas cA/Ps en promover la colocalización celular de sustratos y quinasas. Para esto realizamos un análisis a fin de corroborar si, empleando un subconjunto del GSS conformado por 49 pares quinasa–cA/P de elevada confiabilidad, las proteínas cA/P son propensas a interaccionar con un número significativo de los sustratos (*in vivo*) de sus quinasas asociadas. En total, para 10/49 (10.2 %) de los pares analizados encontramos resultados satisfactorios. Estos resultados no nos permiten generalizar — al menos en este set reducido — sobre el rol de las proteínas cA/Ps como reclutadores de sustratos para las quinasas; no obstante, consideramos que los resultados mencionados en el párrafo anterior, aportan evidencia sobre la mencionada función de las proteínas A/P.

En esta tesis hemos estudiado la contribución de las proteínas A/P a determinar la especificidad por sustrato de las quinasas, al promover la colocalización entre dichas quinasas y sus sustratos. Sin embargo, cabría preguntarse si la asociación de dos quinasas distintas a una misma proteína A/P conduciría a que ambas quinasas compartiesen (*in vivo*) un número de sustratos mayor del esperado al azar. Partiendo de nuestra hipótesis sobre la contribución de las proteínas A/P a la especificidad por sustrato de las quinasas, sería de esperar pocos casos de quinasas compartiendo cantidades significativas de sustratos (*in vivo*). Para este análisis contamos con 19 casos de una proteína cA/P que interacciona con al menos dos quinasas distintas las cuales comparten al menos un sustrato *in vivo*. Nuestros resultados muestran que en ninguno de los casos analizados las quinasas comparten una cantidad significativa de sustratos *in vivo*. En nuestra opinión, estos resultados apoyan el modelo en que las proteínas A/P contribuyen de manera efectiva a reducir el conjunto de sustratos potenciales de las quinasas a las cuales se asocian, probablemente mediante el reclutamiento de las quinasas a complejos moleculares o locaciones celulares específicas.

1.5 Conclusiones

- Mediante la integración de datos provenientes de distintos recursos públicos, hemos compilado un conjunto de relaciones quinasa–sitio de fosforilación determinados experimentalmente en humano. Nuestros datos incluyen información correspondiente a 325 (62.7 %) quinasas — en representación del 71.5 % de las familias de quinasas humanas —, 1856 sustratos y 5946 sitios de fosforilación. Mediante la integración de información logramos aumentos del 18 %, 58 % y 59 % en las cantidades de quinasas, sustratos y sitios de fosforilación (respectivamente), en comparación con la media contenida en las bases de datos empleadas como fuente.

- Los patrones observados en los logotipos, mostraron la gran diversidad de motivos de secuencia reconocidos por las quinasas. Además, dichos logotipos sirvieron como guía para la clasificación de las quinasas y las familias de quinasas basados en la composición (*i.e.*, los tipos de residuos) de las secuencias que fosforilan.

- Basados en su puntuación en las PSSMs, hemos clasificado varios residuos como SDRs para varias de las familias de quinasas. Hemos observado que la identidad, la posición en el alineamiento, así como la frecuencia de los SDRs identificados, varía considerablemente entre las familias de quinasas analizadas.

- La significación estadística de las PSSMs generadas fue evaluada tomando en consideración la sensitividad y el contenido de información de las PSSMs. Primeramente, hemos encontrado correlaciones negativas ente la cantidad de sitios semilla y i) la sensitividad de las PSSMs y ii) el contenido de información de las mismas. Basados en el valor de IC de las PSSMs y en la comparación con distribuciones nulas, encontramos que las PSSMs estadísticamente significativas difieren de las no estadísticamente significativas en cuanto al IC, la sensitividad, el número de secuencias semilla y el AUC-ROC.

- Desarrollamos una estrategia para la identificación computacional de proteínas con función conocida como adaptadoras o plataformas de quinasas humanas (cA/P). En total, hemos identificado un grupo de 191 proteínas cA/P, el cual está enriquecido en dominios de proteínas y anotaciones funcionales consistentes con el papel de adaptadoras o plataformas. Las 191 proteínas están asociadas al 55 % de las quinasas humanas, lo cual sugiere que la asociación a proteínas A/P es un fenómeno común entre estas enzimas.

- Nuestros análisis sugieren que, en comparación con proteínas seleccionadas al azar, las 191 proteínas con función adaptadora o plataforma son cinco veces más propensas a interaccionar con cantidades significativas de los sustratos de las quinasas a las cuales están asociadas.

- A partir de un conjunto de 156 quinasas humanas, para 78/156 (50 %) de ellas identificamos un grupo de 279 proteínas con posible función adaptadora o plataforma (pA/P).

Este conjunto está enriquecido en dominios de proteínas y anotaciones funcionales relacionadas con la función predicha y que además sugiere la implicación de estas proteínas en procesos de señalización celular.

- Nuestros análisis de colocalización celular sugieren que, para el 74.6 % de las asociaciones quinasa–pA/P encontradas, la proteína pA/P pueden desempeñar un papel importante en la colocalización de las quinasas y sus correspondientes sustratos.

- Finalmente hemos analizado la relación entre la asociación de quinasas distintas con proteínas A/P en común y la especificidad cruzada entre dichas quinasas (*i.e.*, el número de sustratos compartidos *in vivo*). Nuestros resultados sugieren que quinasas con proteínas A/P en común no comparten más sustratos *in vivo* de lo que cabría esperar al azar. En nuestra opinión, estos resultados sugieren que las proteínas A/P son capaces de reducir el conjunto de sustratos potenciales de las quinasas a las cuales se asocian, probablemente mediante el reclutamiento de dichas enzimas a complejos moleculares o locaciones celulares específicas.

2 General introduction

2.1 Protein phosphorylation

Phosphorylation is the most common post-translational modification (PTM) of proteins, and is of major relevance for the regulation of most cellular processes [1,2]. In a phosphorylation reaction, the gamma phosphate group of an ATP molecule is transferred to an acceptor amino acid in a target protein, the substrate. The transferred phosphate group carries a strong negative charge, that can promote electrostatic interactions in the surroundings of the phosphorylated residue. These new electrostatic interactions can affect the substrate in many aspects such as its conformational state, its interactions with other proteins and its cellular localization among many others [3]. Moreover, protein phosphorylation is a reversible and fast process that have been conserved in evolution as a mechanism for regulating proteins function in a non transcription-dependent manner [4]. The biological relevance of this process is evident by the fact that around 30% of all eukaryotic proteins have been suggested to be target of phosphorylation [1].

2.2 Protein kinases

2.2.1 Kinases are responsible for protein phosphorylation

Kinases are the enzymes responsible for protein phosphorylation. Kinases are capable of catalyzing the addition of the phosphate group from an ATP molecule to the substrate (see Figure 2.1). For human, there is a total of 518 protein kinases described, which account for approximately 2% of all human genes [5]. Several kinases are known to be key players in many cellular processes such a growth, differentiation and apoptosis; and as a consequence, their deregulation have been tightly related to major human pathologies such as cancer [4,6] and diabetes [7, 8]. Moreover, kinases constitute a major class of therapeutic targets for which more than 10'000 patent applications for inhibitors have been received since 2001 in the United States alone [9, 10]. Based on the type of residues that they phosphorylate, kinases can be classified into serine/threonine, tyrosine or dual-specificity kinases. In human, serine/threonine constitute the larger group (69.3%), followed by the tyrosine (16.9%) and the dual-specificity kinases (4.1%). The remaining 9.7% of the 518 human kinases have been suggested to be catalytically inactive due to their lack of canonical catalytic residues [5]. Serine, threonine and tyrosine residues are not phosphorylated with the same frequency. In human, for example, it has been reported that serine residues account for the largest number of known phosphorylations (86.4%), followed by threonines (11.8%) and tyrosines (1.8%) [11].

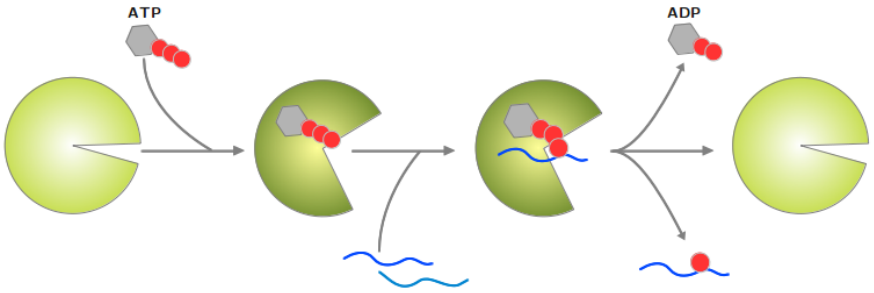


Figure 2.1: Protein phosphorylation reaction.

A protein kinase (green) catalyze the addition of a phosphate group (red circles) from an ATP molecule to a substrate peptide (blue ribbons). The catalytically active kinase is represented in dark green.

2.2.2 The kinase catalytic domain is highly conserved

Due to their functional relevance, protein kinases have been conserved along the evolution from bacteria to metazoan [12, 13]. Moreover, most protein kinases share a common fold of their catalytic domains. In human, most protein kinases (92.3%) share a common fold for the catalytic domain, known as the canonical eukaryotic kinase domain (ePK) [5]. The ePK domain ranges between 250 and 300 amino acids and have a bi-lobular structure (see Figure 2.2). The N-terminal lobe is the smallest of the two and is mainly composed of beta sheets, while the C-terminal lobe is mostly composed of alpha helices. Both lobes are joined by a flexible hinge segment, which allows considerable conformational flexibility to the domain. The ATP binding site, highly conserved among ePK domains, is formed by a deep cleft between the lobes; while the binding site for the substrate peptide lies in a more solvent accessible region, which contains the catalytic residues, that is also between the lobes. Another important structural feature of the ePK domains is the activation loop, a segment that regulates the active state of several kinases upon phosphorylation, [14], and that can influence substrate binding and catalytic efficiency [15]. The remaining 7.7% of human kinases have a catalytic domain that lack sequence similarity to the ePK domain [5]. Due to this characteristic, they are commonly known as atypical protein kinases (aPK) and they have been proposed to have diverged from ePKs early in evolution [12]. Most aPKs have been discovered mainly by biochemical methods or by clear sequence homology to other aPKs. Despite the lack of sequence similarity, for some aPKs a significant structural similarity to the ePK domain has been reported [16–18]. See panels A and B in Figure 2.2.

2.2.3 Phylogeny and diversity of human protein kinases

Starting from genomic sequences, Manning *et al.* [5] cataloged human kinases based on a comparative analysis of the sequence of the catalytic domains, complemented by an analysis of the sequence similarity and domain structure of non-catalytic regions. The authors reported a total of nine major groups, subdivided in 130 different families and 198 subfamilies,

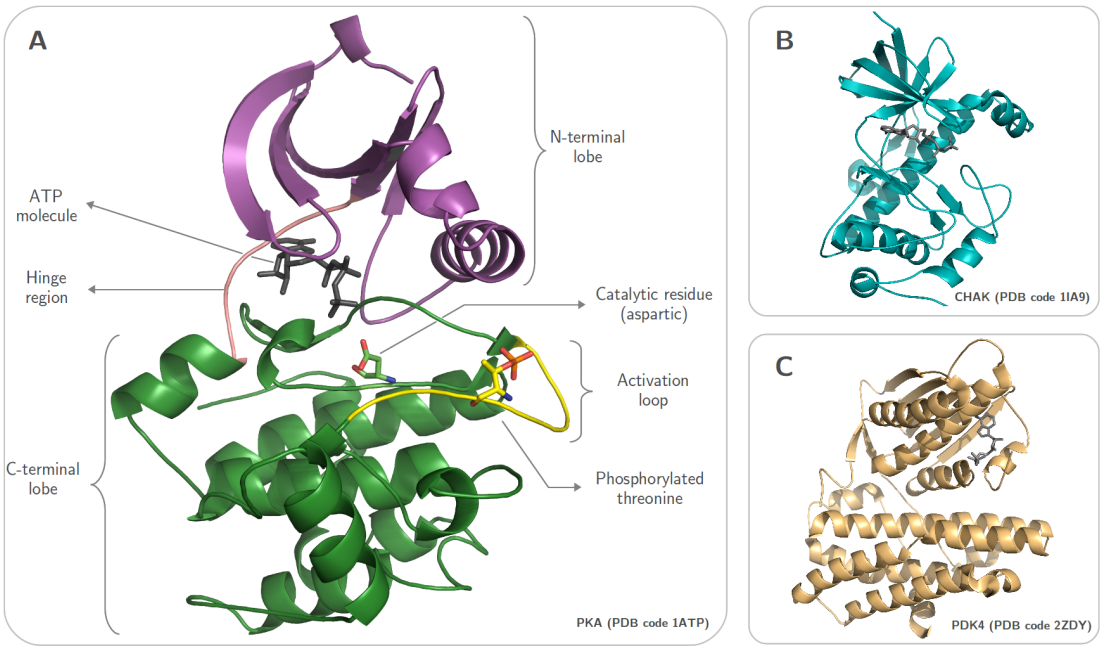


Figure 2.2: Eukaryotic kinase catalytic domain.

A) Catalytic (canonical) domain of protein kinase A (PKA). In magenta, the N-terminal lobe, followed by the hinge region (salmon) and the C-terminal lobe (green). In gray sticks, between the lobes, an ATP molecule. In yellow, the activation loop with a phosphorylated threonine, which stabilizes the kinase in an active conformation. In green sticks the catalytic residue (aspartic), responsible of transferring the phosphate group to the substrate. **B**) Catalytic domain of the atypical kinase CHAK. In gray sticks a molecule of ANP, an analog of ATP. **C**) Catalytic domain of the atypical kinase PDK4. In gray sticks a molecule of ADP.

a finding that highlighted the great sequence diversity of human kinases.

Kinases are an example of great functional diversity that have been achieved through sequence variation and by versatile modular combination of different classes of protein domains [19] (see Figure 2.3). In general, members of the same kinase family tend to retain similar domain composition [5], which leads in some cases to shared functional properties. In protein kinases, only a small group of residues (highly important for ATP binding and for the transfer of the phosphate group) shows remarkable conservation across the entire superfamily [20, 21]. The extreme sequence divergence occurred in kinases over the course of evolution have made them able to phosphorylate a wide variety of targets, to interact with a large range of proteins, and to respond to a myriad of different regulatory mechanisms and cellular signals. This remarkable sequence divergence influence the functional aspects of kinases.

Regarding the *in vivo* substrate specificity, significant differences have been reported between members of the same kinase family, where the sequence identity of the catalytic domain is, on average, close to 60% and the overall domain composition tend to be similar. For example, two recent independent studies showed that, despite their high sequence

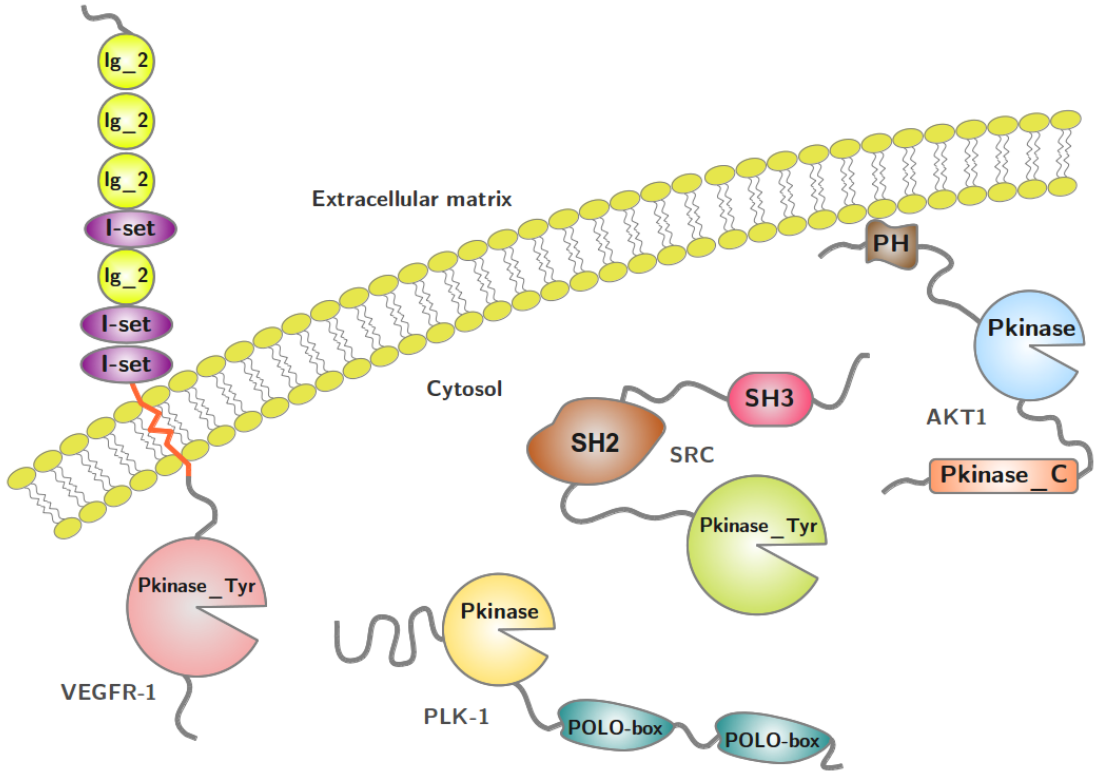


Figure 2.3: Modular domain composition of protein kinases.

Protein kinases have achieved a broad functional diversification thanks to a versatile combination of protein domains. From left to right, Vascular endothelial growth factor receptor 1 (VEGFR-1), a tyrosine-protein kinase that acts as a cell-surface receptor and plays an essential role in angiogenesis, cell survival, cell migration, chemotaxis, and cancer cell invasion [22]; Polo-like kinase 1 (PLK-1), a serine/threonine-protein kinase that performs several key functions throughout M phase of the cell cycle, including the regulation of centrosome maturation, spindle assembly, regulation of mitotic exit and cytokinesis [23]; Proto-oncogene tyrosine-protein kinase Src (SRC), participates in signaling pathways that control a diverse spectrum of biological activities including gene transcription, immune response, cell adhesion, cell cycle progression, apoptosis and migration [24, 25]; RAC-alpha serine/threonine-protein kinase (AKT1), regulate many processes including metabolism, proliferation, cell survival, growth and angiogenesis [26].

identity [a], Aurora kinases A and B (two major mitotic kinases) share few *in vivo* substrates. Both kinases are known to phosphorylate almost identical sequences, however, they do not co-localize during mitosis and therefore they do not phosphorylate the same substrates. The authors showed that, in this case, kinases that share 'motif space' do not share 'localization space' [27, 28]. Similar results were also described for the yeast family of cAMP-dependent protein kinases, homologs of the eukaryotic PKA kinases. Despite being very closely related in sequence [b], the three members of this family (TPK1, TPK2 and TPK3) showed distinct substrate specificities and a very scarce number of common substrates [29]. Together, these

[a] Sequence identities between Aurora kinases A and B. Complete sequence 51%, at the catalytic domain 74.5%.

[b] Sequence identities from multiple sequence alignments. Complete proteins 57.4%; catalytic domains 75.7%.

findings demonstrate that *in vivo* substrate specificity can have a limited relationship with the level of sequence conservation of the kinase catalytic domain, and that the substrate recognition process can be fine-tuned by several other factors.

2.3 Known mechanisms of protein kinases for substrate identification

During the late 80's and early 90's several studies focused on the identification of substrate specificity determinants, and most of those studies were based solely on the analysis of consensus sequences derived from peptides phosphorylated *in vitro* [30–33]. Such analysis provided important insights regarding peptide specificity of kinases and also allowed a general classification of kinases attending the physico-chemical properties of the sequences they tend to phosphorylate. From those, and other more recent studies [29, 34–38] it became evident that even kinases closely related in sequence can have different (although some times overlapping) substrate specificity, but also that consensus sequences were not able to fully explain the substrate specificity observed *in vivo*.

2.3.1 Role of the kinase catalytic site

As mentioned before, the tridimensional (3D) structure of the catalytic domain of protein kinases have been consistently conserved along evolution. However, the catalytic domains can differ in the charge and hydrophobicity of residues at the catalytic and substrate binding sites, as well as in the length of the activation loop; features that are known to be of most importance for substrate specificity and enzymatic regulation [39–41]. Nevertheless, the existence of consensus phosphorylation sequences for many kinases supports the idea of customization to certain type of target sequences. In fact, it is known that there exists a degree of physico-chemical complementarity between residues in the catalytic site of kinases and residues in the vicinity of the phosphorylation site [34, 40]. These complementarities foster substrate recognition on the basis of charge, hydrogen bonding and hydrophobic interactions. Indeed, many kinases can be sub-divided in three classes: i) basophilic, which favor basic residues around the phosphorylation site, ii) acidophilic which favor acidic residues, and iii) proline-directed which require a proline residue immediately N-terminal to the phosphorylation site. Despite the existence of clear consensus sequences for many kinases, these patterns are still too lax to fully explain the complex process of substrate specificity.

2.3.2 Distal docking sites

The next level of substrate specificity is provided by the direct associations of kinases to their substrates via docking sites interactions. These interactions contribute to the recruitment (and to the increased affinity) of substrates by kinases and therefore, to an augmented efficiency of the phosphorylation [42, 43]. Docking sites are often spatially separated from the kinase catalytic cleft and from the phosphorylated residue in the substrate [44]. Moreover, differences in the composition and/or spacing of residues in these docking sites (at either

the kinase and/or the substrate) can modulate the overall selectivity of the interaction [45]. Distal docking sites have been reported for both serine/threonine and tyrosine kinases. In serine/threonine kinases the sites are often part of the catalytic domain, while in tyrosine kinases the docking sites are commonly found in additional domains, away form the catalytic one [46]. Docking-site mediated interactions have been reported for several kinases such as CDK2, ERK, GSK3, JNK, MEK, PDK1 and PHK [46]; a fact that suggests these interactions as a general mechanism for enhancing substrate specificity.

For example, members of the MAPK family (ERK, JNK and p38) are cases of kinases with broad phosphorylation site specificity. They can phosphorylate almost any substrate with the sequence pattern Ser/Thr-Pro, a sequence present in about 90% of all proteins [47]. Therefore, the phosphorylation site specificity of these kinases is not sufficient to explain their substrate selection in the cell. It has been found that MAPKs have a conserved docking site outside the catalytic domain. This docking site, known as CD-domain, is formed by a cluster of negatively charged residues [44]. Also, several activators (MAPKKs), substrates (MAPKAPK) and inactivators (MKPs, phosphatases) of MAPKs contain a complementary docking site, D-domain, formed by positively charged residues [44]. It has been observed that specificity of associations between MAPKs and their partners is achieved, at least partially, by the electrostatic interactions between the CD-domain and the D-domain (see Figure 2.4) [42, 44, 48].

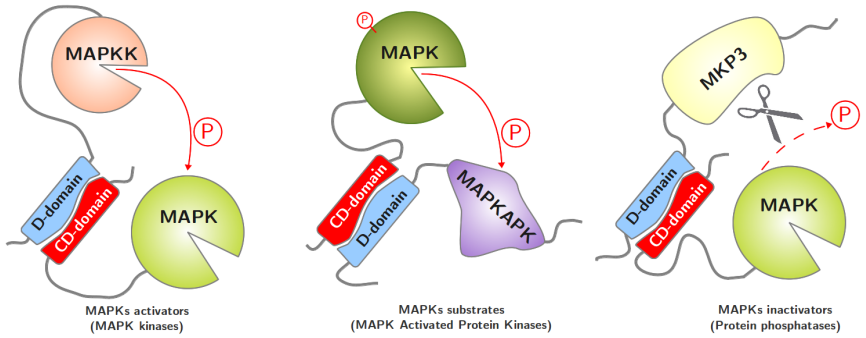


Figure 2.4: Distal docking sites in MAPKs.

MAP kinases have a conserved docking site C-terminal to the catalytic domain (CD-domain, in red). Several activators, substrates and inactivators of MAPKs have a docking site, D-domain (in blue), formed by positively charged residues. The electrostatic interactions between the CD-domain in MAPKs and the D-domain in their partners contribute greatly to MAPKs binding specificity. Activated MAPK is represented in dark green with a phosphate group attached.

2.3.3 Regulatory and targeting domains

As commented previously, non-catalytic domains of protein kinases are central to the biological function of these enzymes. A vast majority of the non-catalytic domains are versatile protein-protein interaction (PPI) elements that can mediate intra and inter molecular associations with other signaling modules. These built-in domains can allosterically modulate

the catalytic activity of the kinase, its association to other proteins and/or its cellular localization. By affecting one or more of the aforementioned properties of kinases, non-catalytic domains can largely influence the substrate specificity of these enzymes.

Among serine/threonine kinases one example of the action of such regulatory elements is the POLO-Box domain (PBD). This domain is exclusively found in the family of Polo-like kinases, composed of five members in human, which are key regulators of several mitotic processes [49]. The PBD performs dual roles in determining subcellular localization and inhibitory regulation of the kinase by intramolecular interactions [50]. By exerting this dual function, PBDs are able to convey a significant deal of target specificity to each member of the Polo-like kinases [50, 51].

Other two canonical examples are the cases of protein kinase A (PKA) and cyclin-dependent kinases (CDKs), although their regulatory elements are not covalently tethered to catalytic domains. PKA is a holoenzyme with several roles in the cell [52–54], which remains inactive while bound to a regulatory subunit (a homodimer). Besides providing allosteric regulation, the regulatory homodimer targets the two bound catalytic domains to macromolecular complexes in the plasma membrane that are involved in cellular signaling [55]. Upon intracellular increased levels of cAMP, the kinase domains are released in an catalytically active state [56, 57]. In the case of the CDKs, they constitute a family present in all known eukaryotes with conserved key roles in the regulation of the cell cycle [58]. Cyclins are the proteins that allosterically regulate CDKs by forming very stable complexes, and without these associations CDKs have very little kinase activity. In addition, cyclins have a docking domain that helps to recruit the CDKs to their correct substrates [59–62] and also to different subcellular compartments [63–65]. In this manner, cyclins are thought to greatly enhance CDKs specificity.

Another case of built-in regulatory domains is found in the family of non-receptor Src tyrosine kinases. Members of this family contain an Src homology-2 (SH2) domain and an Src homology-3 (SH3) domain at the N-terminal position with respect to their kinase catalytic domains. SH2 and SH3 domains are able to promote PPIs by binding — respectively — to phosphorylated tyrosine residues and proline-rich peptides in present in partner proteins. The enzymatic activity of Src kinases is inhibited by intramolecular interactions of the catalytic domain with both SH2 and SH3 domains [66]. It is known that many of the best substrates of Src kinases contain ligands for the SH3 and/or SH2 domains. Such substrates can therefore disrupt the intramolecular interactions in the kinase and activate its catalytic state [66]. These observations suggest that substrate specificity of Src kinases may be dependent on the specificity of its associated SH2 and SH3 domains. Moreover, such mechanism ensures that these kinases are active only upon direct interaction with their proper substrates.

3 Objectives

The global aim of this thesis is the quantification of the contribution of different elements to the observed *in vivo* substrate specificity of human protein kinases. In order to achieve our purposes, we defined the following objectives:

- **Quantification of the contribution of the phosphorylation site and its surrounding residues to the substrate specificity of human kinases.**
- **Quantification of the contribution of adaptor and scaffold proteins to the substrate specificity of human protein kinases.**

4 Sequence logos and position-specific scoring matrices

4.1 Introduction

4.1.1 Contribution of residues neighboring the phosphorylation sites

4.1.1.1 Residues commonly phosphorylated in eukaryotes

Phosphorylation is one of the most common PTM in proteins [1] and serine (Ser), threonine (Thr) and tyrosine (Tyr) are the residues most commonly phosphorylated in eukaryotic organism [42]. These amino acids count with an hydroxyl group in their side chains that is substituted by the terminal phosphate of an ATP molecule in a reaction catalyzed by a protein kinase. In order for the reaction to take place, the kinase must recognize the phospho-acceptor residue in the sequence of the substrate protein (*i.e.*, the phosphorylation site), and also the side chain of the phospho-acceptor residue must be correctly oriented towards the catalytic residues of the enzyme [40]. However, the phospho-acceptor residue is not the only element contributing to the recognition of the phosphorylation site by the kinase; other residues in its close sequence vicinity — generally expanding three to six residues at each side of the phospho-acceptor residue — have been found to play fundamental roles in the interaction with the the substrate binding region in the kinase [27, 33, 67]. The interactions between residues flanking the phospho-acceptor amino acid and residues in the catalytic cleft of the kinase stabilize the kinase-substrate complex and forces the correct orientation of the phospho-acceptor residue into the catalytic site [68]. Generally, the sequence region in the substrate containing the phosphorylation site adopts an extended conformation upon binding to the kinase binding region [67].

4.1.1.2 Positioning and orientation of the phospho-acceptor residue

As stated previously in this section, the phospho-acceptor residue is not the unique responsible for the specificity of the kinase; as there also exist a contribution from residues in its close vicinity. In this sense, the phosphorylation site and its close neighboring residues constitute what is known as a short linear motif (SLiM), an important functional element known to mediate several PPIs [69–71]. Experiments using peptide arrays as well as proteomic studies, have shown that many kinases can be very unspecific in *in vitro* conditions, where they are able to phosphorylate a wide range of different sequences [34, 72]. However, it has also being reported that *in vivo*, kinases can display a high specificity towards particular sequences and substrates [27, 28]. In this sense, it is known that some families of kinases show preference for stretches of sequences in their substrates that are enriched in particular

amino acids. Two examples of this are the MAP kinases and the phosphatidyl inositol 3' kinase-related kinases (PIKK), which for optimal phosphorylation require — respectively —, the presence of a proline (Pro) and a glutamine (Gln) at the first position after the phospho-acceptor residue [73, 74]. There are other available examples of kinases and families of kinases displaying a preference for certain amino acids at particular positions relative to the phospho-acceptor residue [33, 34, 75]. However, due to the lack of experimental data or to their apparently broad specificity patterns, there is still a very large number of human kinases for which the preferences for residues surrounding the phosphorylation site are unknown or ambiguously defined.

4.1.1.3 From phosphorylation sequences to phosphorylation motifs

The sequence preferences displayed by certain kinases, have allowed the identification of consensus sequences (or sequence motifs) that are more commonly phosphorylated by that kinases. A sequence motif is the representation of the relative frequencies of amino acids at given positions in a set of protein sequences that have been previously aligned. The amino acids patterns observed in sequence motifs have, or are hypothesized to have, biological significance. For example, when generating a motif from the sequences known to be phosphorylated by a given kinase (a phosphorylation motif), an alignment of the corresponding phosphorylated sequences is generated keeping the phospho-acceptor residue in the central position of the alignment. Typically, phosphorylation motifs provide information about the residues — in addition to the phospho-acceptor amino acid — that are required at specific positions for the kinase to recognize the phosphorylation site. However, the phosphorylation motifs can also be degenerated at some positions, that is, there may exist levels of uncertainty about what residues are preferred at certain position of the sequence alignment. It is generally accepted that phosphorylation motifs provide the primary specificity [46, 67, 75, 76] while a variety of contextual factors, including co-localization, co-expression and physical interaction of the kinases with their targets, contribute additionally to the *in vivo* substrate specificity [43, 70, 77].

4.1.2 Graphical representation of phosphorylation motifs

Sequence motifs have been widely used by the scientific community, and they are a valuable tool for representing and analyzing patterns in biological sequences [69, 78]. Sequence motifs are usually represented in two ways. One common representation is in the form of a consensus sequence, which is a string of characters where the more frequent and/or infrequent elements of the sequence alignment are shown. The other typically used representation is in the form of a sequence logo [79], where each position in the sequence alignment correspond to a column in the logo. The total height of each column indicates the sequence conservation at that position of the alignment (measured in bits), while the height of each independent symbol indicates the relative frequency of the amino acid at that position. In a sequence logo representation, the height of the Y axis is equal to the Shannon information content [80] calculated for each of the 20 genetically encoded amino acids. When a residue is fully conserved at a given position, the height of the column is $\log_2 20 = 4.32$ bits. On the

contrary, when all residues are equally probable at a position, the height of the column is 0 bits.

In general, compared to the representation by consensus sequences, sequence logos provide a richer and more precise description of sequence motifs; and they can rapidly reveal significant features of the alignment otherwise difficult to perceive [81]. However, they do not provide a mathematical representation of the biological pattern encoded in the sequences aligned. In this sense, position-specific scoring matrices are useful tools to quantitatively encode the information contained in the sequence alignment.

4.1.3 Mathematical representation of phosphorylation motifs

4.1.3.1 Probabilistic models

A position-specific scoring matrix (PSSM) is a valuable tool for the probabilistic representation of signals in a sequence alignment, which allow the scoring of individual sequences based on their binding strength. PSSMs have been extensively used to model approximate patterns in DNA or protein sequences [72, 82, 83]. A PSSM is a tabular numerical representation of sequence motifs displaying their variability as log-likelihood values for each possible residue or nucleotide at each position in a sequence. A PSSM has one row per each symbol of the alphabet (amino acids in our case) and one column per each position in the sequence. In a PSSM the value in each cell of the matrix is the score for a given residue at a given position in the sequence.

There are two methods commonly used to computing the score for each residue at a given position in a PSSM. The first one, based on log-likelihoods, defines the score of a residue at a given position as the natural logarithm of the frequency of that residue in that position in the set of training sequences (*i.e.*, the sequence alignment). The second method, based on log-odds scores, also uses the observed frequency of residues at a given position in the set of training sequences, and includes an additional term that accounts for the observed frequency of that residue in a background model (*e.g.*, a reference proteome) [84]. By using either of the mentioned methods, the overall score of a sequence aligned on a PSSM is defined by the summation of the score of each residue in the sequence at its given position. In the case of phosphorylation motifs, a positive score indicates that a residue is considered to favor the recognition of the sequence as a target for phosphorylation. As opposed, negative scores are considered unfavorable for sequence recognition.

Once a score have been computed for a given sequence on a PSSM, there still remains the issue of assessing the statistical significance of that score. That is, in order to evaluate if the sequence is a statistically significant match to the PSSM, we need to estimate the probability of the background model to achieve a score equal or higher than the one observed. This can be achieved by defining *a priori* the desired level of statistical significance (a p-value) for a match to the PSSM, which will be associated to a threshold value of the score. The efficient identification of matches to PSSMs is a complex problem that has recently attracted the interest of the scientific community [85, 86]. In the Materials and Methods section we will describe in more detail the method we have used for defining the score threshold for our PSSMs.

Another important topic is the assessment of the statistical significance of the PSSM itself. In this regard, it is common to use the information content (IC) of the PSSM as the statistic for the significance test that estimates how different is a PSSM from a background distribution [87]. The IC of a PSSM can be computed assuming equal probabilities for each residue, or it can be computed taking into account the probability of occurrence of each residue in a background model. In this sense, taking into account the frequencies of amino acids in the background model can provide a more accurate weighting of their individual contributions to the overall value of the IC.

PSSMs constitute a valuable tool that can be applied in large-scale analysis of DNA, RNA and protein sequences. Although, the method has limitations such as the assumption of independence between the columns of the matrix, *i.e.*, that each position contributes independently and additively to the functional activity of the sequence [87]. This approximation has generated a lot of controversy about the validity of the model and alternatives to it have been proposed, as for example by using two or more residues as the units of sequence [88,89]. Also, more complex models for motifs analysis have been developed, such as hidden Markov models and words graphs [90, 91]. Regardless of the known limitations, PSSMs are still at the core of several services and methods for identification of sequence motifs in DNA and proteins [92–95].

4.1.3.2 Prediction of phosphorylation sites

During the last decade, several high-throughput experiments have made available thousands of *in vivo* and *in vitro* phosphorylation sites for human [96–100]. At the same time, and using the aforementioned data, several projects have aimed the prediction of *in vivo* phosphorylation sites by developing methods that are able to 'learn' from the sequences of experimentally determined sites. Regarding the use of PSSMs there are two main examples. The first one is the web service `ScanSite`, where the authors use kinase-specific PSSMs — and also allow for user-defined PSSMs — in order to identify phosphorylation sites in protein databases [101]. The other case is `Predikin`, a web server where the authors combine PSSMs with structural information of the phosphorylated sequences in order to predict and filter potential phosphorylation sites [102]. There are several other applications that implement more complex methods such as hidden Markov models (HMM), artificial neural networks (ANN) or expert systems to integrate several sources of information (*e.g.*, structural disorder, sequence conservation, positional correlations of residues) in the prediction algorithm [103–105]. In general, more sophisticated methods of phosphorylation sites identification (*e.g.*, ANN, HMM) perform better in the classification of highly complex and nonlinear sequence patterns [106]. However, in these cases it is more difficult to infer the decisions that support the predictions, as opposed to the cases of PSSMs, where it is much easier to pinpoint the determinants residues of a functional phosphorylation site [103].

In order to evaluate the performance of the prediction methods the sensitivity and specificity must be taken into account. Ideally, the method should be able to identify as many true sites as possible (sensitivity), while ensuring that only true sites are predicted as such (specificity). Unfortunately, current methods suffer of low performance on these parameters [107] and they all have varying levels of success depending on kinase or type of motif targeted

for PTM prediction. This situation is caused mainly by the scarcity of experimental data to train the methods, and also due to the sequence degeneracy of residues surrounding the phosphorylation site; a characteristic that make them likely to appear at random in protein sequences [69, 108]. To overcome this situation, many methods filter the initial predictions by using contextual information such as the structural characteristics, sequence conservation and solvent accessibility of the predicted region, and also by contrasting evidence of shared sub-cellular co-localization and/or PPI between the kinase and the predicted substrate.

4.2 Materials and methods

4.2.1 Phylogenetic classification of human protein kinases

The sequences and classification of the 518 human protein kinases were downloaded from http://www.kinase.com/human/kinome. This is the official web resource containing the supplementary data for the report of the full complement of human protein kinases by Manning *et al.* [5].

4.2.2 Phosphorylation data for human protein kinases

Experimentally determined phosphorylation sites were retrieved from public databases and integrated into a local database (SBNB_PhosphoDB). As the sources for collecting the PTMs we used HPRD [109] (release 9, 13/04/2010), PhosphoSitePlus [110] (as of 03/11/2010) and Phospho-ELM [111] (version 9, 09/2010). We kept only those phosphorylation sites that could be mapped to the sequence of the corresponding substrate and for which the responsible kinase was known. We kept track of the experimental conditions on which the PTM was detected (*in vivo* and/or *in vitro*) as well as the corresponding publication. PTMs with no available supporting publication were filtered out. Table 4.1 shows a summary of the data collected. Our integrated data increases by 18%, 58% and 59% the numbers of kinases, substrates and phosphorylation sites (respectively), with respect to the average contained in the source databases.

Table 4.1: Phosphorylation data for human protein kinases

Database	Kinases	Substrates	P.Sites	P.Events
HPRD	291	938	3382	5896
Phospho-ELM	218	924	3125	2378
PhosphoSitePlus	318	1664	4711	4711
SBNB_PhosphoDB	325 (290)	1856	5946 (3583)	8880 (5171)

P.Sites: refers to the total (non-redundant) number of distinct residues phosphorylated in distinct substrates. **P.Events**: refers to the total number of phosphorylation events. Two distinct kinases may phosphorylate the same residue in the same substrate, these constitute two different phosphorylation events. Within parenthesis, the quantities corresponding to the subset of the data determined in *in vivo* conditions.

4.2.3 Set of 'unphosphorylated' human proteins

In order to assess the performance of the PSSMs generated, it is important to define a set of 'negative' phosphorylation sites. For compiling such set we first selected from the UniProt database [112] (ver. 09/2010) the sequences of human proteins that were not annotated to be phosphorylated. Second, using the program `CD-HIT` [113] we eliminated redundancy – to a 100% of sequence identity – in the resulting set of sequences. Finally, we discarded

all proteins containing an instance of any of the phosphorylated sequences [a] present in our database (*i.e.*, SBNB_PhosphoDB). This procedure produced a final set of 8876 human proteins that constitute our negative test set named 'unphosphorylated' human proteome.

4.2.4 Generation of sequence logos

For generating sequence logos we used the standalone version of the program `WebLogo` [81]. In order to be able to generate the logos, the program needs a sequence alignment as input. For this, we produced a sequence alignment for each kinase and kinase family in SBNB_PhosphoDB using their corresponding phosphorylation sequences. In total, we have generated sequence logos for 325 independent kinases and 93 kinase families.

4.2.5 Position-specific scoring matrices

For generating the position-specific scoring matrices (PSSMs) and assessing their statistical significance and performance, we have developed the in-house software `genpssm`. The program uses as input i) an alignment of phosphorylated sequences, ii) the frequencies of amino acids in human proteins, iii) a *p*-value threshold for assessing the statistical significance of matches to the PSSM and iv) a negative set of phosphorylated sequences (the 'unphosphorylated' human proteome).

4.2.5.1 Generating PSSMs

The algorithm of `genpssm` starts by constructing the PSSM from the sequence alignment. For computing the scores for each residue at each position of the matrix the program uses the Equation 4.1, defined by Claverie and Audic [84]. This equation is based in the log-odds of residues at each position of the alignment and takes into account the frequency of each residue in a background model, which in our case is the human proteome.

$$S_{ip} = \log(\frac{q_{ip}}{f_i}),\ p = 1\ to\ w \tag{4.1}$$

Equation 4.1: Scores representing the propensities of residues. S_{ip} is the score of residue i at position p of the sequence alignment, q is the frequency of residue i at position p, f is the frequency of the residue in the background model and w is the length of the sequence alignment.

4.2.5.2 Selection of a score threshold

Once we have constructed a PSSM, we need to define a method to identify sequences that match to it in a statistically significant manner. For doing this, we need to set the score threshold that will be used to classify matches to the PSSM as statistically significant or not. However, for choosing the value of the score threshold, we first have to define the level of statistical significance of that score, which is the probability that the background model can

[a] We defined as phosphorylated sequence the stretch of residues comprising 4 amino acids in both N and C terminal directions of the phospho-acceptor residue.

achieve a score larger than or equal to the one observed. For computing the aforementioned *p*-value we used a probability ($\alpha \leq 0.05$) of finding a match to any given sequence of nine residues in a protein of length equal to the human average. For accomplishing this we followed the transformations shown in Equation 4.2.

$$
\begin{aligned}
p = 1 - (1 - p - value\ )^{(avgPlen - lenPSite + 1)} \\
p - value\ = 1 - \sqrt[(avgPlen - lenPSite + 1)]{1 - p} \\
p - value\ = 1 - \sqrt[592]{0.95} \\
p - value\ = 8.66e^{-05} \\
p - value\ \approx 1e^{-04}
\end{aligned}
\tag{4.2}
$$

Equation 4.2: Computing the level of statistical significance (*p*-value) for the score threshold. avgPlen: average length of a human protein ($avgPlen = 600$), lenPSite: length of the phosphorylated sequences from which the PSSMs are generated ($lenPSite = 9$).

Once we had defined the general level of statistical significance desired for matches to a PSSM, we determined the corresponding score threshold for each PSSM, using a method by Touzet and Varré based on discretized scores distributions [83].

4.2.5.3 Statistical significance of the PSSMs

Once a PSSM has been generated, we need to evaluate its statistical significance. For this we used the value of the IC and the percent recall achieved by the PSSM. By percent recall we mean the fraction of the seed phosphorylation sequences (*i.e.*, the ones in the input sequence alignment) that a PSSM is able to match with a statistically significant score. For assessing the statistical evaluations of a PSSM, the algorithm in `genpssm` generates a background set of 100'000 PSSMs using random sequences that follow the frequencies of the amino acids in the human proteome. For each of the background PSSMs, the cardinality of seed sequences is kept equal to the one of the PSSM being assessed. For each PSSM of the background, `genpssm` computes the IC and the percent recall, and later uses the background distributions of these statistics to estimate the statistical significance of the corresponding values observed in the PSSM being evaluated. For computing the IC, `genpssm` uses the Kullback–Leibler distance [78, 114], where the IC is the sum of the expected self-information of each element (see Equation 4.3).

$$
IC = -\sum_{i,p} q_{ip} \times \log(\frac{q_{ip}}{f_i}) \tag{4.3}
$$

Equation 4.3 Computing IC of a PSSM. Where q is the frequency of residue i at position p of the sequence alignment and f is the frequency of the residue in the background model.

4.2.5.4 Performance of the PSSMs

The receiver operating characteristic (ROC) curves have been extensively used in biomedicine to illustrate the performance of a classification and prediction model for decision support [115]. A ROC curve is a plot that captures the trade-off of the true positive rate (TPR, or recall) versus the false positive rate (FPR) at various threshold settings (see Equation 4.4). The ROC curve area can vary between 0.5 and 1.0, where an area of 1.0 represents a perfect accuracy, while an area of 0.5 represents an accuracy no better than what would be expected by chance.

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN} \tag{4.4}$$

Equation 4.4. The accuracy of a binary classifier is commonly assessed using TPR and FPR. TP, TN, FP and FN are the counts of true positive, true negative, false positives and false negatives (respectively) when the test is applied to a population.

In our case, we want to evaluate how well the PSSMs are able to distinguish between their corresponding seed sequences (true positives) and sequences in the set of 'unphosphorylated' human proteins (false positives). For generating the ROC curve of each PSSM, we split the whole range of scores of the PSSM into 100 thresholds and for each threshold we computed the TPRs and FPRs. Once we have obtained the TPR and FPR, we computed the area under the ROC curve (AUC-ROC) using the program `CROC` [116].

4.2.6 Quantification of the kinase specificity encoded in the PSSMs

4.2.6.1 Sequence motifs most commonly recognized by kinases

We have attempted the identification of residues in the phosphorylated sequences that are likely to contribute significantly to the substrate recognition by the kinase. For this, we have used the score values in the PSSMs of kinase families. Here, our aim have been to identify those residues that achieve a score equal or higher than half the score of the phospho-acceptor residue (see Equation 4.5). Residues complying with the aforementioned criteria are conserved among the sequences phosphorylated by a given kinase or kinase family, and based on this we classify them as specificity-determinant residues (SDR). The term SDR has been previously used by Kobe *et al.*, although in their work it was applied to substrate-binding residues in the kinase catalytic site [67]. In order to explore more in detail the phosphorylation motifs recognized by kinases, we have analyzed the frequencies of different SDRs among the phosphorylated sequences corresponding to kinase families present in our data. For this, we identified a subset of 22 kinase families for which we have at least 100 phosphorylation events. This strategy allowed us to represent and quantify the relevance of different SDRs for the kinase families analyzed.

$$Score_{SDR} \geq \frac{Score_{PA}}{2} \tag{4.5}$$

Equation 4.5. Identification of SDRs based on their relative scores. PA refers to the phospho-acceptor residue.

4.3 Results and discussion

4.3.1 Motifs recognized by kinases and kinases families

As commented before, sequence logos are valuable tools that have been extensively used to identify patterns in phosphorylated sequences. Therefore, as a previous step to our analysis, we wanted to check that the logos we have generated are in accordance with the ones that have been previously reported in the literature. For this, we have compared by visual inspection our logos to the ones reported by Miller *et al.* [70]. For the cases compared (see Table 4.2 for examples), the patterns in our logos are consistent with the ones previously reported in literature [34,70]. Cases of well known motifs that are characteristic of particular kinases (*e.g.*, CDK and ATM kinases) are distinguishable from our logos. Based on these comparisons, we conclude that the phosphorylation data collected correctly represent the current knowledge about kinase phosphorylation motifs.

4.3.2 Strong specificity-determinant residues from kinase families

We have attempted the quantification of the contribution of residues in the close vicinity of the phospho-acceptor amino acid, to the recognition of the phosphorylation site by the kinases. For this, we analyzed the frequency of the SDRs in the phosphorylation events of 22 kinase families as previously described in Materials and Methods (see section 4.2.6.1). For 19/22 of the families analyzed we have identified at least one SDR in addition to the phospho-acceptor amino acid. For presenting our results, we have classified the kinase families based on the type of phosphorylation sites they recognize. These classes are: proline-directed, glutamine-directed, basophilic and acidophilic.

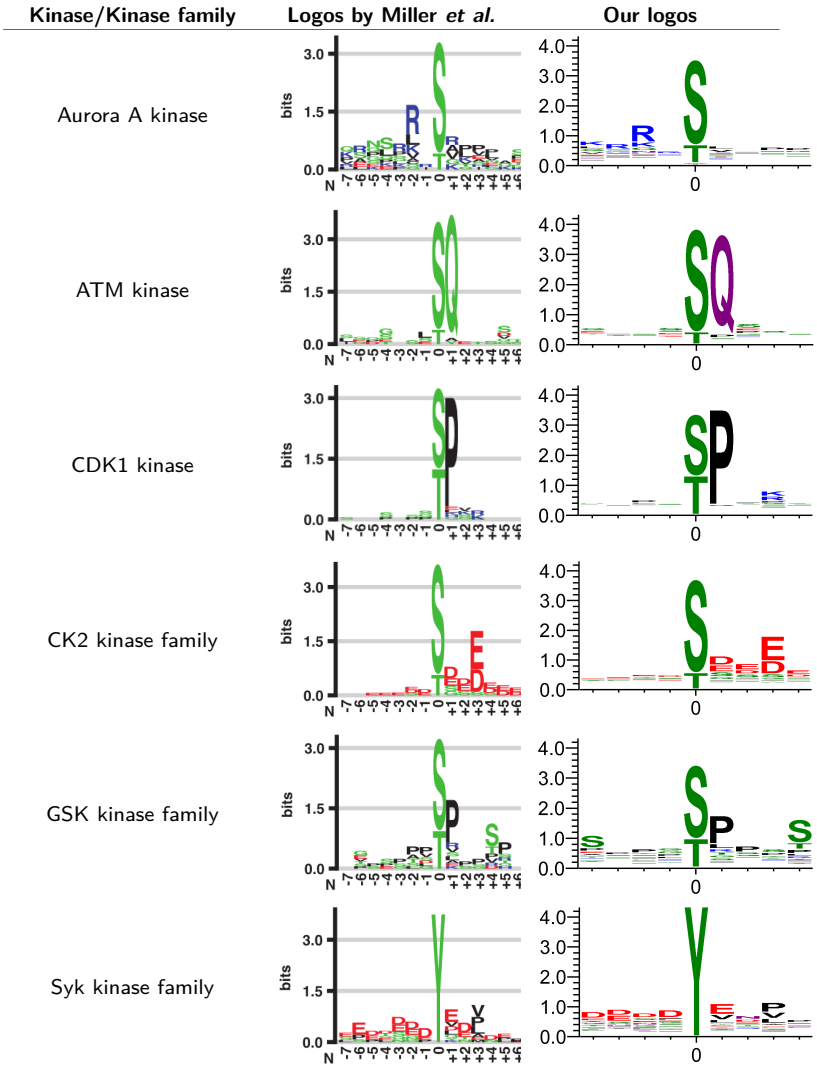
4.3.2.1 Proline-directed kinase families

The class of proline-directed (Pro-directed) is composed by Ser/Thr kinases of the group CMGC (CDK/MAPK/GSK3/CLK). They are involved in a plethora of critical signaling events in the cell [117]. For optimal recognition of their target sequences, Pro-directed kinases require the presence of a proline residue right after the phospho-acceptor amino acid.

By using our method, we have identified six SDRs for three kinase families in the CMGC group. In some cases, the SDRs identified correspond to residues that are known to play important roles in the recognition of the phosphorylation site by these kinases. For quantifying the contribution of the SDRs identified, we have computed their frequencies of occurrence among the phosphorylation events of the Pro-directed kinase families. For comparison purposes, we have also computed the frequencies of occurrence of the SDRs in the phosphorylation events of other kinase families in our data set. The Pro-directed families analyzed here are the ones of the cyclin-dependent kinases (CDKs), mitogen-activated kinases (MAPKs) and glycogen synthase kinases (GSKs). Table 4.3 shows the SDRs identified for them.

In the Table 4.3 can be noted the prevalence of the P+1 residue among sequences phosphorylated by kinases of the Pro-directed class (74.85% on average). A proline residue at position +1 occurs in much less extent (5.95% on average) among the sequences phosphorylated by kinases that do not belong to the current class. Together with P+1, P-2 has

Table 4.2: Sequence logos from phosphorylated sequences



Sequence logos are shown for three independent kinases and three kinase families. The first three examples show cases of basophilic, glutamine and proline -directed kinases (in that order). The fourth and sixth examples show cases of acidophilic kinase families.

also been identified as an SDR for the same families. Although P-2 frequencies among Pro-directed kinases (25.12% on average) are low if compared to P+1, however its frequency among the sequences phosphorylated by other kinase families is still very low (5.83% on average). These two SDRs are well known to have an important contribution to the substrate specificity of most Pro-directed kinases [73, 118].

In addition to P-2 and P+1, for the CDK family we have also identified K+3 as an SDR,

Table 4.3: Specificity-determinant residues from Pro-directed kinase families.

			SDR					
Kinase group	**Kinase family**	**Pevents**	**S-4**	**P-2**	**P+1**	**P+2**	**K+3**	**S+4**
CMGC	CDK	1313		22.77	81.72		21.25	
CMGC	GSK	278	38.49	21.22	53.96	27.70		48.56
CMGC	MAPK	1068		31.37	88.86			
Avg. SDR class			38.49	25.12	74.85	27.70	21.25	48.56
Avg. SDR global			32.39	25.12	65.14	27.70	21.25	38.74
Avg. non-SDR global			14.01	5.83	5.95	8.35	4.16	11.51

SDR: specificity-determinant residues identified. **Pevents**: number of phosphorylation events known for the kinase family. **Avg. SDR class**: Average frequency among the phosphorylation events of current class. **Avg. SDR global**: Average frequency among the phosphorylation events of all kinase families with the SDR (not only the kinases in the current class). **Avg. non-SDR global**: Average frequency among the phosphorylation events of kinase families without the SDR. With exception of the Pevents column, the values in the table represent the percentages of phosphorylation events.

which has been previously reported to be characteristic of this family [70]. This SDR has a frequency comparable to the one of P-2 in the same family, and also has a low occurrence (4.16% on average) in phosphorylation events of other kinase families.

In the case of the GSK family we have identified three additional SDRs (S-4, P+2 and S+4). There is strong evidence in the literature about the role of S+4 as a site for priming phosphorylation, needed by members of this family previous to the phosphorylation of the targeted Ser/Thr residue [119]. It can be noted from Table 4.3 that the frequencies of occurrence of S-4, P+2 and S+4 in phosphorylation events of the kinases that do not have them as SDRs are, on average, relatively low (14.01%, 8.35% and 11.51% respectively, last row Table 4.3). Nevertheless, neither S-4 nor S+4 appear to be SDRs exclusive of the GSK family. Our results suggest that other kinase families also use them as SDRs, although to a lesser extent (see first to last row of Table 4.3). We have not found evidence in the consulted literature that suggest S-4 or P+2 as major players in the substrate specificity of GSK family.

For the MAPK family only P-2 and P+1 where identified as SDRs. However, is interesting to note that for this family both P-2 and P+1 show the highest frequencies of occurrence among the Pro-directed families (31.37% and 88.86% respectively).

For the families CDK and GSK, we provide in Figure 4.1 the corresponding frequencies of occurrence of their SDRs among the phosphorylation events of the 22 kinase families included in the current analysis. It can be noted from Figure 4.1 that the kinase family STE7 (group STE) also shows a relatively strong preference for phosphorylation sites having a proline residue at position +1. Since kinases of the STE7 family are not strict Pro-directed kinases, the SDRs of this family will be discussed further in this document. The bar plots of the SDRs for the families CDK, GSK and MAPK, as well as the sequence logos for each of these families, are provided in the Appendices A1 and A2.

4.3.2.2 Glutamine-directed kinase families

Glutamine-directed (Gln-directed) kinases are represented by the family PIKK, from the group of Atypical kinases. The PIKK family is composed of six Ser/Thr kinases, some of which are

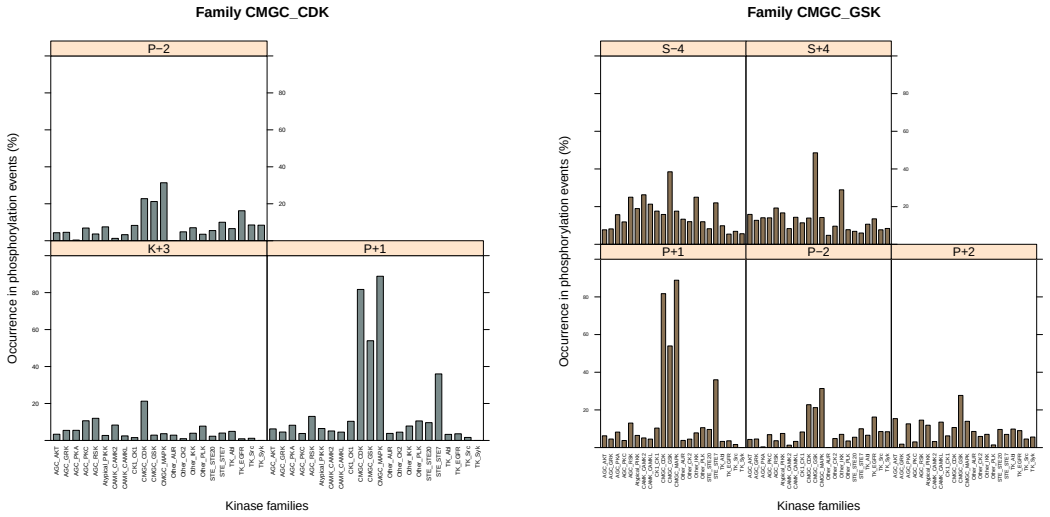


Figure 4.1: Frequencies of the SDRs from two families of Pro-directed kinases.

Frequencies of occurrence of the SDRs of CDK and GSK kinase families among the phosphorylation events of other 22 kinase families. Boxes within each panel represent the SDRs. SDR are represented by the one letter code of the amino acid and its position relative to the phospho-acceptor residue. On the x-axis, the kinase families, on the y-axis the percentage of occurrence of each SDR.

involved in the co-ordination of the cellular response to DNA damage [120–122]. In order to phosphorylate their substrates, most kinases of the PIKK family require a glutamine (Q) residue in the +1 position with respect to the phosphorylation site [74,123]. As we did before for the class of Pro-directed kinases, we have explored the presence of SDRs in the PIKK family and we compared the results to the other 21 kinase families included in the analysis.

For the class of Gln-directed kinases only the already known Q+1 was identified by our method as an SDR. Q+1 is present in a large fraction of the phosphorylation events of PIKK kinases (388/480, 80.83%), and in contrast, its average frequency among the phosphorylation events of the other 21 kinase families analyzed here is rather low (3.98%, last row of Table 4.4). Is also interesting to note that PIKK is the only family, at least of the 22 analyzed here, that uses Q+1 as an SDR (see first to last column of Table 4.4).

From our results, it seems that members of the PIKK appear to rely mostly on the phospho-acceptor and on the Q+1 residues to recognize the phosphorylation site on their substrates. We have not found evidence in the reviewed literature supporting other residues in the close vicinity of the phospho-acceptor residue as major players in the recognition of the phosphorylation site by the PIKK kinases. This is a characteristic that contrasts with most of other kinase families analyzed here, for which we have identified at least two SDRs. The bar plot for the frequency of Q+1 among the 22 kinase families and the sequence logos for the PIKK family are available in the Appendices A1 and A2

Table 4.4: Specificity-determinant residue from a Gln-directed kinase family.

Kinase group	Kinase family	Pevents	SDR Q+1
Atypical	PIKK	480	80.83
Avg. SDR class			80.83
Avg. SDR global			80.83
Avg. non-SDR global			3.98

SDR: specificity-determinant residues identified. **Pevents**: number of phosphorylation events known for the kinase family. **Avg. SDR class**: Average frequency among the phosphorylation events of current class. **Avg. SDR global**: Average frequency among the phosphorylation events of all kinase families with the SDR (not only the kinases in the current class). **Avg. non-SDR global**: Average frequency among the phosphorylation events of kinase families without the SDR. With exception of the Pevents column, the values in the table represent the percentages of phosphorylation events.

4.3.2.3 Basophilic kinase families

The term 'basophilic' has been used to describe kinases that preferentially phosphorylate substrates having basic residues — mainly arginine (R) and lysine (K) — in close vicinity of the phosphorylation site [32]. The class of basophilic kinases is mainly composed of Ser/Thr kinases from the two large groups AGC and CAMK [37, 38], but also kinases from the Other and STE groups share this type of substrate specificity. Here we will discuss our results on the identification of SDRs, focusing on eight families of the basophilic class.

For 5/8 and 8/8 basophilic families in our set we identified arginine at -2 and/or -3 positions (R-2, R-3) as SDRs. These two SDRs are known to play key roles in the identification of the phosphorylation site by the basophilic kinases [34] (see Table 4.5); and in accordance to this, our results show that they have the highest two average frequencies among the phosphorylation events of the basophilic kinases in our set (R-3 = 45.43% and R-2 = 35.67%, see Table 4.5). Moreover, within our set of 22 kinase families, both R-3 and R-2 are identified as SDRs only for the basophilic ones (Table 4.5, first to last row), and also the average occurrence of both SDRs among the phosphorylation events of non-basophilic families are very low (R-3 = 5.14% and R-2 = 5.09%, Table 4.5 last row and Figure 4.2). Together, these results support R-3 and R-2 as very specific SDRs of the basophilic kinases.

In our set of basophilic kinases we count with four families of the AGC (PKA/PKC/PKG) group. For two of them, AKT and PKC (3 and 9 members respectively), we identified other SDRs in addition to the well known R-3 and R-2. Of all eight basophilic families in our set, AKT is the one showing the highest frequency of the SDR R-3 among its phosphorylation events (84.13%). For this family we have also found tryptophan at position +1 (W+1) as an SDR. This is an interesting observation given that tryptophan occurs rarely in the close vicinity of phosphorylation sites, most probably due to it can reduce the often required flexibility of the stretch of sequence targeted for phosphorylation [124]. However, we have found evidence reporting that kinases of the AKT family contribute to the regulation of transcription factors of the FOXO family by phosphorylating sequences containing a conserved W+1 [125]. Although, we have not found evidence in the literature reporting this SDRs

as part of the phosphorylation sequence motifs recognized by PKA kinases. This could be caused by the low occurrence of W+1 among the phosphorylation events of the AKT family (3.85%) (also see Figure 4.2).

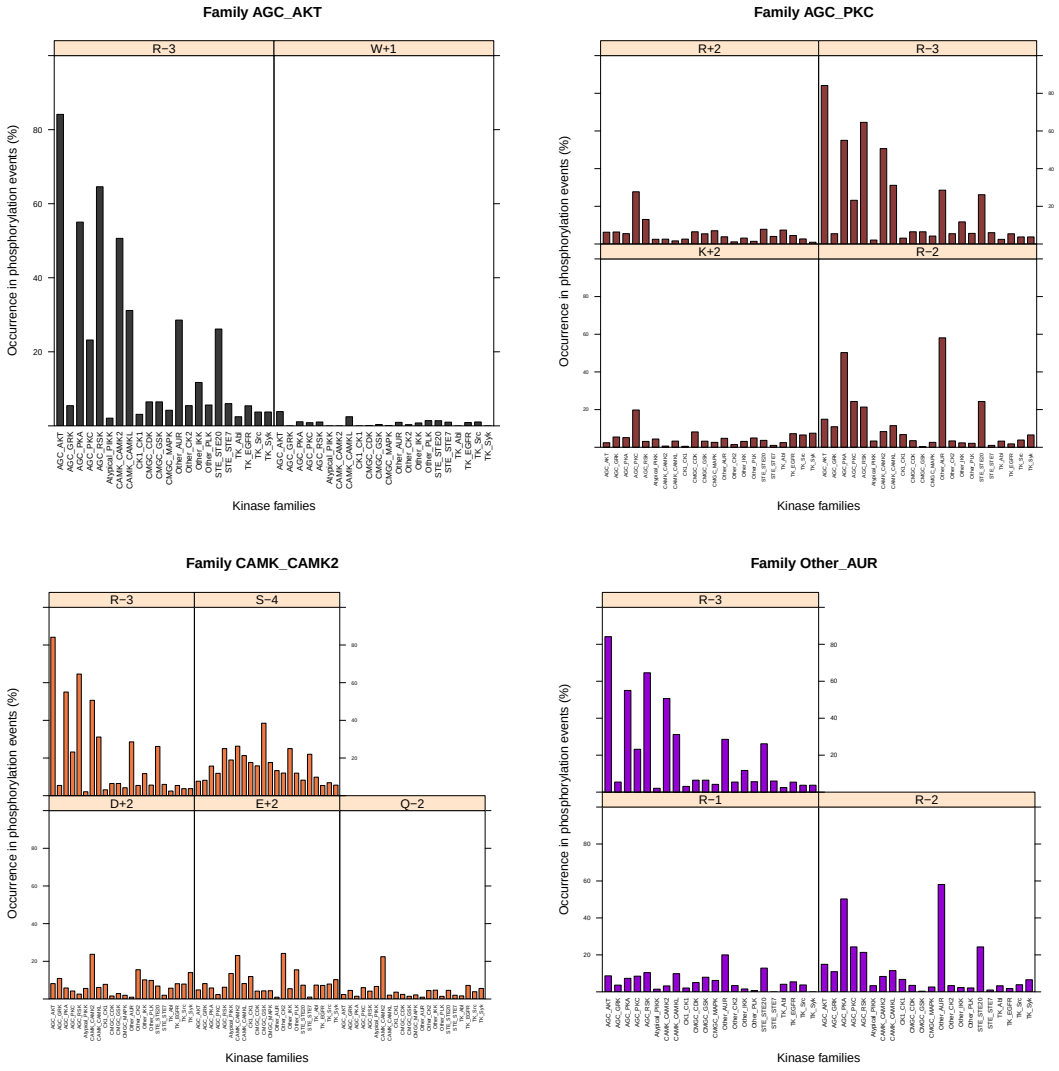


Figure 4.2: Frequencies of SDRs from four families of basophilic kinases.

Frequencies of occurrence of the SDRs of CDK and GSK kinase families among the phosphorylation events of other 22 kinase families. Boxes within each panel represent the SDRs. SDR are represented by the one letter code of the amino acid and its position relative to the phospho-acceptor residue. On the x-axis, the kinase families, on the y-axis the percentage of occurrence of each SDR.

The other family of the AGC group for which we identified additional SDRs is PKC. In this case, we identified R+2 and K+2 as important residues for the recognition of the phosphorylation site by PKC kinases. Both of these residues have been previously reported to

be part of the phosphorylation sequence motifs of members of the PKC family [8, 103, 126]. This requirement for positively charged amino acids at both sides of the phosphor-acceptor residue is, to the best of our knowledge, a characteristic unique to the PKC family. In contrast to the other basophilic families of the AGC group, all the SDRs identified for PKC family have a relatively low frequency of occurrence among the phosphorylation events of the family (23.75% on average) with no single SDR dominating over the others.

After analyzing SDRs on the AGC group, we have focused on CAMK2 and CMKL families (containing 4 and 20 members respectively) of the group CAMK (calcium/calmodulin-regulated kinases). Members of CAMK group are involved in several cellular events such as cell cycle progression, immune and inflammatory responses, signal transduction, gene transcription and synaptic development [127, 128]. Similar to other basophilic families analyzed here, we have identified R-3 as an SDR for both CAMK2 and CMKL. In the case of CAMK2 family we have identified four SDRs in addition to R-3. These SDRs are S-4, Q-2, D+2 and E+2 where the last three have been proposed as relevant for the substrate recognition of CAMK2 kinases [129]. One of the interesting SDRs identified for this family is Q-2, which is present in 22.44% of the phosphorylation events of the family, and in contrast has a low average frequency among the phosphorylation events of the other families in our set (3.7%, see Table 4.5 and Figure 4.2). We have also identified that other two SDRs of this family involve negatively charged residues (aspartic and glutamic acids, D and E respectively) in position +2. The identification of D+2 and E+2 as SDRs for the CAMK2 family contrast with requirement of positively charged SDRs at the same position by members of the previously discussed PKC family. We have also found that D+2 and E+2 are not SDRs exclusive of the CAMK2 family. Also the family CK2, that will be discussed later on, uses them for substrate recognition (see first to last column of Table 4.5, and Figure 4.2).

In the case of the CAMKL, and in contrast with the rest of basophilic families in the analysis, arginine is not the only SDR identified at position -3, but also lysine is identified as SDR at this position, however with a lower frequency if compared to arginine (R-3 = 31.15%, K-2 = 21.31%). Another SDRs identified for CAMKL are cysteines at -2 and +2 positions, however their frequencies among phosphorylation events of this family are relatively low (C-2 = 6.56% and C+2 = 9.84%). Cysteine is a very infrequent amino acid in phosphorylation sites [124], and we have not found evidence in the reviewed literature suggesting a role for neither C-2 nor C+2 in the phosphorylation site recognition by members of the CAMKL family. Given the aforementioned, we feel cautious about both these SDRs and we consider that their actual contribution to the substrate recognition of CAMKL kinases should subject to further investigation. The remaining SDR identified for CAMKL family is asparagine at position +3 (N+3), which appears to be an SDR exclusive of CAMKL kinases, and also has a low average frequency among phosphorylation events of the rest of families in the analysis (3.83%). We have not found evidence in the reviewed literature supporting N+3 as part of the phosphorylation motif identified by members of CAMKL family.

The next family of basophilic kinases analyzed was the one of Aurora kinases (AUR) composed by the members A, B and C, which have been functionally linked to different types of cancer [130]. For Aurora kinases we have identified the SDRs R-3, R-2 and R-1. In contrast to most basophilic families where R-3 is the more frequent SDR, in the case of AUR kinases R-2 is the SDR with the highest frequency among phosphorylation events of

Table 4.5: Specificity-determinant residues from basophilic kinase families.

			SDR													
Kinase group	**Kinase family**	**Pevents**	**S-4**	**R-3**	**K-3**	**R-2**	**Q-2**	**C-2**	**R-1**	**W+1**	**M+1**	**R+2**	**K+2**	**E+2**	**D+2**	**C+2**	**N+3**
AGC	AKT	208		84.13						3.85							
AGC	PKA	547		55.03		50.27											
AGC	PKC	1061		23.19		24.32						27.71	19.79				
AGC	RSK	192		64.58		21.35											
CAMK	CAMK2	156	26.28	50.64			22.44							23.08	23.72		
CAMK	CAMKL	244		31.15	21.31			6.56								9.84	18.03
Other	AUR	105		28.57		58.10			20.00								
STE	STE20	218		26.15	21.56	24.31					7.80						
Avg. SDR class			26.28	45.43	21.44	35.67	22.44	6.56	20	3.85	7.8	27.71	19.79	23.08	23.72	9.84	18.03
Avg. SDR global			32.39	45.43	21.44	35.67	22.44	6.56	20	3.85	7.8	27.71	19.79	23.64	19.63	9.84	18.03
Avg. non-SDR global			14.01	5.14	5.65	5.09	3.5	1.4	5.04	0.66	2.25	4.57	3.88	6.85	6.17	1.89	3.83

SDR: specificity-determinant residues identified. **Pevents**: number of phosphorylation events known for the kinase family. **Avg. SDR class**: Average frequency among the phosphorylation events of current class. **Avg. SDR global**: Average frequency among the phosphorylation events of all kinase families with the SDR (not only the kinases in the current class). **Avg. non-SDR global**: Average frequency among the phosphorylation events of kinase families without the SDR. With exception of the Pevents column, the values in the table represent the percentages of phosphorylation events.

the family (see Table 4.5 and Figure 4.2). Moreover, in AUR kinases R-2 has the highest value among all basophilic families in the analysis, which points to the relevance of this SDR for the family. The SDRs here identified for the AUR family have been previously reported as part of the phosphorylation sequence motifs identified by this family [131–133].

The last family analyzed in the class of basophilic kinases is STE20. STE20 is a family of 30 members from the group STE (homologs of yeast Sterile 7, Sterile 11, and Sterile 20 kinases). The family contains members of the MAPK cascade which are involved in a myriad of signaling processes in the cell [134, 135]. For STE20 we have identified the typical basophilic SDRs R-3 and R-2, and in addition K-3 and M+1 (see Table 4.5). In the case of M+1, its frequency among the phosphorylation events of the family is not particularly high (7.80%) and due to this we decided to take a closer look at it. We have noted that only five kinases from two sub-families out of the 11 that compose the STE20 group are the only ones responsible for the phosphorylation of sequences containing M+1. Moreover, these two subfamilies (MST and PAKA) are the mostly studied ones among the STE20 family, accounting for up to 65.5% of the phosphorylation events. Therefore, the presence of M+1 as an SDR of STE20 kinases seems to be produced by a bias of the data towards kinases of MST and PAK sub-families. However, we have not found evidence in the reviewed literature regarding M+1 as a key residue for neither MST not PAK kinases.

4.3.2.4 Acidophilic kinase families

Protein kinases in the acidophilic class are able to recognize phosphorylation sites surrounded by acidic residues (*i.e.*, aspartic and glutamic) [34,37,38]. This class is represented by several Ser/Thr kinases of the groups CK1, CMGC, Other and also by tyrosine kinases (TK group). Here we analyze the SDRs identified for five kinase families (CK1, CK2, IKK, PPLK and Syk) from the aforementioned groups.

The first family in our analysis is CK1 (casein kinase 1). Human CK1 kinases conform a small branch of seven proteins with diverse biological roles [136,137], that belong to a group of the same name. Initially, CK1 kinases were thought to phosphorylate only sites containing a previously phosphorylated Ser/Thr residue at position -3 (a 'primed' site) [138,139]. Further experiments showed that these kinases could also phosphorylate sites with negatively charged residues at the N-terminal positions. CK1 kinases are usually classified as an acidophilic family even tough negatively charged substrates remain much poorer than those containing phosphate groups [140, 141]. For CK1 kinases we have identified S-3 and S+3 as SDRs, but not E-3 or D-3. As stated before, S-3 is the priming residue required by CK1 kinases, and once the targeted Ser/Thr residue is phosphorylated, it serves as the 'priming' position for the next phosphorylation at S+3 [142]. From our data, S-3 and S+3 seem to be SDRs exclusive of CK1 kinases (see Table 4.6, first to last column), although their frequencies in phosphorylation events of other families do not seem negligible (see Table 4.6, last column and Figure 4.3).

The next family in our analysis is CK2 from the CMGC group. CK2 is a family of only two members (CK2a1 and CK2a2) that have been intensively studied due to their roles in cell growth, cell death, and cell survival [143, 144]. For CK2 kinases we have identified a total of eight SDRs, all of them aspartic or glutamic acid residues located mainly at C-terminal position with respect to the phosphorylation site (see Table 4.6 and Figure 4.3); a sequence pattern that have been previously reported for these kinases [34, 145]. From this sequence motif, is interesting to note the prevalence of glutamic residues in those positions where both aspartic or glutamic acids are identified as SDR, an observation that is particularly evident at position +3. We have identified E+3 and D+1 as the more relevant SDRs for CK2 kinases, a result that supports early findings by L. Pinna [146]. Half of the eight SDRs identified for CK2 kinases are shared by another families in our set. D-1 and D+3 are shared with the family IKK (see Table 4.6), while the other two (E+2 and D+2) are shared with the basophilic family CAMK2 (see Table 4.6 first to last row and Figure 4.3). We find that, on average, the occurrence of SDRs of the CK2 family among the phosphorylation events of other families in our set remains at relatively low values (5.77%).

We have also analyzed the family of IKK kinases from the Other group. The IKK family is composed of four members that play an important role in innate immunity as an essential part of the NF-kappaB signaling pathway [147, 148]. To the best of our knowledge, IKK kinases have not been previously classified as acidophilic. However, based on the available data, our results suggest that these kinases have a preference for phosphorylation sites surrounded by acidic residues. For IKK family we have identified the SDRs D-1, D+3 and S+4, of which D-1 is an typical acidophilic SDR present in four out of the five acidophilic families in our analysis and D+3 is also shared by the CK2 family (see Table 4.6 and Figure 4.3). In the

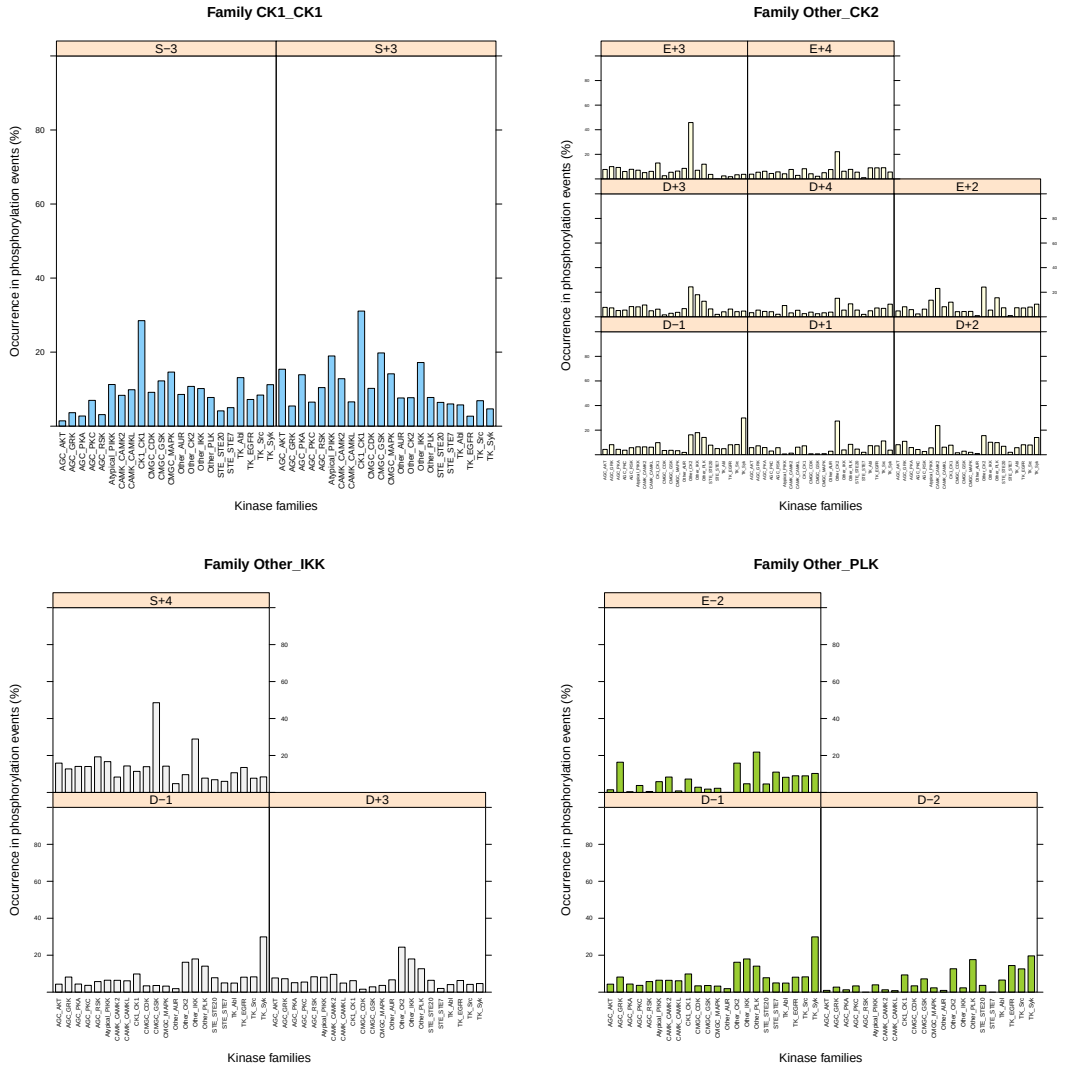


Figure 4.3: Frequencies of SDRs from four families of basophilic kinases.

Frequencies of occurrence of the SDRs of CDK and GSK kinase families among the phosphorylation events of other 22 kinase families. Boxes within each panel represent the SDRs. SDR are represented by the one letter code of the amino acid and its position relative to the phospho-acceptor residue. On the x-axis, the kinase families, on the y-axis the percentage of occurrence of each SDR.

case of S+4, there is evidence in the literature suggesting that some members of the IKK family are responsible for the multiple phosphorylation of the substrate NF-kappaB1 p105 by recognizing consecutive motifs with the consensus sequence DSXX[DE]S [149, 150].

Table 4.6: Specificity-determinant residues from acidophilic kinase families.

Kinase group	Kinase family	Pevents	SDR												
			S-3	E-2	D-2	D-1	D+1	E+2	D+2	S+3	E+3	D+3	S+4	E+4	D+4
CK1	CK1	193	28.50							31.09					
CMGC	CK2	624				16.19	27.4	24.2	15.54		45.83	24.36		22.12	15.06
Other	IKK	128				17.97						17.97	28.91		
Other	PLK	142		21.83	17.61	14.08									
TK	Syk	107				29.91									
Avg. SDR class			28.50	21.83	17.61	19.54	27.40	24.20	15.54	31.09	45.83	21.17	28.91	22.12	15.06
Avg. SDR global			28.50	21.83	17.61	19.54	27.40	23.64	19.63	31.09	45.83	21.17	38.74	22.12	15.06
Avg. non-SDR global			8.08	5.91	5.22	5.64	4.60	6.85	6.17	9.85	6.15	5.90	11.51	5.76	5.05

SDR: specificity-determinant residues identified. **Pevents**: number of phosphorylation events known for the kinase family. **Avg. SDR class**: Average frequency among the phosphorylation events of current class. **Avg. SDR global**: Average frequency among the phosphorylation events of all kinase families with the SDR (not only the kinases in the current class). **Avg. non-SDR global**: Average frequency among the phosphorylation events of kinase families without the SDR. With exception of the Pevents column, the values in the table represent the percentages of phosphorylation events.

The next family analyzed was the one of Polo-like kinases (PLK). PLKs is a highly conserved family of Ser/Thr kinases [151] that are involved in the regulation of the cell cycle progression [27, 152]. PLKs have paramount roles in mitosis, and therefore, it is not unexpected that their deregulation has been associated to cancer and oncogenesis [151, 153]. For this family we have identified the SDRs E-2, D-2 and D-1. In the cases of E-2 and D-2, we have found previous reports supporting the relevance of these SDRs as main players in the identification of substrates by PLKs [27, 28, 34, 97, 154]. Regarding D-1, we have not found previous reports suggesting this position as an SDR for members of the PLK family. However, D-1 is shared by most of the acidophilic families in our analysis and we consider that is plausible a potential contribution to the substrate specificity of PLKs, although to a lower extent if compared to E-2 and D-2 (see Table 4.6 and Figure 4.3). Most studies regarding the phosphorylation motifs identified by PLK kinases refer also to a preference for hydrophobic residues in position $+1$ [27, 28, 34, 97, 154]. We consider that we have not been able to identify such SDRs most likely due to the cut-off criteria used for the definition of SDRs.

The last family analyzed among the acidophilic class is the one of Spleen tyrosine kinases (Syk), composed of SYK and ZAP-70, which play crucial roles in the adaptive immune response [155, 156]. Several reports suggest that tyrosine kinases display preference for acidophilic sequences [33, 157, 158] and Syk kinases are no exception to this. For the Syk family previous studies have reported phosphorylation sequences enriched in aspartic and/or glutamic residues, where the position -1 plays a fundamental role [141, 159]. We have identified D-1 as the only SDR for Syk kinases and is interesting to note that, in this case, D-1 shows the highest frequency of occurrence among all acidophilic kinases in our analysis (29.91%, see see Table 4.6 and Figure 4.3). Nevertheless, due to that the number of phosphorylation events explained by D-1 is relatively low (29.91%) we consider that other SDRs, not detected by our method, might be contributing to the substrate specificity in this family.

4.3.3 Statistical analysis of the PSSMs

In this section we discuss the evaluation of the statistical significance and performance of the PSSMs previously generated. In our data set, the number of phosphorylated sequences available for a given kinase can span from only one to several hundreds. Due to this, we need to understand how the number of seed phosphorylation sites available for a kinase could affect the statistical significance and the performance of its PSSM. The main purpose here have been the identification of statistically significant PSSMs and for this we conducted statistical evaluations for both the PSSMs from independent kinases and for those from kinase families, setting the statistical significance level at $\alpha < 0.01$.

4.3.3.1 Independent kinases

We have started by analyzing the statistical significance of the PSSMs based on the percent recall that they achieve on the set of seed phosphorylation sites (see section 4.2.5.3 of Materials and methods). Here we have found a negative correlation between the number

of seed phosphorylation sites of a kinase and the percent recall of its PSSM (R = -0.59, *p*-value = $2.38e^{-31}$. See Figure 4.4). This relationship suggests that an increasing number of seed phosphorylation sites can degenerate the signal contained in the PSSM, which will directly affect the ability of a PSSM to identify those sites at a given significance threshold. However, by using the percent recall as a statistic, we have not found a relationship between the number of seed phosphorylation sites and the statistical significance of a PSSM (Pearson R = -0.093, *p*-value = 0.095. See Figure 4.4). In deed, all the PSSMs in the analysis (325) were statistically significant.

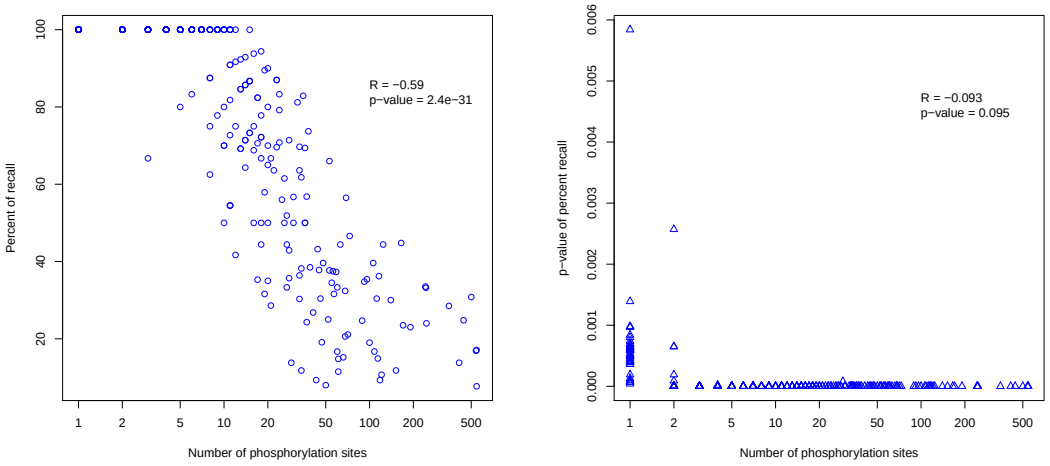


Figure 4.4: Percent recall and statistical significance of PSSMs from independent kinases.

On the left, the relationship between number of seed phosphorylation sites and the percent recall of the PSSMs. On the right, the relationship between the number of seed phosphorylation sites and the *p*-value of the PSSM, based on the percent recall as the test statistic. The x-axes are shown in logarithmic scale.

As previously mentioned, we have also analyzed the statistical significance of the PSSMs based on their IC. Here we have found a negative correlation, although not particularly strong, between the number of seed phosphorylation sites and the IC of a PSSM (R = -0.4, *p*-value = $9.8e^{-14}$ (see Figure 4.5). This result suggests that the sequence degeneracy caused by the increase of the number of seed phosphorylation sites can affect not only the percent recall of a PSSM, as previously seen, but can also decrease the information contained in the PSSM model. As can be expected, from two previous analysis, we have found a strong correlation between the IC of a PSSM and its percent recall (R = 0.8, *p*-value = 0 (see Figure 4.5).

Nevertheless, by using the IC as a statistic, we were able to identify a relationship between the number of seed phosphorylation sites and the statistical significance of the PSSMs. In this sense, our results show that PSSMs with a statistically significant IC were generated from sets of phosphorylation sites significantly larger (10.2 times larger) than the PSSMs with non statistically significant IC (see Table 4.7 and Figure 4.6). We also observe a notable and statistically significant difference regarding the percent recall that the two sets of PSSMs

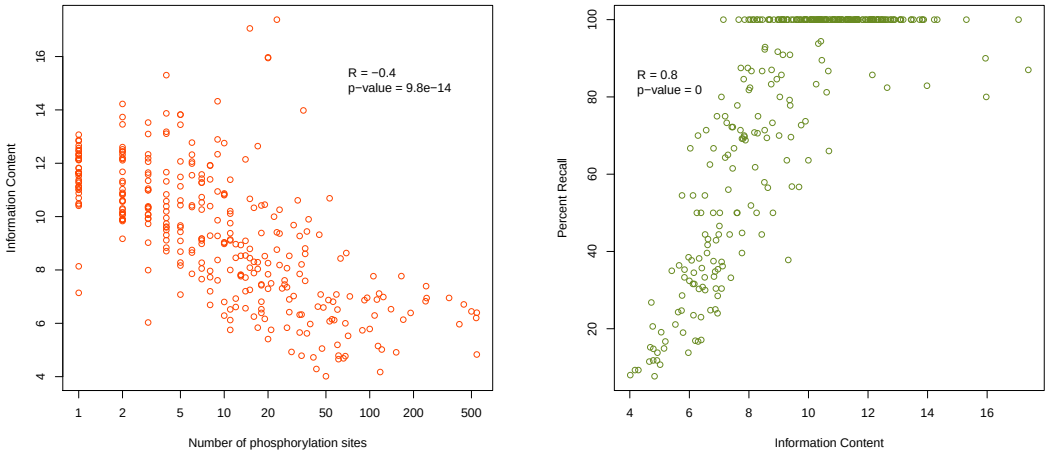


Figure 4.5: IC and seed phosphorylation sites of PSSMs from independent kinases.

On the left, the relationship between number of seed phosphorylation sites and the IC of the PSSMs. On the right, the relationship between the IC and the percent recall of the PSSMs.

achieve on their sets of seed phosphorylation sites. In this case, the non statistically significant PSSMs achieve a mean percent recall 1.4 times larger than the one of statistically significant PSSMs. As previously discussed, this comes as a result of the difference between the number of seed phosphorylation sites used for generating the PSSMs. The sets of PSSMs were also compared based on two additional parameters, the IC and the percent recall on the test set of 'unphosphorylated' human proteins. On both cases we also found statistically significant differences between the two sets of PSSMs (see Table 4.7 and Figure 4.6).

We have also used the area under the curve of the receiver operating characteristic (AUC-ROC) for comparing the performances of the two sets of PSSMs. Here we have found statistically significant differences between the medians of the AUC-ROC for the two sets (see Table 4.7 and Figure 4.7). Additionally, we have also found a negative correlation (R = -0.63, p-value = $6.2e^{-37}$) between the AUC-ROC and the number of seed phosphorylation sites of a PSSM (see Figure 4.7).

Here we have conducted a statistical and performance analysis on the set of 325 PSSM generated for independent kinases. We have been able to assess the significance of the PSSMs based on their IC values and we have identified that half of the PSSMs in our set (163/325, 50.2%) are statistically significant. The sets of statistically significant and non statistically significant PSSMs show significant differences on several parameters evaluated here.

Table 4.7: Statistically and non statistically significant PSSMs from independent kinases.

		IC		Psites		%Recall		%Recall TS		AUC-ROC	
	No. PSSMs	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**
Statistically significant	163	8.46	8.99	23	57.63	69.7	64.82	3.9	4.35	0.9	0.85
Non statistically significant	162	10.09	9.77	3	5.65	100	90.95	3	3.14	1	0.97
Mann-Whitney U test p-value		$6.16e^{-04}$		$7.55e^{-36}$		$1.09e^{-18}$		$2.11e^{-14}$		$2.80e^{-20}$	

The table shows the sets of statistically and non statistically significant PSSMs and the five parameters on which they were compared. The last row shows the results of the Mann-Whitney U test for estimating differences between the two populations of PSSMs. **IC**, information content, **Psites**, number of seed phosphorylation sites, **%Recall**, percent recall achieved on the set of seed phosphorylation sites, **%Recall TS**, percent recall achieved on the test set of 'unphosphorylated' proteins, **AUC-ROC**, area under the receiving operating characteristic curve.

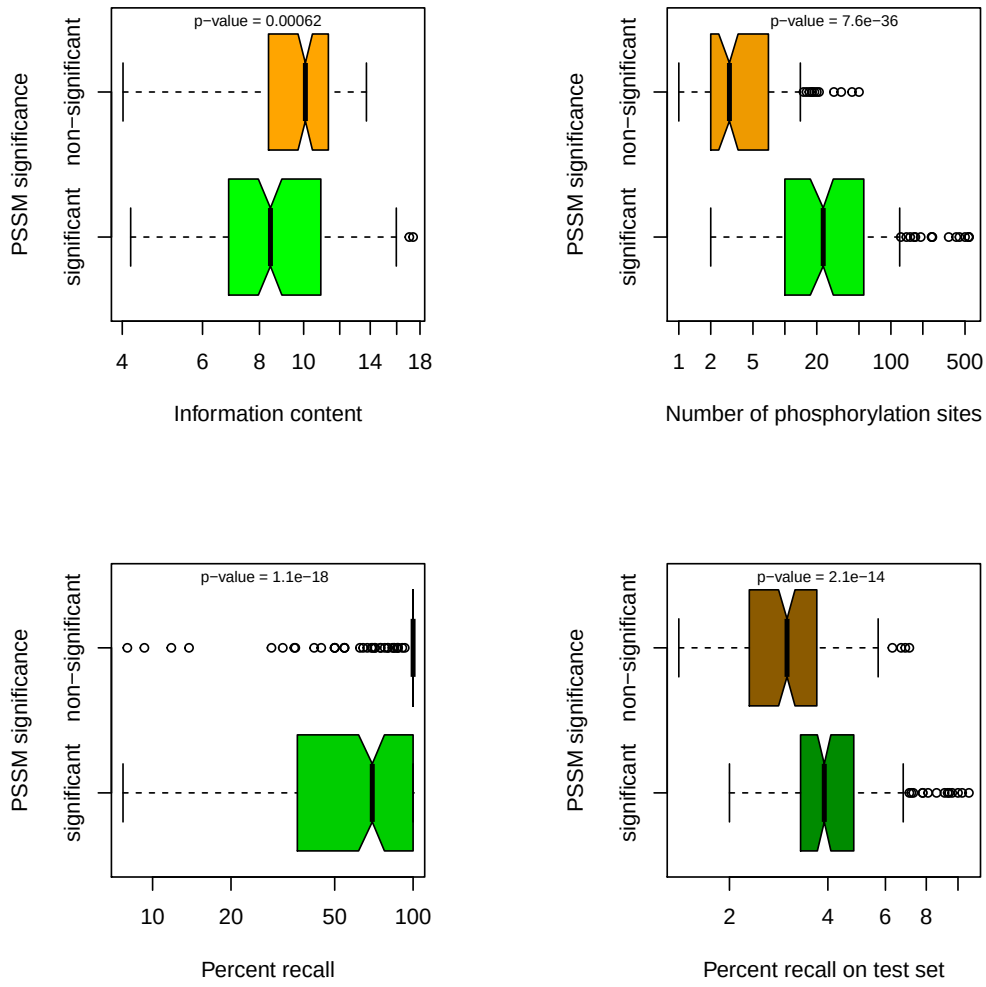


Figure 4.6: Statistically and non statistically significant PSSMs from independent kinases.

The PSSMs sets were compared based on the four parameters, i) information content, ii) number of seed phosphorylation sites, iii) percent recall on the seed phosphorylation sites and iv) percent recall on the set of 'unphosphorylated' human proteins. The Mann-Whitney test was used to estimate the statistical significance of the differences between the median values of the parameters. The vertical bar within each box represents the median. If the notches in the boxes do not overlap, this is 'strong evidence' that the medians differ. The x-axis are shown in logarithmic scale.

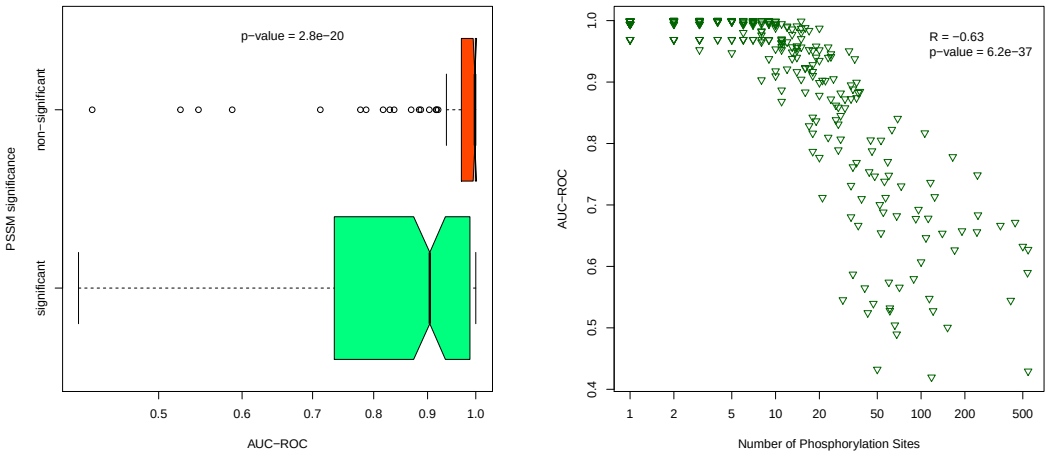


Figure 4.7: AUC-ROC for PSSMs from independent kinases.

On the left, the distributions of the AUC-ROC for both sets of PSSMs. On the right, the correlation between the AUC-ROC and the number of seed phosphorylation sites.

4.3.3.2 Kinase families

We have also evaluated the statistical significance and performance of the PSSMs generated for the 93 kinase families in our set. For this, we followed the same procedures used for the analysis of the PSSMs from independent kinases. In global terms, the relationships observed here between the parameters under study were very similar to the ones observed for independent kinases. This is an expected result, given that the only parameter affected when merging kinases by families is the number of seed phosphorylation sites, and all other parameters depending on it will vary accordingly.

Again, when analyzing the statistical significance of PSSMs based on the *p*-value of their percent recall, we found a negative correlation between the number of phosphorylation sites and the percent recall (R = -0.48, *p*-value = $1.2e^{-06}$). As previously explained, this negative correlation is caused by the degeneracy of the signal in the PSSM as the number of seed phosphorylation sites increases. Regarding the use of the percent recall as a statistic for the classification of significant PSSMs, we have not found a significant correlation between the number of phosphorylation sites and the *p*-value of the percent recall (R = -0.089, *p*-value = 0.4), with all the PSSMs from the 93 families being classified as statistically significant. Therefore, the percent recall is not a useful statistic for identifying significant PSSMs. See plots on Figure 4.8.

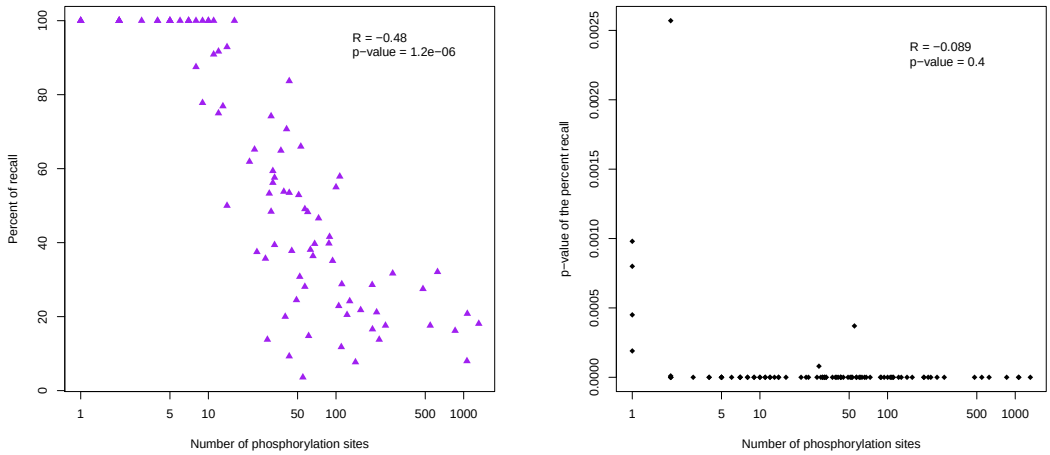


Figure 4.8: Percent recall and statistical significance of PSSMs from kinase families.

On the left, the relationship between number of seed phosphorylation sites and the percent recall of the PSSMs. On the right, the relationship between the number of seed phosphorylation sites and the *p*-value of the PSSM based on the percent recall. The x-axes are shown in logarithmic scale.

When analyzing the statistical significance of the PSSMs, this time based on their IC value, we have found a negative correlation between the number of phosphorylation sites and the IC (R = -0.33, *p*-value = 0.0013, see Figure 4.9) and, as expected, we also find a

positive correlation between the percent recall and the IC (R = 0.85, p-value = 0, see Figure 4.9). Similar to the analysis of PSSMs of independent kinases, this result shows that the sequence degeneracy of the phosphorylation motif caused by the increment of the number of seed phosphorylation site affects not only the percent recall, but also the IC of the PSSM.

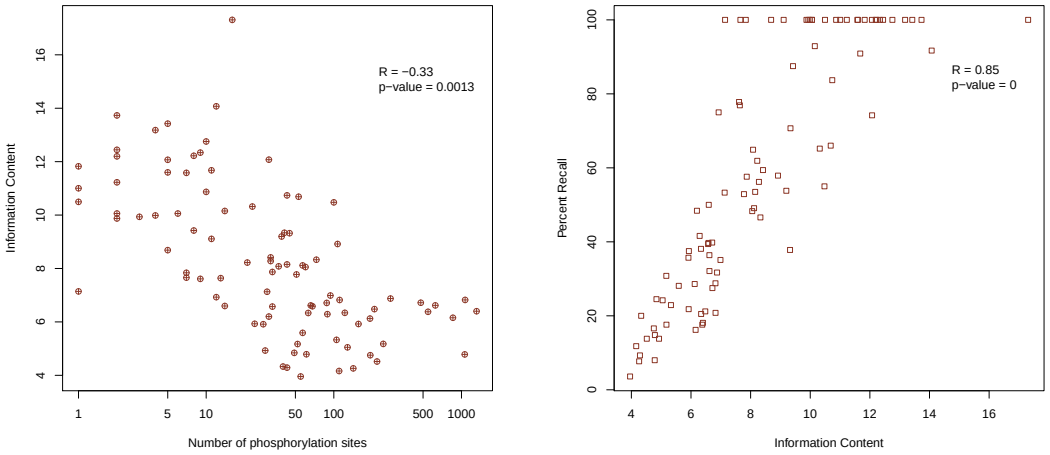


Figure 4.9: IC and seed phosphorylation sites of PSSMs from kinase families.

On the left, the relationship between number of seed phosphorylation sites and the IC of the PSSMs. On the right, the relationship between the IC and the percent recall of the PSSMs.

By using the IC as a statistic for classifying PSSMs into statistically and not statistically significant, it was again possible to distinguish two sets of PSSMs. Also, we observed differences between the sets when compared based on the percent recall of seed phosphorylation sites, the percent recall on the test set and the AUC-ROC. In contrast to what was previously found for the PSSMs of independent kinases, when comparing the sets based on the median IC values, we do not observe statistically significant difference between the sets. For the other remaining comparison criteria — percent recall and AUC-ROC — we found that the two sets of PSSMs differ significantly (see Table 4.8 and Figures 4.10 and 4.11).

Table 4.8: Statistically and non statistically significant PSSMs from kinase families.

		IC		Psites		%Recall		%Recall TS		AUC-ROC	
	No. PSSMs	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**
Statistically significant	69	7.13	8.12	52.00	145.60	46.60	50.80	4.00	4.46	0.77	0.76
Non statistically significant	24	9.49	8.83	4.50	10.88	100.00	81.23	2.80	3.19	1.00	0.90
Mann-Whitney U test p-value		$1.77e^{-01}$		$8.57e^{-09}$		$1.89e^{-04}$		$1.29e^{-04}$		$1.36e^{-04}$	

The table shows the sets of statistically and non statistically significant PSSMs and the five parameters on which they were compared. The last row shows the results of the Mann-Whitney U test for estimating differences between the two populations of PSSMs. **IC**, information content, **Psites**, number of seed phosphorylation sites, **%Recall**, percent recall achieved on the set of seed phosphorylation sites, **%Recall TS**, percent recall achieved on the test set of 'unphosphorylated' proteins, **AUC-ROC**, area under the receiving operating characteristic curve.

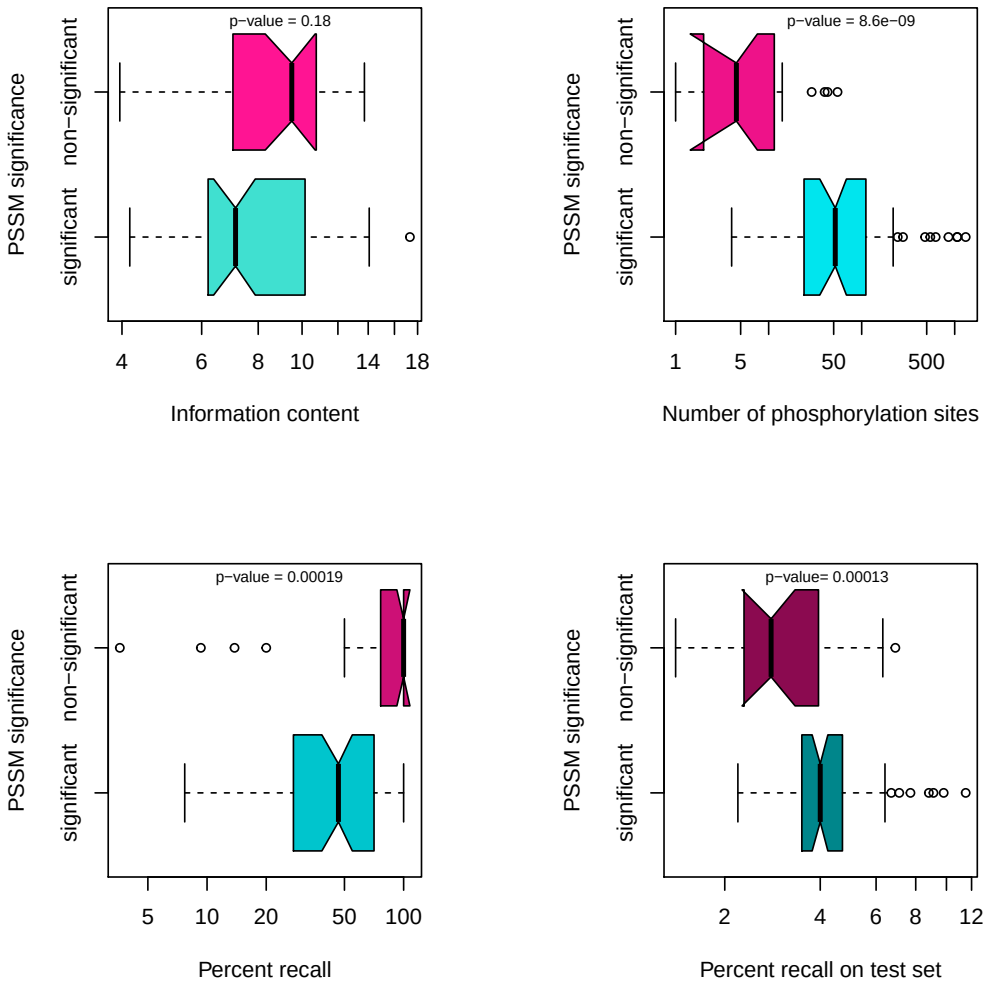


Figure 4.10: Statistically and non statistically significant PSSMs from kinase families.

The PSSMs sets were compared based on the four parameters, i) information content, ii) number of seed phosphorylation sites, iii) percent recall on the seed phosphorylation sites and iv) percent recall on the set of 'unphosphorylated' human proteins. The Mann-Whitney test was used to estimate the statistical significance of the differences between the median values of the parameters. The vertical bar within each box represents the median. If the notches in the boxes do not overlap, this is 'strong evidence' that the medians differ. The x-axis are shown in logarithmic scale.

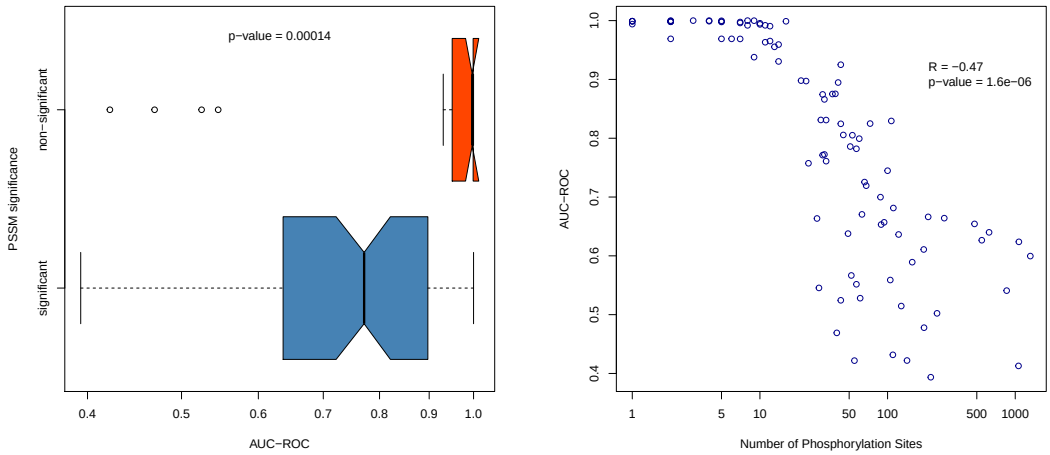


Figure 4.11: AUC-ROC for PSSMs from kinase families.

On the left, the distributions of AUC-ROC for both sets of PSSMs. On the right, the correlation between the AUC-ROC and the number of seed phosphorylation sites.

Here we have analyzed the statistical significance and performance of PSSMs derived from phosphorylation sites of 93 kinase families. We have identified a total of 69/93 (74.2%) PSSMs to be statistically significant in terms of their IC. We also observe that the sets of statistically significant and non statistically significant PSSMs differ in most of the comparison criteria that we have used in the analysis.

4.4 Concluding remarks

- By integrating data from different public resources, we have collected a set of experimentally determined kinase–phosphorylation sites relationships in human. The data comprise 325 (62.7%) human kinases — accounting for 71.5% of the kinase families —, 1856 substrates and 5946 phosphorylation sites. Our integrated data increase by 18%, 58% and 59% the numbers of kinases, substrates and phosphorylation sites (respectively), if compared to the averages from the source databases.

- By using sequence logos, we have graphically represented the phosphorylation motifs recognized by kinases and kinase families in our data. The patterns in sequence logos showed the great diversity of sequences phosphorylated by the kinases, and also guided the classification of kinases and kinase families based on the residue composition of the sequences that they target.

- We have used PSSMs as the probabilistic models for representing the phosphorylation motifs targeted by kinases and kinase families in our data. Based on their scores in the PSSMs, we classified several residues as SDRs for some of the kinase families in our set. We have observed that the identity, the position on the sequence and the frequency of the SDRs identified, vary considerably among the different families analyzed.

- Kinases from the MAPK and CDK families rely mostly on the SDR P+1. This is evident in our data given the high frequencies of P+1 among phosphorylation events of the two families ($\text{MAPK}_{\text{P+1}} = 88.86\%, \text{CDK}_{\text{P+1}} = 81.72\%$). In comparison, the contribution of P+1 is smaller for the family GSK ($\text{GSK}_{\text{P+1}} = 53.96\%$), for which the presence of the SDRs $\text{GSK}_{\text{S}-4} = 38.49\%$ and $\text{GSK}_{\text{S+4}} = 48.56\%$ seems to be relevant.

- For the family PIKK, the high frequency of the SDR Q+1 among the phosphorylation events of the family ($\text{PIKK}_{\text{Q+1}} = 80.83\%$) suggests it as a major requirement for the recognition of cognate phosphorylation sites. Moreover, given the low frequency (3.98%) of this SDR among the phosphorylation events of the other kinase families in our analysis, we consider that Q+1 is used almost exclusively by PIKK kinases.

- In the case of the basophilic kinases, we observe that the frequencies of the SDRs identified vary greatly from one family to the other. Regarding the family AKT, we observe a high frequency for R-3 ($\text{AKT}_{\text{R}-3} = 84.13\%$) among the phosphorylation events of the family. This suggests R-3 as a major SDR for AKT kinases. In comparison, other basophilic families such as PKA, PKC, RSK, CAMKL and AUR appear to rely on either R-3, K-3 or R-2 with frequencies that can range between 21.31% to 58.10% among the phosphorylation events of each family.

- In the case of acidophilic kinases, we also observe a large variability regarding the SDRs identified as well as on their relative frequencies. For the CK1 family we identified the non-acidic SDRs $\text{CK1}_{\text{S}-3} = 28.50\%$ and $\text{CK1}_{\text{S+3}} = 31.09\%$, which upon phosphorylation by upstream kinases, acquire the negative charge required for the recognition by the CK1 kinases. For the CK2 family we identified eight different SDRs — all of them

Asp or Glu — with a rather low average frequency of 23.8%. To our opinion, given their low frequencies, many of these SDRs might co-exist in the same target sequence and therefore contribute in an cooperative manner to the recognition of the target site.

- For the AKT family we identified the SDR AKT_{W+1} that, to the best of our knowledge, have not been previously identified as part of the sequence motif recognized by this family. However, we found evidences in the literature linking W+1 to the recognition of target sequences in the family of FOXO transcription factors.

- On average, the identified SDRs have a rather low frequency of occurrence (6.01%) among the complementary phosphorylation events. That is, the phosphorylation events corresponding to those kinase families that do not count with the given SDR. To our opinion, this suggests that the identified SDRs work not only as a positive selection element for cognate target sequences, but also as negative selection factors for non-cognate phosphorylation sites.

- Regarding the analysis of PSSMs, we have found negative correlations between the number of seed phosphorylation sites and i) the percent recall of the PSSM (R = -0.59, *p*-value = $2.4e^{-31}$) and ii) the information content of the PSSM (R = -0.4, *p*-value = $9.8e^{-14}$). These results show the effect that the sequence degeneracy caused by the increase of the seed phosphorylation sites can exert on the performance of the PSSM and on its level of self-information.

- Based on their values of IC, and on the comparison to random backgrounds, we have estimated the statistical significance of PSSMs from both independent kinases and kinases families. We observe that, in most cases, statistical and non-statistically significant PSSMs differ not only in their values of the IC but also in their percent recall, their AUC-ROC and their numbers of seed phosphorylation sites.

5 Contribution of adaptor and scaffold proteins

5.1 Introduction

In the cell, several processes can take place at the same time and sometimes in overlapping locations. In order to efficiently coordinate these processes, the cells have developed several mechanisms that are responsible for the spatial and temporal organization of the biological events. One of these mechanisms consists in the recruitment of the proteins required for a given process to specific locations within the cell, or their assembly in larger functional macromolecular complexes. This recruitment process is carried out in many cases by adaptor and/or scaffold proteins, which are macromolecules able to bind more than one partner at the time and promote in this way their mutual interaction and regulation [160]. Initially, adaptors and scaffolds were regarded as passive platforms for protein recruitment or assembly of signaling complexes [161–164]. However, recent studies have made clear that they can play rather active regulatory functions. There is an increasing interest in understanding the molecular mechanisms that govern these proteins given the evidence accumulated about their roles in several processes such as signal transduction, cell-cell communication and cell structural organization [165], and their potential implication in human pathologies [166–168].

5.1.1 Adaptors and scaffolds are multidomain spatio-temporal regulators

The boundaries for clearly distinguishing between adaptor and scaffold proteins have been largely debated by the community. However, it is accepted that they are able to bind to more than one protein at once and that they generally contain multiple domains and motifs for protein-protein interaction (*e.g.*, SH2, SH3, PH and WW) [165, 169]. In this regard, Buday and Tompa defined adaptors as proteins that are able to link to functional members of a catalytic pathway; and scaffolds as proteins of higher molecular mass — if compared to adaptors — that target and regulate at the same time at least two signaling enzymes and promote their communication by proximity [170]. This ability to bind to two or more signaling proteins at once, provides the cell with a mechanism to regulate the fidelity of signaling events in space and time, to propagate information by proximity, and to determine the specificity of information flow in intracellular networks [171–173]. It has also been shown that these proteins can exert an inhibitory effect on signal transmission. If their concentrations exceed that of its partners, incomplete complexes will form and this could effectively dissipate signaling [174–176]. Therefore, optimal expression levels of adaptors and scaffolds are required to achieve maximal response of the pathways they are involved in.

Besides promoting their interaction by proximity, scaffolds and adaptors can localize sig-

naling proteins at specific regions or compartments of the cell, such as the plasma membrane, the cytoplasm, the nucleus, the Golgi, the endosomes and the mitochondria which might be important for the local production of signaling intermediates [173].

Finally, it has been suggested that *bona fide* adaptors and scaffolds should not have intrinsic catalytic activity [177]. However, cases such as the kinase suppressor of Ras (KSR) and the focal adhesion kinase (FAK) have been reported to play, under certain conditions, the roles of enzymes and scaffolds [167, 178].

5.1.1.1 Adaptors and scaffolds of the MAPK/ERK cascade

Two hallmarks of scaffolding and adaptor function in multicellular organisms are the scaffold kinase suppressor of RAS1 (KSR) and the adaptor growth factor receptor-bound protein 2 (GRB2). Both GRB2 and KSR play central roles in the MAP kinase cascade of ERK, a pathway present in all multicellular organisms [179] and for which a similar mechanism is also present in yeast [161, 180]. The ERK cascade is activated by extracellular factors (*e.g.*, VGF and VEGF), culminates in transcriptional response affecting several processes such as cell proliferation and differentiation. The deregulation of this cascade is tightly related to human cancers [176, 181, 182]. In this cascade, the activation of receptor tyrosine kinases (RTK) by extracellular mitogens (*e.g.*, EGF and VEGF) results in the recruitment of the adaptor GRB2, that binds to son of sevenless homolog (SOS), which then interacts with and activates RAS. This leads to the activation of RAF and thereby, the initiation of the sequential phosphorylation steps of the MAPK cascade [183]. Activated ERK, the last element of the three-tiered kinase cascade (RAF-MEK-ERK), can later phosphorylate either cytosolic or nuclear substrates (see Figure 5.1).

In more detail, the adaptor GRB2 recruits SOS to the proximity of the activated RTKs via an interaction of an SH3 domain of GRB2 with a proline-rich region in SOS; and later, via its SH2 domain, GRB2 binds to a phosphorylated tyrosine in the activated RTK [184, 185]. In this way the GRB2-SOS complex gains access and activates its membrane-bound target RAS.

Once RAS has been activated, the scaffold KSR migrates from the cytoplasm to the plasma membrane carrying a constitutively bound MEK — a dual-specificity protein kinase — to the encounter with RAF [162, 186]. The binding of KSR to RAF promotes conformational changes in the alpha C helices of both proteins, which allosterically turn the kinase domains into active conformations [183, 187]. Upon binding to RAF, KSR promotes a conformational change in MEK — by releasing its otherwise inaccessible activation segment — and that allows RAF to activate MEK by phosphorylation [178, 188]. RAF, KSR and MEK interact via their kinase domains in a non mutually exclusive manner [178, 189]. However, the RAF molecule bound to KSR is sterically unable to phosphorylate MEK in *cis*, an action that must be performed (in *trans*) by another RAF molecule [178]. Despite KSR has been considered as a inactive pseudokinase due to the lack of canonical catalytic residues, recent *in vitro* experiments suggest that KSR is a *bona fide* kinase whose catalytic activity is required to cooperate with RAF for the activation of MEK [178, 190]. Regarding ERK, its binding to KSR is induced by RAS activation and mediated by the sequence motif FXFP; a conserved docking site present in KSR (far in sequence from the kinase domain) that specifically binds

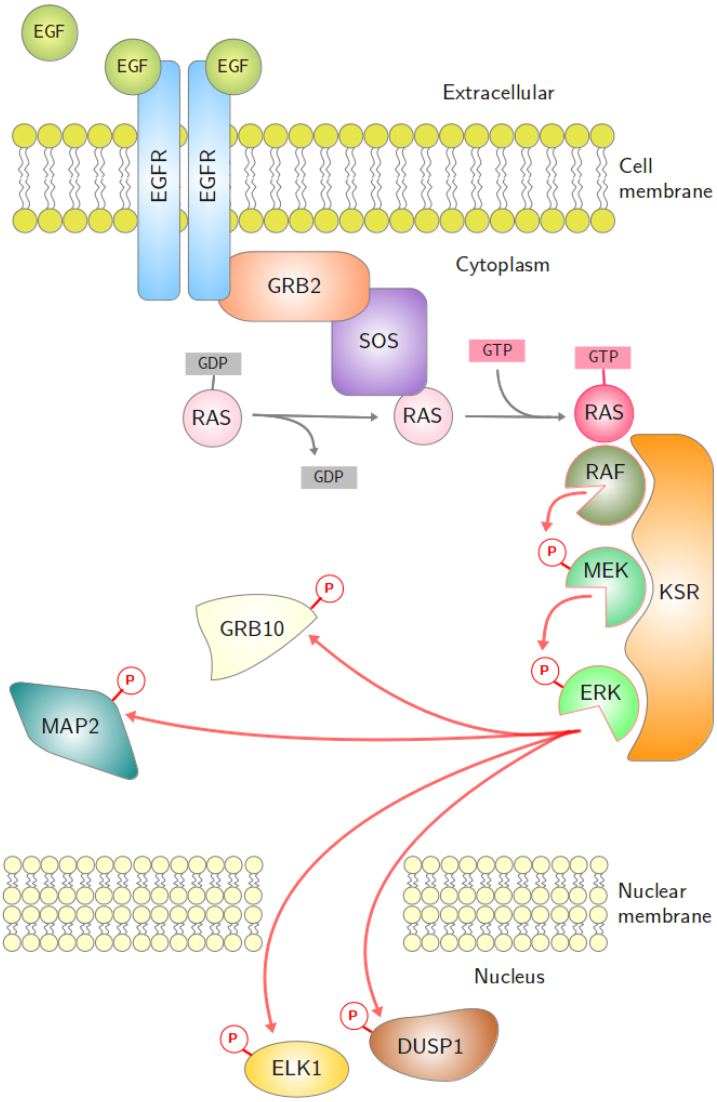


Figure 5.1: ERK MAP kinase cascade.

Upon stimulus from extracellular mitogens, the adaptor GRB2 binds to activated RTKs and promotes the activation of RAS. Activated RAS contribute to the transduction of extracellular signal by activating the three-tiered MAPK/ERK cascade. The three kinases of the MAPK/ERK cascade are held together by the scaffold KSR. Activated ERK phosphorylates several substrates in the cytoplasm and in the nucleus.

to ERK [189,191]. Once ERK is bound to the RAF-KSR-MEK complex, MEK promotes ERK activation by phosphorylating residues in the activation segment of ERK [192]. The activation of ERK leads to the transmission of the cellular signal by the further phosphorylation of several substrates in different cellular compartments [175]. However, ERK is not only responsible for the transmission of the extracellular signal to other effectors in the cell, but it also regulates

the MAPK cascade using a feedback loop that involves KSR and RAF. It has been recently shown that phosphorylation of KSR and RAF by ERK promotes the dissociation of the KSR-RAF complex and the release of KSR from the plasma membrane with the subsequent attenuation of the pathway output [193]. This feedback phosphorylation of KSR by ERK demonstrates a relevance of the scaffold in the dynamic regulation of the pathway, a role that goes beyond the tethering and co-localization of the cascade elements. By the mechanisms described here, KSR is able to assure an efficient propagation of the signal along the three-tier group of kinases and to actively contribute to the transmission of the signal to effector proteins in the cell.

Besides KSR, other scaffolds such as β-arrestin, HOMER, IQGAP1, MP1 and paxillin participate in several pathways. For reviews on these and other scaffolds involved in signaling events the reader can refer to excellent reviews by Brown and Sacks [179], Shaw and Filbert [173] and Pan *et al.* [194].

5.1.1.2 The IQGAP scaffolds regulate cytoskeleton dynamics

The IQGAP proteins have been identified in a wide spectrum of organisms that range from yeast to human. In human, there have been identified three members (IQGAP1, IQGAP2 and IQGAP3) which have considerable sequence identity (around 43%) but they differ in aspects such as function, tissue distribution and cellular localization [195]. IQGAP1 is the most widely studied member of this family in human and is known to be ubiquitously expressed and to play fundamental roles in several cellular processes such as signaling, microtubules regulation, cell polarization, migration and motility among many others.

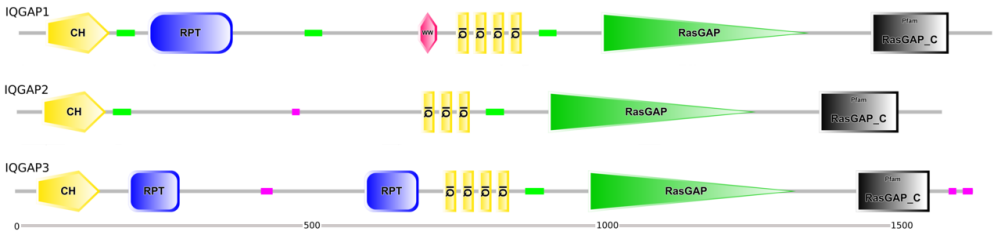


Figure 5.2: The human IQGAP protein family.

CH, calponin homology domain (a.k.a CHD); RPT, region of internal repeat; WW, poly-proline-binding domain; IQ, calmodulin-binding motif; RasGAP, GTPase-activator domain for Ras-like GTPases; RasGAP_C, carboxy-terminal sequence found in members of the IQGAP family; magenta bars denote potential coiled coils; green bars denote stretches of low structural complexity. The lowermost bar indicates approximate residues numbers. Diagram adapted from the web resource SMART [196].

IQGAP1 contains several domains and motifs for protein-protein binding that allow it to interact with a large number of partners (see Figure 5.2). Starting from its N-terminal, IQGAP1 contain a calponin homology domain (CHD) which is responsible for IQFGAP1 binding to F-actin [197]. Following is a tandem of multiple IQGAP coiled-coil repeats (IRs) that cause dimerization of proteins containing them [198]. Next, is a WW motif, a small protein-protein interaction module responsible of the binding of IQGAP1 to ERK kinases.

In this case, is interesting to note that the binding via the WW motif differs from the FXFP motif that is usually recognized as binding site by ERK kinases [191]. Further in sequence there are four IQ motifs, responsible of the binding to calmodulin, MEK kinases and myosin [199]. Following the IQ motifs is the GTPase activating protein related domain (GRD), that is involved in the binding to the Rho GTPases CDC42 and RAC1. The last domain is a RasGAP C-terminal domain (RGCT), that is responsible for the binding to β-catenin, E-cadherin and CLIP-170 [200]. The acronym for these proteins is derived from the presence of the IQ motifs and the GAP related domains. Despite their family name and the structural similarity to GTPse activating proteins (GAPs), there is no evidence of GAP catalytic activity for IQGAP proteins [200].

IQGAP play fundamental role in cytoskeletal architecture. By binding to F-actin, the microfilament formed out of actine polymerization, IQGAP1 promotes cross-linking of actin filaments [197, 198] and also stimulates and regulates the assembly of branched actin filaments [201, 202]. Both of these processes are part of the dynamic organization that that orchestrate important mechanical functions, including cell motility and adhesion.

Regarding its role in cellular signaling, IQGAP1 is known to be a scaffold of the MAPK/ERK cascade [203]. IQGAP1 binds directly to both MEK1/2 and ERK1/2 modulating their activation and response to EGF and CD44 and localizing them to the plasma membrane [166, 179]. It has been shown that IQGAP1 binds to B-RAF [204] and that also integrates Ca2+/calmodulin and B-RAF signaling, uncovering a novel mechanism that links Ca2+ to the MAPK/ERK pathway [205] and the pathway to the cytoskeletal dynamics.

Binding of IQGAP1 to Ca2+/calmodulin, an intermediate messenger protein that transduces calcium signals, is regulated by Ca2+; and this interaction has been reported to produce conformational changes in IQGAP1 that abrogate the ability of the scaffold to bind to any other partner [198, 206–208]. In this way, Ca2+/calmodulin functions as a master regulator of IQGAP1 binding. For example, the binding of Ca2+/calmodulin to IQGAP1 positively regulates cell-cell adhesion by inhibiting the interaction of the scaffold with β-catenin and E-cadherin.

The Rho GTPases CDC42 and RAC1 are among the best characterized binding partners of IQGAP1. The Rho GTPases is a family of small G proteins that regulate intracellular actin dynamics and are involved in cellular functions such as cell morphology and cell motility [200, 209]. IQGAP1 binds to CDC42 and RAC1 only in their GTP-bound state (active state) and stabilizes them by inhibiting their GTPase activity [210]. CDC42 and RAC1 function as positive regulators of cell-cell adhesion processes, which use IQGAP1 as an effector of for the regulation mechanism [211, 212]. The binding of these Rho GTPases to IQGAP1 inhibit the interaction of the scaffold with β-catenins, a process that leads to the weakening of cell-cell attachments [197, 213].

β-catenin plays a key regulatory roles in cell-cell adhesion processes by linking the cytoplasmic domain of E-cadherin to α-catenin, which in turn binds to the actin filaments of the cytoskeleton [207]. When not bound to Rho GTPases or Ca2+/calmodulin, IQGAP1 is able to interact with β-catenin and dissociate the E-cadherin—β-catening complex, thereby weakening cell-cell attachment [214, 215] (see Figure 5.3). In this way, IQGAP1 negatively regulates cell-cell adhesion and contributes to cell motility and invasion [208].

Links between the plus-ends of microtubules and cortical regions are essential for cell

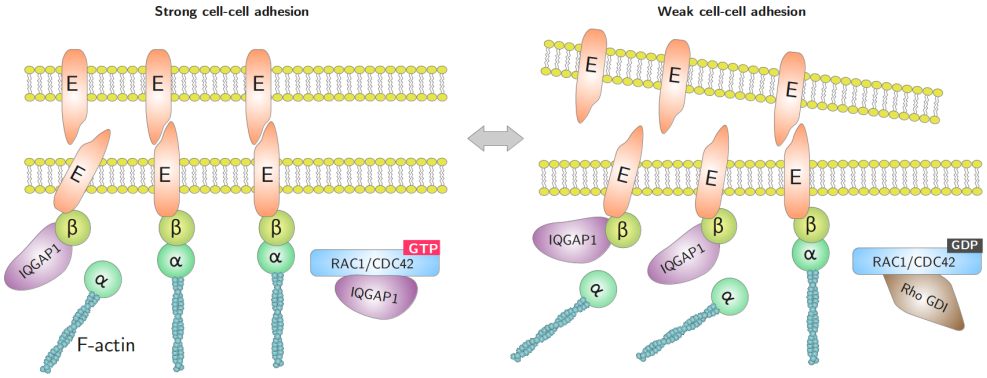


Figure 5.3: Role of IQGAP1 in the regulation of cell-cell adhesion.

At sites of cell-cell contact exists a dynamic equilibrium between the complexes E-cadherin—β-catenin—α-catenin and E-cadherin—β-catenin—IQGAP1. The ratio between these complexes determine the strength of E-cadherin-mediated cell-cell adhesion. RAC1/CDC42 and IQGAP1 can serve as positive and negative regulators of cadherin activity, respectively. E, E-cadherin; α, α-catenin; β, β-catenin.

polarity and migration. IQGAP1 has been shown to interact with CLIP-170 [216], a protein that binds to the plus-end of microtubules, regulates their dynamics and is required for recruiting microtubules at cortical regions of the cell [217, 218]. It has been suggested that IQGAP1—CLIP-170 interaction is enhanced by activated CDC42/RAC1 [216] and that RAC1/CDC42 marks cortical spots to which the IQGAP1-CLIP-170 complex is targeted, leading to formation of polarized microtubule arrays and to further cell polarization [213]. In this manner, IQGAP1 is involved in the regulation of the dynamics and in the capture of growing microtubules at the leading edge of migrating cells, which results in cell polarization and directional migration [216, 218].

IQGAP1 plays crucial roles in several cellular processes such as signaling, via the MAPK/ERK cascade, and in cell-cell adhesion, cell polarization and directional cell migration by linking Rho-family GTPases with the actin cytoskeleton and microtubules. Many of these processes have been related to neoplasia, tumor progression and metastasis [219, 220]. It is therefore not surprising that recent evidence links IQGAP1 expression [221–223] and localization [224] to neoplasia. For an excellent review on this topic the reader can refer to White *et al.* 2009 [166].

5.1.2 Adaptors and scaffolds contribute to kinase substrate specificity

Efficient signals transmission from the plasma membrane to specific intracellular sites is an essential process in living organisms. Reversible phosphorylation of proteins is an important element of internal cellular communication however, many protein kinases and protein phosphatases have relatively broad substrate specificities and may be used in varying combinations to achieve distinct biological responses. Evidence collected over the past 15 years clearly shows that specificity of signal transduction events can be modulated at the molecular level by scaffold and/or adaptor proteins, which position signaling enzymes at proper cellu-

lar localization [170, 225–227]. This allows their efficient catalytic activation and accurate substrate selection. For the intracellular second-messenger, cyclic AMP-dependent protein kinase A (PKA), the effect of cellular compartmentalization on signaling specificity has been widely studied. In the following section we present examples of how PKA kinases achieve substrate specificity aided by the A-kinase anchoring proteins (AKAP) family of scaffolds.

5.1.2.1 The biological roles of A-kinase anchoring proteins

Protein kinase A (PKA), also known as cAMP-dependent protein kinase, is a family of Ser/Thr kinases of broad specificity whose activity is regulated by the cellular levels of cyclic AMP (cAMP). PKAs are involved in several cellular functions including regulation of metabolism [53], learning and memory [52] and exocytosis [54]. The activity of PKAs is regulated by two regulatory (R) subunits, which form a dimer that binds two catalytic (C) subunits, forming a tetrameric holoenzyme. There are four isoforms of the regulatory subunit (RIα, RIβ, RIIα, RIIβ) and three types of catalytic subunits (Cα, Cβ, Cγ), each of which display different patterns of cellular localization and tissue expression [228, 229]. Two main PKA subtypes (I and II) are defined by the identity of their regulatory subunits, RI and RII. Binding of cAMP to the regulatory dimers cause the dissociation and further activation of the catalytic subunits [56, 57] (see Figure 5.4).

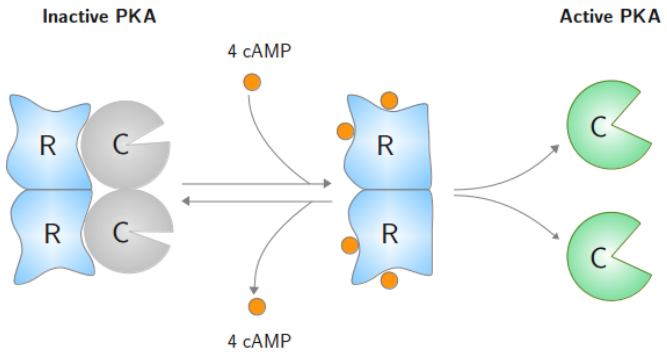


Figure 5.4: cAMP-dependent protein kinase (PKA) activation.

PKA is a tetramer composed of two catalytic subunits and two regulatory subunits, that repress the catalytic units when they are bound together. The binding of cAMP to the regulatory subunits release the catalytic subunits, releasing also their inhibition.

PKAs are targeted to discrete subcellular environments by AKAPs, a diverse family of scaffold proteins [55, 230]. Recruiting a signaling enzyme to a specific subcellular location not only ensures that the enzyme is near to its relevant targets, but also can prevent indiscriminate phosphorylation of other proteins. AKAPs not only target PKAs to subcellular locations, but at the same time they can also bind to other signaling molecules (*e.g.*, protein phosphatases (PP2B/calcineurin) [231], protein kinase C (PKC) [232] and phosphodiesterase PDE4D3 [233] to form multi-protein complexes (transduceosomes) that integrate cAMP signaling with other pathways [227, 234]. Next, we provide examples of different AKAPs and their

roles in PKA localization and cellular signaling integration.

5.1.2.2 mAKAP assembles a signalosome at the nuclear envelope

In striated myocytes and neurons, the muscle-selective anchoring protein (mAKAP) organizes several proteins to control signaling events that occur close to the nuclear membrane. There, mAKAP functions as a scaffold for assembling a signalosome that is responsive to cAMP, Ca2+, and MAP kinase signaling. Besides interacting with PKA, mAKAP binds also to other signaling proteins such as the Rap guanine nucleotide exchange factor (Epac1) [233], the phosphodiesterase PDE4D3 [235], the MAP kinase ERK5 [233], the calcium release channel ryanodine receptor (RyR) [236] and the phosphatase PP2B (calcineurin) [237] (see Figure 5.5). An increased local concentration of cAMP activates PKA, which in turn phosphorylates and activates the mAKAP–associated PDE4D3, to enhance the reduction of cAMP concentration [238]. In this complex scenario, PDE4D3 also functions as an adaptor that recruits Epac1 to enable the cAMP-dependent downregulation of PDE4D3-associated ERK5 [239]. When associated to mAKAP, calcineurin dephosphorylates and therefore inactivates ERK5 [240], while ERK5 contributes in this context to the suppression of PDE4D3 activity [233]. As stated previously, the mAKAP signalosome also includes the channel RyR, which is a major cardiac ion channel responsible for calcium release from intracellular stores [236]. At the nuclear envelope, a subset of RyR is bound to mAKAP and via this association RyR can be regulated by PKA-mediated phosphorylation [239].

The signalosome assembled by the scaffold mAKAP constitutes a a highly structured network in which pathways such as the ones of cAMP, calcium and mitogens can be integrated and regulated by bringing relevant signaling molecules into close proximity. This macromolecular complex is has been shown to be involved in the regulation of key processes such as cardiac contractility [239, 241].

5.1.3 Computational identification of signaling scaffold proteins

Scaffolds are extremely diverse proteins which structure and function may significantly overlap with those of other protein classes and they also lack common sequence signature motifs, similar to the ones found in enzymes. Therefore, the identification of scaffolds based only on sequence is currently not possible. However, scaffolds often contain frequently occurring PPI domains or motifs (*e.g.*, SH2, SH3, PD, WW) that can be easily identified from sequence analysis. The presence of multiple PPI elements on their sequence allow the scaffolds to interact at the same time with more than one protein, and therefore to assemble macromolecular complexes in a modular way [194]. In some cases, these protein complexes can be assembled in combinatorial process, where the biological function of the resulting complex will depend on the associated elements [242].

A common feature of signaling scaffolds is that they usually interact with at least two signaling proteins that are involved in the transduction of information in pathways [160, 165]. Based on this property, Zeke *et al.* proposed a general definition for the identification of scaffold proteins in interactomes [177]. They suggested that signaling scaffolds can be defined as proteins that: i) lack intrinsic catalytic activity relevant for signaling; ii) have at

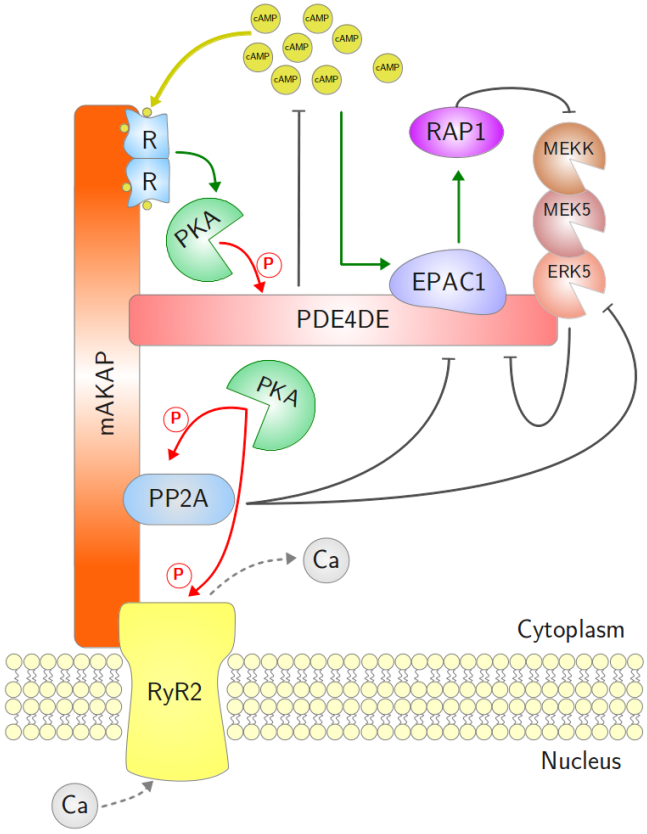


Figure 5.5: mAKAP assembles a signalosome at the nuclear membrane.

Increased levels of cAMP activates mAKAP-anchored PKA at submicromolar concentrations. Activated PKA phosphorylates the phosphodiesterase PDE4D3, leading to its activation which increases cAMP degradation, creating a classic negative feedback loop for PKA. Anchored PKA also phosphorylates and regulates the activity of phosphatase PP2A, which promotes deactivation of PDE4D3 and its associated ERK5 kinase. High concentrations of cAMP also stimulate Epac1, which exerts an inhibitory effect on the MEK5-ERK5 pathway trough the activation of the GTPase Ras-related protein 1 (Rap1). In turn, ERK5 is able to promote PDE4D3 deactivation by phosphorylation. PKA also activates the mAKAP-associated RyR2 receptor, which enhances Ca2+ mobilization from intracellular stores. By organizing such a macromolecular complex, mAKAP plays a prominent role in the regulation of cardiac function.

least two binding partners with catalytic activity relevant for signaling and iii) have binding partners that interact with each other in a direct or indirect way. Based on the former definition, Ramirez and Albrecht conducted an identification of potential human signaling scaffolds proteins (PSPs) using an integrated dataset of human PPIs [243]. The authors identified a total of 250 PSPs, which were enriched in Pfam domains that mediate PPIs and also on Gene Ontology terms related to protein binding, protein localization and cellular component organization and formation.

As commented previously on this chapter, scaffold proteins were initially thought to only provide tethering to other signaling proteins. However, it has been systematically shown that

some scaffolds retain catalytic activity which could be involved in the active regulation of the pathways they are involved in [170]. Therefore, although it serves the simple purpose of separating two basic functions in signal transduction — enzymes as main effectors and scaffolds as temporal-spatial coordinators [172] — it is clear that the exclusion of proteins with enzymatic activity as potential scaffolds constitutes a significant limitation.

5.2 Materials and methods

5.2.1 Protein-protein interactions for human

For identifying reliable binding partners for human protein kinases and their substrates we have used the in-house database PPI-DB. This database integrates the data available from several public protein-interaction resources such as Intact [244], MINT [245], DIP [246] and HPRD [109] among others. PPI-DB contains almost 45'000 high confidence binary interactions experimentally determined in human.

5.2.2 Statistical enrichment of Pfam families

We have used Pfam (ver. 24.0) [247] as the resource for the annotation of proteins based on the functional domains they contain. For computing the enrichment of proteins sets on particular functional domains we used an hypergeometric test and we have corrected for multiple testing using both Bonferroni's and Benjamini-Hochberg's methods [248]. We have defined a statistical significance threshold $\alpha < 0.001$. As the background set we have used the human proteome, defined from canonical Swiss-Prot entries in the UniProt database (ver 2011_11) [112] with evidence of existence at either protein level, transcript level or inferred from homology.

5.2.3 Statistical enrichment of Gene Ontology terms

We have used the Gene Ontology database [249] to functionally annotate proteins in our analysis based their associated biological processes, cellular components and molecular functions. For conducting a statistical enrichment analysis of GO terms in our proteins sets we have used the `R` package `GOstats` (ver. 2.22.0) [250]. We have corrected for multiple testing by using Bonferroni's method and we have defined a statistical significance threshold $\alpha < 0.05$. As the background we defined a set of 4656 genes comprising all PPI partners and substrates of human protein kinases.

5.2.4 Automatic collection of known adaptor/scaffold proteins

As the source for the automatic collection of known adaptor/scaffold proteins of human protein kinases we used the Swiss-Prot database [112]. We selected human proteins containing at least one of the terms adaptor and/or scaffold in the Function field of the entry. We then filtered the resulting list of proteins keeping only those that are known to have a binary interaction with at least one human protein kinase. We denote this set as known adaptor/scaffold (kAS) proteins.

5.2.5 Gold standard set of kinase–adaptor/scaffold pairs

In order to compile a 'gold standard set' (GSS) of kinase–adaptor/scaffold associations, we filtered the pairs collected in section 5.2.4. As the filtering criteria we used i) the experiment type used for detecting the interaction and ii) the evidence in the literature supporting a

biological role for the kinase–adaptor/scaffold complex. As for the first criteria, we kept only those kinase–adaptor/scaffold pairs whose interactions have been experimentally detected using *in vivo* conditions. For the second criteria, we reviewed the relevant literature and kept only those pairs where the adaptor/scaffold have been shown to provide that function for its associated kinase.

5.2.6 Identification of potential adaptors/scaffolds from substrates partners

The purpose of this method is the identification of proteins (P_i) in the human interactome that interacts with a large fraction (statistically significant) of the substrates of a given kinase (K_k). Proteins that met this criteria were classified as potential adaptors/scaffolds (pAS) of K_k. The classification was based on a statistical test, where we used as the statistic the number of substrates (of a given kinase) interacting with P_i. Here, we used a strategy of PPI network randomizations to generate a series of backgrounds distributions of the statistic. These backgrounds distributions were used later to estimate the likeliness of a protein to interact with a large fraction of a kinase substrates. We conducted the current analysis only for kinases for which we known at least five substrates, and we prevented for the possibility of a kinase being classified as its own adaptor/scaffold.

5.2.6.1 Data

For this analysis we used PPI-DB as the source for human binary interactions and we used the kinase-substrates associations in our database SBNB_PhoshpoDB.

5.2.6.2 Collection of PIN for randomization

We collected a first-level sub-network of the human interactome by using as seeds the substrates of kinases having at least five substrates in SBNB_PhoshpoDB. The total number of seeds used was 1807, corresponding to 156 kinases. The final sub-network contains 6528 proteins and 17633 unique interactions.

5.2.6.3 Generating background distributions

The probability of any P_i to be classified as a pAS for a given K_k, might largely depend on the number of substrates known for K_k. In order to conduct fair evaluations, we generated several sets of background distributions of the statistic according to the different number of substrates (S) known for the kinases in our data set. For generating the background sets we proceeded as follows. We started from the human sub-network previously mentioned and we randomly rewired it by randomly switching edges. Then, we randomly select a node (K) in the network having at least S partners (excluding self-interactions), and for a number S of these partners we identified their first neighbors (P_i) in the randomized network. Finally, for each P_i we counted with how many of the S partners it interacts. As mentioned earlier, these counts represent the statistic used for the background distributions. We repeated this procedure for each of the distributions (according to the different number of substrates) until 10'000 values of the statistic were stored.

5.2.6.4 Testing the classification method

We conducted a Fisher's exact test (right tale) for assessing whether the known adaptors/scaffolds are more likely to interact with a significantly large fraction of kinases substrates than any random protein (P_i) in the human interactome. For the test we used two categories, i) whether P_i is a known adaptor/scaffold protein and ii) whether P_i is known to interact with a significantly large fraction of its kinases substrates. For the first category we used the set of known adaptors/scaffolds collected in section 5.2.4; while for the second category we have assessed the statistical significance by computing empirical *p*-values based on the background distributions generated in section **??**.

5.2.6.5 Algorithm for the identification of potential adaptors/scaffolds

Here, we included only those kinases for which we known at least five substrates (both *in vivo* and *in vitro*). Then, for every protein (P_i) that interacts with at least one of the substrates of kinase K_k, we estimated its probability of being a potential adaptor/scaffold for K_k. This probability was based on the computation of an empirical *p*-value from the background distributions corresponding to the number of substrates of K_k (see section **??**). If P_i is found to interact with a significantly large fraction of the substrates of K_k, then, P_i is classified as a potential adaptor/scaffold for K_k. The statistical significance threshold was set to $\alpha < 0.01$ and we corrected for multiple testing using the Benjamini-Hochberg's method.

5.2.6.6 Evaluating the performance of the method

We evaluated the performance of the method based on its precision, its recall and its F1 score. The F1 score can be interpreted as a weighted average of the precision and the recall, and is a measure of the method's accuracy. For the computation of these parameters we used as reference the set of known adaptor/scaffold proteins identified in section 5.2.4. The evaluation of performance of the method was done for every kinase included in the analysis. This is, for each kinase with at least one pAS and one kAS we counted how many of its kASs are re-identified by our strategy (*i.e.*, the recall) and we also computed the ratio of the kASs re-identified over the total of proteins classified as pASs (*i.e.*, the precision). The F1 score was computed as shown in Equation 5.1.

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall} \tag{5.1}$$

Equation 5.1 Computing the F1 score.

5.2.6.7 Co-annotation of substrates and potential adaptors/scaffolds

Here, we first computed the enrichments in cellular compartment (CC) terms of the Gene Ontology (GO) for the set of substrates of each kinase. Later, for each kinase, we identified those potential adaptors/scaffolds that were annotated with the same CC terms found to be enriched in the set of substrates of the corresponding kinase.

We used the `R` package `GOstats` for both the functional annotation of substrates and potential adaptors/scaffolds and for the computation of the enrichments. As the background we used a set of 4656 genes comprising all PPIs and substrates of human protein kinases and we used a statistical significance threshold of $\alpha < 0.05$.

5.2.7 Identification of known adaptors/scaffolds interacting with significantly large sets of substrates

Here we have focused on the identification of known adaptors/scaffolds that interact with a significantly large fraction of the substrates of a given kinase. In this analysis, we considered only the pairs of kinase-adaptor/scaffold present in the gold standard set, excluding those pairs for which we have less than five *in vivo* substrates for the kinase. For estimating the statistical significance, we have computed an empirical *p*-value using as the statistic the number of substrates (of a given kinase) that interacts with a given adaptor/scaffold of the same kinase. Similar to previous analysis, each empirical *p*-value is computed from a background distribution that takes into account the cardinality of the set of *in vivo* substrates available for the kinase.

5.2.7.1 Data

Here we have used the gold standard set of kinase-kAS associations. We have considered only kinase-kAS pairs for which we have at least five *in vivo* substrates for the kinase. In total, we count with 51 kinase-kAS pairs comprising 31 kinases and 36 known adaptor/scaffold proteins. For compiling a kinase-related subnetwork of the human interactome, we used as seeds all human kinases for which we have at least five *in vivo* substrates (111 kinases in total), plus their corresponding substrates. Later, using the human interactome in PPI-DB, we retrieved the PPI neighbors of these seeds up to the second level. The final subnetwork is composed of 5939 unique proteins and contains a total of 15'046 unique PPIs.

5.2.7.2 Generating background distributions

Here we describe the algorithm followed for generating the background distributions used in the current analysis. These distributions will be used for estimating the statistical significance of the number of substrates of a given kinase that interact with a given adaptor/scaffold of the same kinase. Several distributions have been generated, in accordance to the cardinality of the sets of substrates in our data. For this procedure, we have used the previously described kinase-related subnetwork of the human interactome.

For generating the random distribution for a set of substrates of cardinality S, the algorithm proceeds as follows. The first step is the random selection of a (hypothetical) 'kinase' node from the subnetwork having at least $S+1$ neighbors (excluding self-interactions). Next, the neighbors are randomly split into the groups 'substrates' — of cardinality S — and 'adaptors/scaffolds' which contain the remaining neighbors of the 'kinase' node. Then, for each 'adaptor/scaffold', we count with how many of the 'substrates' it does interact. These counts are the statistics and therefore they are kept in the background distribution. At

last, the subnetwork is randomized by randomly assigning edges between the nodes. The complete procedure is repeated until 10'000 values of the statistic have been collected for each distribution. This algorithm was implemented in `R` statistical environment using the library `igraph` [251].

5.2.7.3 Estimating the statistical significance

We use an empirical *p*-value to estimate the statistical significance of the number of substrates of a given kinase that interact with an adaptor/scaffold of the same kinase. The *p*-value is computed based on the background distributions described in section 5.2.7.2 and we set a statistical significance threshold $\alpha < 0.05$. We conducted this test for each of the 51 kinase-kAS pairs described in section 5.2.7.1.

5.2.7.4 Statistical likeliness substrate-adaptor/scaffold interactions

We need to test the hypothesis of whether the known adaptors/scaffolds in the gold standard set are more likely to interact with a significantly large fraction of the substrates of their corresponding kinases than any random protein (P_i) of the human interactome. For this, we conducted a right tale Fisher's exact test using as categories i) whether P_i is an adaptor/scaffold in the gold standard set and ii) whether P_i interacts with a significantly large fraction of the substrates. The statistical significance used in category ii) was based on the empirical *p*-values computed in section 5.2.7.3.

5.2.8 Identification of kinases sharing significantly large sets of substrates

Here we approached the identification of kinases sharing at least one known adaptor/scaffold and also sharing a significantly large number of substrates. The number of substrates shared by the kinases (*i.e.*, the overlap) was used as the statistic. For estimating the statistical significance of this overlap, we computed an empirical *p*-value based on background distributions. These background distributions were generated by randomly sampling sets of kinases of different cardinality [a] and later computing the overlap between their corresponding sets of substrates. The backgrounds distributions are cardinality-specific and are composed of 10'000 values of the statistic. Only kinases having at least five *in vivo* substrates were considered in this analysis.

5.2.8.1 Data

We have considered only kinases for which we know at least five *in vivo* substrates (111 kinases in total). Regarding the known adaptors/scaffolds, we have used both the set of automatically collected ones (see section 5.2.4) and the manually revised gold standard set (see section 5.2.5).

[a] The cardinality of a set is a measure of the number of elements of the set. In this case, it refers to the number of kinases.

5.2.8.2 Generating background distributions

Here, we first determined the cardinalities of kinase sets for which the background distributions must be generated. For this, we computed the number of kinases shared by each pair of known adaptors/scaffolds in our dataset. In our data, known adaptors/scaffolds are shared solely by sets of either two or three kinases. Therefore, generated background distributions corresponding to these sets cardinalities. For generating any of the background distributions, we randomly sampled sets of two or three kinases from the set of 111 kinases with at least five *in vivo* substrates. Next, we computed the overlap between their corresponding sets of substrates and we kept this value as the statistic for the background distribution. Each distribution contains 1000 values of the statistic.

5.2.8.3 Estimating the statistical significance

We use an empirical *p*-value to estimate the statistical significance of the number of substrates shared by a set of kinases. The *p*-value is computed based on the background distributions described before and we set a statistical significance threshold $\alpha < 0.05$. As stated previously, kinases in the set must share at least one known adaptor/scaffold protein. We conducted separate analysis for each of the two sets of known adaptors/scaffolds (*i.e.*, the automatically collected kAS and the gold standard set).

5.3 Results and discussion

5.3.1 Set of known adaptors/scaffolds collected for human kinases

We have previously described a strategy for the automatic collection of known adaptor/scaffold proteins of human protein kinases (see section 5.2.4). By using this strategy, we compiled a list of 191 proteins comprising 109 adaptor and 82 scaffold proteins (see the Appendices section A3 for the complete set of results). We refer to this set as the known adaptors/scaffolds (kAS). The 191 kAS proteins associate with 287 (55.4%) human kinases via 1281 binary PPIs. In total, 94 (72.3%) kinase families are represented by the aforementioned 287 human kinases. These findings suggest that the association with adaptors/scaffolds is a probably a widespread mechanism among human kinases, comprising a large fraction of the kinase families and all nine major kinase groups. As commented previously, the association of kinases to adaptors/scaffolds can affect the function of these enzymes in aspects such as substrate specificity, allosteric modulation and sub-cellular localization.

Following the procedure described in section 5.2.2, we have analyzed the enrichment of Pfam domain families among the set of kAS proteins. We have found 23 domains to be significantly enriched among proteins in the set, from which 14/23 (60.8%) are known to be directly involved in promoting PPIs (see Table 5.1). The over-representation of PPI–promoting domains such as PDZ, SH2 and SH3 supports the idea of a role as adaptors/scaffolds for the proteins in the set. Moreover, for the kAS set we have also computed the enrichment in terms of the Molecular Function (MF) category of the Gene Ontology database. Our results show that 100% of the terms found to be enriched in the set are related to protein binding, adaptor or scaffolding functions (see Table 5.2). These results further support the biological role as adaptors/scaffolds of kAS proteins.

In order to contrast our results, we have compared them to the ones obtained by Ramírez and Albrecht [243], who developed a computational strategy and identified a total of 250 potential signaling scaffold proteins in the human interactome. The overlap between our set and the one from Ramírez and Albrecht is of 67 proteins, corresponding to 35% and almost 30% of each set respectively. Regarding the enrichment in Pfam families, our set contains a considerable fraction (9/14, 64.3%) of the domains found to be enriched among the 250 potential scaffolds by Ramírez and Albrecht. These nine common Pfam families are known to be involved in promoting PPIs (see Table 5.1). The kAS set is also enriched in other domains such as pleckstrin homology domain (PH), which is known to recruit proteins to regions of the plasma membrane where signalosomes are assembled upon extracellular activation. When comparing the two proteins sets based on their enrichments on MF annotations, we find that terms related to protein binding and scaffolding are overrepresented in both sets. However, for the kAS set, we identified two times more MF enriched terms; all of them related to either adaptor, scaffolding, or protein binding functions (see Table 5.2).

We consider that, compared to ours, the larger number of scaffolds identified by Ramírez and Albrecht can be explained by the fact that they allow for indirect interactions between the potential scaffold and signaling molecules. In contrast, have applied a more stringent selection criteria by including in our results only those proteins with evidence of adaptor/scaffolding function and that also interacts in a binary mode with at least one ki-

Table 5.1: Pfam families enriched among kAS proteins.

Pfam ID	Pfam family name	Proteins	Domain instances	Enrichment ratio	Raw *p*-value	Adjusted *p*-value	References
14-3-3	**14-3-3 protein**	7	7	6.42	$6.94e^{-13}$	$1.65e^{-11}$	A
Adaptin_N	Adaptin N terminal region	4	4	4.33	$4.58e^{-05}$	$3.46e^{-04}$	B
AMPKBI	AMP kinase beta subunit, interation domain	2	2	6.83	$1.02e^{-04}$	$6.29e^{-04}$	C
BACK	BTB And C-terminal Kelch	5	5	3.38	$1.23e^{-04}$	$7.27e^{-04}$	
BPS	**BPS (Between PH and SH2)**	3	3	7	$7.84e^{-07}$	$8.89e^{-06}$	
Caveolin	**Caveolin**	3	3	6.42	$3.89e^{-06}$	$3.52e^{-05}$	D
Clat_adaptor_s	Clathrin adaptor complex Small chain	4	4	4.72	$1.53e^{-05}$	$1.23e^{-04}$	E
Cullin	Cullin family	4	4	5.42	$1.95e^{-06}$	$2.04e^{-05}$	F
Cullin_Nedd8	Cullin protein neddylation domain	4	4	6.1	$2.32e^{-07}$	$2.86e^{-06}$	
EBP50_C-term	EBP50, C-terminal	2	2	6.83	$1.02e^{-04}$	$6.29e^{-04}$	
GKAP	Guanylate-kinase-associated protein (GKAP) protein	4	4	6.25	$1.40e^{-07}$	$2.11e^{-06}$	
Guanylate_kin	Guanylate kinase	5	5	4.04	$1.35e^{-05}$	$1.15e^{-04}$	
IBB	Importin beta binding domain	3	3	5.3	$5.39e^{-05}$	$3.86e^{-04}$	
IRS	**Phosphotyrosine-binding domain (IRS-1 type)**	8	8	6.33	$2.74e^{-14}$	$9.33e^{-13}$	G
PB1	PB1 domain	5	5	5.28	$1.58e^{-07}$	$2.14e^{-06}$	
PDZ	**PDZ domain**	24	70	4.1	$1.29e^{-22}$	$1.75e^{-20}$	H
PH	Pleckstrin homology domain	16	17	3.23	$2.14e^{-11}$	$4.15e^{-10}$	I
Phe_ZIP	**Phenylalanine zipper**	2	2	6.83	$1.02e^{-04}$	$6.29e^{-04}$	J
PID	**Phosphotyrosine interaction domain (PTB/PID)**	7	8	4.55	$1.95e^{-08}$	$3.32e^{-07}$	K
SAPS	SIT4 phosphatase-associated protein	3	6	6.42	$3.89e^{-06}$	$3.52e^{-05}$	
SH2	**SH2 domain**	21	23	4.45	$3.37e^{-22}$	$2.29e^{-20}$	L
SH3_1	**SH3 domain**	18	36	3.61	$1.22e^{-14}$	$5.52e^{-13}$	M
SH3_2	Variant SH3 domain	13	15	4.12	$7.30e^{-13}$	$1.65e^{-11}$	N

In bold, the Pfam families found to be enriched among the potential scaffolds identified by Ramírez and Albrecht [243]. Last column contain references to the publications supporting the role of each domain in promoting PPIs. **A)** [252, 253], **B)** [254, 255], **C** [256], **D** [257, 258], **E** [259, 260], **F** [261, 262], **G** [263, 264], **H** [173, 265, 266], **I** [160, 267], **J** [268], **K** [269, 270], **L** [271, 272], **M** [273–275], **N** [273–276].

Table 5.2: Molecular function terms enriched among kAS proteins.

GO Id	GO description	Enrichment ratio	Raw *p*-value	Adjusted *p*-value
GO:0005078	MAP-kinase scaffold activity	328.9	$1.09e^{-07}$	$4.21e^{-05}$
GO:0005168	neurotrophin TRKA receptor binding	328.9	$1.09e^{-07}$	$4.21e^{-05}$
GO:0005068	transmembrane receptor protein tyrosine kinase adaptor activity	109.62	$7.49e^{-07}$	$2.89e^{-04}$
GO:0005165	neurotrophin receptor binding	109.62	$7.49e^{-07}$	$2.89e^{-04}$
GO:0060090	**binding, bridging**	95.68	$3.14e^{-08}$	$1.21e^{-05}$
GO:0005158	insulin receptor binding	52.26	$2.02e^{-14}$	$7.79e^{-12}$
GO:0005159	insulin-like growth factor receptor binding	45.92	$4.84e^{-07}$	$1.87e^{-04}$
GO:0030159	receptor signaling complex scaffold activity	41.55	$5.19e^{-08}$	$2.00e^{-05}$
GO:0005070	SH3/SH2 adaptor activity	37.49	$1.88e^{-17}$	$7.26e^{-15}$
GO:0032947	protein complex scaffold	31.26	$2.04e^{-06}$	$7.87e^{-04}$
GO:0051219	phosphoprotein binding	18.65	$5.39e^{-08}$	$2.08e^{-05}$
GO:0005080	protein kinase C binding	16.69	$7.05e^{-07}$	$2.72e^{-04}$
GO:0031267	small GTPase binding	9.24	$3.60e^{-08}$	$1.39e^{-05}$
GO:0005515	protein binding	8.07	$1.72e^{-13}$	$6.63e^{-11}$
GO:0019904	**protein domain specific binding**	6.01	$2.80e^{-08}$	$1.08e^{-05}$
GO:0019899	**enzyme binding**	5.82	$5.07e^{-16}$	$1.96e^{-13}$

In bold, the terms that were also found to be enriched among the potential human scaffolds identified by Ramírez and Albrecht. Enrichments were computed following the procedure described in section 5.2.3, but this time using $\alpha < 0.001$ and as the background set the human proteome as described in section 5.2.2.

nase. Regarding the differences in the Pfam families and the MF terms enriched in each set, we consider that the differences arise due to the protein composition of each set, as well as due to the the background sets used for computing the enrichments.

The set of 191 kAS proteins will be used for assessing the performance of a computational strategy, which we have developed for identifying potential adaptors/scaffolds of human kinases.

5.3.2 Gold standard set of known adaptors/scaffolds

We have previously described a set of kinase-kAS associations automatically collected; and we have used this set for evaluating the performance of a computational strategy for the identification of potential adaptors/scaffolds (see 5.2.6). However, we are aware that our automatically collected set of kinase-kAS pairs may contain several false positives. In fact, evidence of a binary interaction between a kAS and a kinase does not necessarily implies an adaptor/scaffolding function of the former towards the latter. Therefore, we decided to compile a gold standard set (GSS) of manually curated kinase-kAS pairs for which *in vivo* PPIs and adaptor/scaffolding functions have been explicitly reported in the literature (see 5.2.5). The GSS will serve as a reference for testing our working hypothesis, which assume that the specificity of kinases is, at least partially, promoted by the association of the kinases to specific adaptor/scaffold proteins.

The GSS here collected is composed of 75 kinase-kAS pairs comprising 46 kAS proteins and 47 kinases (see Figure 5.6). Among the 47 kinases we find 28 tyrosine kinases, 16 serine/threonine kinases and 3 dual specificity kinases. Within the GSS we have identified a subset kinase-kAS pairs for which we known at least five *in vivo* substrates for the corresponding kinase. This subset, comprising 49 kinase-kAS pairs, 31 kinases and 36 kAS proteins, will be used in analysis that will be discussed further in this document.

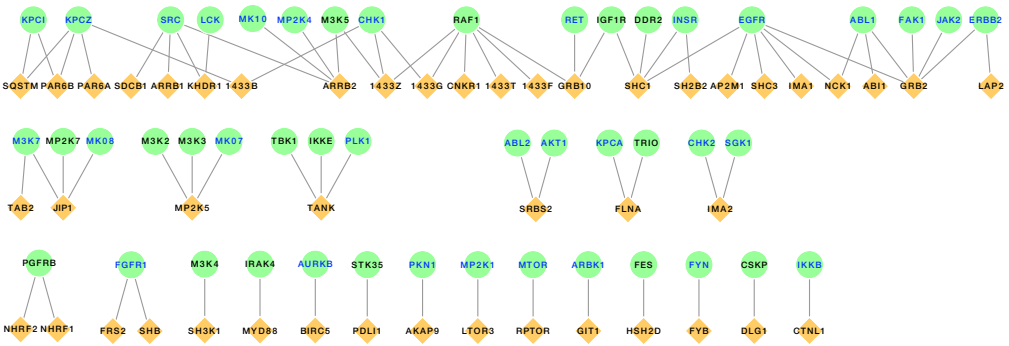


Figure 5.6: Gold standard set of kinase-kAS pairs.

Kinases are represented by circles and adaptors/scaffolds are represented by diamonds. Edges represent binary PPIs. Kinases with at least five *in vivo* substrates are labeled in blue. All labels correspond to UniProt Ids.

5.3.3 Potential adaptors/scaffolds identified from substrates partners

Here we present the results of a computational strategy for the identification of potential adaptors/scaffolds (pAS) of human protein kinases. In this strategy we ran under the assumption that, in contrast to any random protein in the human interactome, an adaptor/scaffold interacts with a significantly large fraction of the substrates of a given kinase. In this way, the adaptor/scaffold would function as a platform to which the kinase could bind and therefore gain spatial proximity to its relevant set of substrates. First, we tested whether our assumption holds for the set of kinase-known adaptor/scaffold (PK-kAS) pairs previously collected. For this, we conducted a Fisher's exact test (see section 5.2.6.4) using only those PK-kAS pairs for which we known at least five substrates for the kinase (comprising 156 kinases in total). The result of the test suggests that kAS proteins are five times more likely to interact with a large fraction of the substrates of their corresponding kinases than any random kinase partner (p-value $= 1.08e^{-15}$, odds ratio $= 5.04$). Given that this result supports our classification criteria, we therefore proceeded to the identification of pAS proteins for human protein kinases in the human interactome. Figure 5.7 shows a schematic representation of the strategy, which is described in section 5.2.6 of Materials and methods.

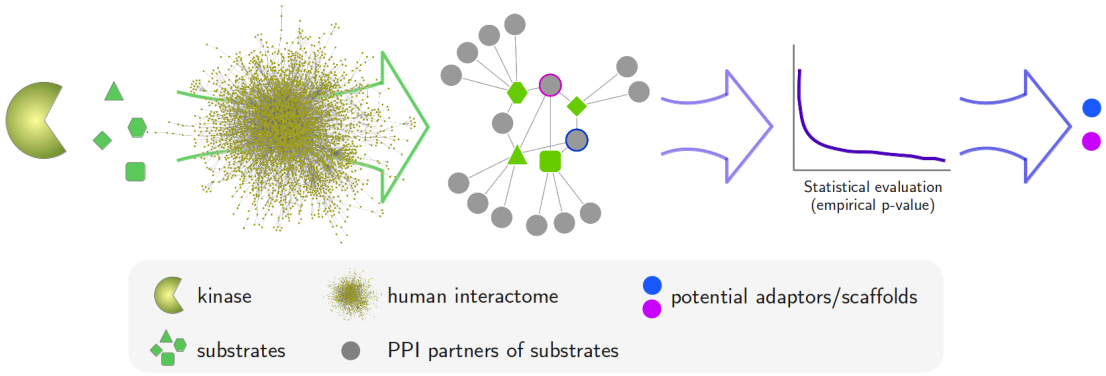


Figure 5.7: Identification of PPI partners overrepresented among kinase substrates.

The strategy classifies as potential adaptors/scaffolds those human proteins that are known to interact with a significantly large fraction of the substrates of a given kinase.

By using this strategy, we identified 706 relationships kinase–potential adaptor/scaffold (PK–pAS). In total, we classified 279 proteins as pASs, of which 71 (25.4%) are present in the kAS set. For 78 (50%) of the 156 analyzed kinases we identified at least one pAS. These 78 kinases cover 44 out of 130 (33.8%) of all kinase families. The complete set of pairs are available in Appendices (see A4). We have conducted an analysis of enrichment in Pfam domain families for the complete set of 279 pASs. The results of this analysis show enrichments in a total of 10 Pfam families, all of them known to mediate PPIs or to be present in proteins involved in cellular signaling (see Table 5.3). Five out of the ten Pfam families found to be enriched among pASs were also enriched among the set kAS set, a finding that supports the hypothesis of common biological functions for proteins in both sets.

Table 5.3: Pfam families enriched among pAS proteins.

Pfam ID	Pfam family name	Proteins	Domain instances	Enrichment ratio	Raw *p*-value	Adjusted *p*-value
14-3-3	**14-3-3**	6	6	5.65	$1.04e^{-09}$	$7.52e^{-08}$
Furin-like	Furin-like cysteine rich region	6	6	5.46	$2.74e^{-09}$	$1.58e^{-07}$
GKAP	**Guanylate-kinase-associated protein (GKAP)**	4	4	5.7	$6.35e^{-07}$	$2.29e^{-05}$
Pkinase	Protein kinase domain	23	24	2.06	$1.00e^{-08}$	$4.12e^{-07}$
Pkinase_Tyr	Tyrosine kinase	17	18	2.94	$1.23e^{-10}$	$1.18e^{-08}$
Recep_L_domain	Recep_L_domain	6	12	5.37	$4.21e^{-09}$	$2.02e^{-07}$
SH2	**SH2 domain**	15	17	3.42	$1.37e^{-11}$	$1.97e^{-09}$
SH3_1	**SH3 domain**	21	26	3.29	$6.59e^{-15}$	$1.90e^{-12}$
SH3_2	**Variant SH3 domain**	8	9	2.87	$1.51e^{-05}$	$4.36e^{-04}$
Y_phosphatase	Protein tyrosine phosphatase	7	10	3.34	$6.24e^{-06}$	$2.00e^{-04}$

In bold, the Pfam families that were also found to be enriched among the set of known adaptors/scaffolds.

To further support the role proposed for the pAS proteins, we conducted an analysis of the enrichment in Molecular Function (MF) terms. We found a total of 39 MF terms to be overrepresented in the set (see Table 5.4). A considerable fraction of the terms (24/39, 61.5%) refer to 'protein binding' functions of signaling-related molecules such as receptors, kinases, phosphatases and transcription factors. These results suggest that, taken as a set, the pAS proteins could be able to mediate PPIs for different classes of signaling-related proteins. However, in contrast with the kAS set, for the pASs we do not find enrichments in MF terms directly related to adaptor nor scaffolding functions. These results suggest that, as a set, the pAS proteins have not been previously associated to either adaptor or scaffolding functions. Therefore, we consider that our classification strategy is able identify a set of previously unknown adaptor/scaffold proteins of human kinases, whose functional annotations are in agreement with the proposed biological roles.

Table 5.4: Molecular function terms enriched among pAS proteins.

GO Id	GO description	Enrichment ratio	Raw *p*-value	Adjusted *p*-value
GO:0030235	nitric-oxide synthase regulator activity	inf	$2.03e^{-09}$	$1.37e^{-06}$
GO:0046965	retinoid X receptor binding	76.68	$4.82e^{-10}$	$3.26e^{-07}$
GO:0004861	cyclin-dependent protein kinase inhibitor activity	65.48	$1.57e^{-08}$	$1.06e^{-05}$
GO:0004716	receptor signaling protein tyrosine kinase activity	54.36	$4.74e^{-07}$	$3.21e^{-04}$
GO:0051879	Hsp90 protein binding	36.37	$1.59e^{-07}$	$1.08e^{-04}$
GO:0051059	NF-kappaB binding	33.82	$1.95e^{-09}$	$1.32e^{-06}$
GO:0042169	SH2 domain binding	26.35	$1.18e^{-10}$	$8.01e^{-08}$
GO:0050681	androgen receptor binding	22.62	$5.15e^{-11}$	$3.49e^{-08}$
GO:0001102	RNA polymerase II activating transcription factor binding	22.53	$1.74e^{-07}$	$1.18e^{-04}$
GO:0030331	estrogen receptor binding	19.15	$4.27e^{-07}$	$2.89e^{-04}$
GO:0042826	histone deacetylase binding	16.51	$2.46e^{-11}$	$1.66e^{-08}$
GO:0019903	protein phosphatase binding	15.76	$1.23e^{-09}$	$8.32e^{-07}$
GO:0004860	protein kinase inhibitor activity	15.49	$4.41e^{-08}$	$2.99e^{-05}$
GO:0008013	beta-catenin binding	15.19	$3.19e^{-10}$	$2.16e^{-07}$
GO:0035257	nuclear hormone receptor binding	15.14	$2.14e^{-12}$	$1.45e^{-09}$
GO:0002039	p53 binding	13.77	$1.05e^{-07}$	$7.08e^{-05}$
GO:0005057	receptor signaling protein activity	10.15	$1.11e^{-09}$	$7.50e^{-07}$
GO:0019901	protein kinase binding	10.1	$8.59e^{-20}$	$5.81e^{-17}$
GO:0017124	SH3 domain binding	9.71	$3.82e^{-11}$	$2.59e^{-08}$
GO:0031625	ubiquitin protein ligase binding	9.49	$1.88e^{-10}$	$1.27e^{-07}$
GO:0047485	protein N-terminus binding	8.58	$2.93e^{-07}$	$1.98e^{-04}$
GO:0005515	**protein binding**	7.76	$2.76e^{-18}$	$1.87e^{-15}$
GO:0008134	transcription factor binding	7.17	$1.81e^{-12}$	$1.23e^{-09}$
GO:0019899	**enzyme binding**	6.87	$2.47e^{-13}$	$1.67e^{-10}$
GO:0004672	protein kinase activity	6.86	$1.65e^{-09}$	$1.12e^{-06}$
GO:0003690	double-stranded DNA binding	6.81	$1.37e^{-08}$	$9.28e^{-06}$
GO:0001067	regulatory region nucleic acid binding	6.62	$1.15e^{-14}$	$7.82e^{-12}$
GO:0003682	chromatin binding	6.58	$6.35e^{-11}$	$4.30e^{-08}$
GO:0060090	**binding, bridging**	6.3	$2.49e^{-07}$	$1.69e^{-04}$
GO:0008022	protein C-terminus binding	6.08	$1.45e^{-07}$	$9.83e^{-05}$
GO:0044212	transcription regulatory region DNA binding	5.8	$2.53e^{-07}$	$1.71e^{-04}$
GO:0042802	identical protein binding	5.42	$4.08e^{-09}$	$2.76e^{-06}$
GO:0019904	**protein domain specific binding**	4.93	$9.83e^{-10}$	$6.66e^{-07}$
GO:0000989	transcription factor binding transcription factor activity	4.63	$1.46e^{-11}$	$9.90e^{-09}$
GO:0016772	transferase activity, transferring phosphorus-containing groups	4.18	$2.86e^{-16}$	$1.94e^{-13}$
GO:0004674	protein serine/threonine kinase activity	3.92	$6.42e^{-08}$	$4.35e^{-05}$
GO:0005524	ATP binding	2.56	$8.31e^{-09}$	$5.63e^{-06}$
GO:0030554	adenyl nucleotide binding	2.55	$7.48e^{-09}$	$5.07e^{-06}$
GO:0032555	purine ribonucleotide binding	2.3	$5.73e^{-08}$	$3.88e^{-05}$

In bold, the terms found to be enriched in the set of known adaptors/scaffolds. Enrichments were computed as described in section 5.2.3, but this time using $\alpha < 0.001$ and the as background set the human proteome as described in section 5.2.2.

In order to compare the similarities among the sets of adaptors/scaffolds here commented (*i.e.*, kAS, pAS and the set identified by Ramírez and Albrecht), we have generated a series of Venn diagrams that show the overlaps in terms of the protein composition, as well as the Pfam families and the molecular function terms enriched in each set (see Figure 5.8). We have found that, in terms of their shared proteins, the average overlap between each set is of 18.4%, which highlights the low consensus between the three strategies for the identification of adaptor or scaffold proteins. To our opinion, the low protein overlap arise from the different criteria applied by each identification method. In this regard, one of the main differences between our methods and the one implemented by Ramírez and Albrecht is that we allow enzymes to be classified as potential adaptors or scaffolds. In contrast, Ramírez and Albrecht follow the criteria that scaffold proteins lack enzymatic activity [177]. To our opinion, this criteria is inaccurate, given that proteins with well known scaffolding roles are also known to have enzymatic activity, such as the cases of the focal adhesion kinase (FAK) [167] and the kinase suppressor of Ras (KSR) [178, 190]. Moreover, differences in the enriched Pfam families as well as in the molecular function terms, are likely to arise given that different sets of proteins were considered as background when computing the enrichments. Finally, differences in the definition of the set of PPIs used as the human interactome can also influence the results of the identification strategies.

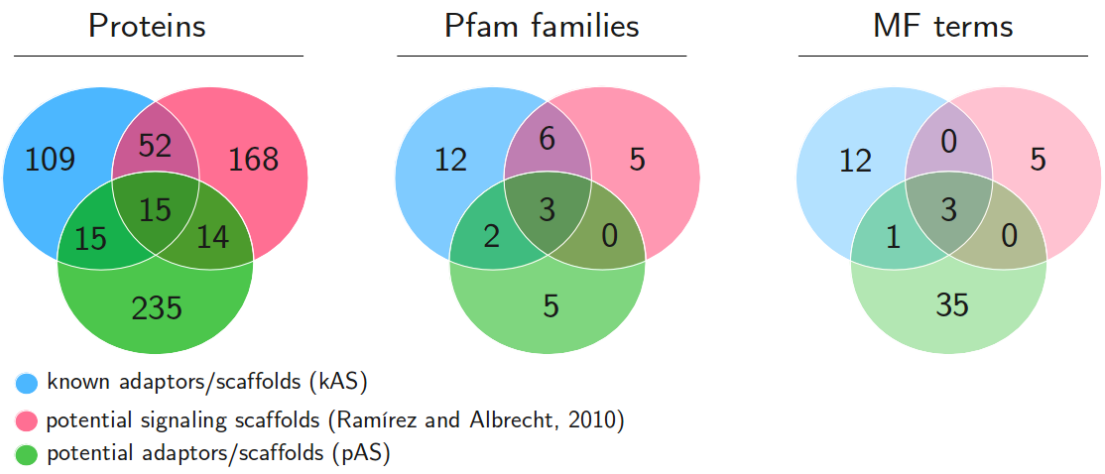


Figure 5.8: Comparison of three different sets of adaptor/scaffold proteins.

We have used our computational strategy for identifying a list of 279 pAS proteins from the human interactome. As the next step, we assessed the performance of our strategy based on the computation of its precision and recall as described in section 5.2.6.6. For a total of 22 kinases we could compute the performance of the classification strategy. This group pf 22 kinases is composed by those for which we i) count with at least five substrates, ii) know at least one kAS protein and iii) have identified at least one pAS protein. The results show low values of the evaluated parameters with averages of 0.22, 0.20 and 0.17 for the recall, the precision and the F1 score respectively (see Table 5.5 and Figure 5.9). However,

these results are not unexpected, given that only a small fraction (37/706, 5.24%) of the kinase–pAS relationships identified in the strategy are also present among the reference set of kinase–kAS pairs.

To our opinion, the performance of the current strategy can also be negatively influenced by the incompleteness of the current human binary interactome. This situation could be affecting the recall due to missing evidence of PPIs between kASs and substrates. In general, missing PPIs constitute a mayor hurdle for the current strategy, given its basis on the analysis of partners shared by the substrates of given kinase.

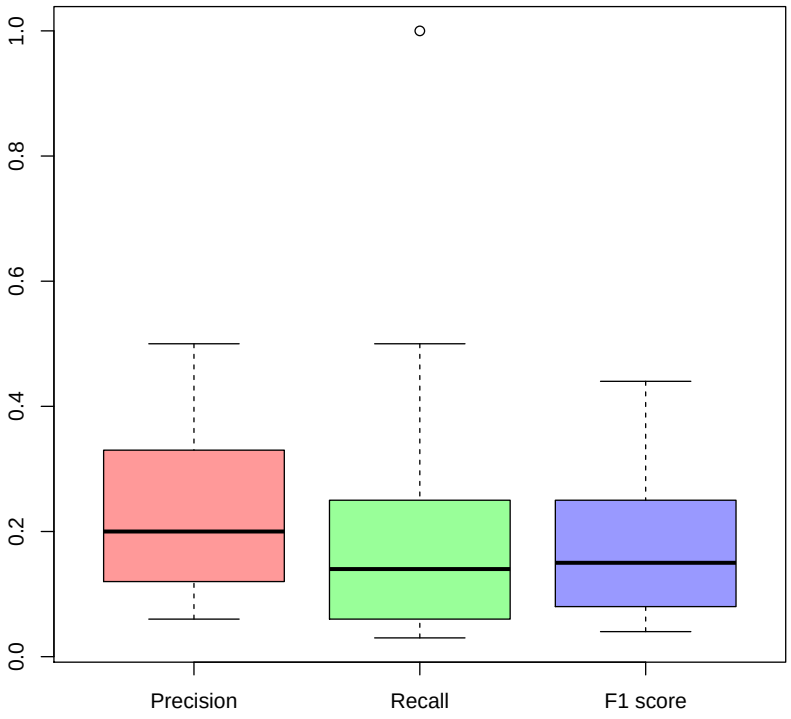


Figure 5.9: Performance of the computational strategy for the identification of pAS proteins.

Considering the aforementioned parameters — particularly the F1 score — the overall performance of the classification strategy can be considered low. However, these parameters can not account for the chances of the pASs identified to play such a role for their respective kinases.

Table 5.5: Performance of the strategy for identification of pAS proteins.

Kinase	pAS	kAS recalled	Recall	Precision	F1 score
AKT1	34	2/15	0.13	0.06	0.08
KPCD	23	1/8	0.12	0.04	0.06
KPCE	3	1/5	0.2	0.33	0.25
KS6A1	12	1/5	0.2	0.08	0.11
KS6A5	4	1/3	0.33	0.25	0.28
PRKDC	21	1/10	0.1	0.05	0.07
CHK1	8	2/8	0.25	0.25	0.25
MARK2	4	2/5	0.4	0.5	0.44
MAPK2	16	1/2	0.5	0.06	0.11
KC1A	17	1/6	0.17	0.06	0.09
E2AK2	6	1/3	0.33	0.17	0.22
PAK2	1	1/5	0.2	1	0.33
MP2K4	12	2/4	0.5	0.17	0.25
ABL1	33	4/28	0.14	0.12	0.13
EGFR	68	2/33	0.06	0.03	0.04
INSR	13	1/13	0.08	0.08	0.08
JAK2	13	3/9	0.33	0.23	0.27
PGFRB	70	3/18	0.17	0.04	0.06
FYN	15	2/25	0.08	0.13	0.1
HCK	12	2/10	0.2	0.17	0.18
KSYK	13	2/10	0.2	0.15	0.17
ZAP70	3	1/9	0.11	0.33	0.17
	18.23	-	0.22	0.20	0.17

Kinase, kinase UniProt Id; **pAS**, number of potential adaptors/scaffolds identified for the corresponding kinase; **kAS recalled**, number of known adaptors/scaffolds of the corresponding kinase that were recalled by our strategy; **Recallb**, **Precision** and **F1 score**, values of the performance parameters achieved by the strategy. Last row shows the average values for the corresponding columns.

In an additional analysis, we have taken into account of the role that co-localization with the substrates can play in the specificity of kinases. Here, we attempted the identification of cellular component terms (CC) shared by the pAS proteins and the substrates of the corresponding kinases. More in detail, we evaluated whether the pAS proteins of a given kinase, are annotated to the same CC terms that have been found to be enriched in the set of substrates of that kinase (see section 5.2.6.7 of Materials and methods).

For 527/706 (74.6%) of the kinase-pAS pairs, we found evidence of co-localization between the pAS and the substrates. This set of 527 kinase-pAS pairs account for 41 kinases, 156 pASs — corresponding to 52.6% and 55.9% (respectively) of the ones in the initial 706 kinase-pAS pairs— and 35 unique CC terms. The Figure 5.10 shows a pie chart representation of the CC terms shared by the pAS proteins and the sets of substrates; while in the Table 5.6 we show cases of pAS proteins that are found to co-localize with substrates of their corresponding kinases. For example, the pair formed by the β-adrenergic receptor kinase 1 (ARBK1) and the Na(+)/H(+) exchange regulatory cofactor NHE-RF (NHRF1), where the later it has been reported to be involved in the scaffolding of β-adrenergic receptors — substrates of ARBK1 — at the plasma membrane [277]. Another example is the case of the checkpoint kinase-1 (CHK1) and the 14-3-3 protein zeta (1433Z), where the later it has been reported to be required for the nuclear retention of CHK1 [278].

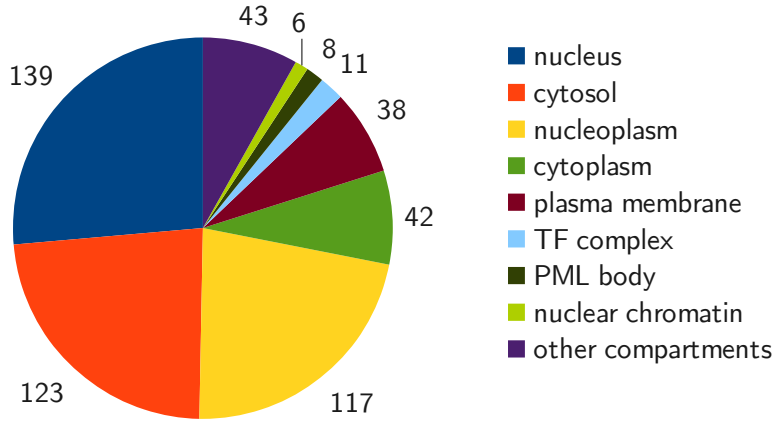


Figure 5.10: Cellular component terms shared by substrates and pAS proteins.

The slices represent the number of kinase-pAS pairs where the pAS is annotated to the the given CC term. TF and PML stand for transcription factor and nuclear bodies respectively.

Table 5.6: Cellular component terms shared by pAS proteins and substrates.

Kinase	GO description	Enrichment ratio	Adjusted *p*-value	pAS co-annotated
ARBK1	apical plasma membrane	15.7	$4.08e^{-02}$	NHRF1
CDK4	chromatin	16.12	$3.76e^{-02}$	EP300
CDK4	transcription factor complex	20.92	$1.67e^{-02}$	E2F4,EP300
CDK9	PML body	61.38	$4.19e^{-03}$	PIAS4
CHK1	nucleoplasm	10.16	$1.30e^{-04}$	1433Z,CHK2,EP300,MDM2,UBC
CHK2	PML body	29.86	$3.25e^{-03}$	RB,SIRT1,SUMO1
CSK	membrane raft	33.87	$2.74e^{-03}$	ERBB2
EGFR	endosome	8.47	$6.48e^{-04}$	FYN,GRB2,NTRK1
FYN	cell junction	3.98	$3.15e^{-02}$	PTN12
INSR	cytosol	12.11	$5.41e^{-04}$	ABI1,GRB2,IRS1,P85A,SRC,UBC
KC1A	lateral plasma membrane	38.75	$3.46e^{-02}$	CTNB1
KC1A	APC-Axin-1-beta-catenin complex	275.94	$4.18e^{-02}$	CTNB1
KCC2G	vesicle membrane	19.29	$7.14e^{-04}$	GRB2,NCK1
PDPK1	mitochondrion	6.81	$2.48e^{-02}$	1433Z,CASP3,MAD1,PDK1
PLK1	nucleus	4.33	$7.49e^{-04}$	ABL1,ANDR,GRB2,P53,VHL

Kinase, UniProt ID of the kinase; **GO description**, description of the Cellular Compartment term of the Gene Ontology enriched in the set of substrates of the kinase; **Enrichment ratio**, ratio of enrichment of the GO term; **Adjusted *p*-value**, multiple test correction by Bonferroni's method; **pAS co-annotated**, pASs of the current kinase that are annotated to the corresponding GO term. For the complete set of results see the section A5 in Appendices.

In our opinion, these results suggest that the association to pAS proteins might play an important role in the co-localization of the analyzed kinases with their cognate sets of substrates. Nevertheless, we are aware that in many cases, the cellular component shared by the substrates and the pAS proteins are too broad (*e.g.*, cytosol, nucleoplasm, cytoplasm) and can not fully justify, based on spatial constrains, the substrate specificity of the kinases.

The complete set of results on co-localization are provided in the section A5 of the Appendices.

5.3.4 Adaptors/scaffolds binding to significantly large numbers of substrates

In this section we continued with the analysis on the contribution of adaptors/scaffolds to the observed *in vivo* substrate specificity of human protein kinases. In the previous section, we showed results suggesting that kAS proteins are five times more likely to interact with a statistically significant number of substrates than any random protein in a sub-network of the human interactome. To explore further on this result, we tested whether it would hold for a reduced set of manually curated kinase–kAS associations. For this, we used a subset of kinase–kAS pairs from the previously described 'gold standard set' (GSS), where we known at least five *in vivo* substrates for each kinase. The section 5.2.7 contains the complete description of the strategy, and the Figure 5.11 shows its schematic representation.

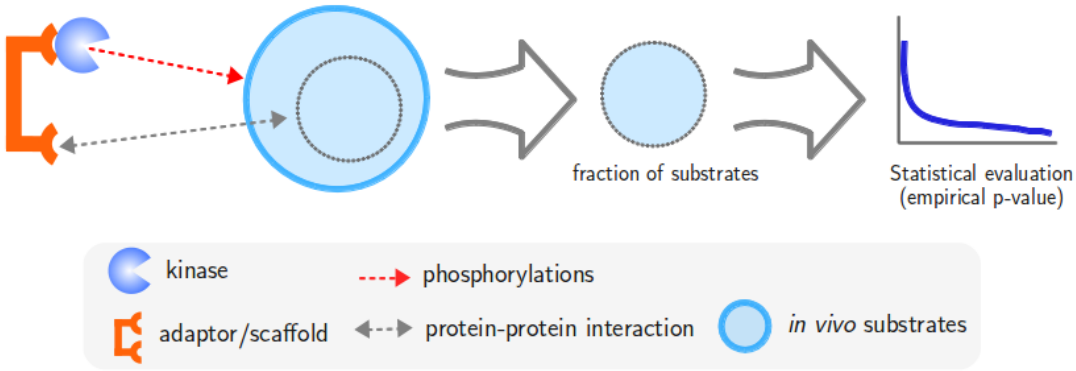


Figure 5.11: Adaptors/scaffolds interacting with a significant number of substrates.

Based on an empirical *p*-value we test whether a known adaptor/scaffold interacts with a significantly large number of the substrates of an associated kinase.

Table 5.7: Adaptors/scaffolds that interact with a significant fraction of substrates.

Kinase	Adaptor or scaffold	Substrate fraction	Raw *p*-value	Adjusted *p*-value	Substrates interacting with the adaptor/scaffold
ABL1	GRB2	14/46	$4.00e^{-04}$	$1.00e^{-02}$	ABI1, ABL2, BCAP, BCR, BTK, CDN1B, EGFR, JAK2, M4K1, MUC1, PTN6, RPGF1, UFO, WASL
ABL1	NCK1	6/46	$2.50e^{-03}$	$1.95e^{-02}$	ABL2, EGFR, JAK2, M4K1, RPGF1, WASL
CHK1	1433B	4/17	$3.90e^{-03}$	$1.95e^{-02}$	MDM4, MPIP1, MPIP2, MPIP3
CHK1	1433Z	4/17	$3.90e^{-03}$	$1.95e^{-02}$	MDM4, MPIP1, MPIP2, P53
EGFR	GRB2	6/27	$2.80e^{-03}$	$1.95e^{-02}$	CBL, EPS15, ERBB2, GAB1, MUC1, PLD2

Proteins are represented by their UniProt Ids. **Substrate fraction** represents the fraction of the substrates that are known to interact with the adaptor/scaffold.

From our analysis, we could identify only 5/49 (10.2%) successful cases of kinase-kAS pairs where the adaptor/scaffold interacts with a significantly large number of the *in vivo*

substrates of its associated kinase (see Table 5.7). Due to the limited number of successful cases, we can not derive generalizations — at least based on this dataset — about a potential role of the adaptors/scaffolds as substrate recruiters for the kinases. However, we have previously shown results suggesting a relevant role for adaptors/scaffolds in providing cellular co-localization for kinases and substrates.

5.3.5 Association to adaptors/scaffolds diminish cross-specificity of kinases

Previously, we have used cellular component annotations to explore the role that the potential adaptors/scaffolds may play in kinase specificity by promoting spatial proximity between the kinases and their cognate sets of substrates. However, being aware of the fact that different kinases may associate to a common adaptor/scaffold, we wanted to test whether kinases that share an adaptor/scaffold also share a number of *in vivo* substrates larger than what would be expected by chance. Stated in other words, we tested whether the association to common adaptors/scaffolds would promote significant substrate cross-specificity between kinases. In this analysis, under the assumption that adaptors/scaffolds are major players in kinase substrate specificity, we would expect to find few (or none) cases of significant substrate cross-specificity. We have applied the current analysis to both the kAS and the GSS sets, considering only *in vivo* substrates (see section 5.2.8 for a description of the strategy and Figure 5.12 for a schematic representation).

We have first analyzed the set of PK-kAS associations where the kinases have at least five *in vivo* substrates and for which the kAS is known to interact with at least two kinases. In total, we count with 195 PK-kAS pairs — involving 54 kinases and 57 kAS proteins — , which represents 15.2% of the complete set of PK-kAS associations. In total, we have identified 19 cases where a kAS is known to interact with at least two kinases who share at least one *in vivo* substrate. These cases involve 19 different kAS proteins and 21 different kinases. For none of the cases we found kinases sharing a statistically significant number of *in vivo* substrates (see Table 5.8).

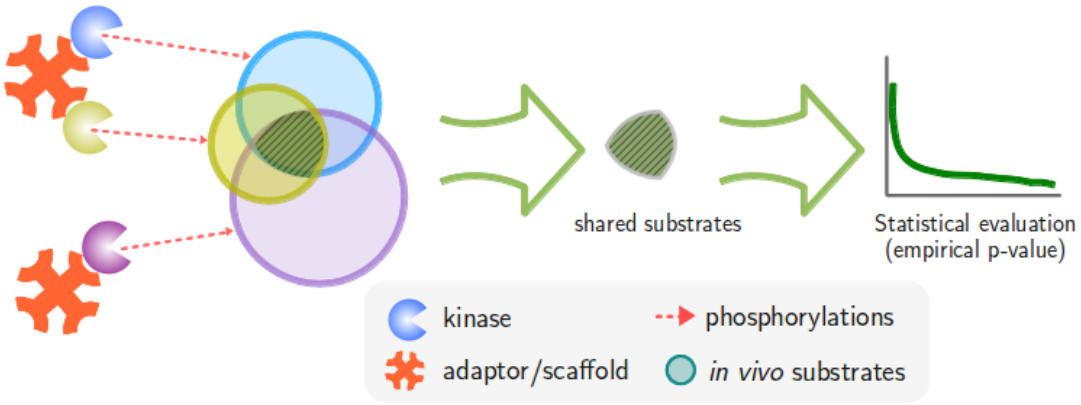


Figure 5.12: Association to adaptors/scaffolds and substrate cross-specificity of kinases.

Secondly, we analyzed the 'gold standard set' (GSS) of kinase-adaptor/scaffold associations, where we count with six adaptors/scaffolds associated to kinases sharing at least one *in vivo* substrate. A total of nine kinases were associated to the aforementioned adaptors/scaffolds. Again, none of the kinase pairs were found to share a statistically significant number of *in vivo* substrates (see Table 5.9).

As a result, from our strategy and available data, we have not found any case of two

Table 5.8: Statistical significance of the number of shared substrates for PK-kAS pairs.

Adaptor/scaffold	Associated kinases	Shared substrates	*p*-value
CD2AP	ABL1,FYN	1	1.00
FRS3	MK01,FGFR1	1	1.00
FYB	ABL1,FYN	1	1.00
TGFI1	FAK1,FAK2	1	1.00
APBB1	EGFR,ERBB2	2	1.00
DOK4	EGFR,ERBB2	2	1.00
DOK6	EGFR,ERBB2	2	1.00
JIP2	EGFR,ERBB2	2	1.00
KHDR1	HCK,LCK,SRC	2	0.11
SHC2	EGFR,ERBB2	2	1.00
SHC3	EGFR,ERBB2	2	1.00
PAR6A	KPCI,KPCZ	3	1.00
PAR6B	KPCI,KPCZ	3	1.00
BIRC5	AURKA,AURKB	4	1.00
SH2B1	EGFR,INSR	4	1.00
PKHO1	AKT1,CSK21	5	1.00
ELP1	GSK3B,MK08	7	0.78
DAG1	FYN,SRC	13	0.52
SCRIB	MK01,MK03	73	0.10

Proteins are represented by their UniProt Ids. **Shared substrates**, number of *in vivo* substrates shared by the kinases; ***p*-value**, statistical significance of the number of substrates shared by the kinases.

Table 5.9: Statistical significance of the number of shared substrates in the GSS set.

Adaptor/scaffold	Associated kinases	Shared substrates	*p*-value
IMA2	SGK1,CHK2	1	1.00
NCK1	ABL1,EGFR	3	1.00
PAR6B	KPCI,KPCZ	3	1.00
SQSTM	KPCI,KPCZ	3	1.00
SHC1	EGFR,INSR	4	1.00
KHDR1	LCK,SRC	14	0.43

Proteins are represented by their UniProt Ids. **Shared substrates**, number of *in vivo* substrates shared by the kinases; ***p*-value**, statistical significance of the number of substrates shared by the kinases.

(or more) kinases that, sharing a common adaptor/scaffold, have a statistically significant large number of common *in vivo* substrates (*i.e.*, a large substrate cross-specificity). To our opinion, this results reinforce the concept of the adaptors/scaffolds as major contributors to the *in vivo* substrate specificity of kinases. From these results we could hypothesize that, even if two kinases bind to common adaptors/scaffolds, these adaptors/scaffolds are able to mediate the recruitment of each kinase to its corresponding location in the cell.

5.4 Concluding remarks

- We have collected a set of 191 known adaptor/scaffold proteins, that associate to a total of 287 (55%) human kinases, which in turn represent 94 (72.3%) of all human kinase families. These data suggest that association to adaptors and scaffolds is a common mechanism of human kinases.

- We have found that, in comparison to random proteins in the human interactome, known adaptors/scaffolds are five times more likely to interact with a large fraction of the substrates of the kinases they are associated (p-value $= 1.08e^{-15}$). This result suggest a role for adaptors/scaffolds in facilitating the encounter of the kinases with their cognate substrates.

- In an attempted to further support the aforementioned result, we tested whether it holds for a reduced set of only 49 (manually curated) kinase–known adaptor/scaffold pairs (a subset of the GSS). We found that, for 10/49 (10.2%) of the cases, the known adaptor/scaffold interacts with a significantly large number of the *in vivo* substrates. Although this result does not allow us to generalize — at least on this reduced set — about the role of adaptors/scaffolds as substrate recruiters, other experiments in this document suggested their role in promoting colocalization of kinases and substrates.

- Starting from a set of 156 kinases (representing 47% of the kinase families), we have identified a set of 279 potential adaptors/scaffolds for 78 (50%) of the kinases, covering 44 (33.8%) of the kinase families. The set of potential adaptors/scaffolds is enriched in functional annotations and domain families that support their association to proteins involved in cellular signaling. To our opinion, these results support the role intended for the potential adaptors/scaffolds and also suggest that associations to this type of molecules is not infrequent among kinases in our set.

- For 527/706 (74.6%) of the kinase-potential adaptor/scaffold associations previously identified, we have found evidence of colocalization between the potential adaptors and scaffolds and the substrates of the corresponding kinase. In total we have identified colocalizations at 35 different cellular compartments. To our opinion, these results suggest that for most of the associations found, the adaptors/scaffolds might play a fundamental role in the colocalization of the kinase and its substrates.

- Regarding the substrate cross-specificity of kinases, we have not found any case of two or more kinases that, having an adaptor/scaffold in common, also share a number of *in vivo* substrates larger than what would be expected by chance. To our opinion, these results reinforce the idea that adaptors/scaffolds effectively contribute to constrain the set of potential substrates available for a kinase, most probably by recruiting the enzymes to particular locations or macromolecular complexes.

6 General discussion

In the current work, we have approached the identification and the quantification of the contribution of different elements to the substrate specificity of human protein kinases. For this, we have analysed: i) the residues in the close neighbourhood of the the phosphorylation site, ii) the association of kinases to adaptors or scaffold proteins and iii) the cellular co-localization of kinases and their substrates. We have identified residues in the close vicinity of the phosphorylation sites, that function as positive (or negative) elements for the substrate recognition by the kinases. Our results regarding the association of kinases to adaptor/scaffold proteins, suggest that these interactions may play important roles in the localization of the enzymes with their set of cognate substrates and also in diminishing substrate cross-specificity *in vivo*.

6.1 Analysis of phosphorylation sites and their adjacent residues

6.1.1 Sequence logos

Initially, we have generated logos from the stretches of sequences phosphorylated by the different kinases and kinases families in our data. The logos here generated, provided us an initial grasp on hte sequence diversity and the main differences between the phosphorylation motifs targeted by the different kinases. Also, the logos allowed us to identify several residues — neighboring the phosphorylation sites — that are likely to play important roles for the specificity of the kinases. Additionally, by comparing our logos to other previously published [70], we verified that our integrated data agree — in general terms — with the phosphorylation motifs already known for the kinases in our set. Finally, aided by the logos, classified the kinases based on the residue composition of the stretch of sequences surrounding the phospho-acceptor residue (*i.e.*, acidophilic, basophilic, proline-directed and glutamine-directed). We have observed that, kinases of the same class often belong to the same or closely related families. However, there are exceptions such as the kinases of the SYK family (group of tyrosine kinases) and the serine/threonine kinases from the families CK1, CK2 and PLK — groups CK1, CMGC and Other, respectively — which are all classified as acidophilic.

6.1.2 Specificity-determinant residues

Using the in-house program `genpssm`, we have constructed PSSMs for performing a quantitative analysis of the contribution to the kinase specificity of the residues flanking the phosphorylation sites. For this analysis, we selected a set of 22 kinase families for which we

count with at least 100 phosphorylation events. For 19 (86.4%) of the families studied we identified at least one SDR. For all these 19 families, we correctly classified as SDR residues that have been reported to play important roles in the specificity of the corresponding kinases. These are the cases, for example, of $MAPK_{P+1}$, $PIKK_{Q+1}$, AKT_{R-3} and $CK2_{E+3}$. The quantification of the relevance of the SDRs — based on their frequency of occurrence among the phosphorylation events of each family — shows a wide variation across the different families. For example, the four SDRs previously mentioned have relatively high frequencies that range between 88.86% and 45.83%; however, other SDRs identified by our method show much lower frequencies (*e.g.*, PKC_{K+2} = 19.79%, $CAMKL_{N+3}$ = 18.03% and $CK2_{D+2}$ = 15.54%). Based on our data, we hypothesize that the combination of multiple SDRs of low frequencies contribute in an additive way to the recognition of the phosphorylation site by the kinase. In contrast, we consider that SDRs of high frequencies — which are highly required for the identification of the phosphorylation site — are responsible of a larger contribution to the kinase specificity. Moreover, we have noted that the frequency of any given SDR is low — 6.0% on average — among the phosphorylation events of the kinase families that do not count with that SDR. Our interpretation of this observation is that, SDRs may also function as elements of negative selection to avoid the phosphorylation of non-cognate sequences.

By our analysis we have identified SDRs that, to the best of our knowledge, have not been previously reported as determinants of the specificity for the corresponding kinase families. These are the cases, for example, of $CAMKL_{N+3}$ and AKT_{W+1} — both from basophilic kinases —, with frequencies of 18.03% and 3.85% respectively. The SDR N+3, is present in the sequences targeted by the microtubule affinity-regulating kinases (MARK) — CAMKL family members — within the repeat regions of the human TAU protein, which is implicated in Alzheimer's disease [279]. Besides, N+3 have a low frequency (3.83%) among the phosphorylation events of the other 21 kinase families in the analysis. Given that the repeat regions of TAU are responsible for the binding to the microtubules [280]; we consider that the presence of N+3 in these regions is an important element for the recognition by MARK kinases, and therefore for the regulation of the association of TAU to the microtubules. The case of W+1 — identified as an SDR for the AKT family — is an interesting result, given that tryptophan is rarely found in the close sequence vicinity of phosphorylation sites — 0.66% among the phosphorylation events of non AKT kinase families —. W+1 was identified as an SDR even when occurring at low frequency (3.85%) among the phosphorylation events of AKT kinases, which prompted us to research further about the biological relevance of the finding. Interestingly we found reports in the literature showing that, by phosphorylating sequences containing a conserved W+1, some AKT kinases are implicated in the regulation of transcription factors of the FOXO family [125]. To our opinion, this result supports the utility of our approach for the identification of SDRs, even for residues that occur at low frequency among the phosphorylation sites of the kinase of interest.

6.1.3 Position-specific scoring matrices

In this work we have also evaluated the performance and the statistical significance of the PSSMs generated for kinases and kinases families in our data. Using the recall and the

IC values as statistics for the comparison to random backgrounds, we have estimated the statistical significance of our PSSMs. We observe that, for most cases, statistical and non-statistically significant PSSMs differ in their values of the IC, the percent recall, the AUC-ROC and the numbers of seed phosphorylation sites. Regarding the performance, we have found negative correlations between the number of seed phosphorylation sites and i) the percent recall of the PSSM and ii) the IC of the PSSM. These results show that the combined increase of i) the sequence diversity and ii) the number of seed phosphorylation sites from which a PSSM is generated, can exert a negative effect in both the performance and the level of self-information of that PSSM. From this result we hypothesize that, for some kinases the substrate specificity might be represented best by multiple PSSMs, a concept that have been previously applied in the analysis of DNA recognition by transcription factors [281]. Although not covered in the work here presented, we consider that in such cases, multiple PSSMs could be useful for modeling fairly different phosphorylation motifs that are targeted by the same kinase.

6.2 Analysis of the association of kinases to adaptors and scaffolds

6.2.1 Identification of adaptors and scaffolds

We have compiled a set of 191 human proteins that are known to play roles of adaptors or scaffolds of human protein kinases. We based our selection criteria on the functional annotation of these proteins in the UniProt database and also on the evidence of binary PPIs between them and the kinases. From the analysis of this set of 191 proteins in the context of the human interactome we extracted two main messages, i) that the association to adaptor/scaffold proteins is extended among human protein kinases — 55% kinases (72.3% kinase families) bind to at least one adaptor/scaffold — and ii) that the adaptor/scaffold associate to significantly large numbers of the substrates of human protein kinases. Many adaptors and scaffolds count with one or multiple functional domains that confer them the capability of interacting at the same time with two or more proteins [160]. Our findings support the hypothesis that adaptors and scaffolds may play important roles in the specificity of a large number of kinases, probably by contributing to the encounter of the kinases and their substrates.

Based on the evidence of association to large numbers of the substrates of human protein kinases and using the PPIs in the human interactome, we approached the identification of potential adaptors/scaffolds of kinases. The enrichment in functional terms and proteins domains, agrees with the role intended for the set of 279 proteins identified as potential adaptors or scaffolds. Moreover, the functional enrichment suggest that the 279 are involved in processes of cellular signalling. Only 71 (25%) of these proteins are among the set of known adaptors and scaffolds previously described, a result that supports the ability of the method to uncover potential functions for more than 200 proteins.

In 2010, Ramírez and Albrecht developed a computational strategy for the identification of signaling scaffold proteins in the human interactome [243]. In their work, the authors

suggested a total of 250 potential scaffold proteins. The comparison of this set of 250 scaffolds with the two sets identified by us, show little overlap in terms of the proteins, the molecular function terms and the protein domains enriched. The differences found between these sets can be justified by several factors. First, our definition of the human interactome comprise only high confidence binary PPIs; while the interactome used by Ramírez and Albrecht is not restricted to binary interactions and therefore contains PPIs of much lower confidence. In this sense, our definition of the human interactome reduces the chances of false positives, although at the expense of missing some potentially good cases. Also, following the definition of signaling scaffolds proposed by Zecke *et al.* [177], the authors Ramírez and Albrecht did not consider enzymes as potential scaffolds. However, is known that proteins with scaffolding functions have also enzymatic activity (*e.g.*, the focal adhesion kinase (FAK) [167] and the kinase suppressor of Ras (KSR) [178,190]). To avoid missing cases as the two previously mentioned, we did not impose any filtering criteria based on molecular functions. In general, the computational identification of adaptor and scaffold proteins is still a challenging task, given that these proteins do not share a common evolutionary origin, neither they share sequence signature motifs.

6.2.2 Colocalization of adaptors and scaffolds with substrates

In the present work, based on their annotations to cellular component terms, we investigated the evidence of colocalization between the 279 potential adaptors/scaffolds and the sets of substrates of the associated kinases. For 527 (74.6%) of the kinase–potential adaptor/scaffold associations we found evidence of colocalization to at least one cellular component term, accounting a total of 35 unique terms. For approximately 80% of the colocalization cases, the cellular component terms were — to our opinion — too general to support the hypothesis of the potential adaptor/scaffold as an element promoting spatial proximity between the kinases and their substrates (*e.g.*, cytoplasm and cytosol). However, among the remaining 20% of the cases, we found interesting examples that may contribute to support the afore mentioned hypothesis. For example, for the case of the checkpoint kinase-1 (CHK1) and the 14-3-3 protein zeta (1433Z, a known adaptor protein), it has been reported that 1433Z retains CHK1 at the nucleus, where the kinase regulates the mitotic progression in response to DNA damage [278]. As a second example, we also correctly predicted the Na(+)/H(+) exchange regulatory cofactor NHE-RF (NHRF1) to be a scaffold for the β-adrenergic receptor kinase 1 (ARBK1) at the plasma membrane. Reports have shown that indeed NHRF1 serves as a scaffold for the substrates of ARBK1 [277]. Another example is the casein kinase α-1 (KC1A), for which we identified the catenin β-1 (CTNB1) as a potential adaptor/scaffold. KC1A phosphorylates CTNB1 at serine 45, they are both components of the canonical Wnt signaling pathway and they are also part of the large APC–Axin-1–β-catenin complex [282]. Interestingly, CTNB1 contains 12 repeats of the Armadillo (ARM) domain, which is implicated in mediating PPIs. It has been recently suggested that proteins containing ARM repeats, constitute an attractive modular system as scaffolds for peptide-mediated PPIs [283]. Therefore, we consider that CTNB1 may constitute a plausible scaffold that may promote spatial proximity between KC1A and its substrates.

6.2.3 Adaptors and scaffolds diminish kinase cross-specificity

Finally, we have investigated the relationship between the association to common adaptors or scaffolds, and the substrate cross-specificity of kinases. In total we analyzed 23 cases of two or more kinases that associate to a common adaptor or scaffold, and for non of the cases the kinases shared a number of *in vivo* substrates larger than what would be expected due to chance. Nevertheless, we found the case of the kinases MK01 and MK03 — ERK2 and ERK1 MAP kinases, respectively — which share 73 *in vivo* substrates. Even when it was not statistically significant, the number substrates in common was very large when compared to other sets of kinases in out analysis, and therefore we decided to explore this particular case in more detail. In fact, ERK1 and ERK2 are very closely related kinases, with 82% and 89% of identity in their full and catalytic domain sequences. ERK1 and ERK2 share many if not all functions [284] and despite numerous efforts to establish differences, the detection of such distinctive functions it has been difficult to pinpoint [285]. Therefore, we consider that their large sequence identity, together with their almost identical functions can explain the large substrate overlap reflected in our data. To our opinion, these results support the hypothesis that adaptors and scaffolds are able to diminish *in vivo* substrate cross-specificity by recruiting the kinases to specific macromolecular complexes or cellular locations.

7 General conclusions

Protein kinases constitute one of the largest and more diverse superfamilies of proteins in human, where they account for nearly 2% of the genes. Many kinases play fundamental roles in several cellular processes such as signalling, replication and growth. Due to their involvement in key functions, many kinases have been associated to important human pathologies such as cancer and diabetes. Despite that most of the kinases share a highly conserved catalytic domain, the observed *in vivo* substrate specificity of these enzymes show little correlation with their sequences. In this sense is known that, *in vivo*, the substrate specificity of these protein kinases is a complex phenomena that is tightly regulated by several factors.

The specific objectives of the present thesis are the quantification of the contribution — to the kinase specificity — of two elements: i) the amino acids neighboring the phospho-acceptor residue in the sequence of the substrate and ii) the association of protein kinases to adaptor or scaffold proteins.

For the first objective, we started by collecting a set of experimentally determined kinase–phosphorylation sites relationships in human, which accounted for 62.7% of human kinases — representing 71.5% of the human kinase families — related to almost 6000 different phosphorylation sites in more than 1800 distinct substrates. We used these data for constructing sequence logos and PSSMs from the sets of sequences targeted in the substrates by the different kinases and kinases families in our data.

The analysis of the sequence logos allowed us i) to obtain a general grasp on the diversity of sequences targeted by the different kinases, ii) to confirm that the sequence patterns recognized by kinases in our set were comparable to the ones reported in the literature and iii) to guide the classification of kinases and kinase families based on the residue composition of the stretch of sequences surrounding the phospho-acceptor residue (*e.g.*, acidophilic, basophilic, proline-directed and glutamine-directed).

Regarding the PSSMs, we have used them to identify SDRs for 22 kinase families. The analysis of the SDRs showed that the type of amino acids recognized as SDR, their positions around the phospho-acceptor residue, as well as their frequencies, vary greatly among the different kinase families. Some kinase families display a strong preference for particular SDRs, which occur in more than 80% of the corresponding phosphorylation events (*e.g.*, $\mathrm{MAPK}_{\mathrm{P}+1} = 88.86\%$, $\mathrm{CDK}_{\mathrm{P}+1} = 81.72\%$, $\mathrm{PIKK}_{\mathrm{Q}+1} = 80.83\%$ and $\mathrm{AKT}_{\mathrm{R}-3} = 84.13\%$). Other families show mild preferences for the SDRs identified, which occur at lower frequencies among the corresponding phosphorylation events (*e.g.*, $\mathrm{GSK}_{\mathrm{S}-4} = 38.5\%$, $\mathrm{GSK}_{\mathrm{P}+1} = 53.96\%$, $\mathrm{CK1}_{\mathrm{S}-3} = 28.5\%$, $\mathrm{CK1}_{\mathrm{S}+3} = 31.09\%$, $\mathrm{CK2}_{\mathrm{D}-1} = 16.19\%$ and $\mathrm{CK2}_{\mathrm{E}+3} = 45.83\%$). However, we have observed that multiple SDRs are generally identified in families for which the frequencies of the SDRs tend to be relatively low. We consider that in these cases, the presence of multiple SDRs might contribute cooperatively to the recognition of the target site in the sequence of the substrate. Moreover, we have noted that the SDRs occur at low

frequency (6.01% on average) among the complementary phosphorylation events. That is, the phosphorylation events corresponding to those kinase families that do not count with the given SDR. To our opinion, this suggests that the SDRs here identified might function not only as elements for the positive selection of cognate target sequences, but also as negative selection factors for non-cognate phosphorylation sites.

Continuing with the analysis of the PSSMs, in this work we have also studied their performance and their statistical significance using as the test statistics the recall and the IC. Regarding the performance, we have found negative correlations between the number of seed phosphorylation sites and i) the percent recall of the PSSM (R = -0.59, p-value = $2.4e^{-31}$) and ii) the IC of the PSSM (R = -0.4, p-value = $9.8e^{-14}$). To our opinion, these results show the effect that the sequence degeneracy caused by the increase of the seed phosphorylation sites can exert on the performance of the PSSM and on its level of self-information. Moreover, based on the values of IC and on the comparison to random backgrounds, we have estimated the statistical significance of PSSMs from both independent kinases and kinases families. We observe that, in most cases, statistical and non-statistically significant PSSMs differ not only in their values of the IC but also in their percent recall, their AUC-ROC and their numbers of seed phosphorylation sites.

With respect to our second objective, we started by collecting from UniProt a set of 191 proteins with known adaptor or scaffolding function, and that are known to physically interact with at least one human kinase. These 191 proteins associate to 55% of the human kinases, which in turn account for 72.3% of all human kinase families. Besides, we have tested the functional enrichments of the group, and we have found over-representation of functional terms and protein domains related to 'protein binding' and to 'protein scaffolding' functions. To our opinion, these data suggest that i) the association to adaptors or scaffolds is a common mechanism among human kinases and ii) that the set of 191 known adaptor/scaffold proteins is functionally coherent with the intended biological role. Moreover, if compared to random proteins in the human interactome, the 191 proteins (as a set) are five times more likely to interact with a large fraction of the substrates of the human kinases they are associated to (p-value = $1.08e^{-15}$). We consider that this result suggests a role for adaptor and scaffold proteins in facilitating the encounter of the kinases with their cognate substrates.

In the current thesis we approached the identification of potential adaptor/scaffold proteins for human kinases. Based on our previous findings, we devised a strategy that aimed the identification of human proteins that interact with significant fractions of the substrates of kinases. We executed this strategy on a group of 156 kinases (accounting for 47% of the human kinase families) for which we have at least five substrates. For 50% of the initial kinases (covering 33.8% of all human kinase families), we identified a set of 279 potential adaptors/scaffolds. The set of potential adaptors/scaffolds is enriched in functional terms and in domain families that suggest a tight link to protein-protein binding functions in processes of cellular signalling and that also, to our opinion, support the biological role intended for this group of proteins. Moreover, we have found that for 74.6% of the kinase–potential adaptor/scaffold associations previously identified, the adaptor/scaffold is annotated under cellular component terms found to be enriched among the set of substrates of the associated kinase. We consider that these results put forward a role for the potential adaptors/scaffolds in promoting the co-localization of the kinases and their sets of substrates.

Finally, taking into consideration the fact that adaptors and scaffold may promote co-localization of kinases and their substrates; we analyzed whether the association of different kinases to common adaptors/scaffolds, might have a relationship with the *in vivo* substrate cross-specificity of the kinases. In this sense, we have not found any case of two or more kinases that, having an adaptor/scaffold in common, also share a significant number of *in vivo* substrates. To our opinion, this result reinforce the concept that the association of kinases to adaptors/scaffolds plays a fundamental role in targeting the kinases to their corresponding substrates, most probably by recruiting the enzymes to particular locations or macromolecular complexes.

Bibliography

[1] Cohen, P. May 2002 *Nature cell biology* **4(5)**, E127–30.

[2] Lothrop, A. P., Torres, M. P., and Fuchs, S. M. April 2013 *FEBS letters* **587(8)**, 1247–57.

[3] Forrest, A., Ravasi, T., Taylor, D., Huber, T., Hume, D., and Grimmond, S. (2003) *Genome research* **13(6b)**, 1443.

[4] Arena, S., Benvenuti, S., and Bardelli, a. September 2005 *Cellular and molecular life sciences : CMLS* **62(18)**, 2092–9.

[5] Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. December 2002 *Science (New York, N.Y.)* **298(5600)**, 1912–34.

[6] Torkamani, A., Verkhivker, G., and Schork, N. J. August 2009 *Cancer letters* **281(2)**, 117–27.

[7] Cohen, P. (2001) *Eur J Biochem.* **268(19)**, 5001–5010.

[8] Pearce, L. R., Komander, D., and Alessi, D. R. January 2010 *Nature reviews. Molecular cell biology* **11(1)**, 9–22.

[9] Akritopoulou-Zanze, I. and Hajduk, P. J. March 2009 *Drug discovery today* **14(5-6)**, 291–7.

[10] Zhang, C., Habets, G., and Bollag, G. November 2011 *Nature Biotechnology* **29(11)**, 981–983.

[11] Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. November 2006 *Cell* **127(3)**, 635–48.

[12] Scheeff, E. D. and Bourne, P. E. October 2005 *PLoS computational biology* **1(5)**, e49.

[13] Suga, H., Dacre, M., deMendoza, a., Shalchian-Tabrizi, K., Manning, G., and Ruiz-Trillo, I. May 2012 *Science Signaling* **5(222)**, ra35–ra35.

[14] Lochhead, P. a. January 2009 *Science signaling* **2(54)**, pe4.

[15] Steichen, J. M., Iyer, G. H., Li, S., Saldanha, S. A., Deal, M. S., Woods, V. L., and Taylor, S. S. February 2010 *The Journal of biological chemistry* **285(6)**, 3825–32.

[16] Leonard, C. J., Aravind, L., and Koonin, E. V. October 1998 *Genome research* **8(10)**, 1038–47.

[17] Ryazanov, A. G. March 2002 *FEBS letters* **514(1)**, 26–9.

[18] Yamaguchi, H., Matsushita, M., Nairn, a. C., and Kuriyan, J. May 2001 *Molecular cell* **7(5)**, 1047–57.

[19] Deshmukh, K., Anamika, K., and Srinivasan, N. January 2010 *Progress in biophysics and molecular biology* **102(1)**, 1–15.

[20] Hanks, S. K. and Hunter, T. May 1995 *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **9(8)**, 576–96.

[21] Taylor, S. S. and Radzio-Andzelm, E. May 1994 *Structure (London, England : 1993)* **2(5)**, 345–55.

[22] Koch, S., Tugues, S., Li, X., Gualandi, L., and Claesson-Welsh, L. July 2011 *The Biochemical journal* **437(2)**, 169–83.

[23] Strebhardt, K. August 2010 *Nature reviews. Drug discovery* **9(8)**, 643–60.

[24] Parsons, S. J. and Parsons, J. T. October 2004 *Oncogene* **23(48)**, 7906–9.

[25] Thomas, S. M. and Brugge, J. S. January 1997 *Annual review of cell and developmental biology* **13**, 513–609.

[26] Hers, I., Vincent, E. E., and Tavaré, J. M. October 2011 *Cellular signalling* **23(10)**, 1515–27.

[27] Alexander, J., Lim, D., Joughin, B. a., Hegemann, B., Hutchins, J. R. a., Ehrenberger, T., Ivins, F., Sessa, F., Hudecz, O., Nigg, E. a., Fry, A. M., Musacchio, A., Stukenberg, P. T., Mechtler, K., Peters, J.-M., Smerdon, S. J., and Yaffe, M. B. January 2011 *Science signaling* **4(179)**, ra42.

[28] Kettenbach, A. N., Schweppe, D. K., Faherty, B. K., Pechenick, D., Pletnev, A. a., and Gerber, S. a. January 2011 *Science signaling* **4(179)**, rs5.

[29] Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R., McCartney, R. R., Schmidt, M. C., Rachidi, N., Lee, S.-J., Mah, A. S., Meng, L., Stark, M. J. R., Stern, D. F., De Virgilio, C., Tyers, M., Andrews, B., Gerstein, M., Schweitzer, B., Predki, P. F., and Snyder, M. December 2005 *Nature* **438(7068)**, 679–84.

[30] House, C., Wettenhall, R. E., and Kemp, B. E. January 1987 *The Journal of biological chemistry* **262(2)**, 772–7.

[31] Kennelly, P. J. and Krebs, E. G. August 1991 *The Journal of biological chemistry* **266(24)**, 15555–8.

[32] Pinna, L. A. and Ruzzene, M. December 1996 *Biochimica et biophysica acta* **1314(3)**, 191–225.

[33] Kreegipuu, A., Blom, N., Brunak, S., and Ja, J. (1998) *FEBS letters* **430**, 45–50.

[34] Mok, J., Kim, P. M., Lam, H. Y. K., Piccirillo, S., Zhou, X., Jeschke, G. R., Sheridan, D. L., Parker, S. a., Desai, V., Jwa, M., Cameroni, E., Niu, H., Good, M., Remenyi, A., Ma, J.-L. N., Sheu, Y.-J., Sassi, H. E., Sopko, R., Chan, C. S. M., De Virgilio, C., Hollingsworth, N. M., Lim, W. a., Stern, D. F., Stillman, B., Andrews, B. J., Gerstein, M. B., Snyder, M., and Turk, B. E. January 2010 *Science signaling* **3(109)**, ra12.

[35] Hegemann, B., Hutchins, J. R. a., Hudecz, O., Novatchkova, M., Rameseder, J., Sykora, M. M., Liu, S., Mazanek, M., Lenart, P., Heriche, J.-K., Poser, I., Kraut, N., Hyman, a. a., Yaffe, M. B., Mechtler, K., and Peters, J.-M. November 2011 *Science Signaling* **4(198)**, rs12–rs12.

[36] Zhu, H., Klemic, J. F., Chang, S., Bertone, P., Casamayor, a., Klemic, K. G., Smith, D., Gerstein, M., Reed, M. a., and Snyder, M. November 2000 *Nature genetics* **26(3)**, 283–9.

[37] Zhu, G., Fujii, K., Belkina, N., Liu, Y., James, M., Herrero, J., and Shaw, S. March 2005 *The Journal of biological chemistry* **280(11)**, 10743–8.

[38] Zhu, G., Liu, Y., and Shaw, S. January 2005 *Cell cycle (Georgetown, Tex.)* **4(1)**, 52–6.

[39] Eswaran, J. and Knapp, S. March 2010 *Biochimica et biophysica acta* **1804(3)**, 429–32.

[40] Johnson, L. N. January 2001 *Ernst Schering Research Foundation workshop* **430(34)**, 47–69.

[41] Kornev, A. P., Haste, N. M., Taylor, S. S., and Eyck, L. F. T. November 2006 *Proceedings of the National Academy of Sciences of the United States of America* **103(47)**, 17783–8.

[42] Biondi, R. M. and Nebreda, A. R. May 2003 *The Biochemical journal* **372(Pt 1)**, 1–13.

[43] Holland, P. M. and Cooper, J. a. May 1999 *Current biology : CB* **9(9)**, R329–31.

[44] Tanoue, T., Adachi, M., Moriguchi, T., and Nishida, E. February 2000 *Nature cell biology* **2(2)**, 110–6.

[45] Reményi, A., Good, M. C., Bhattacharyya, R. P., and Lim, W. A. December 2005 *Molecular cell* **20(6)**, 951–62.

[46] Ubersax, J. A. and Ferrell, J. E. July 2007 *Nature reviews. Molecular cell biology* **8(7)**, 530–41.

[47] Fantz, D. a., Jacobs, D., Glossip, D., and Kornfeld, K. July 2001 *The Journal of biological chemistry* **276(29)**, 27256–65.

[48] Gavin, a. C. and Nebreda, a. R. March 1999 *Current biology : CB* **9(5)**, 281–4.

[49] vanVugt, M. A. T. M. and Medema, R. H. April 2005 *Oncogene* **24(17)**, 2844–59.

[50] Elia, A. E. H., Rellos, P., Haire, L. F., Chao, J. W., Ivins, F. J., Hoepker, K., Mohammad, D., Cantley, L. C., Smerdon, S. J., and Yaffe, M. B. October 2003 *Cell* **115(1)**, 83–95.

[51] van deWeerdt, B. C. M., Littler, D. R., Klompmaker, R., Huseinovic, A., Fish, A., Perrakis, A., and Medema, R. H. June 2008 *Biochimica et biophysica acta* **1783(6)**, 1015–22.

[52] Abel, T. and Nguyen, P. V. January 2008 *Progress in brain research* **169**, 97–115.

[53] McKnight, G. S., Cummings, D. E., Amieux, P. S., Sikorski, M. A., Brandon, E. P., Planas, J. V., Motamed, K., and Idzerda, R. L. January 1998 *Recent progress in hormone research* **53**, 139–59; discussion 160–1.

[54] Szaszák, M., Christian, F., Rosenthal, W., and Klussmann, E. April 2008 *Cellular signalling* **20(4)**, 590–601.

[55] Colledge, M. and Scott, J. D. June 1999 *Trends in cell biology* **9(6)**, 216–21.

[56] Kim, C., Xuong, N.-H., and Taylor, S. S. February 2005 *Science (New York, N.Y.)* **307(5710)**, 690–6.

[57] Kim, C., Cheng, C. Y., Saldanha, S. A., and Taylor, S. S. September 2007 *Cell* **130(6)**, 1032–43.

[58] Bloom, J. and Cross, F. R. February 2007 *Nature reviews. Molecular cell biology* **8(2)**, 149–60.

[59] Cheng, K.-Y., Noble, M. E. M., Skamnaki, V., Brown, N. R., Lowe, E. D., Kontogiannis, L., Shen, K., Cole, P. A., Siligardi, G., and Johnson, L. N. August 2006 *The Journal of biological chemistry* **281(32)**, 23167–79.

[60] Jeffrey, P. D., Russo, A. A., Polyak, K., Gibbs, E., Hurwitz, J., Massagué, J., and Pavletich, N. P. July 1995 *Nature* **376(6538)**, 313–20.

[61] Miller, M. E. and Cross, F. R. May 2001 *Journal of cell science* **114(Pt 10)**, 1811–20.

[62] Schulman, B. A., Lindstrom, D. L., and Harlow, E. September 1998 *Proceedings of the National Academy of Sciences of the United States of America* **95(18)**, 10453–8.

[63] Jackman, M., Lindon, C., Nigg, E. A., and Pines, J. March 2003 *Nature cell biology* **5(2)**, 143–8.

[64] Jackman, M., Firth, M., and Pines, J. April 1995 *The EMBO journal* **14(8)**, 1646–54.

[65] Hagting, A., Jackman, M., Simpson, K., and Pines, J. July 1999 *Current biology : CB* **9(13)**, 680–9.

[66] Miller, W. T. June 2003 *Accounts of chemical research* **36(6)**, 393–400.

[67] Kobe, B., Kampmann, T., and Forwood, J. K. (2005) *Biochimica et biophysica acta* **1754**, 200 – 209.

[68] Brinkworth, R. I., Breinl, R. A., and Kobe, B. (2002) *Proceedings of the National Academy of Sciences of the United States of America* **100(1)**, 74–79.

[69] Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., and Gibson, T. J. January 2012 *Molecular bioSystems* **8(1)**, 268–81.

[70] Miller, M. L., Jensen, L. J., Diella, F., Jø rgensen, C., Tinti, M., Li, L., Hsiung, M., Parker, S. a., Bordeaux, J., Sicheritz-Ponten, T., Olhovsky, M., Pasculescu, A., Alexander, J., Knapp, S., Blom, N., Bork, P., Li, S., Cesareni, G., Pawson, T., Turk, B. E., Yaffe, M. B., Brunak, S. r., and Linding, R. January 2008 *Science Signaling* **1(35)**, ra2.

[71] Stein, A. and Aloy, P. May 2010 *PLoS Computational Biology* **6(5)**, e1000789.

[72] Newman, R. H., Hu, J., Rho, H.-S., Xie, Z., Woodard, C., Neiswinger, J., Cooper, C., Shirley, M., Clark, H. M., Hu, S., Hwang, W., Seop Jeong, J., Wu, G., Lin, J., Gao, X., Ni, Q., Goel, R., Xia, S., Ji, H., Dalby, K. N., Birnbaum, M. J., Cole, P. a., Knapp, S., Ryazanov, A. G., Zack, D. J., Blackshaw, S., Pawson, T., Gingras, A.-C., Desiderio, S., Pandey, A., Turk, B. E., Zhang, J., Zhu, H., and Qian, J. April 2013 *Molecular Systems Biology* **9(655)**, 1–12.

[73] Gonzalez, F. a., Raden, D. L., and Davis, R. J. November 1991 *The Journal of biological chemistry* **266(33)**, 22159–63.

[74] Kim, S. T., Lim, D. S., Canman, C. E., and Kastan, M. B. December 1999 *The Journal of biological chemistry* **274(53)**, 37538–43.

[75] Songyang, Z. and Cantley, L. C. (1995) *Trends in biochemical sciences* **20(11)**, 470–475.

[76] Hjerrild, M. and Gammeltoft, S. September 2006 *FEBS letters* **580(20)**, 4764–70.

[77] Yaffe, M. B., Leparc, G. G., Lai, J., Obata, T., Volinia, S., and Cantley, L. C. April 2001 *Nature biotechnology* **19(4)**, 348–53.

[78] D'haeseleer, P. April 2006 *Nature biotechnology* **24(4)**, 423–5.

[79] Schneider, T. D. and Stephens, R. M. October 1990 *Nucleic acids research* **18(20)**, 6097–100.

[80] Shannon, C. E. (1948) *Bell System Technical Journal* **27**, 379–423, 623–656.

[81] Crooks, G. E., Hon, G., Chandonia, J.-m., and Brenner, S. E. (2004) *Genome research* **14(6)**, 1188–1190.

[82] Hertz, G. Z. and Stormo, G. D. (1999) *Bioinformatics (Oxford, England)* **15(7-8)**, 563–77.

[83] Touzet, H. and Varré, J.-S. January 2007 *Algorithms for molecular biology : AMB* **2**, 15.

[84] Claverie, J. M. and Audic, S. Oct 1996 *Computer applications in the biosciences : CABIOS* **12(5)**, 431–9.

[85] Beckstette, M., Homann, R., Giegerich, R., and Kurtz, S. January 2006 *BMC bioinformatics* **7**, 389.

[86] Korhonen, J., Martinmäki, P., Pizzi, C., Rastas, P., and Ukkonen, E. December 2009 *Bioinformatics (Oxford, England)* **25(23)**, 3181–2.

[87] Schneider, T. R., Stormo, G. D., Gold, L., Ehrenfeuch, A., and Schneider T.D, Stormo G.D, Gold D, E. A. (1986) *Journal of molecular biology* **188**, 415–431.

[88] Bulyk, M. L., Huang, X., Choo, Y., and Church, G. M. June 2001 *Proceedings of the National Academy of Sciences of the United States of America* **98(13)**, 7158–63.

[89] Stormo, G. D. April 2011 *Genetics* **187(4)**, 1219–24.

[90] Marinescu, V. D., Kohane, I. S., and Riva, A. January 2005 *BMC bioinformatics* **6**, 79.

[91] Naughton, B., Fratkin, E., Batzoglou, S., and Brutlag, D. Oct 2006 *Nucleic acids research* **34(20)**, 5730–9.

[92] Attwood, T. K. September 2002 *Briefings in bioinformatics* **3(3)**, 252–63.

[93] Portales-Casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W. W., and Sandelin, A. January 2010 *Nucleic acids research* **38(Database issue)**, D105–10.

[94] Sigrist, C. J. A., Cerutti, L., deCastro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., and Hulo, N. January 2010 *Nucleic acids research* **38(Database issue)**, D161–6.

[95] Wingender, E., Dietze, P., Karas, H., and Knüppel, R. January 1996 *Nucleic acids research* **24(1)**, 238–41.

[96] Beausoleil, S. a., Jedrychowski, M., Schwartz, D., Elias, J. E., Villén, J., Li, J., Cohn, M. a., Cantley, L. C., and Gygi, S. P. August 2004 *Proceedings of the National Academy of Sciences of the United States of America* **101(33)**, 12130–5.

[97] Dephoure, N., Zhou, C., Villén, J., Beausoleil, S. a., Bakalarski, C. E., Elledge, S. J., and Gygi, S. P. August 2008 *Proceedings of the National Academy of Sciences of the United States of America* **105(31)**, 10762–7.

[98] Daub, H., Olsen, J. V., Bairlein, M., Gnad, F., Oppermann, F. S., Körner, R., Greff, Z., Kéri, G., Stemmann, O., and Mann, M. August 2008 *Molecular cell* **31(3)**, 438–48.

[99] Olsen, J. V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M. L., Jensen, L. J., Gnad, F., Cox, J., Jensen, T. S., Nigg, E. a., Brunak, S. r., and Mann, M. January 2010 *Science signaling* **3(104)**, ra3.

[100] Pan, C., Olsen, J. V., Daub, H., and Mann, M. December 2009 *Molecular & cellular proteomics : MCP* **8(12)**, 2796–808.

[101] Obenauer, J. C. July 2003 *Nucleic Acids Research* **31(13)**, 3635–3641.

[102] Saunders, N. F. W., Brinkworth, R. I., Huber, T., Kemp, B. E., and Kobe, B. (2008) *BMC Bioinformatics* **11**, 1–11.

[103] Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. r. June 2004 *Proteomics* **4(6)**, 1633–49.

[104] Diella, F., Cameron, S., Gemünd, C., Linding, R., Via, A., Kuster, B., Sicheritz-Pontén, T., Blom, N., and Gibson, T. J. June 2004 *BMC bioinformatics* **5**, 79.

[105] Xue, Y., Ren, J., Gao, X., Jin, C., Wen, L., and Yao, X. September 2008 *Molecular & cellular proteomics : MCP* **7(9)**, 1598–608.

[106] Presnell, S. R. and Cohen, F. E. January 1993 *Annual review of biophysics and biomolecular structure* **22**, 283–98.

[107] Gao, J., Thelen, J. J., Dunker, a. K., and Xu, D. December 2010 *Molecular & cellular proteomics : MCP* **9(12)**, 2586–600.

[108] Edwards, R. J., Davey, N. E., and Shields, D. C. January 2007 *PloS one* **2(10)**, e967.

[109] Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. January 2009 *Nucleic acids research* **37(Database issue)**, D767–72.

[110] Hornbeck, P. V., Chabra, I., Kornhauser, J. M., Skrzypek, E., and Zhang, B. June 2004 *Proteomics* **4(6)**, 1551–61.

[111] Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., and Diella, F. November 2010 *Nucleic acids research* **39(November 2010)**, 261–267.

[112] The UniProt Consortium January 2012 *Nucleic acids research* **40(Database issue)**, D71–5.

[113] Li, W. and Godzik, A. July 2006 *Bioinformatics (Oxford, England)* **22(13)**, 1658–9.

[114] Stormo, G. D. January 2000 *Bioinformatics (Oxford, England)* **16(1)**, 16–23.

[115] Lasko, T. a., Bhagwat, J. G., Zou, K. H., and Ohno-Machado, L. October 2005 *Journal of biomedical informatics* **38(5)**, 404–15.

[116] Swamidass, S. J., Azencott, C.-A., Daily, K., and Baldi, P. May 2010 *Bioinformatics (Oxford, England)* **26(10)**, 1348–56.

[117] Krishna, M. and Narang, H. November 2008 *Cellular and molecular life sciences : CMLS* **65(22)**, 3525–44.

[118] Clark-Lewis, I., Sanghera, J. S., and Pelech, S. L. August 1991 *The Journal of biological chemistry* **266(23)**, 15180–4.

[119] Fiol, C. J., Mahrenholzs, A. M., Wangg, Y., Roeske, R. W., and Roach, P. J. (1987) *The Journal of biological chemistry* **262**, 14042–14048.

[120] Abraham, R. T. (2004) *DNA repair* **3(8-9)**, 883–7.

[121] Abraham, R. T. and Tibbetts, R. S. April 2005 *Science (New York, N.Y.)* **308(5721)**, 510–1.

[122] Chiang, G. G. and Abraham, R. T. October 2007 *Trends in molecular medicine* **13(10)**, 433–42.

[123] Mordes, D. A. and Cortez, D. September 2008 *Cell cycle (Georgetown, Tex.)* **7(18)**, 2809–12.

[124] Blom, N., Gammeltoft, S., and Brunak, S. December 1999 *Journal of molecular biology* **294(5)**, 1351–62.

[125] Matsuzaki, H., Ichino, A., Hayashi, T., Yamamoto, T., and Kikkawa, U. October 2005 *Journal of biochemistry* **138(4)**, 485–91.

[126] Kielbassa, K., Müller, H. J., Meyer, H. E., Marks, F., and Gschwendt, M. March 1995 *The Journal of biological chemistry* **270(11)**, 6156–62.

[127] Racioppi, L. and Means, A. R. December 2008 *Trends in immunology* **29(12)**, 600–7.

[128] Wayman, G. a., Tokumitsu, H., Davare, M. a., and Soderling, T. R. July 2011 *Cell calcium* **50(1)**, 1–8.

[129] Songyang, Z., Lu, K. P., Kwon, Y. T., Tsai, L. H., Filhol, O., Cochet, C., Brickey, D. a., Soderling, T. R., Bartleson, C., Graves, D. J., DeMaggio, a. J., Hoekstra, M. F., Blenis, J., Hunter, T., and Cantley, L. C. November 1996 *Molecular and cellular biology* **16(11)**, 6486–93.

[130] Mahadevan, D. and Beeck, S. July 2007 *Expert opinion on drug discovery* **2(7)**, 1011–26.

[131] Ferrari, S., Marin, O., Pagano, M. a., Meggio, F., Hess, D., El-Shemerly, M., Krystyniak, A., and Pinna, L. a. August 2005 *The Biochemical journal* **390(Pt 1)**, 293–302.

[132] Johnson, L. N. June 2011 *Science Signaling* **4(179)**, pe31–pe31.

[133] Sardon, T., Pache, R. a., Stein, A., Molina, H., Vernos, I., and Aloy, P. December 2010 *EMBO reports* **11(12)**, 977–84.

[134] Brown, J. L., Stowers, L., Baer, M., Trejo, J., Coughlin, S., and Chant, J. May 1996 *Current biology : CB* **6(5)**, 598–605.

[135] Yao, Z., Zhou, G., Wang, X. S., Brown, A., Diener, K., Gan, H., and Tan, T. H. January 1999 *The Journal of biological chemistry* **274(4)**, 2118–25.

[136] Cheong, J. K. and Virshup, D. M. April 2011 *The international journal of biochemistry & cell biology* **43(4)**, 465–9.

[137] Knippschild, U., Gocht, A., Wolff, S., Huber, N., Löhler, J., and Stöter, M. June 2005 *Cellular signalling* **17(6)**, 675–89.

[138] Flotow, H., Graves, P. R., Wang, a. Q., Fiol, C. J., Roeske, R. W., and Roach, P. J. August 1990 *The Journal of biological chemistry* **265(24)**, 14264–9.

[139] Zhai, L., Graves, P. R., Robinson, L. C., Italiano, M., Culbertson, M. R., Rowles, J., Cobb, M. H., DePaoli-Roach, A. A., and Roach, P. J. May 1995 *The Journal of biological chemistry* **270(21)**, 12717–24.

[140] Pulgar, V., Marin, O., Meggio, F., Allende, C. C., Allende, J. E., and Pinna, L. a. March 1999 *European journal of biochemistry / FEBS* **260(2)**, 520–6.

[141] Brunati, A. M., Donella-Deana, A., Ruzzene, M., Marin, O., and Pinna, L. a. June 1995 *FEBS letters* **367(2)**, 149–52.

[142] Rena, G., Woods, Y. L., Prescott, A. R., Peggie, M., Unterman, T. G., Williams, M. R., and Cohen, P. May 2002 *The EMBO journal* **21(9)**, 2263–71.

[143] Trembley, J. H., Chen, Z., Unger, G., Slaton, J., Kren, B. T., Van Waes, C., and Ahmed, K. (2010) *BioFactors (Oxford, England)* **36(3)**, 187–95.

[144] Sarno, S. and Pinna, L. a. September 2008 *Molecular bioSystems* **4(9)**, 889–94.

[145] Meggio, F. and Pinna, L. a. March 2003 *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **17(3)**, 349–68.

[146] Pinna, L. A. October 2002 *Journal of Cell Science* **115(20)**, 3873–3878.

[147] Häcker, H. and Karin, M. October 2006 *Science's STKE : signal transduction knowledge environment* **2006(357)**, re13.

[148] Chau, T., Gioia, R., Gatot, J., Patrascu, F., Carpentier, I., Chapelle, J., O'Neill, L., Beyaert, R., Piette, J., and A, C. April 2008 *Trends in biochemical sciences* **33(4)**, 171–80.

[149] Heissmeyer, V., Krappmann, D., Hatada, E. N., and Scheidereit, C. February 2001 *Molecular and cellular biology* **21(4)**, 1024–35.

[150] Lang, V., Janzen, J., Fischer, G. Z., Soneji, Y., Beinke, S., Salmeron, A., Allen, H., Hay, R. T., Ben-Neriah, Y., and Ley, S. C. January 2003 *Molecular and cellular biology* **23(1)**, 402–13.

[151] Takai, N., Hamanaka, R., Yoshimatsu, J., and Miyakawa, I. January 2005 *Oncogene* **24(2)**, 287–91.

[152] Malumbres, M. July 2011 *Physiological reviews* **91(3)**, 973–1007.

[153] Luo, J. and Liu, X. March 2012 *Protein & cell* **3(3)**, 182–97.

[154] Nakajima, H., Toyoshima-Morimoto, F., Taniguchi, E., and Nishida, E. July 2003 *The Journal of biological chemistry* **278(28)**, 25277–80.

[155] Fischer, A., Picard, C., Chemin, K., Dogniaux, S., leDeist, F., and Hivroz, C. June 2010 *Seminars in immunopathology* **32(2)**, 107–16.

[156] Mócsai, A., Ruland, J., and Tybulewicz, V. L. J. June 2010 *Nature reviews. Immunology* **10(6)**, 387–402.

[157] Donella-Deana, A., Marin, O., Brunati, A. M., Cesaro, L., Piutti, C., and Pinna, L. A. September 1993 *FEBS Letters* **330(2)**, 141–145.

[158] Huang, H.-D., Lee, T.-Y., Tzeng, S.-W., Wu, L.-C., Horng, J.-T., Tsou, A.-P., and Huang, K.-T. July 2005 *Journal of computational chemistry* **26(10)**, 1032–41.

[159] Kim, J.-Y., Huh, K., Jung, R., and Kim, T. J. March 2011 *Immunology letters* **135(1-2)**, 151–7.

[160] Pawson, T. and Scott, J. D. December 1997 *Science* **278(5346)**, 2075–2080.

[161] Choi, K. Y., Satterberg, B., Lyons, D. M., and Elion, E. A. August 1994 *Cell* **78(3)**, 499–512.

[162] Therrien, M., Michaud, N. R., Rubin, G. M., and Morrison, D. K. November 1996 *Genes & Development* **10(21)**, 2684–2695.

[163] Tsunoda, S., Sierralta, J., Sun, Y., Bodner, R., Suzuki, E., Becker, A., Socolich, M., and Zuker, C. S. July 1997 *Nature* **388(6639)**, 243–9.

[164] Zhang, W., Sloan-Lancaster, J., Kitchen, J., Trible, R. P., and Samelson, L. E. January 1998 *Cell* **92(1)**, 83–92.

[165] Good, M. C., Zalatan, J. G., and Lim, W. a. May 2011 *Science* **332(6030)**, 680–686.

[166] White, C. D., Brown, M. D., and Sacks, D. B. June 2009 *FEBS letters* **583(12)**, 1817–24.

[167] Cance, W. G., Kurenova, E., Marlowe, T., and Golubovskaya, V. March 2013 *Science Signaling* **6(268)**, pe10–pe10.

[168] Zhang, H., Photiou, A., Arnhild, G., Stebbing, J., and Giamas, G. February 2012 *Cellular signalling* **24(6)**, 1173–1184.

[169] Flynn, D. C. October 2001 *Oncogene* **20(44)**, 6270–2.

[170] Buday, L. and Tompa, P. November 2010 *The FEBS journal* **277(21)**, 4348–55.

[171] Morrison, D. K. and Davis, R. J. January 2003 *Annual review of cell and developmental biology* **19**, 91–118.

[172] Bhattacharyya, R. P., Reményi, A., Yeh, B. J., and Lim, W. A. January 2006 *Annual review of biochemistry* **75**, 655–80.

[173] Shaw, A. S. and Filbert, E. L. January 2009 *Nature reviews. Immunology* **9(1)**, 47–56.

[174] Ferrell, J. E. October 2000 *Science's STKE : signal transduction knowledge environment* **2000(52)**, pe1.

[175] Kolch, W. November 2005 *Nature reviews. Molecular cell biology* **6(11)**, 827–37.

[176] Dhanasekaran, D. N., Kashef, K., Lee, C. M., Xu, H., and Reddy, E. P. May 2007 *Oncogene* **26(22)**, 3185–202.

[177] Zeke, A., Lukács, M., Lim, W. a., and Reményi, A. August 2009 *Trends in cell biology* **19(8)**, 364–74.

[178] Brennan, D. F., Dar, A. C., Hertz, N. T., Chao, W. C. H., Burlingame, A. L., Shokat, K. M., and Barford, D. April 2011 *Nature* **472(7343)**, 366–9.

[179] Brown, M. D. and Sacks, D. B. April 2009 *Cellular Signalling* **21(4)**, 462–469.

[180] Elion, E. A. November 2001 *Journal of cell science* **114(Pt 22)**, 3967–78.

[181] Vial, E., Sahai, E., and Marshall, C. J. July 2003 *Cancer cell* **4(1)**, 67–79.

[182] Lo, H.-W. December 2010 *Current cancer drug targets* **10(8)**, 840–8.

[183] Udell, C. M., Rajakulendran, T., Sicheri, F., and Therrien, M. February 2011 *Cellular and molecular life sciences : CMLS* **68(4)**, 553–65.

[184] Nimnual, A. and Bar-Sagi, D. August 2002 *Science's STKE : signal transduction knowledge environment* **2002(145)**, pe36.

[185] Giubellino, A., Burke, T. R., and Bottaro, D. P. August 2008 *Expert opinion on therapeutic targets* **12(8)**, 1021–33.

[186] Michaud, N. R., Therrien, M., Cacace, a., Edsall, L. C., Spiegel, S., Rubin, G. M., and Morrison, D. K. November 1997 *Proceedings of the National Academy of Sciences of the United States of America* **94(24)**, 12792–6.

[187] Rajakulendran, T., Sahmi, M., Lefrançois, M., Sicheri, F., and Therrien, M. September 2009 *Nature* **461(7263)**, 542–5.

[188] Roy, F., Laberge, G., Douziech, M., Ferland-McCollough, D., and Therrien, M. February 2002 *Genes & development* **16(4)**, 427–38.

[189] Morrison, D. K. May 2001 *Journal of cell science* **114(Pt 9)**, 1609–12.

[190] Hu, J., Yu, H., Kornev, A. P., Zhao, J., Filbert, E. L., Taylor, S. S., and Shaw, A. S. April 2011 *Proceedings of the National Academy of Sciences of the United States of America* **108(15)**, 6067–72.

[191] Jacobs, D., Glossip, D., Xing, H., Muslin, a. J., and Kornfeld, K. January 1999 *Genes & development* **13(2)**, 163–75.

[192] Roskoski, R. January 2012 *Biochemical and biophysical research communications* **417(1)**, 5–10.

[193] McKay, M. M., Ritt, D. a., and Morrison, D. K. July 2009 *Proceedings of the National Academy of Sciences of the United States of America* **106(27)**, 11022–7.

[194] Pan, C. Q., Sudol, M., Sheetz, M., and Low, B. C. November 2012 *Cellular signalling* **24(11)**, 2143–65.

[195] Shannon, K. B. January 2012 *International journal of cell biology* **2012**, 894817.

[196] Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. May 1998 *Proceedings of the National Academy of Sciences of the United States of America* **95(11)**, 5857–64.

[197] Fukata, M., Kuroda, S., Fujii, K., Nakamura, T., Shoji, I., Matsuura, Y., Okawa, K., Iwamatsu, A., Kikuchi, A., and Kaibuchi, K. November 1997 *The Journal of biological chemistry* **272(47)**, 29579–83.

[198] Mateer, S. C., McDaniel, A. E., Nicolas, V., Habermacher, G. M., Lin, M.-J. S., Cromer, D. A., King, M. E., and Bloom, G. S. April 2002 *The Journal of biological chemistry* **277(14)**, 12324–33.

[199] White, C. D., Erdemir, H. H., and Sacks, D. B. April 2012 *Cellular signalling* **24(4)**, 826–34.

[200] Mateer, S. C., Wang, N., and Bloom, G. S. (2003) *Cell Motility and the Cytoskeleton* **55(February)**, 147–155.

[201] Benseñor, L. B., Kan, H.-M., Wang, N., Wallrabe, H., Davidson, L. A., Cai, Y., Schafer, D. A., and Bloom, G. S. February 2007 *Journal of cell science* **120(Pt 4)**, 658–69.

[202] Pelikan-Conchaudron, A., Le Clainche, C., Didry, D., and Carlier, M.-F. October 2011 *The Journal of biological chemistry* **286(40)**, 35119–28.

[203] Roy, M., Li, Z., and Sacks, D. B. September 2005 *Molecular and cellular biology* **25(18)**, 7940–52.

[204] Ren, J.-G., Li, Z., and Sacks, D. B. August 2008 *The Journal of biological chemistry* **283(34)**, 22972–82.

[205] Ren, J.-G., Li, Z., Crimmins, D. L., and Sacks, D. B. October 2005 *The Journal of biological chemistry* **280(41)**, 34548–57.

[206] Briggs, M. W. and Sacks, D. B. May 2003 *FEBS Letters* **542(1-3)**, 7–11.

[207] Briggs, M. W. and Sacks, D. B. June 2003 *EMBO reports* **4(6)**, 571–4.

[208] Mataraza, J. M., Briggs, M. W., Li, Z., Entwistle, A., Ridley, A. J., and Sacks, D. B. October 2003 *The Journal of biological chemistry* **278(42)**, 41237–45.

[209] David, M., Petit, D., and Bertoglio, J. August 2012 *Cell cycle (Georgetown, Tex.)* **11(16)**, 3003–10.

[210] Ho, Y.-D. January 1999 *Journal of Biological Chemistry* **274(1)**, 464–470.

[211] Wang, S., Watanabe, T., Noritake, J., Fukata, M., Yoshimura, T., Itoh, N., Harada, T., Nakagawa, M., Matsuura, Y., Arimura, N., and Kaibuchi, K. February 2007 *Journal of cell science* **120(Pt 4)**, 567–77.

[212] Owen, D., Campbell, L. J., Littlefield, K., Evetts, K. A., Li, Z., Sacks, D. B., Lowe, P. N., and Mott, H. R. January 2008 *The Journal of biological chemistry* **283(3)**, 1692–704.

[213] Noritake, J., Watanabe, T., Sato, K., Wang, S., and Kaibuchi, K. May 2005 *Journal of cell science* **118(Pt 10)**, 2085–92.

[214] Kuroda, S., Fukata, M., Nakagawa, M., Fujii, K., Nakamura, T., Ookubo, T., Izawa, I., Nagase, T., Nomura, N., Tani, H., Shoji, I., Matsuura, Y., Yonehara, S., and Kaibuchi, K. August 1998 *Science (New York, N.Y.)* **281(5378)**, 832–5.

[215] Fukata, M., Nakagawa, M., Itoh, N., Kawajiri, A., Yamaga, M., Kuroda, S., and Kaibuchi, K. March 2001 *Molecular and cellular biology* **21(6)**, 2165–83.

[216] Fukata, M., Watanabe, T., Noritake, J., Nakagawa, M., Yamaga, M., Kuroda, S., Matsuura, Y., Iwamatsu, A., Perez, F., and Kaibuchi, K. June 2002 *Cell* **109(7)**, 873–85.

[217] Schuyler, S. C. and Pellman, D. May 2001 *Cell* **105(4)**, 421–4.

[218] Gundersen, G. G. October 2002 *Current biology : CB* **12(19)**, R645–7.

[219] Ridley, A. J. August 2001 *Journal of cell science* **114(Pt 15)**, 2713–22.

[220] Sahai, E. and Marshall, C. J. February 2002 *Nature reviews. Cancer* **2(2)**, 133–42.

[221] Drugan, J. K., Rogers-Graham, K., Gilmer, T., Campbell, S., and Clark, G. J. November 2000 *The Journal of biological chemistry* **275(45)**, 35021–7.

[222] Sugimoto, N., Imoto, I., Fukuda, Y., Kurihara, N., Kuroda, S., Tanigami, A., Kaibuchi, K., Kamiyama, R., and Inazawa, J. January 2001 *Journal of human genetics* **46(1)**, 21–5.

[223] Nabeshima, K., Shimao, Y., Inoue, T., and Koono, M. February 2002 *Cancer letters* **176(1)**, 101–9.

[224] Takemoto, H., Doki, Y., Shiozaki, H., Imamura, H., Utsunomiya, T., Miyata, H., Yano, M., Inoue, M., Fujiwara, Y., and Monden, M. March 2001 *International journal of cancer. Journal international du cancer* **91(6)**, 783–8.

[225] Alexa, A., Varga, J., and Reményi, A. November 2010 *The FEBS journal* **277(21)**, 4376–82.

[226] Brummer, T., Schmitz-Peiffer, C., and Daly, R. J. November 2010 *The FEBS journal* **277(21)**, 4356–69.

[227] Logue, J. S. and Scott, J. D. November 2010 *The FEBS journal* **277(21)**, 4370–5.

[228] Brandon, E. P., Idzerda, R. L., and McKnight, G. S. June 1997 *Current opinion in neurobiology* **7(3)**, 397–403.

[229] Taskén, K., Skå lhegg, B. S., Taskén, K. A., Solberg, R., Knutsen, H. K., Levy, F. O., Sandberg, M., Orstavik, S., Larsen, T., Johansen, A. K., Vang, T., Schrader, H. P., Reinton, N. T., Torgersen, K. M., Hansson, V., and Jahnsen, T. January 1997 *Advances in second messenger and phosphoprotein research* **31**, 191–204.

[230] Lester, L. B. and Scott, J. D. January 1997 *Recent progress in hormone research* **52**, 409–29; discussion 429–30.

[231] Coghlan, V. M., Perrino, B. A., Howard, M., Langeberg, L. K., Hicks, J. B., Gallatin, W. M., and Scott, J. D. January 1995 *Science (New York, N.Y.)* **267(5194)**, 108–11.

[232] Klauck, T. M., Faux, M. C., Labudda, K., Langeberg, L. K., Jaken, S., and Scott, J. D. March 1996 *Science (New York, N.Y.)* **271(5255)**, 1589–92.

[233] Dodge-Kafka, K. L., Soughayer, J., Pare, G. C., Carlisle Michel, J. J., Langeberg, L. K., Kapiloff, M. S., and Scott, J. D. September 2005 *Nature* **437(7058)**, 574–8.

[234] Feliciello, A., Gottesman, M. E., and Avvedimento, E. V. April 2001 *Journal of molecular biology* **308(2)**, 99–114.

[235] Dodge, K. L., Khouangsathiene, S., Kapiloff, M. S., Mouton, R., Hill, E. V., Houslay, M. D., Langeberg, L. K., and Scott, J. D. April 2001 *The EMBO journal* **20(8)**, 1921–30.

[236] Kapiloff, M. S., Jackson, N., and Airhart, N. September 2001 *Journal of cell science* **114(Pt 17)**, 3167–76.

[237] Pare, G. C., Bauman, A. L., McHenry, M., Michel, J. J. C., Dodge-Kafka, K. L., and Kapiloff, M. S. December 2005 *Journal of cell science* **118(Pt 23)**, 5637–46.

[238] Carlisle Michel, J. J., Dodge, K. L., Wong, W., Mayer, N. C., Langeberg, L. K., and Scott, J. D. August 2004 *The Biochemical journal* **381(Pt 3)**, 587–92.

[239] Diviani, D., Dodge-Kafka, K. L., Li, J., and Kapiloff, M. S. November 2011 *American journal of physiology. Heart and circulatory physiology* **301(5)**, H1742–53.

[240] Carnegie, G. K., Means, C. K., and Scott, J. D. April 2009 *IUBMB life* **61(4)**, 394–406.

[241] Dodge-Kafka, K. L. and Kapiloff, M. S. July 2006 *European journal of cell biology* **85(7)**, 593–602.

[242] Pawson, T. April 2007 *Current opinion in cell biology* **19(2)**, 112–6.

[243] Ramírez, F. and Albrecht, M. January 2010 *Trends in cell biology* **20(1)**, 2–4.

[244] Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., and Hermjakob, H. January 2012 *Nucleic acids research* **40(Database issue)**, D841–6.

[245] Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A. P., Santonico, E., Castagnoli, L., and Cesareni, G. January 2012 *Nucleic acids research* **40(Database issue)**, D857–61.

[246] Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. January 2004 *Nucleic acids research* **32(Database issue)**, D449–51.

[247] Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., and Finn, R. D. January 2012 *Nucleic acids research* **40(Database issue)**, D290–301.

[248] Huang, D. W., Sherman, B. T., and Lempicki, R. a. January 2009 *Nucleic acids research* **37(1)**, 1–13.

[249] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. May 2000 *Nature genetics* **25(1)**, 25–9.

[250] Falcon, S. and Gentleman, R. January 2007 *Bioinformatics (Oxford, England)* **23(2)**, 257–8.

[251] Csardi, G. and Nepusz, T. (2006) *InterJournal* **Complex Sy**, 1695.

[252] Gardino, A. K. and Yaffe, M. B. September 2011 *Seminars in cell & developmental biology* **22(7)**, 688–95.

[253] Obsil, T. and Obsilova, V. September 2011 *Seminars in cell & developmental biology* **22(7)**, 663–72.

[254] Boehm, M. and Bonifacino, J. S. March 2002 *Gene* **286(2)**, 175–86.

[255] Happel, N., Höning, S., Neuhaus, J.-M., Paris, N., Robinson, D. G., and Holstein, S. E. H. March 2004 *The Plant journal : for cell and molecular biology* **37(5)**, 678–93.

[256] Hardie, D. G. October 2007 *Nature reviews. Molecular cell biology* **8(10)**, 774–85.

[257] Lajoie, P., Goetz, J. G., Dennis, J. W., and Nabi, I. R. May 2009 *The Journal of cell biology* **185(3)**, 381–5.

[258] Serra, M. and Scotlandi, K. November 2009 *Cancer letters* **284(2)**, 113–21.

[259] McMahon, H. T. and Mills, I. G. August 2004 *Current opinion in cell biology* **16(4)**, 379–91.

[260] Touz, M. C., Kulakova, L., and Nash, T. E. July 2004 *Molecular biology of the cell* **15(7)**, 3053–60.

[261] Sarikas, A., Xu, X., Field, L. J., and Pan, Z.-Q. October 2008 *Cell cycle (Georgetown, Tex.)* **7(20)**, 3154–61.

[262] Zimmerman, E. S., Schulman, B. A., and Zheng, N. December 2010 *Current opinion in structural biology* **20(6)**, 714–21.

[263] Reiss, K., Del Valle, L., Lassak, A., and Trojanek, J. August 2012 *Journal of cellular physiology* **227(8)**, 2992–3000.

[264] Shaw, L. M. June 2011 *Cell cycle (Georgetown, Tex.)* **10(11)**, 1750–6.

[265] Roberts, S., Delury, C., and Marsh, E. October 2012 *The FEBS journal* **279(19)**, 3549–58.

[266] Thomson, T. M., Benjamin, K. R., Bush, A., Love, T., Pincus, D., Resnekov, O., Yu, R. C., Gordon, A., Colman-Lerner, A., Endy, D., and Brent, R. December 2011 *Proceedings of the National Academy of Sciences of the United States of America* **108(50)**, 20265–70.

[267] Scheffzek, K. and Welti, S. August 2012 *FEBS letters* **586(17)**, 2662–73.

[268] Dhe-Paganon, S., Werner, E. D., Nishi, M., Hansen, L., Chi, Y.-I., and Shoelson, S. E. October 2004 *Nature structural & molecular biology* **11(10)**, 968–74.

[269] Bork, P. and Margolis, B. March 1995 *Cell* **80(5)**, 693–4.

[270] Schlessinger, J. and Lemmon, M. A. July 2003 *Science's STKE : signal transduction knowledge environment* **2003(191)**, RE12.

[271] Filippakopoulos, P., Müller, S., and Knapp, S. December 2009 *Current opinion in structural biology* **19(6)**, 643–9.

[272] Liu, B. a., Shah, E., Jablonowski, K., Stergachis, a., Engelmann, B., and Nash, P. D. December 2011 *Science Signaling* **4(202)**, ra83–ra83.

[273] Mayer, B. J. April 2001 *Journal of cell science* **114(Pt 7)**, 1253–63.

[274] Reebye, V., Frilling, A., Hajitou, A., Nicholls, J. P., Habib, N. A., and Mintz, P. J. February 2012 *Cellular signalling* **24(2)**, 388–92.

[275] Saksela, K. and Permi, P. August 2012 *FEBS letters* **586(17)**, 2609–14.

[276] Kami, K., Takeya, R., Sumimoto, H., and Kohda, D. August 2002 *The EMBO journal* **21(16)**, 4268–76.

[277] Karthikeyan, S., Leung, T., and Ladias, J. A. A. May 2002 *The Journal of biological chemistry* **277(21)**, 18973–8.

[278] Jiang, K., Pereira, E., Maxfield, M., Russell, B., Goudelock, D. M., and Sanchez, Y. July 2003 *The Journal of biological chemistry* **278(27)**, 25207–17.

[279] Matenia, D. and Mandelkow, E.-M. July 2009 *Trends in biochemical sciences* **34(7)**, 332–42.

[280] Alonso, A., Zaidi, T., Novak, M., Grundke-Iqbal, I., and Iqbal, K. June 2001 *Proceedings of the National Academy of Sciences of the United States of America* **98(12)**, 6923–8.

[281] Badis, G., Berger, M. F., Philippakis, A. a., Talukder, S., Gehrke, A. R., Jaeger, S. a., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C.-F., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R., and Bulyk, M. L. June 2009 *Science (New York, N.Y.)* **324(5935)**, 1720–3.

[282] Penman, G. A., Leung, L., and Näthke, I. S. October 2005 *Journal of cell science* **118(Pt 20)**, 4741–50.

[283] Reichen, C., Hansen, S., and Plückthun, A. August 2013 *Journal of structural biology* **In Press**, 0–0.

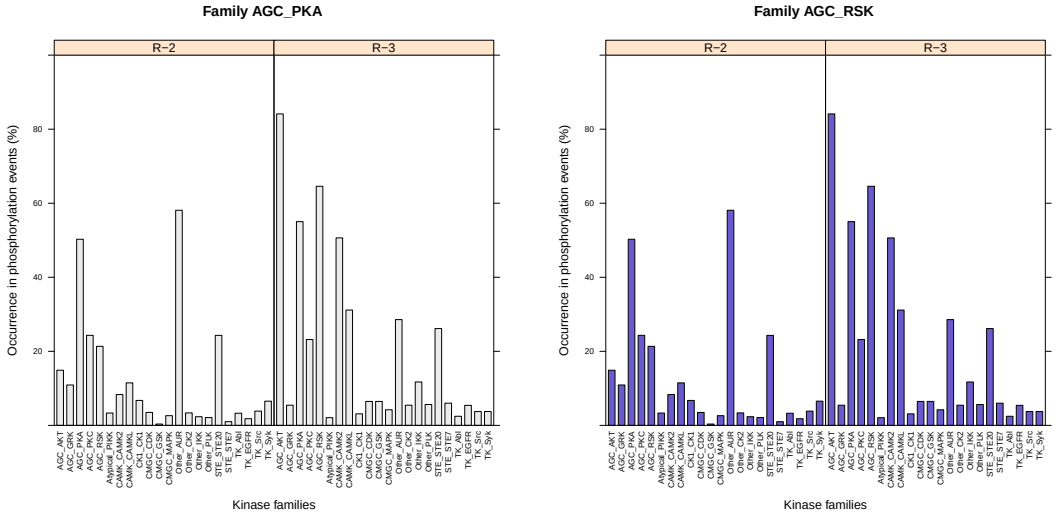
[284] Roskoski, R. August 2012 *Pharmacological research : the official journal of the Italian Pharmacological Society* **66(2)**, 105–43.

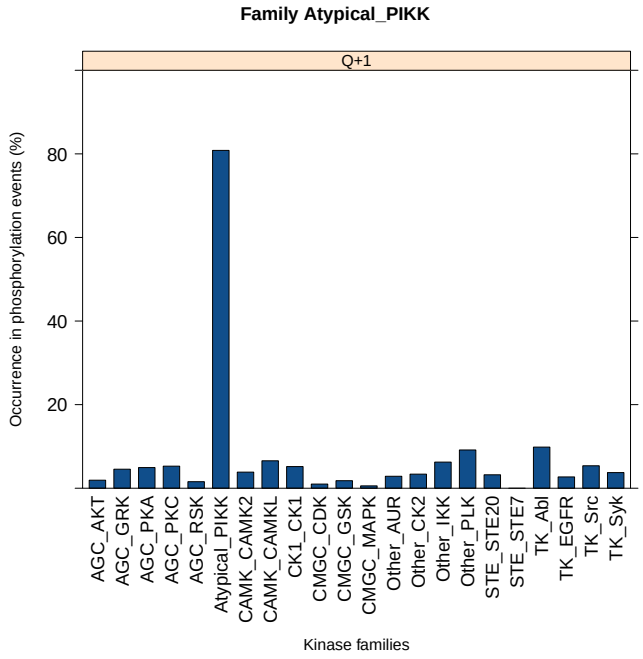
[285] Lloyd, A. C. January 2006 *Journal of Biology* **5(5)**, 13.

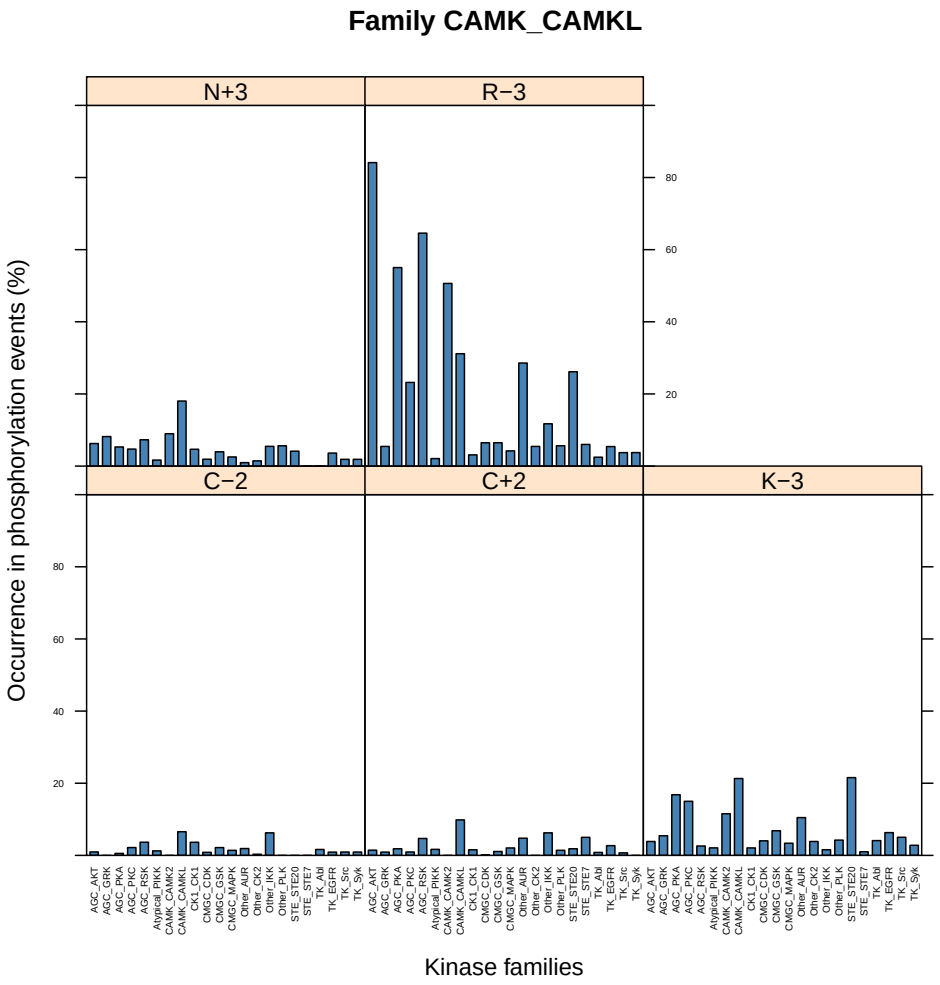
Appendices

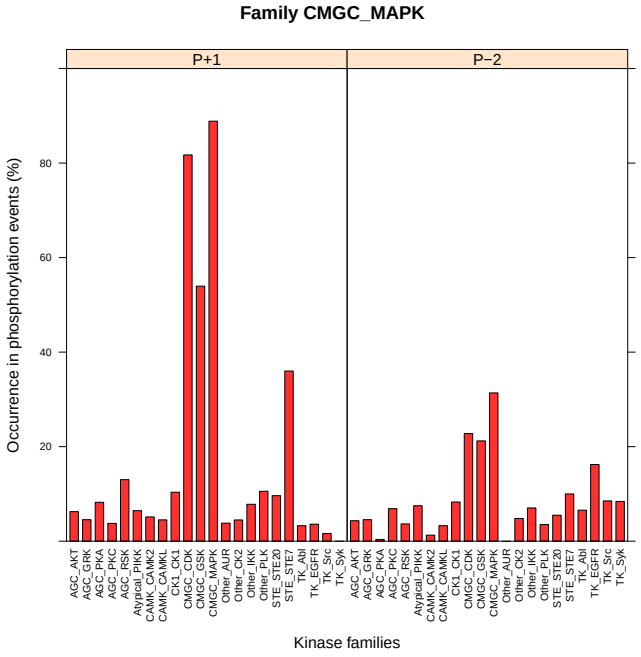
A1 SDRs from kinase families

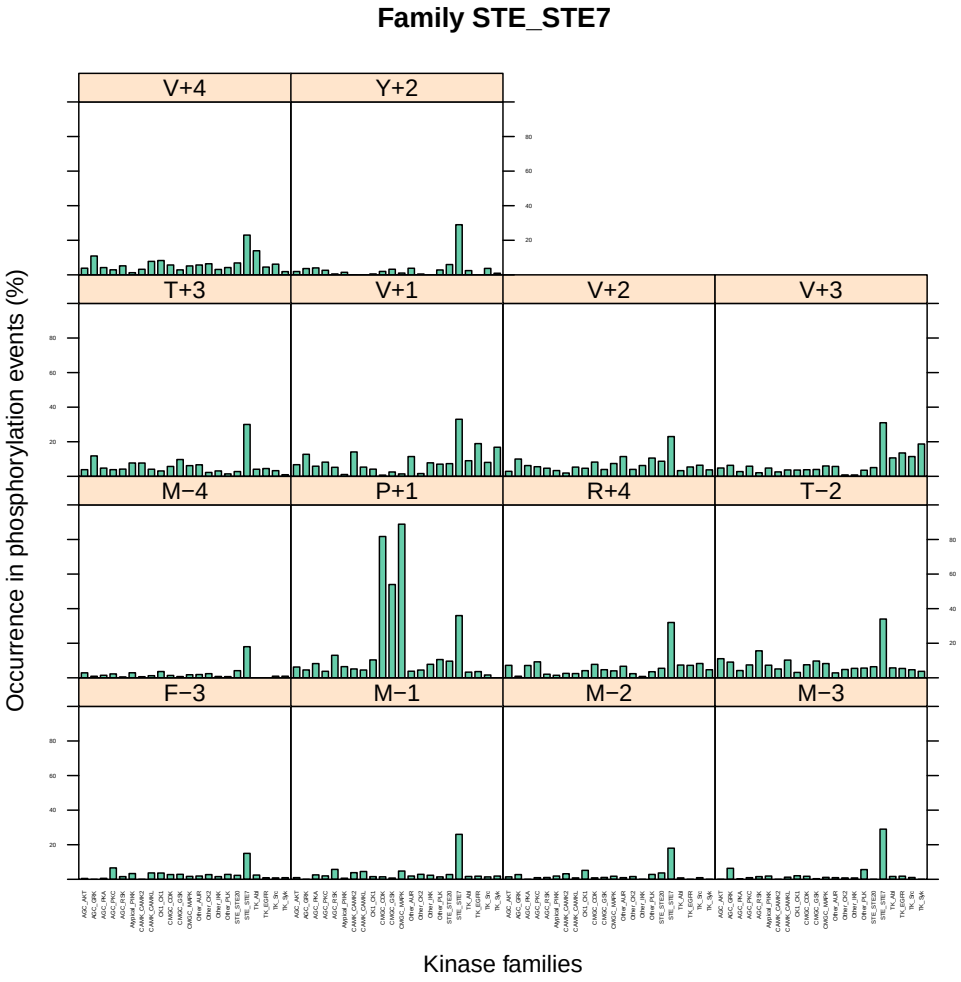
Descriptive legend. This section shows the frequencies of occurrence of the SDRs of nine kinase families. Boxes within each panel represent the SDRs. SDRs are represented by the one letter code of the amino acid and its position relative to the phospho-acceptor residue. On the x-axis, the kinase families, on the y-axis the percentage of occurrence of each SDR across the 22 families included in the analysis.

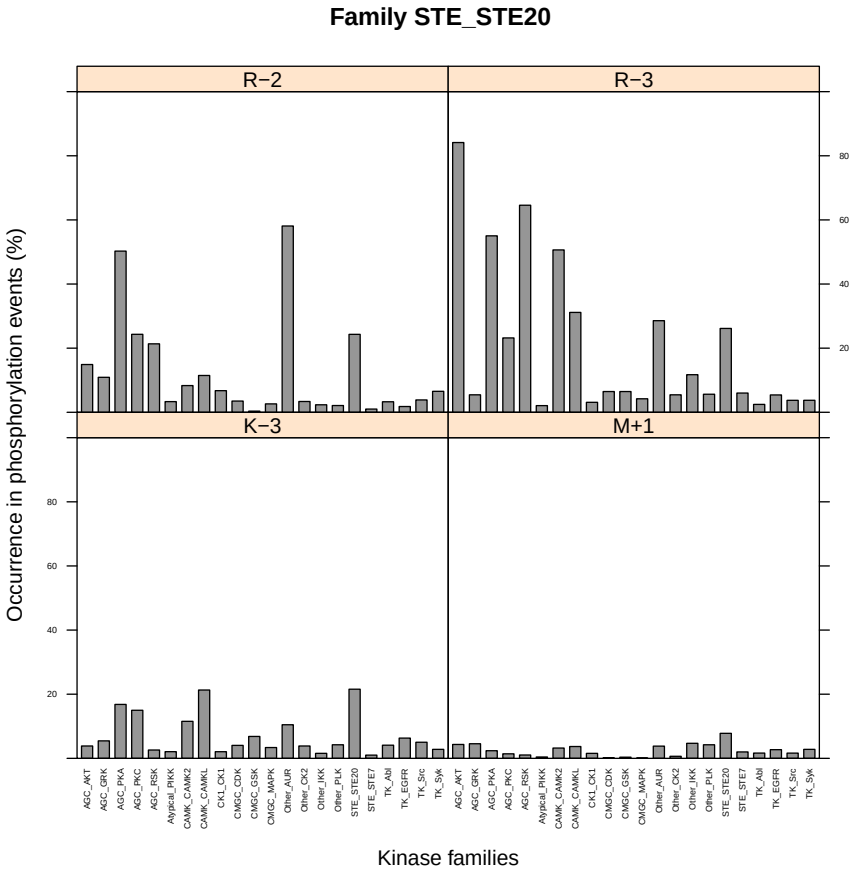


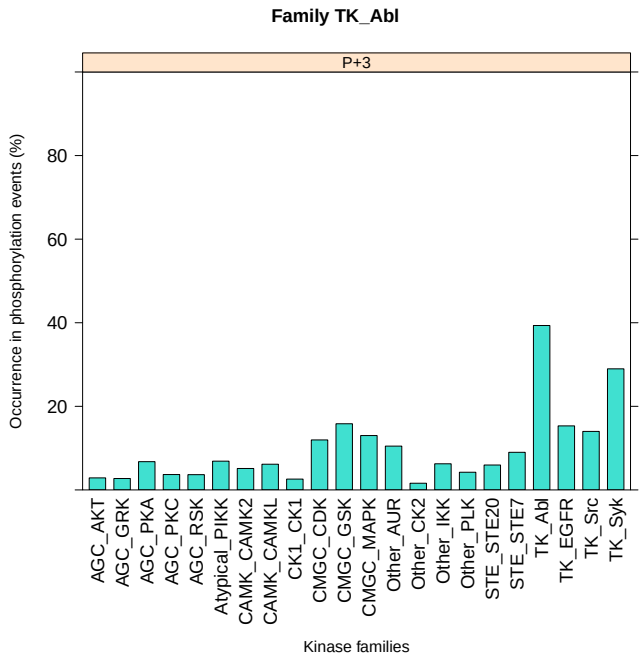


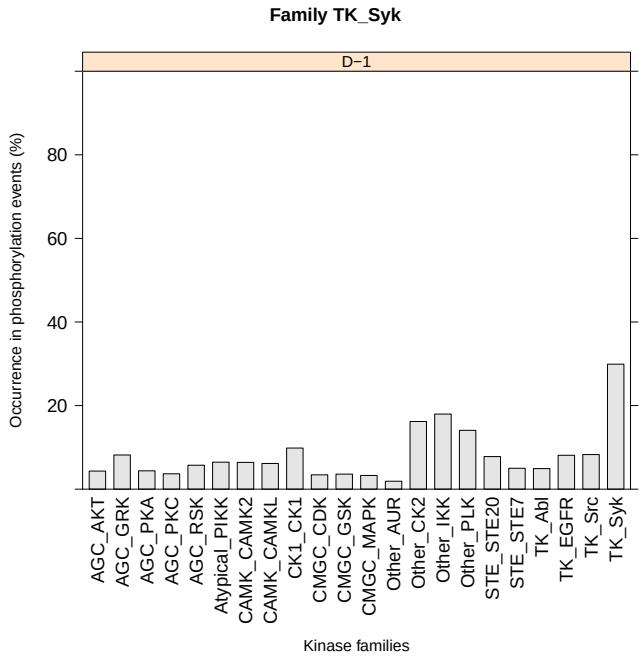

Family CAMK_CAMKL
N+3
R−3
C−2
C+2
K−3
Occurrence in phosphorylation events (%)
Kinase families
AGC_AKT
AGC_GRK
AGC_PKA
AGC_PKC
AGC_RSK
Atypical_PIKK
CAMK_CAMK2
CAMK_CAMKL
CK1_CK1
CMGC_CDK
CMGC_GSK
CMGC_MAPK
Other_AUR
Other_CK2
Other_IKK
Other_PLK
STE_STE20
STE_STE7
TK_Abl
TK_EGFR
TK_Src
TK_Syk


Family CMGC_MAPK
P+1
P−2
Occurrence in phosphorylation events (%)
Kinase families
AGC_AKT
AGC_GRK
AGC_PKA
AGC_PKC
AGC_RSK
Atypical_PIKK
CAMK_CAMK2
CAMK_CAMKL
CK1_CK1
CMGC_CDK
CMGC_GSK
CMGC_MAPK
Other_AUR
Other_CK2
Other_IKK
Other_PLK
STE_STE20
STE_STE7
TK_Abl
TK_EGFR
TK_Src
TK_Syk

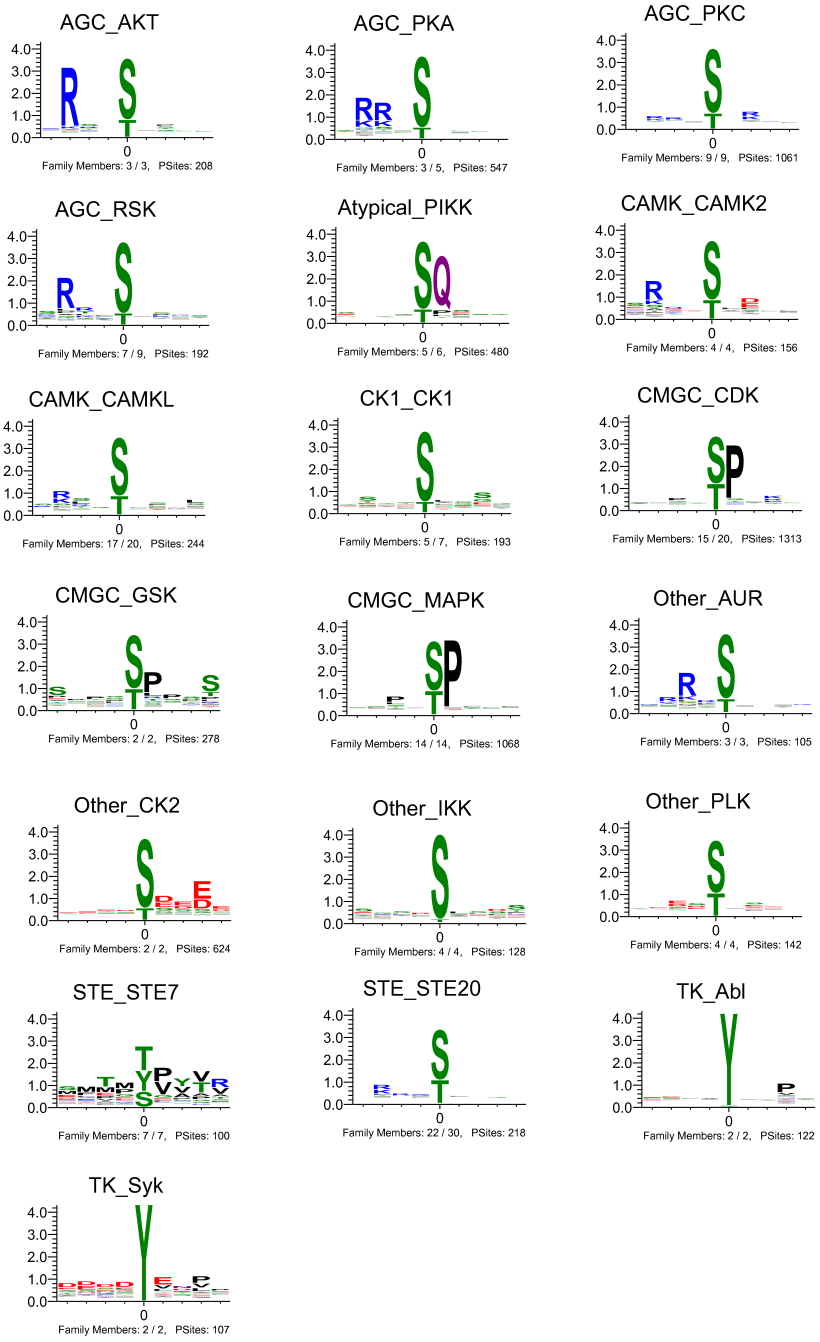

Family STE_STE7
V+4
Y+2
T+3
V+1
V+2
V+3
M−4
P+1
R+4
T−2
F−3
M−1
M−2
M−3
Occurrence in phosphorylation events (%)
Kinase families


Family STE_STE20
R−2
R−3
K−3
M+1
Occurrence in phosphorylation events (%)
20
40
60
80
AGC_AKT
AGC_GRK
AGC_PKA
AGC_PKC
AGC_RSK
Atypical_PRKK
CAMK_CAMK2
CAMK_CAMKL
CK1_CK1
CMGC_CDK
CMGC_GSK
CMGC_MAPK
Other_AUR
Other_CK2
Other_JKK
Other_PLK
STE_STE20
STE_STE7
TK_Abl
TK_EGFR
TK_Src
TK_Syk
Kinase families


Family TK_Abl
P+3
Occurrence in phosphorylation events (%)
20
40
60
80
AGC_AKT
AGC_GRK
AGC_PKA
AGC_PKC
AGC_RSK
Atypical_PIKK
CAMK_CAMK2
CAMK_CAMKL
CK1_CK1
CMGC_CDK
CMGC_GSK
CMGC_MAPK
Other_AUR
Other_CK2
Other_IKK
Other_PLK
STE_STE20
STE_STE7
TK_Abl
TK_EGFR
TK_Src
TK_Syk
Kinase families


Family TK_Syk
D−1
Occurrence in phosphorylation events (%)
20
40
60
80
AGC_AKT
AGC_GRK
AGC_PKA
AGC_PKC
AGC_RSK
Atypical_PIKK
CAMK_CAMK2
CAMK_CAMKL
CK1_CK1
CMGC_CDK
CMGC_GSK
CMGC_MAPK
Other_AUR
Other_CK2
Other_IKK
Other_PLK
STE_STE20
STE_STE7
TK_Abl
TK_EGFR
TK_Src
TK_Syk
Kinase families

A2 Sequence logos from kinase families



Sequence logos from the kinase families for which we identified at least one SDR. On the x-axis, the phospho-acceptor residue is shown in the central position. On the y-axis, in bits, the total sequence conservation at that position of the alignment. For each family we provide de fraction of the members for which we count with at least one phosphorylation site, and the total number of phosphorylation sites available for the family.

A3 List of known adaptors/scaffolds

Table 1: Known adaptor or scaffold proteins of human kinases

Kinase family	Kinase	Known adaptor or scaffold
AGC_AKT	AKT1	1433S, 1433Z, 2AAA, BCL10, GAB2, GRB10, IKKB, JIP1, P85A, PKHO1, PRKDC, RANB3, SH2B2, SH3R1, SRBS2
AGC_AKT	AKT2	PKHO1, SH3R1, SRBS2
AGC_AKT	AKT3	PKHO1
AGC_DMPK	MRCKB	MP2K5
AGC_DMPK	MRCKG	1433S
AGC_DMPK	ROCK1	GAB1
AGC_GRK	ARBK1	ARRB1, CAV1, GIT1, PK3CG
AGC_GRK	ARBK2	GIT1
AGC_GRK	GRK5	CAV1, DLGP2
AGC_GRK	GRK6	NHRF1
AGC_GRK	RK	CAV1
AGC_MAST	MAST1	AP1M2, ECSIT
AGC_NDR	LATS2	JUB
AGC_NDR	ST38L	1433B
AGC_NDR	STK38	1433B, 1433Z, ARRB1, ARRB2
AGC_PKA	KAPCA	1433Z, AKAP9, AKP13, BIRC5, CAV1, CUL5, FLNA, NEB2, RANB9, RGS14
AGC_PKA	KAPCB	FLNA
AGC_PKB	PDPK1	1433F, 1433T, GIT1, NHRF2, P85A
AGC_PKC	KPCA	1433G, 1433Z, AFAP1, AKP13, CBL, FLNA, IKKB, IQGA1, NHRF1, RBP1, SHC1
AGC_PKC	KPCB	1433G, AFAP1, IKKB
AGC_PKC	KPCD	1433B, 1433G, 1433Z, AFAP1, IKKB, PK3CB, PRKDC, SHC1
AGC_PKC	KPCE	1433Z, AFAP1, AKAP9, IQGA1, PK3CB
AGC_PKC	KPCG	1433G, AFAP1, PICK1
AGC_PKC	KPCI	1433F, 1433Z, FRS3, IKKB, MP2K5, PAR6A, PAR6B, PAR6G, PARD3, SQSTM
AGC_PKC	KPCL	AKP13, PAR6A
AGC_PKC	KPCT	1433G, BCL10, IKKB
AGC_PKC	KPCZ	1433B, 1433F, 1433G, 1433T, 1433Z, FADD, GRB14, IKKB, JUB, MP2K5, P85A, PAR6A, PAR6B, PARD3, PK3CG, SQSTM
AGC_PKN	PKN1	AKAP9, HOME3
AGC_PKN	PKN2	NCK1, NCK2
AGC_RSK	KS6A1	1433B, FLNA, GRB2, RANB3, RPTOR
AGC_RSK	KS6A3	RANB3
AGC_RSK	KS6A5	1433Z, ITSN1, PDLI1
AGC_RSK	KS6B1	TRAF4
AGC_RSK	KS6B2	JIP2
AGC_SGK	SGK1	IKKB, IMA2, NHRF2
AGC_YANK	ST32C	GRB2
Atypical_BCR	BCR	1433B, 1433E, 1433F, 1433G, 1433S, 1433T, 1433Z, CBL, DLG1, GAB2, GRB10, GRB2, LAP2, NHRF3, PK3CG
Atypical_BRD	BRD4	GRB2, NCK1, P85A
Atypical_PDHK	BCKD	TRAF4
Atypical_PDHK	PDK1	1433T
Atypical_PIKK	ATM	AP1B1, AP3B1, AP3B2, IKKB, MDC1, PRKDC
Atypical_PIKK	ATR	CLSPN, DAXX, MDC1, PRKDC
Atypical_PIKK	MTOR	1433T, 1433Z, RPTOR, SEPT2
Atypical_PIKK	PRKDC	1433B, 1433G, 1433Z, BIRC5, ELP1, IKKB, PP6R1, PP6R2, PP6R3, PRKDC
Atypical_RIO	RIOK1	1433B
Atypical_TAF1	TAF1L	PPIP1
Atypical_TIF1	TIF1B	1433Z, AAKB1, BCAR3
CAMK_CAMK1	KCC1D	CUL3
CAMK_CAMK1	KCC1G	ARRB2

Continued on next page

Table 1 – continued from previous page

Kinase family	Kinase	Known adaptor or scaffold
CAMK_CAMK2	KCC2A	1433B, 1433T, DLG1, FLNA, GIT1, SQSTM
CAMK_CAMK2	KCC2B	1433B, SQSTM
CAMK_CAMK2	KCC2D	SQSTM
CAMK_CAMK2	KCC2G	FLNA, SQSTM
CAMK_CAMKL	AAPK1	AAKB1, AAKB2, ABI1, IMA2, RPTOR
CAMK_CAMKL	AAPK2	AAKB1
CAMK_CAMKL	CHK1	1433B, 1433G, 1433S, 1433Z, CLSPN, CUL1, CUL4A, PRKDC
CAMK_CAMKL	HUNK	2AAA
CAMK_CAMKL	MARK1	1433G, 1433S
CAMK_CAMKL	MARK2	1433B, 1433F, 1433S, 1433Z, PAR6G
CAMK_CAMKL	MARK3	1433B, 1433G, 1433S, 1433Z, IKKB
CAMK_CAMKL	MARK4	1433F, PAR6G
CAMK_CAMKL	SIK1	1433Z
CAMK_CAMKL	SIK3	1433Z
CAMK_CAMKL	STK11	AP2M1, PARD3
CAMK_CASK	CSKP	CD2AP, CSKP, DLG1, LAP2
CAMK_DAPK	DAPK1	1433T, CUL3, FADD, KLH20, PDCD6
CAMK_DAPK	DAPK3	DAXX, GRB14
CAMK_MAPKAPK	MAPK2	1433Z, SHC1
CAMK_MAPKAPK	MKNK1	PRKDC
CAMK_MLCK	MYLK	GRB2, NCK1, P85A
CAMK_MLCK	MYLK2	ELP1
CAMK_MLCK	TITIN	FLNA, SQSTM
CAMK_PKD	KPCD1	1433T, 1433Z, AKP13
CAMK_PKD	KPCD2	1433F, CSKP, GRIP1, PDLI7
CAMK_PKD	KPCD3	IMA2
CAMK_RAD53	CHK2	IMA2, MDC1
CAMK_Trbl	TRIB3	GIT1
CAMK_Trio	TRIO	FLNA
CK1_CK1	KC1A	1433T, 1433Z, AP3B2, BCL10, FADD, IMA1
CK1_CK1	KC1D	AKAP9
CK1_CK1	KC1E	1433F
CK1_CK1	KC1G2	NCK1
CK1_VRK	VRK1	RAN
CK1_VRK	VRK2	KSR1, RAN
CK1_VRK	VRK3	RAN
CMGC_CDK	CD11B	1433B, 1433E, 1433G, 1433T, AP2M1
CMGC_CDK	CDK1	1433S, BIRC5, DEDD, DIAP1, DLG1, FLNA, IL16, KHDR1, ODFP2, RPTOR
CMGC_CDK	CDK14	1433B, 1433E, 1433F, 1433T
CMGC_CDK	CDK16	1433F, 1433G, 1433T, 1433Z
CMGC_CDK	CDK17	1433G, 1433Z
CMGC_CDK	CDK18	1433Z
CMGC_CDK	CDK2	BIRC5, C2D1A, CNKR2, CUL1, DBNL, DLG1, DTL, JIP4, MDC1
CMGC_CDK	CDK4	BIRC5, DBNL
CMGC_CDK	CDK5	DAB1
CMGC_CDK	CDK9	BCL10, CLSPN, CSKP, CUL1, IMA2
CMGC_CDKL	CDKL5	GRB2
CMGC_CLK	CLK1	1433G
CMGC_CLK	CLK2	1433G
CMGC_CLK	CLK3	1433G
CMGC_DYRK	DYR1A	1433B, 1433E, 1433G, 1433S
CMGC_DYRK	DYR1B	RANB9
CMGC_DYRK	DYRK4	CUL3
CMGC_DYRK	HIPK1	DAXX
CMGC_DYRK	HIPK2	CUL1, DAXX, RANB9
CMGC_DYRK	HIPK3	ARRB2, FADD, GRB2, TGFI1

Continued on next page

Table 1 – continued from previous page

Kinase family	Kinase	Known adaptor or scaffold
CMGC_DYRK	PRP4B	1433G, ARRB1, ARRB2
CMGC_GSK	GSK3A	1433G
CMGC_GSK	GSK3B	1433Z, BEX1, ELP1, MP2K5, SLAP1
CMGC_MAPK	MK01	APBB1, ARRB1, ARRB2, CAV1, FRS2, FRS3, GAB1, GAB2, GRB10, GRB2, IQGA1, KHDR1, LTOR3, NEB2, SCRIB, SH2B1, SHC1, SQSTM
CMGC_MAPK	MK03	ARRB1, ARRB2, CAV1, GAB1, GAB2, GRB10, LTOR3, SCRIB, SH2B1, SQSTM
CMGC_MAPK	MK04	GAB1, GAB2
CMGC_MAPK	MK06	2AAA, AAKB1, ITSN1, PDLI1, RANB9, SHC1
CMGC_MAPK	MK07	1433B, DLG1, MP2K5, SH22A
CMGC_MAPK	MK08	1433B, 1433S, 1433Z, CBL, ELP1, FADD, JIP1, JIP3, JIP4, P85A, PRKDC
CMGC_MAPK	MK09	ARRB1, ARRB2, GRB2, JIP1, JIP2, JIP3, PRKDC
CMGC_MAPK	MK10	ARRB1, ARRB2, JIP3
CMGC_MAPK	MK12	DLG1, DLG2, INADL, LAP2
CMGC_MAPK	MK13	JIP2
CMGC_MAPK	MK14	ARRB1, FLNA, JIP4, KHDR1, SHC1, SMAD7
CMGC_MAPK	MK15	TGFI1
CMGC_MAPK	NLK	CUL1
CMGC_SRPK	SRPK1	1433B, 1433G
CMGC_SRPK	SRPK2	1433B
Other_AUR	AURKA	BIRC5, IKKB, PARD3
Other_AUR	AURKB	BIRC5, KLH13, KLH21, KLHL9, PARD3
Other_AUR	AURKC	BIRC5
Other_BUB	BUB1	AP1B1, AP3B1
Other_BUB	BUB1B	AP1B1, AP3B1, DNMBP
Other_CAMKK	KKCC1	1433F
Other_CDC7	CDC7	CLSPN
Other_CK2	CSK21	1433B, 1433S, 1433T, ARRB1, ARRB2, CAV1, CAV2, IL16, MDC1, P85A, PKHO1, SEPT2, TAF1
Other_CK2	CSK22	ARRB2, CAV1, IL16
Other_IKK	IKKA	ARRB1, ARRB2, BCL10, ELP1, IKKB, PRKDC, TANK
Other_IKK	IKKB	1433B, ARRB1, ARRB2, BCL10, CTNL1, CUL1, CUL3, DOK1, ELP1, FLNA, IKKB, IQGA1, PRKDC, TCAM1
Other_IKK	IKKE	AP1S1, GRB2, IKKB, SEPT2, TANK, TBKB1, TCAM1
Other_IKK	TBK1	1433E, IKKB, NCK1, TANK, TBKB1, TCAM1
Other_MOS	MOS	HOME3
Other_NAK	AAK1	AP1M1, AP2M1
Other_NAK	GAK	AP1M2, AP2M1
Other_NEK	NEK1	1433F, CTNL1
Other_NEK	NEK2	GIT1
Other_NEK	NEK6	ARRB1
Other_NEK	NEK8	GRB2, NCK1
Other_NEK	NEK9	RAN
Other_NKF3	SG223	1433S
Other_NKF4	STK35	PDLI1
Other_Other-Unique	RN5A	IQGA1
Other_PEK	E2AK2	DBNL, IKKB, TIRAP
Other_PLK	PLK1	BIRC5, CENPU, CLSPN, FADD, IKKB, ITSN1, RAN, TANK
Other_PLK	PLK2	CSKP, DLGP4, ELP1
Other_PLK	PLK4	1433S
Other_Slob	PXK	P85A
Other_TLK	TLK1	1433E
Other_TOPK	TOPK	DLG1
Other_ULK	ULK2	SQSTM
Other_ULK	ULK4	1433T
Other_VPS15	PI3R4	1433B, 1433G
Other_WEE	WEE1	1433B, 1433S, 1433T, 1433Z
Other_Wnk	WNK1	1433E, 1433G, 1433Z
RGC_RGC	GUC2C	NHRF3
STE_STE-Unique	M3K14	ARRB1, ARRB2, ELP1, GRB10, GRB14, GRB7, IKKB, TRAF1, TRAF5

Continued on next page

Table 1 – continued from previous page

Kinase family	Kinase	Known adaptor or scaffold
STE_STE-Unique	M3K8	KSR2
STE_STE11	M3K1	1433E, FADD, FLNA, GRB2, IKKB
STE_STE11	M3K2	1433B, 1433G, 1433S, MP2K5, SH22A
STE_STE11	M3K3	1433B, 1433E, 1433F, 1433G, 1433T, 1433Z, 2AAA, 2AAB, DLG1, FLNA, GAB1, IKKB, IQGA1, JIP4, MP2K5, PRKDC, RMP, SCRIB, TM1L1
STE_STE11	M3K4	1433Z, SH3K1, TRAF4
STE_STE11	M3K5	1433B, 1433E, 1433F, 1433S, 1433T, 1433Z, ARRB1, ARRB2, BIRC5, DAXX, JIP3, PDCD6, TRAF1, TRAF5
STE_STE11	M3K6	1433B, 1433E, 1433F, 1433G, 1433S, 1433T, 1433Z
STE_STE20	M4K1	DAPP1, DBNL, GRB2, NCK1, P85A
STE_STE20	M4K3	GRB2, ITSN2, NCK1, SH3R1
STE_STE20	M4K5	GRB2, NCK1
STE_STE20	MINK1	1433B, 1433Z, ABI1, CSKI1, NCK1
STE_STE20	PAK1	1433G, 1433Z, FLNA, GIT1, GRB2, NCK1, NCK2, SH3K1, SRBS2
STE_STE20	PAK2	GIT1, GRB2, NCK1, SH3R1, SRBS2
STE_STE20	PAK3	NCK1
STE_STE20	PAK4	1433G, 1433S, 1433Z, GRB2, RAN
STE_STE20	PAK6	1433T
STE_STE20	SLK	NHRF3
STE_STE20	STK24	STRN, T3JAM
STE_STE20	STK25	1433Z, CCM2, STRN
STE_STE20	STK3	CNKR1
STE_STE20	STK4	1433G, CNKR1
STE_STE20	STRAB	GRB2
STE_STE20	TNIK	AKAP9, CNKR2, NCK1
STE_STE7	MP2K1	GRB10, JIP3, KSR1, KSR2, LTOR3
STE_STE7	MP2K2	DLG1, KSR1
STE_STE7	MP2K3	ARRB1, JIP2
STE_STE7	MP2K4	ARRB1, ARRB2, JIP3, JIP4
STE_STE7	MP2K5	1433Z, GRB2
STE_STE7	MP2K7	CNKR1, FADD, JIP1, JIP2, JIP3
TKL_IRAK	IRAK1	FADD, IKKB, MYD88, SQSTM, TAB2, TRAF4
TKL_IRAK	IRAK2	MYD88, TIRAP
TKL_IRAK	IRAK3	MYD88
TKL_IRAK	IRAK4	MYD88, TIRAP
TKL_LISK	LIMK1	1433Z
TKL_LISK	LIMK2	PARD3
TKL_MLK	ILK	NCK2, SHC1
TKL_MLK	M3K10	1433E, CNKR1, JIP1, JIP2
TKL_MLK	M3K11	IKKB, JIP1, JIP2, JIP3, TRAF5
TKL_MLK	M3K12	JIP1, JIP2
TKL_MLK	M3K13	IKKB, JIP1
TKL_MLK	M3K7	1433E, BCL10, ELP1, FLNA, IKKB, JIP1, SMAD7, TAB2, TAB3, TRAF5
TKL_MLK	M3KL4	1433B
TKL_MLK	MLTK	1433G, 1433S, 1433Z, ITSN1
TKL_RAF	ARAF	1433E, 1433G, 1433S, 1433Z, KLH12, P85A
TKL_RAF	BRAF	1433B, 1433G, 1433S, 1433T, 1433Z
TKL_RAF	RAF1	1433B, 1433E, 1433F, 1433G, 1433S, 1433T, 1433Z, ARRB2, CNKR1, CNKR2, GRB10, JIP3, KSR2
TKL_RIPK	RIPK1	CRADD, FADD, IKKB, SQSTM, TAB2, TANK, TCAM1, TRAF1
TKL_RIPK	RIPK2	TAB2, TCAM1, TRAF1, TRAF5
TKL_RIPK	RIPK3	1433E, FADD, FLNA, IQGA1, NEB2, PRKDC
TKL_RIPK	RIPK4	TRAF1, TRAF4, TRAF5
TKL_STKR	ACV1B	SMAD7
TKL_STKR	ACVL1	CAV1
TKL_STKR	ACVR1	AP2B1, IKKB, JUB, SQSTM
TKL_STKR	AVR2A	DAXX, LAP2, MAGI2
TKL_STKR	AVR2B	LAP2
TKL_STKR	BMR1A	SMAD7

Continued on next page

Table 1 – continued from previous page

Kinase family	Kinase	Known adaptor or scaffold
TKL_STKR	BMR1B	RAN, SASH3, SH3K1, SMAD7, SQSTM
TKL_STKR	TGFR1	1433Z, AP2B1, CAV1, CUL5, IKKB, P85A, P85B, PAR6A, RAN, SCRIB, SMAD7, SQSTM
TKL_STKR	TGFR2	AP2B1, DAXX, GRB2, P85A, P85B, SCRIB, SHC1, SMAD7, TGFI1
TK_Abl	ABL1	1433S, 1433Z, ABI1, AKAP6, CAV1, CBL, CD2AP, DAAM1, DLGP1, DLGP2, DLGP3, DLGP4, DOK1, FLNA, FYB, G3BP2, GRB10, GRB2, GRIP2, NCK1, P85A, PPIP1, PRKDC, RAN, SHB, SHC1, SHD, SRBS2
TK_Abl	ABL2	ABI1, GRB2, NCK1, P85A, SRBS2
TK_Ack	ACK1	GRB2, HSH2D, MAGI3, NCK1
TK_Ack	TNK1	1433S
TK_Alk	ALK	GRB2, SHC1, SHC3
TK_Alk	LTK	P85A
TK_Axl	MERTK	GRB2
TK_Axl	TYRO3	P85A, RANB9
TK_Axl	UFO	CBL, GRB2, NCK2, P85A, P85B, RANB9, SHC1
TK_Csk	CSK	ARRB1, CAV1, DAG1, DOK1, DOK3, PARD3, SHC1, TGFI1
TK_DDR	DDR1	NCK2, SHC1
TK_DDR	DDR2	SHC1
TK_EGFR	EGFR	1433T, 1433Z, ALDOA, AP2M1, APBB1, CAV1, CAV3, CBL, DOK2, DOK4, DOK5, DOK6, GAB1, GRB10, GRB14, GRB2, GRB7, IMA1, JIP1, JIP2, NCK1, NCK2, P85A, P85B, SH22A, SH2B1, SH3K1, SHC1, SHC2, SHC3, SLAP1, SRBS2, TM1L1
TK_EGFR	ERBB2	APBB1, CAV1, CBL, CUL5, DAB1, DOK1, DOK4, DOK6, GAB2, GRB2, GRB7, JIP1, JIP2, LAP2, NCK2, P85A, P85B, SH22A, SH2B2, SHC1, SHC2, SHC3, SLAP1, SLAP2
TK_EGFR	ERBB3	DAB1, DAPP1, GRB2, GRB7, NCK1, NCK2, P85A, P85B, RASL2, SHC1, SHC3
TK_EGFR	ERBB4	DLG1, DLG2, GRB2, NCK2, P85B, SHC1
TK_Eph	EPHA2	BACD2, CBL, GRB2, IMA3, SHC1
TK_Eph	EPHA7	SDCB1
TK_Eph	EPHA8	CBL, PK3CG
TK_Eph	EPHB1	CBL, GRB10, GRB2, GRB7, NCK1
TK_Eph	EPHB2	ITSN1, KSR1, PICK1
TK_Eph	EPHB6	CBL, GRB2
TK_FGFR	FGFR1	FRS2, FRS3, GRB2, NCK2, P85A, P85B, SHB, SLAP1
TK_FGFR	FGFR2	1433Z, GRB2, P85A, SHC1
TK_Fak	FAK1	GIT1, GRB2, GRB7, JIP3, NCK1, NCK2, P85A, SHC1, TGFI1
TK_Fak	FAK2	DLGP3, GIT1, GRB2, P85A, SHC1, SRBS2, TGFI1
TK_Fer	FER	ABI1
TK_Fer	FES	ABI1, HSH2D, P85A
TK_InsR	IGF1R	1433B, 1433E, 1433G, 1433Z, ARRB2, CBL, GRB10, GRB14, P85A, P85B, SHC1
TK_InsR	INSR	1433B, CAV1, CAV3, CBL, DOK1, GAB1, GRB10, GRB14, GRB7, P85A, SH2B1, SH2B2, SHC1
TK_JakA	JAK1	GRB2, SH2B2
TK_JakA	JAK2	2AAB, GAB2, GRB10, GRB2, NCK1, P85A, SH2B1, SH2B2, SHC1
TK_JakA	JAK3	KHDR1, P85A, SH2B2
TK_JakA	TYK2	CBL, KHDR2, P85A
TK_Met	MET	1433Z, CBL, GAB1, GRB2, P85A, RANB9, RBP10, SHC1
TK_Met	RON	1433T, 1433Z, CBL, GAB1, GRB2, SHC1
TK_PDGFR	CSF1R	CBL, P85A, SHC1
TK_PDGFR	FLT3	GRB2, IKKB, NCK1, P85A, SHC1
TK_PDGFR	KIT	CBL, DOK1, GRB2, GRB7, P85A, P85B, PK3CG, SH2B2
TK_PDGFR	PGFRA	CAV1, CAV3, CBL, GRB2, P85A
TK_PDGFR	PGFRB	CAV1, CAV3, CBL, GAB1, GRB10, GRB14, GRB2, GRB7, NCK1, NCK2, NHRF1, NHRF2, P85A, P85B, SH2B2, SHB, SHC1, SLAP1
TK_Ret	RET	CBL, DOK1, DOK5, DOK6, FRS2, GRB10, GRB2, GRB7, P85A, PDLI7, SHC1, SHC3
TK_Ryk	RYK	ABI1, PRKDC
TK_Src	BLK	CBL
TK_Src	FGR	ABI1, ARRB1, CBL, DOK1, SH3K1
TK_Src	FRK	ABI1
TK_Src	FYN	ABI1, AKAP6, CAV1, CBL, CD2AP, DAG1, DLGP1, DLGP2, DLGP3, DLGP4, DOK1, DOK3, DOK4, FLNA, FYB, GRB10, KHDR1, NCK1, P85A, P85B, SH2B2, SH3K1, SHC1, TGFI1, TM1L1
TK_Src	HCK	ABI1, ARRB1, CBL, CD2AP, DOK1, DOK2, KHDR1, P85A, P85B, SH3K1
TK_Src	LCK	ABI1, CBL, DAPP1, DLG1, DOK1, DOK3, KHDR1, P85A, PK3CG, SH22A, SH3K1, SHC1, SQSTM

Continued on next page

Table 1 – continued from previous page

Kinase family	Kinase	Known adaptor or scaffold
TK_Src	LYN	ABI1, CBL, DAPP1, DOK1, DOK2, DOK3, GAB2, KHDR1, PRKDC, SH2B2, SH3K1, SHC1
TK_Src	PTK6	KHDR1
TK_Src	SRC	ABI1, AFAP1, AKAP6, AP2B1, ARRB1, ARRB2, CAV1, CAV2, CBL, DAAM1, DAB1, DAG1, DAPP1, DLGP1, DLGP2, DLGP3, DLGP4, DOK1, DOK2, DOK4, FLNA, GAB1, GAB2, GRB10, GRB2, IKKB, JIP3, KHDR1, P85A, SDCB1, SH22A, SH3K1, SHB, SHC1, SHC3, TRAF1
TK_Src	YES	ABI1, CBL, DOK1, NHRF1
TK_Syk	KSYK	CBL, DBNL, GRB2, MYD88, NHRF1, P85A, P85B, SH2B2, SHC1, SLAP1
TK_Syk	ZAP70	CBL, DBNL, GAB2, GRB2, PK3CG, SHB, SHC1, SLAP1, SLAP2
TK_Tec	BMX	CAV1, ELP1
TK_Tec	BTK	ARRB1, CAV1, CBL, DAAM1, DAPP1, GRB2, KHDR1, MYD88, SH2B2, TIRAP
TK_Tec	ITK	ABI1, GRB2, IMA2, KHDR1, SH22A
TK_Tec	TEC	DOK1, P85B
TK_Tec	TXK	SH22A
TK_Tie	TIE2	GRB2, SHC1
TK_Trk	NTRK1	CAV1, CBL, FRS2, FRS3, GRB2, P85A, RUSC1, SH2B1, SH2B2, SHC1, SHC2, SHC3, SQSTM
TK_Trk	NTRK2	AP1B1, DOK5, FRS2, FRS3, NCK2, P85A, SH2B1, SHC1, SHC2, SHC3, SQSTM
TK_Trk	NTRK3	DOK5, FRS2, SHC1, SHC2
TK_VEGFR	VGFR1	CBL, GRB2, NCK1, P85A
TK_VEGFR	VGFR2	CAV1, CBL, FRS2, GRB10, GRB2, IQGA1, NCK1, SH22A, SHB, SHC1, SHC2
TK_VEGFR	VGFR3	GRB2, SHC1, SHC3

A4 List of potential adaptors/scaffolds

Table 2: Potential adaptor or scaffold proteins of human kinases

Kinase family	Kinase	Number of substrates	Potential adaptor or scaffold
AGC_AKT	AKT1	119	1433B, 1433E, 1433G, 1433T, 1433Z, ABL1, ACTB, BCL2, CBP, CCND1, CTNB1, EP300, ESR1, FHL2, GRB2, GSK3B, NCOA1, NCOR2, NEMO, P53, P85A, PAK2, RB, SF3A2, SIRT1, SMAD2, SMAD3, SRC, TRAF2, TRAF6, UB2D1, UBC, UBC9, ZHX1
AGC_AKT	AKT2	14	CBP, UBC9
AGC_DMPK	ROCK1	24	DPYL1, GRB2, IKKE, TRAF6, TSC1
AGC_DMPK	ROCK2	6	DPYL4
AGC_GRK	ARBK1	15	NHRF1, TSC1
AGC_PKB	PDPK1	21	1433Z, CASP3, GSK3B, MAD1, PDK1, RAF1, TAU, VIME
AGC_PKC	KPCB	45	1B42, EGFR, GBLP, P85A, PASK
AGC_PKC	KPCD	72	1433Z, ABL1, CBL, CSK2B, DAXX, EGFR, ERBB2, ERBB3, ESR1, GRB2, IGF1R, M3K5, MDM2, NEMO, P53, P85A, PASK, PIAS1, SRC, TF65, UBC, XRCC6, YBOX1
AGC_PKC	KPCE	26	1433Z, FHL2, PASK
AGC_PKC	KPCG	27	EGFR, P85A
AGC_PKC	KPCT	19	NEMO, TSC1
AGC_PKC	KPCZ	29	1433E, EP300, NCF1, TF65
AGC_PKG	KGP1	29	PLS1, UBC, ZYX
AGC_PKN	PKN1	6	1433Z, CBX5
AGC_RSK	KS6A1	34	AKT1, CHD3, EP300, GRB2, NCOA1, NCOR2, P53, RXRA, SMAD3, SRC, TF65, UBC9
AGC_RSK	KS6A3	24	1433G, 1433Z, CHD3, EP300, PASK, UBC
AGC_RSK	KS6A5	20	1433G, 1433Z, EP300, NCOR2
AGC_RSK	KS6B1	17	AKT1, KS6A1
AGC_SGK	SGK1	17	1433G, 1433Z, KS6A1, MP2K1
Atypical_PDHK	PDK1	6	AKT1, CASP3, CENPR, MAD1, PDPK1, TAU
Atypical_PIKK	MTOR	9	TAU
Atypical_PIKK	PRKDC	27	1433Z, ACTB, ANDR, APLF, CBP, CEBPB, CHD1L, CSK21, DAXX, EP300, GRB2, GSK3B, HS90A, IMA2, MDM2, NCOR2, NR4A1, PNKP, RFA1, SIRT1, UBC
CAMK_CAMK1	KCC4	9	ESR1, HIF1A, UBC9
CAMK_CAMK2	KCC2G	14	GRB2, NCK1, P85A, PLCG1
CAMK_CAMKL	AAPK1	27	1433B, 1433G, 2AAA, AKT1, CDC37, CHK2, GRB2, HNRL1, IMMT, P53, SIRT1, UBC9
CAMK_CAMKL	CHK1	19	1433B, 1433E, 1433Z, CHK2, EP300, MDM2, UBC, UBC9
CAMK_CAMKL	MARK2	6	1433S, 1433Z, HIF1A, UBC9
CAMK_DAPK	DAPK1	6	B2CL1, P63
CAMK_DAPK	DAPK3	8	SNAI1, UBB
CAMK_MAPKAPK	MAPK2	23	1433B, 1433Z, BHE40, CHK1, CHK2, DAXX, EP300, GSK3B, MDM2, MDM4, P53, PIAS1, PLK1, SETB1, SRBS2, ZHX1
CAMK_PIM	PIM1	13	1433B, 1433Z, P53, SMAD3, UBC
CAMK_PKD	KPCD1	28	1433S, ABL1, EGFR, EP300, SMAD3, UBC9
CAMK_RAD53	CHK2	17	1433B, 1433Z, ANDR, CBP, CHK1, COM1, FHL2, HDAC1, MDM2, MDM4, RB, SIRT1, SUMO1
CK1_CK1	KC1A	32	1433E, 1433Z, ANDR, B2CL1, CBP, CHK1, CTNB1, EP300, FHL2, FLNA, M3K5, PIAS1, PIN1, PSN1, SETB1, UBC9, ZHX1
CK1_CK1	KC1D	15	ANDR, CBP, FOXO3, GSK3B, HDAC1, KAT2B, PSN1
CMGC_CDK	CDK3	6	EP300, ESR1
CMGC_CDK	CDK4	11	ANDR, COM1, DGKZ, E2F4, EP300, MYC
CMGC_CDK	CDK5	43	1433Z, ABL1, ANDR, DYN2, EGFR, ERBB2, GSK3B, RAC1, SRC
CMGC_CDK	CDK6	8	CCND1, CCND3, CDK2, HDAC1, MCM7
CMGC_CDK	CDK7	17	CBP, CCND1, CCND3, CD2A1, CDN1A, CDN2C, MED1, NCOA1, NCOA3, NCOA6, NCOR1, NCOR2, NR0B2, NRIP1, PIAS1, PML, PNRC2, SP1
CMGC_CDK	CDK9	7	ANDR, EP300, HDAC1, NCOA6, PIAS4, PIN1, PSD11, ZBTB3
CMGC_DYRK	DYR1A	7	DYN2
CMGC_DYRK	DYRK2	7	GSK3B, KITH
CMGC_DYRK	HIPK2	8	CBP, EP300, PIAS1, UBC9
CMGC_GSK	GSK3A	16	CHIP, LEF1, RB
CMGC_MAPK	MK03	119	1433B, 1433Z, ANDR, CBP, CHIP, CSK2B, DAXX, EGFR, EP300, ESR1, GCR, GRB2, HDAC1, JUN, MDM2, MED25, MK01, MK14, NCOA1, NCOA3, NCOA6, NCOR1, NCOR2, NR0B2, NR1H2, NR4A1, P53, P85A, PIAS1, PPARG, RB, RXRA, SMAD3, SMAD4, SP1, SRC, SUMO1, TF65, UBC9

Continued on next page

Table 2 – continued from previous page

Kinase family	Kinase	Number of substrates	Potential adaptor or scaffold
CMGC_MAPK	MK09	27	A4, ASPP2, ATF3, B2CL1, BCL2, GSK3B, IMA2, MK08, PASK, PLK1, SIRT1, UBC, UBC9
CMGC_MAPK	MK11	11	RBL2
CMGC_MAPK	MK12	6	A4
Other_AUR	AURKA	20	1433Z, ANXA7, DDX5, H4, KITH, LMO4, MDM2, SNAI1, TP53B
Other_AUR	AURKB	21	CBX5
Other_IKK	IKKA	14	ANDR, CBP, ESR1, FBW1A, FOXO1, IKBA, NCOA3, NFKB1, P53, RS3, SMAD4, TF65, UBC
Other_IKK	IKKB	15	CBP, SMAD3, TF65, UBC
Other_NEK	NEK6	6	TAU
Other_PEK	E2AK2	6	1433Z, 2AAA, CHK2, GSK3B, IKKB, UBC
Other_PLK	PLK1	45	ABL1, ANDR, GRB2, P53, VHL
Other_PLK	PLK3	11	CHK1, HDAC1, IMA2, PLK1, UBC, XPO1
Other_TTK	TTK	7	P53, UBC9
STE_STE-Unique	M3K8	17	1B42, AAKB1, IKKE, MPIP3, TRAF6, TSC1
STE_STE20	PAK2	10	GRB2
STE_STE7	MP2K1	9	MP2K2
STE_STE7	MP2K4	6	ARRB1, ARRB2, CD2A1, DUS1, DUS10, JUN, MKNK2, NCOA3, P85A, SSU72, TF65, ZEP1
TKL_MLK	ILK	7	ZHX1
TKL_MLK	M3K7	7	1433S
TK_Abl	ABL1	55	ANDR, ARRB1, ASPP2, CBL, CRK, EGFR, EP300, ERBB2, ERBB3, ESR1, FBW1A, FYN, GRB2, LCK, MDM2, MED28, MK06, NCK1, NR0B2, P53, P85A, PLCG1, PLS1, PTN12, PTPRB, PTPRC, PTPRG, PTPRJ, PTPRO, SRC, TF65, UBC, UBC9
TK_Csk	CSK	11	ABI1, ABL1, ABL2, ADA15, ASAP1, CBL, CDN1B, ERBB2, FAK1, KHDR1, MED28, PAK2, PTPRZ, RL10, SOS1, SPR2A
TK_EGFR	EGFR	34	ABI1, ABL1, ABL2, AFF2, AIRE, AKAP2, ARHGB, ASAP1, ASAP2, DLG4, DLGP1, DLGP3, DLX4, DNJA3, DUS15, EFS, ERBB2, ERBB3, ERBB4, EXTL3, FANCA, FCG2B, FCG2C, FLNA, FLNB, FLNC, FYN, GHR, GRB2, GSCR1, I20L2, ICAL, ID4, IKKE, INSR, JAK2, LCP2, M4K1, MEPE, MINT, MLL4, MYH9, NTRK1, NTRK2, P85A, PAR3, PDIA2, PHAR2, PTN12, PTPRB, PTPRC, PTPRG, PTPRJ, PTPRO, RIN3, RRAS, SHAN2, SHAN3, SNX17, SNX3, SNX7, SOS1, SOS2, SRC, STF1, SUV92, TGON2, UBC
TK_InsR	INSR	18	ABI1, ASAP1, EGFR, ERBB2, ERBB3, GRB2, IGF1R, IRS1, NTRK1, P85A, PGFRB, SRC, UBC
TK_JakA	JAK1	9	ERBB2
TK_JakA	JAK2	19	1433Z, ERBB2, FINC, GRB10, GRB2, GSTK1, HS90B, MP2K1, MP2K2, P85A, PTPRB, PTPRC, PTPRJ
TK_PDGFR	PGFRB	6	ABI1, ABL1, ABL2, ADA15, AIRE, AKAP6, APOL5, ASAP1, ASAP2, ASB16, BCAR1, CBL, CKAP5, CP4F2, DAG1, DLGP1, DLGP2, DLGP3, DLGP4, DLX4, DOCK1, DOCK3, DPOD1, DUS15, E41L3, EGFR, ERBB2, ERBB3, EXTL3, FAK1, FANCA, FCG2B, FCG2C, FLNA, GASR, GNS, GRB2, HCN2, HXC8, I20L2, ICAL, KI67, LCP2, M4K1, MED28, MEPE, NKX21, P85A, PAK2, PAR3, PAX3, PAX7, PDIA2, PRIC3, RIN3, RPGF1, RPP38, RRAS, RTN4, SELN, SHAN3, SHRM2, SNX17, SNX3, SOS1, SOS2, SP1, SRC, TAU, TULP4
TK_Ret	RET	10	DUS1, ERBB2, KS6B1
TK_Src	FYN	50	EGFR, ERBB2, ERBB3, ERBB4, GRB2, NCK1, P85A, PSN1, PTN12, PTPRB, PTPRC, PTPRG, PTPRJ, SH21A, SRC
TK_Src	HCK	13	ABI1, ABL1, ASAP1, CRKL, EGFR, ERBB2, FYN, GRB2, LCK, NCK1, P85A, SRC
TK_Src	LYN	33	ABL1, CRK, EGFR, ERBB2, ERBB3, FYN, GRB2, JAK2, LCK, NCK1, P85A, PLCG1, SRC
TK_Syk	KSYK	25	ABI1, EGFR, ERBB2, FLNA, FYN, GRB2, MED28, NCK1, P85A, PLCG1, SH3K1, SRC, UBC
TK_Syk	ZAP70	9	EGFR, GRB2, SRC
TK_Tec	BTK	9	ERBB2

A5 Cellular compartment annotation of potential adaptors/scaffolds

Table 3: Cellular compartment annotation of potential adaptor or scaffold proteins

Kinase	GO description	Enrichment ratio	Adjusted *p*-value	Potential adaptor or scaffold
AAPK1	cytosol	3.89	$3.93e^{-02}$	1433B, 1433G, 2AAA, AKT1, CDC37, GRB2, P53
ABL1	cytosol	3.24	$3.74e^{-03}$	ARRB1, CBL, CRK, FBW1A, FYN, GRB2, LCK, MDM2, NCK1, P53, P85A, PLCG1, PLS1, PTN12, SRC, TF65, UBC
ABL1	perikaryon	18.7	$3.73e^{-02}$	ESR1
AKT1	intracellular	8.5	$5.85e^{-03}$	CCND1, NEMO, SMAD2
AKT1	nucleus	3.11	$2.20e^{-06}$	1433B, 1433T, ABL1, BCL2, CBP, CCND1, CTNB1, EP300, ESR1, FHL2, GRB2, GSK3B, NCOR2, NEMO, P53, PAK2, RB, SIRT1, SMAD2, SMAD3, SRC, TRAF6, UBC9, ZHX1
AKT1	nucleoplasm	2.74	$1.95e^{-04}$	1433Z, CBP, CCND1, EP300, ESR1, FHL2, NCOA1, NCOR2, P53, RB, SF3A2, SIRT1, SMAD2, SMAD3, UB2D1, UBC
AKT1	transcription factor complex	3.86	$3.66e^{-02}$	CBP, CTNB1, EP300, SMAD2, SMAD3
AKT1	cytoplasm	2.84	$3.85e^{-03}$	1433B, 1433E, 1433T, 1433Z, ABL1, ACTB, BCL2, CBP, CTNB1, EP300, GRB2, GSK3B, NEMO, P53, SIRT1, SMAD2, SMAD3, TRAF2, TRAF6, UB2D1, UBC9
AKT1	cytosol	2.6	$8.80e^{-05}$	1433B, 1433E, 1433G, 1433Z, ABL1, ACTB, BCL2, CCND1, CTNB1, GRB2, GSK3B, NEMO, P53, P85A, PAK2, SMAD2, SMAD3, SRC, TRAF2, TRAF6, UB2D1, UBC
ARBK1	apical plasma membrane	15.7	$4.08e^{-02}$	NHRF1
CDK4	chromatin	16.12	$3.76e^{-02}$	EP300
CDK4	nucleus	inf	$3.68e^{-02}$	ANDR, DGKZ, EP300, MYC, RBBP8
CDK4	nucleoplasm	12.43	$9.20e^{-03}$	ANDR, E2F4, EP300, MYC
CDK4	transcription factor complex	20.92	$1.67e^{-02}$	E2F4, EP300
CDK5	cytosol	4.15	$9.67e^{-04}$	1433Z, ABL1, DYN2, GSK3B, RAC1, SRC
CDK5	cytoskeleton	8	$4.50e^{-07}$	SRC
CDK5	basolateral plasma membrane	8.77	$4.22e^{-04}$	EGFR, ERBB2
CDK5	axon	13.27	$3.09e^{-06}$	ANDR
CDK7	cyclin-dependent protein kinase holoenzyme complex	67.98	$3.74e^{-03}$	CCND1, CCND3, CDN1A
CDK7	nucleus	17.82	$6.33e^{-03}$	CBP, CCND1, CCND3, CD2A1, CDN1A, CDN2C, MED1, NCOA3, NCOA6, NCOR1, NCOR2, NR0B2, NRIP1, PIAS1, PML, PNRC2, SP1
CDK7	nucleoplasm	16.44	$5.40e^{-06}$	CBP, CCND1, CDN1A, MED1, NCOA1, NCOA3, NCOA6, NCOR1, NCOR2, NR0B2, PIAS1, PML
CDK7	transcription factor complex	20.79	$4.41e^{-04}$	CBP, NCOA6, NCOR1
CDK9	nucleoplasm	inf	$7.33e^{-05}$	ANDR, EP300, HDAC1, NCOA6, PIN1, PSD11
CDK9	PML body	61.38	$4.19e^{-03}$	PIAS4
CHK1	nucleus	inf	$2.06e^{-04}$	1433B, EP300, MDM2, UBC9
CHK1	nucleoplasm	10.16	$1.30e^{-04}$	1433Z, CHK2, EP300, MDM2, UBC
CHK1	PML body	27.87	$2.91e^{-03}$	CHK2, UBC9
CHK2	nucleus	17.82	$6.02e^{-03}$	1433B, ANDR, CBP, CHK1, FHL2, HDAC1, MDM2, MDM4, RB, RBBP8, SIRT1, SUMO1
CHK2	nucleoplasm	9.36	$7.16e^{-04}$	1433Z, ANDR, CBP, CHK1, FHL2, HDAC1, MDM2, RB, SIRT1, SUMO1
CHK2	PML body	29.86	$3.25e^{-03}$	RB, SIRT1, SUMO1
CHK2	Rb-E2F complex	553.88	$5.46e^{-03}$	RB
CSK	plasma membrane	12.52	$9.45e^{-03}$	CBL, ERBB2, PAK2, SOS1
CSK	membrane raft	33.87	$2.74e^{-03}$	ERBB2
DYRK2	cytosol	inf	$1.30e^{-02}$	GSK3B, KITH
EGFR	cytoplasm	inf	$1.28e^{-03}$	ABL1, AIRE, ARHGB, ASAP1, DNJA3, DUS15, EFS, ERBB4, FANCA, FLNA, FLNB, FLNC, GRB2, ID4, IKKE, JAK2, MYH9, PTN12, SHAN2, SNX17, SNX3
EGFR	endosome	8.47	$6.48e^{-04}$	FYN, GRB2, NTRK1
EGFR	cytosol	4.78	$1.90e^{-03}$	ABI1, ABL1, ABL2, ARHGB, DNJA3, ERBB4, FLNA, FLNB, FLNC, FYN, GRB2, IKKE, INSR, JAK2, LCP2, MYH9, P85A, PTN12, SNX17, SOS1, SOS2, SRC, UBC

Continued on next page

Table 3 – continued from previous page

Kinase	GO description	Enrichment ratio	Adjusted *p*-value	Potential adaptor or scaffold
FYN	plasma membrane	7.48	$1.25e^{-10}$	EGFR, ERBB2, ERBB3, ERBB4, GRB2, NCK1, P85A, PSN1, PTN12, PTPRC, PTPRJ, SRC
FYN	cell junction	3.98	$3.15e^{-02}$	PTN12
HCK	cytosol	9.49	$1.01e^{-02}$	ABI1, ABL1, CRKL, FYN, GRB2, LCK, NCK1, P85A, SRC
IKKA	nucleus	inf	$5.87e^{-03}$	ANDR, CBP, ESR1, FBW1A, FOXO1, IKBA, NCOA3, NFKB1, P53, RS3, SMAD4, TF65
IKKA	nucleoplasm	11.68	$1.97e^{-03}$	ANDR, CBP, ESR1, FOXO1, NCOA3, NFKB1, P53, SMAD4, TF65, UBC
IKKA	I-kappaB/NF-kappaB complex	806.18	$3.07e^{-03}$	IKBA, NFKB1
IKKB	cytosol	11.24	$1.19e^{-03}$	SMAD3, TF65, UBC
INSR	cytosol	12.11	$5.41e^{-04}$	ABI1, GRB2, IRS1, P85A, SRC, UBC
JAK2	cytosol	5.61	$3.98e^{-02}$	1433Z, GRB10, GRB2, HS90B, MP2K1, MP2K2, P85A
KC1A	lateral plasma membrane	38.75	$3.46e^{-02}$	CTNB1
KC1A	APC-Axin-1-beta-catenin complex	275.94	$4.18e^{-02}$	CTNB1
KC1D	axon	25.5	$2.77e^{-04}$	ANDR, PSN1
KC1D	growth cone	45.05	$1.48e^{-03}$	PSN1
KC1D	Axin-APC-beta-catenin-GSK3B complex	170.42	$3.30e^{-02}$	GSK3B
KCC2G	vesicle membrane	19.29	$7.14e^{-04}$	GRB2, NCK1
KCC4	nucleoplasm	18.65	$2.78e^{-03}$	ESR1, HIF1A
KCC4	transcription factor complex	20.48	$1.69e^{-02}$	HIF1A
KPCD	integral to plasma membrane	3.18	$4.68e^{-02}$	ERBB3, IGF1R
KS6A5	nucleus	8.9	$2.51e^{-02}$	EP300, NCOR2
KS6B1	cytosol	9.74	$7.05e^{-04}$	AKT1, KS6A1
KSYK	plasma membrane	7.24	$5.40e^{-04}$	EGFR, ERBB2, FLNA, FYN, GRB2, NCK1, P85A, PLCG1, SH3K1, SRC, UBC
LYN	plasma membrane	5.21	$3.18e^{-04}$	CRK, EGFR, ERBB2, ERBB3, FYN, GRB2, LCK, NCK1, P85A, PLCG1, SRC
M3K8	nucleus	16.76	$1.37e^{-02}$	AAKB1, IKKE, MPIP3, TRAF6
M3K8	cytosol	8.43	$6.26e^{-03}$	AAKB1, IKKE, MPIP3, TRAF6, TSC1
MAPK2	cytosol	5.52	$5.79e^{-03}$	1433B, 1433Z, CHK1, DAXX, GSK3B, MDM2, P53, PLK1
MK03	nuclear chromosome	3.7	$3.17e^{-02}$	JUN
MK03	nuclear chromatin	6.24	$9.16e^{-04}$	ANDR, CBP, NCOA1, NCOR1, P53, RXRA
MK03	insoluble fraction	15.79	$3.25e^{-03}$	MK01
MK03	nucleus	2.95	$5.63e^{-06}$	1433B, ANDR, CBP, CSK2B, DAXX, EGFR, EP300, ESR1, GCR, GRB2, HDAC1, MDM2, MK01, MK14, NCOA3, NCOA6, NCOR1, NCOR2, NR0B2, NR1H2, NR4A1, P53, PIAS1, PPARG, RB, RXRA, SMAD3, SMAD4, SP1, SRC, SUMO1, TF65, UBC9
MK03	nucleoplasm	2.93	$1.86e^{-05}$	1433Z, ANDR, CBP, EP300, ESR1, GCR, HDAC1, JUN, MDM2, MED25, MK01, MK14, NCOA1, NCOA3, NCOA6, NCOR1, NCOR2, NR0B2, NR1H2, NR4A1, P53, PIAS1, PPARG, RB, RXRA, SMAD3, SMAD4, SUMO1, TF65
MP2K1	soluble fraction	16.06	$4.68e^{-02}$	MP2K2
MP2K1	late endosome	51	$7.89e^{-03}$	MP2K2
MP2K1	focal adhesion	34.16	$2.92e^{-03}$	MP2K2
MP2K4	nucleoplasm	27.92	$7.06e^{-03}$	DUS1, JUN, NCOA3, TF65
PAK2	cytosol	12.94	$1.35e^{-02}$	GRB2
PDPK1	soluble fraction	9.11	$2.71e^{-02}$	1433Z
PDPK1	mitochondrion	6.81	$2.48e^{-02}$	1433Z, CASP3, MAD1, PDK1
PKN1	histone deacetylase complex	81.8	$3.63e^{-02}$	CBX5
PLK1	nucleus	4.33	$7.49e^{-04}$	ABL1, ANDR, GRB2, P53, VHL
PLK1	nucleoplasm	5.34	$2.47e^{-06}$	ANDR, P53
PLK1	cytosol	3.44	$2.83e^{-03}$	ABL1, GRB2, P53, VHL
PLK3	nucleus	inf	$4.60e^{-02}$	CHK1, HDAC1, IMA2, PLK1
PLK3	nucleoplasm	12.43	$1.15e^{-02}$	CHK1, HDAC1, IMA2, PLK1, UBC, XPO1
PRKDC	nucleus	30.56	$1.84e^{-06}$	ANDR, APLF, CBP, CEBPB, CHD1L, CSK21, DAXX, EP300, GRB2, GSK3B, IMA2, MDM2, NCOR2, NR4A1, PNKP, RFA1, SIRT1
PRKDC	nucleoplasm	9.45	$4.99e^{-07}$	1433Z, ANDR, CBP, EP300, IMA2, MDM2, NCOR2, NR4A1, RFA1, SIRT1, UBC

Continued on next page

Table 3 – continued from previous page

Kinase	GO description	Enrichment ratio	Adjusted *p*-value	Potential adaptor or scaffold
RET	cytosol	11.64	$2.71e^{-02}$	KS6B1
TTK	nucleolus	17.51	$4.41e^{-02}$	P53
TTK	PML body	61.38	$4.75e^{-03}$	P53, UBC9

A6 Submitted article

Article submitted to Molecular BioSystems.

CREATED USING THE RSC LaTeX PCCP ARTICLE TEMPLATE – SEE www.rsc.org/electronicfiles FOR DETAILS

ARTICLE TYPE www. rsc.org/xxxxxx | XXXXXXXX

Specificity-determinants in human protein kinases †

M.A. Alonso-Tarajano,[a] R. Mosca,[a] and P. Aloy*[ab]

Received DD/MM/YYYY, Accepted DD/MM/YYYY
First published on the web DD/MM/YYYY
DOI: 10.1039/xyz

Protein kinases participate in a myriad of cellular processes of major biomedical interest. The *in vivo* substrate specificity of these enzymes is a process determined by several factors, and despite several years of research on the topic, is still far from being totally understood. In the present work, we have quantified the contributions to the kinase substrate specificity of the phosphorylation sites and their surrounding residues in the sequence and of the association of kinases to adaptor or scaffold proteins. We have used position-specific scoring matrices (PSSMs) to represent the stretches of sequences phosphorylated by different families of kinases. From these PSSMs we have identified specificity determinant residues (SDRs) for several kinase families, and we have also identified relationships between the number of phosphorylation sites from which a PSSM is generated and the statistical significance and the performance of that PSSM. Additionally, we have found that proteins with known function as adaptors or scaffolds (kAS), tend to interact with a large fraction of the substrates of the kinases. Based on this characteristic, we have used the human interactome to identify a set of potential adaptors/scaffolds (pAS) for human kinases. Our results suggest that pAS proteins tend to co-localize with the substrates of the kinases they are associated to, and that these associations may contribute significantly to diminish crossed-specificity of protein kinases. In general, our results indicate the relevance of several SDRs for both the positive and negative selection of phosphorylation sites by kinase families and also suggest that the association of kinases to pAS proteins may be an important factor for the localization of the enzymes with their set of substrates.

1 Introduction

Phosphorylation is the most common post-translational modification of proteins, and is also an important mechanism for the regulation of protein function.[1] Protein phosphorylation is a reversible and fast reaction that have been conserved in evolution as a mechanism for regulating proteins in a non transcription-dependent manner.[2] The addition (or removal) of a phosphate group, can regulate different characteristics and properties of the affected protein such as its conformation, its activation state, its interactions with other proteins or its cellular localization.[3]

Protein kinases are the enzymes that catalyze the phosphorylation reaction. In human there have been described 518 protein kinases, which constitutes one of the largest families of proteins and accounts for nearly 2% of our genes.[4] Kinases are key players in several cellular processes and their deregulation have been tightly related to pathologies such as cancer[2,5] and diabetes[6,7]. Most protein kinases share a common fold of the catalytic domain, but despite their similarities at the catalytic region, kinases have achieved a remarkable sequence diversity by combining different classes of protein domains.[4,8] Indeed, this diversity plays a major role in the substrate specificity and functional aspects observed *in vivo* for these enzymes.[9–11] In general, the *in vivo* substrate specificity observed in kinases, is known to be determined by several contextual factors such as the sequence vicinity of the phosphorylation site, cellular localization, cell-type specific coexpression and interactions of kinases and their substrates with adaptor or scaffold proteins.[12,13]

Advances in high-throughput phosphoproteomic methodologies, have provided valuable data of experimentally determined phosphorylation sites for hundreds of kinases from yeast, human and other organisms.[11,14–17] Based on the aforementioned data, several authors have studied the kinase specificity by analyzing different sequence motifs that are targeted by the kinases in their substrates.[9,13,18,19] These motifs — often termed phosphorylation motifs — have been generally represented in position-specific scoring matrices (PSSMs), which allow the probabilistic modeling of signals in sequence alignments.[20] PSSMs have been previously used for the prediction of novel phosphorylation sites and for the assignment of experimentally determined phosphorylation sites to kinases.[21,22] Other more sophisticated methods for the prediction of phosphorylation sites implement complex algorithms such as hidden Markov models, artificial neural networks or expert systems to integrate several sources of information (*e.g.*, struc-

† Electronic Supplementary Information (ESI) available: DOI: 10.1039/xyz

[a] *Joint IRB-BSC Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), Baldiri Reixac 10-12, Barcelona 08028, Spain. Tel: +34 93 40 39690; E-mail: patrick.aloy@irbbarcelona.org*
[b] *Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona 08010, Spain*

tural disorder, sequence conservation, positional correlations of residues).[23] However, in those cases it is more difficult to infer the decisions that support the predictions, as opposed to the cases of PSSMs, where it is much easier to pinpoint the determinant residues of a functional phosphorylation site.[23]

Adaptors and scaffolds are multidomain proteins involved in the dynamic spatio-temporal organization of large signaling complexes and cellular structures.[24] Due to their roles in cellular signaling, some of these proteins have be implicated in cancer and tumorigenesis.[25,26] The specificity of many signal transduction events is modulated by adaptors and/or scaffolds, which can recruit signaling enzymes to proper cellular locations[27,28]. Indeed, the associations of kinases with adaptors and scaffolds can enhance efficient catalytic activation and accurate substrate selection. This is the case of the PKA kinase, which is targeted to discrete cellular environments by the A-kinase anchoring protein (AKAP)[29]. Other two examples are the kinase suppressor of Ras (KSR) and IQGAP, which function as platforms and regulators of the mitogen-activated protein kinase (MAPK) pathway.[25,30] Adaptors and scaffolds are extremely diverse proteins which lack common sequence signature motifs. Therefore, their identification based only on sequence is currently not possible. Nevertheless, these proteins often contain protein-protein interaction domains (*e.g.*, SH2, SH3 and PD) and it has been suggested that some scaffolds interact with at least two signaling proteins[24]. Based on these characteristics, Ramirez and Albrecht devised a computational method from which they identified 250 potential human signaling scaffolds[31]. However, in their analysis the authors excluded proteins with intrinsic catalytic activity as potential scaffolds, a criteria that may constitute a limitation of their method[32,33].

In this article we explored two elements that contribute to the substrate specificity of human protein kinases. First, we focused on the identification of SDRs in the sequences phosphorylated by several kinase families families, and we quantified their contribution to the specificity of those families. Second, we studied how the association of kinases to adaptor and scaffold proteins may influence the cellular colocalization of kinases and their substrates, and also how these associations may diminish the substrate crossed-specificity of kinases.

2 Materials and methods

2.1 Integration of human phosphorylation data

We compiled a local database of experimentally determined phosphorylation sites by integrating data from the public resources HPRD[34], PhosphoSitePlus[35] and Phospho-ELM[36]. We kept only those phosphorylation sites for which the responsible kinase was known and we filtered out those without a supporting publication. Our integrated set increases by 18%, 58% and 59% the numbers of kinases, substrates and phosphorylation sites (respectively), with respect to the average contained in the three source databases (see section 1 of supplementary material).

2.2 Construction of the position-specific scoring matrices

For generating the PSSMs and estimating their statistical significance and performance, we have developed the program `genpssm`. As input `genpssm` takes i) a multiple sequence alignment of nine residues long peptides with the phosphorylation site in the central position ii) the frequencies of amino acids in the human proteome and iii) a cut-off p-value ($1e^{-04}$) for selecting matches to the PSSM.The scores of the PSSM are computed using the equation 1, which is based in the log-odds of residues at each position of the alignment and also considers the frequencies of residues in the human proteome.[37]

$$S_{rp} = \log(\frac{q_{rp}}{f_r}),\ p = 1\ to\ w \tag{1}$$

S: score of residue r at position p; q: frequency of residue r at position p; f: frequency of residue r in the reference proteome and w: length of the sequence alignment.

2.3 Evaluation of the position-specific scoring matrices

We have used the information content (IC) the percent of recall (recall) and the area under the receiver operating characteristic curve (AUC-ROC) to evaluate the statistical significance and the performance of the PSSMs. By percent recall we mean the fraction of seed phosphorylation sequences that match the cognate PSSM with a statistically significant score. The statistical significance tests were based on empirical p-values for both the IC and the recall. For this, we used sets of 100'000 PSSMs that have been generated from random sequences, and that also respect the cardinality of seed sequences of the PSSM being assessed. For computing the IC we have used the KullbackLeibler distance[38] (see equation 2), where the IC is the sum of the expected self-information of each element.

$$IC = -\sum_{r,p} q_{rp} \times \log(\frac{q_{rp}}{f_r}) \tag{2}$$

IC: information content; q: frequency of residue r in position p of the sequence alignment; f: frequency of the residue r in the reference proteome.

2.4 Identification of specificity-determinant residues

We have selected 22 kinase families with at least 100 phosphorylation events. For each family, we have attempted the identification of residues that could contribute significantly to the specificity of the corresponding kinases (*i.e.*, the SDRs).

For this, based on the corresponding PSSMs, we have classified as SDRs those residues with a score equal or higher than half the score of the phospho-acceptor residue. Finally, we computed the frequency of each SDR across the phosphorylation events of each family in the experiment.

2.5 Identification and analysis of known adaptors and scaffolds

We collected from UniProtKB/Swiss-Prot all human proteins annotated with either adaptor or scaffold terms in their Function field. We filtered out the cases without evidence of binary interaction with at least one human protein kinase. For this, we used the high confidence human interactome from Interactome3D[39], a resource developed by our group. From this, we obtained a set of 191 known adaptor or scaffold (kAS) proteins, which associate to 287 human protein kinases.

Some adaptors and scaffolds associate to both the kinases and their substrates. Based on this, we tested whether the kAS proteins interact with a statistically significant number of the substrates of the kinases to which they are associated. For this we have used as the statistic the number of interactions of proteins in a subnetwork of the interactome. For constructing the backgrounds for the statistical test, we first selected the kinases with at least five substrates (156 kinases in total) and using those substrates as seeds we generated a first level subnetwork of the human interactome. We generated different backgrounds depending on the cardinality of substrates (S) of each kinase. For generating the backgrounds we started by randomly selecting a node (K) having at least S partners. Later, for a number S of randomly selected K's partners, we identified the first neighbors (P). Finally, we counted the number of interactions between each P and all K's partners. While randomly rewiring the subnetwork, we repeated the process 10'000 times for each background set. For testing the initial hypothesis, we conducted a right tale Fisher's exact test.

2.6 Identification of kinases sharing a significant number of substrates

We have investigated if there exist a relationship between the association to common adaptors or scaffolds and the substrate cross-specificity of kinases. For this, we have approached the identification of kinases sharing at least one kAS protein and also sharing a significant number of *in vivo* substrates. The size of the overlap between the sets of substrates was used as a test statistic. For estimating the statistical significance of the overlaps, we computed empirical p-values for sets of kinases with cardinality two and three. We performed the analysis only for the 111 kinases for which we known at least five *in vivo* substrates.

3 Results and discussion

3.1 Position-specific scoring matrices

Here we have analyzed the performance and statistical significance of the PSSMs corresponding to the 93 kinase families for which we count with at least one phosphorylation site. Regarding their performance, we have found negative correlations between the number of seed phosphorylation sites and i) the recall (R = –0.48, p-value = $1.2e^{-06}$), ii) the IC (R = –0.33, p-value = 0.0013) and iii) the AUC-ROC (R = –0.47, p-value = $1.6e^{-06}$). These results suggest that in our data, the increase of the sequence diversity generated by the increase of the number of seed phosphorylation sites, can exert a negative effect in both the performance and the level of self-information of a PSSM (see Figure 1). We suggest that the substrate specificity of some kinases and kinase families might be represented best by multiple PSSMs, a concept that have been previously applied in the analysis of DNA recognition by transcription factors[40]. Although not covered in the work here presented, we consider that in such cases, multiple PSSMs could be useful for modeling fairly different phosphorylation motifs that are targeted by the same kinase or kinase family.

The IC can be used as a statistic to estimate how different is that PSSM from a uniform distribution. From our analysis, 69/93 (74.2%) of the PSSMs were found to be statistically significant; and the two sets of PSSMs — significant and not significant — differ in their median values of the percent recall, the AUC-ROC and the number of seed phosphorylation sites. Our results show that PSSMs with a statistically significant IC were generated from sets of seed phosphorylation sites larger than the ones from not statistically significant PSSMs. In this sense, and in agreement to what was previously mentioned, significant PSSMs show significantly lower values of recall and AUC-ROC (see Table 1 and Figure 2). Surprisingly, we have not found significant differences between the two sets of PSSMs based on their median IC values. However, in an equivalent comparison using PSSMs from independent kinases, we have found significant differences if the median IC values between sets of significant and non significant PSSMs (Mann-Whitney U test p-value = $6.2e^{-04}$).

Based on the results of the current analysis, we selected a subset of significant PSSMs to conduct the identification of specificity-determinant residues (SDRs) for the corresponding families of kinases.

3.2 Specificity-determinant residues

From the previously identified group of significant PSSMs, we selected 22 for which we count with at least 100 phosphorylation events. For 19/22 (86.4%) of the families analyzed we identified at least one SDR. For these 19 families we have successfully classified as SDRs residues that have been reported

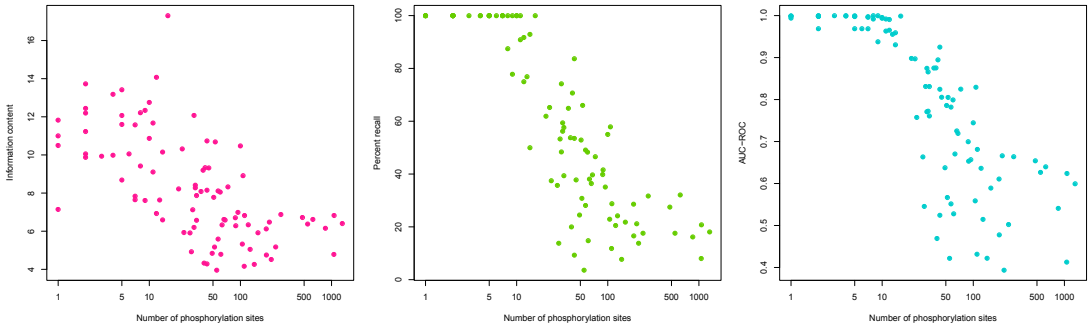


Fig. 1 Correlations of measurements with the number of seed phosphorylation sites. The IC, percent recall and AUC-ROC display negative correlations with the number of seed phosphorylation sites. X-axes are displayed in logarithmic scale.

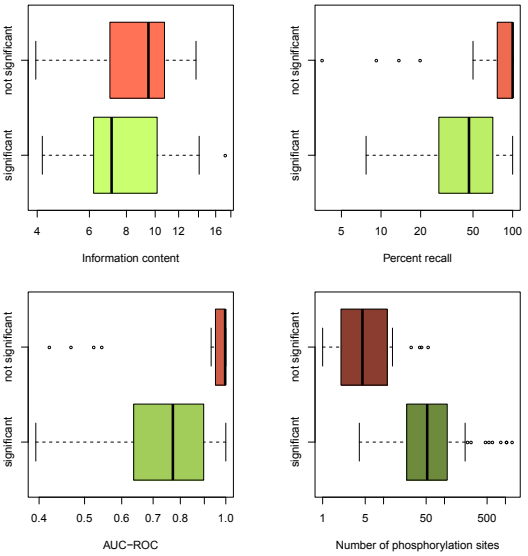


Fig. 2 The PSSMs were classified based on the significance of their IC. The two groups of PSSMs were later compared based on their IC, percent recall, AUC-ROC and number of seed phosphorylation sites. The thick lines in the boxes represent the medians.

Table 1 Comparison of significant and not significant sets of PSSMs.

	Total PSSMs	IC	% recall	AUC-ROC	Psites
Significant	69	7.13	46.60	0.77	52.00
Not significant	24	9.49	100.00	1.00	4.50
p-value		0.177	$1.89e^{-04}$	$1.36e^{-04}$	$8.57e^{-09}$

The table shows the median values of the parameters used for comparing the two sets of PSSMs. The last row shows the results of the Mann-Whitney *U* test, which is based on the differences of the medians (significance level $\alpha < 0.05$). Psites stands for seed phosphorylation sites.

to play important roles in the specificity of the corresponding kinases (*e.g.*, $MAPK_{P+1}$, $PIKK_{Q+1}$, AKT_{R-3} and $CK2_{E+3}$ ⋆). The quantification of the relevance of the SDRs — based on their frequency among the phosphorylation events of each family — shows a wide variation across the different families. For example, the four SDRs previously mentioned have relatively high frequencies that range between 88.86% and 45.83%; however, other SDRs show much lower frequencies (*e.g.*, PKC_{K+2} = 19.79%, $CAMKL_{N+3}$ = 18.03% and $CK2_{D+2}$ = 15.54%, see Table 2).

Based on our data, we hypothesize that the combination of multiple SDRs of low frequencies contribute in an additive way to the recognition of the phosphorylation sites by the kinases. In contrast, we consider that SDRs of high frequencies have a larger contribution to the kinase specificity. Moreover, we have noted that the frequency of any given SDR is low — 6.0% on average — among the phosphorylation events of the kinase families that do not count with that SDR. To our opinion, this suggests that SDRs may also function as elements of negative selection to avoid the phosphorylation of non-cognate sequences.

⋆ SDRs are represented by the acronym of the kinase family, followed by the residue (one letter code) and its position relative to the phosphorylation site.

Table 2 Specificity-determinant residues of kinase families.

Kinase family	SDR	% freq.	% cross-freq.
CDK	P+1	81.72	5.95
GSK	S-4	38.49	14.01
GSK	P+1	53.96	5.95
GSK	S+4	48.56	11.51
MAPK	P-2	31.37	5.83
MAPK	P+1	88.86	5.95
PIKK	Q+1	80.83	3.98
AKT	R-3	84.13	5.14
AKT	W+1	3.85	0.66
CAMKL	R-3	31.15	5.14
CAMKL	K-3	21.31	5.65
CAMKL	N+3	18.03	3.83
PKC	R-3	23.19	5.14
PKC	R-2	24.32	5.09
PKC	R+2	27.71	4.57
PKC	K+2	19.79	3.88
CK1	S-3	28.5	8.08
CK1	S+3	31.09	9.85
CK2	D+2	15.54	6.17
CK2	E+3	45.83	6.15

In the table, **% freq.**: frequency of the SDR among the phosphorylation events of current kinase family. **% cross-freq.**: frequency of the SDR among the complementary phosphorylation events, that is, the ones from kinase families without the current SDR.

We have identified SDRs that, to the best of our knowledge, have not been previously reported as determinants of the specificity for the corresponding kinase families. These are the cases of $CAMKL_{N+3}$ and AKT_{W+1}, with frequencies of 18.03% and 3.85% respectively. The SDR N+3, is present in the sequences targeted by the microtubule affinity-regulating kinases (MARK) — CAMKL family members — within the repeat regions of the human TAU protein, which is implicated in Alzheimer's disease[41]. Besides, N+3 have a low frequency (3.83%) among the phosphorylation events of the other 21 kinase families in the analysis. Given that the repeat regions of TAU are responsible for the binding to the microtubules[42]; we consider that the presence of N+3 in these regions is an important element for the recognition by MARK kinases, and therefore for the regulation of the association of TAU to the microtubules. The identification of W+1 as an SDR for the AKT family is an interesting result, given that tryptophan is rarely found in the close sequence vicinity of phosphorylation sites — 0.66% among the phosphorylation events of non AKT kinase families —. W+1 was identified as an SDR even when occurring at low frequency (3.85%) among the phosphorylation events of AKT kinases, which prompted us to research further about the biological relevance of the finding. Interestingly we found reports in the literature showing that, by phosphorylating sequences containing a conserved W+1, some AKT kinases are implicated in the regulation of transcription factors of the FOXO family[43]. To our opinion, this result supports the utility of our approach for the identification of SDRs, even for residues that occur at low frequency among the phosphorylation sites of the kinase of interest.

3.3 Association of kinases to known adaptors and scaffolds

As previously described, we have compiled a set (kAS) of 191 human proteins that are known to function as adaptors or scaffolds (see section 6 in supplementary material). These 191 kAS proteins associate to 287 (55.4%) — via 1281 binary PPIs —- protein kinases, which represent a total of 94 (72.3%) kinase families and also comprise the nine major groups in which human kinases are classified. To our opinion, these findings suggest that the association to adaptors or scaffolds is a widespread mechanism among human protein kinases. The results from the analysis of enrichments in Pfam domain families[44] and molecular function terms (MF) of the Gene Ontology[45] show that 14/23 (60.8%) of the enriched Pfam domains are known to be directly involved in promoting PPIs (*e.g.*, PDZ, SH2 and SH3); and that 100% of the enriched MF terms are related to protein binding, adaptor or scaffolding functions (see sections 4 and 5 in supplementary material). Together, these results support the biological role as adaptors or scaffolds of the proteins in the kAS set.

Adaptors and scaffolds can recruit kinases to cellular compartments where the enzymes gain spatial proximity to its relevant set of substrates. In this manner, adaptors and scaffolds can function as linking elements between the kinases and their substrates. Based on this, we searched for evidence supporting that kAS proteins could interact with a large number of the substrates of the kinases to which they associate. The result of our analysis suggests that, compared to any random kinase partner, kAS proteins are five times more likely to interact with a significantly large number of the substrates of their corresponding kinases (p-value = $1.08e^{-15}$). This result supports our initial assumption and therefore we decided to use this property of kAS proteins to identify potential adaptors and scaffolds of proteins kinases in the human interactome.

3.4 Potential adaptors and scaffolds of protein kinases

We have identified a total of 706 associations kinase–potential adaptor/scaffold (K–pAS). These include 279 pAS proteins — 25.4% of them is present in the kAS set — that are known to interact with 78 (50%) of the 156 kinases initially considered for the experiment. The 78 kinases cover 44 (33.8%) of all human kinase families. Analysis of Pfam domains composition show enrichment 10 Pfam families, all of them known to mediate PPIs or to be present in proteins involved in cellular signaling (see section 7 in supplementary material). Half

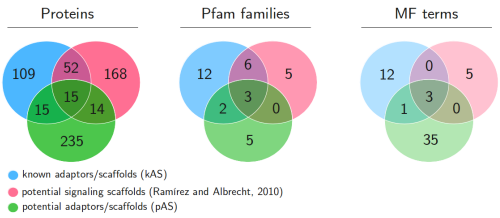


Fig. 3 Comparison of the sets of adaptors and scaffolds.

of the ten Pfam families enriched in the set of pAS proteins were also enriched in the kAS set, a finding that supports the hypothesis of common biological functions. Additionally, we have found 39 MF terms to be overrepresented in the pAS set (see section 6 in supplementary material). A considerable fraction of these terms (24/39, 61.5%) refer to 'protein binding' functions of signaling-related molecules such as receptors, kinases, phosphatases and transcription factors. This suggests pAS proteins could be able to mediate PPIs for different classes of signaling-related proteins. In contrast with the kAS set, for the pAS proteins we do not find enrichments in MF terms directly related to adaptor nor scaffolding functions; a result that supports the pAS proteins as a novel set of potential adaptors and scaffolds.

We have compared the three sets of adaptors and scaffolds commented in this work (*i.e.*, kAS, pAS and the set identified by Ramírez and Albrecht) in terms of their protein composition and enriched MF terms and Pfam domains (see 3). We have found a relatively low average overlap of proteins between the three sets (18.4%), which highlights the lack of a consensus criteria for the computational identification of adaptors and scaffolds. In contrast to our methods, Ramrez and Albrecht considered that scaffolds lack intrinsic enzymatic activity[46]. We consider this criteria to be inaccurate, given the cases of the focal adhesion kinase (FAK)[33] and the kinase suppressor of Ras (KSR)[32], which are both scaffolds with reported catalytic activity. Differences in the Pfam families and the MF terms enriched can be partially attributed to differences in sets of proteins defined as the backgrounds. Nevertheless, for all the three sets the Pfam domains and MF terms enriched support the hypothesis of adaptor or scaffolding roles. Finally, differences in the definition of human interactome can also influence the results of the identification strategies.

Taken together, we consider that our strategy have been able to suggest a set of potential adaptor and scaffold of human human protein kinases, whose functional annotations are in agreement with the proposed biological roles.

3.5 Cellular colocalization of kinases, adaptors, scaffolds and substrates

Adaptors and scaffolds can play a fundamental role in the *in vivo* specificity of protein kinases by promoting the cellular colocalization of these enzymes with their cognate substrates. Here we have searched for evidence of colocalization of the pAS proteins with the substrates of the associated kinases. For this, we have used the 706 K–pAS relations previoulsy identified, and we have evaluated whether a given pAS is annotated to a cellular component term (CC) — from the Gene Ontology database — that have been previously found to be enriched in the set of substrates of its associated kinase.

For 527/706 (74.6%) of the K–pAS pairs, we found evidence of colocalization between the pAS and the substrates. This set of 527 K–pAS pairs accounts for 41 kinases, 156 pAS proteins — corresponding to 52.6% and 55.9% (respectively) of the ones in the initial 706 kinase-pAS pairs— and 35 unique CC terms. In 4 we show a pie chart representation of the CC terms shared by the pAS proteins and the sets of substrates; while in the Table 3 we show cases of pAS proteins that are found to colocalize with substrates of their corresponding kinases. For example, the pair formed by the β-adrenergic receptor kinase 1 (ARBK1) and the Na(+)/H(+) exchange regulatory cofactor NHE-RF (NHRF1), where the later it has been reported to be involved in the scaffolding of β-adrenergic receptors — substrates of ARBK1 — at the plasma membrane[47]. Another example is the case of the checkpoint kinase-1 (CHK1) and the 14-3-3 protein zeta (1433Z), where the later it has been reported to be required for the nuclear retention of CHK1[48]. A third case is casein kinase α-1 (KC1A), for which we identified the catenin β-1 (CTNB1) as a pAS. KC1A phosphorylates CTNB1 at serine 45, both proteins are components of the canonical Wnt signaling pathway and they are also part of the large APC–Axin-1–β-catenin complex[49]. Interestingly, CTNB1 contains 12 repeats of the Armadillo (ARM) domain, which is implicated in mediating PPIs. It has been recently suggested that proteins containing ARM repeats, constitute an attractive modular system as scaffolds for peptide-mediated PPIs[50]. Therefore, we consider that CTNB1 may constitute a plausible scaffold that may promote spatial proximity between KC1A and its substrates.

To our opinion, these results suggest that the association to pAS proteins might play an important role in the colocalization of the analyzed kinases with their cognate sets of substrates. Nevertheless, we are aware that in many cases, the CC shared by the substrates and the pAS proteins are too broad (*e.g.*, cytosol, nucleoplasm, cytoplasm) and can not fully justify, based on spatial constrains, the substrate specificity of the kinases.

Table 3 Cellular component terms shared by pAS proteins and substrates.

Kinase	CC description	Enrichment ratio	Adj. *p*-value	pAS co-annot.
ARBK1	apical plasma membrane	15.7	$4.08e^{-02}$	NHRF1
CDK4	chromatin	16.12	$3.76e^{-02}$	EP300
CDK4	transcription factor complex	20.92	$1.67e^{-02}$	E2F4,EP300
CDK9	PML body	61.38	$4.19e^{-03}$	PIAS4
CHK1	nucleoplasm	10.16	$1.30e^{-04}$	1433Z,CHK2,EP300,MDM2,UBC
CHK2	PML body	29.86	$3.25e^{-03}$	RB,SIRT1,SUMO1
CSK	membrane raft	33.87	$2.74e^{-03}$	ERBB2
EGFR	endosome	8.47	$6.48e^{-04}$	FYN,GRB2,NTRK1
FYN	cell junction	3.98	$3.15e^{-02}$	PTN12
INSR	cytosol	12.11	$5.41e^{-04}$	ABI1,GRB2,IRS1,P85A,SRC,UBC
KC1A	lateral plasma membrane	38.75	$3.46e^{-02}$	CTNB1
KC1A	APC-Axin-1-beta-catenin complex	275.94	$4.18e^{-02}$	CTNB1
KCC2G	vesicle membrane	19.29	$7.14e^{-04}$	GRB2,NCK1
PDPK1	mitochondrion	6.81	$2.48e^{-02}$	1433Z,CASP3,MAD1,PDK1
PLK1	nucleus	4.33	$7.49e^{-04}$	ABL1,ANDR,GRB2,P53,VHL

CC description, description of the CC term enriched in the set of substrates of the kinase; **Enrichment ratio**, ratio of enrichment of the CC term; **Adjusted *p*-value**, multiple test correction by Bonferroni's method; **pAS co-annot.**, pAS proteins associated to the current kinase, that are annotated to the corresponding CC term. Kinases and pAS proteins are represented by their UniProt IDs. See full table of results in **SuppMat**.

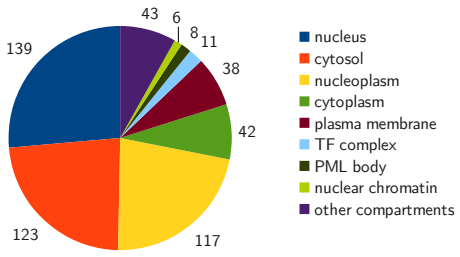


Fig. 4 Cellular component terms shared by substrates and pAS proteins. The slices represent the number of K–pAS pairs where the pAS protein is annotated to the the given CC term. TF and PML stand for transcription factor and nuclear bodies respectively.

3.6 Association to adaptors and scaffolds diminish substrate cross-specificity of kinases

We have analyzed the role that potential adaptors and scaffolds may play in kinase specificity by promoting spatial proximity between the enzymes and their substrates. However, different kinases may associate to the same adaptors and scaffolds and this could lead to substrate cross-specificity. Here we have tested whether the association to common kAS proteins would promote significant substrate cross-specificity between kinases. For this, we have used the subset of K–kAS associations where the kinases have at least five substrates and for which the kAS in the analysis are known to interact with at least two kinases. In total we analyzed 23 cases of two or more kinases that associate to a common adaptor or scaffold, and for non of the cases the kinases shared a number of *in vivo* substrates larger than what would be expected due to chance (see Table 4). Nevertheless, we found the case of the kinases MK01 and MK03 — ERK2 and ERK1 MAP kinases, respectively — which share 73 *in vivo* substrates. Even when it was not statistically significant, the number substrates in common was very large when compared to other sets of kinases in out analysis, and therefore we decided to explore this particular case in more detail. In fact, ERK1 and ERK2 are very closely related kinases, with 82% and 89% of identity in their full and catalytic domain sequences. ERK1 and ERK2 share many if not all functions[51] and despite numerous efforts to establish differences, the detection of such distinctive functions it has been difficult to pinpoint[52]. Therefore, we consider that their

large sequence identity, together with their almost identical functions can explain the large substrate overlap reflected in our data. To our opinion, these results support the hypothesis that adaptors and scaffolds are able to diminish *in vivo* substrate cross-specificity by recruiting the kinases to specific macromolecular complexes or cellular locations.

Table 4 Statistical significance of the number of shared substrates for K–kAS pairs.

Adap./Scaff.	Assoc. kinases	Shared subst.	*p*-value
APBB1	EGFR, ERBB2	2	1.00
BIRC5	AURKA, AURKB	4	1.00
CD2AP	ABL1, FYN	1	1.00
DAG1	FYN, SRC	13	0.52
DOK4	EGFR, ERBB2	2	1.00
DOK6	EGFR, ERBB2	2	1.00
ELP1	GSK3B, MK08	7	0.78
FRS3	MK01, FGFR1	1	1.00
FYB	ABL1, FYN	1	1.00
IMA2	SGK1, CHK2	1	1.00
JIP2	EGFR, ERBB2	2	1.00
KHDR1	LCK, SRC	14	0.43
NCK1	ABL1, EGFR	3	1.00
PAR6A	KPCI, KPCZ	3	1.00
PAR6B	KPCI, KPCZ	3	1.00
PKHO1	AKT1, CSK21	5	1.00
SCRIB	MK01, MK03	73	0.10
SH2B1	EGFR, INSR	4	1.00
SHC1	EGFR, INSR	4	1.00
SHC2	EGFR, ERBB2	2	1.00
SHC3	EGFR, ERBB2	2	1.00
SQSTM	KPCI, KPCZ	3	1.00
TGFI1	FAK1, FAK2	1	1.00

Proteins are represented by their UniProt Ids. **Shared substrates**, number of *in vivo* substrates shared by the kinases; ***p*-value**, statistical significance of the number of substrates shared by the kinases.

4 Conclusions

Protein kinases constitute one of the largest and more diverse superfamilies of proteins in human and they are implicated in several cellular processes and pathologies [1,6]. Despite most kinases share a highly conserved catalytic domain, the observed *in vivo* substrate specificity of these enzymes show little correlation with their primary sequences. In this sense, is known that the *in vivo* specificity of protein kinases is regulated by several factors. In the current work we have approached the identification and the quantification of the contribution of different elements to the substrate specificity of human protein kinases. For this we have analysed the residues in the close neighbourhood of the the phosphorylation site, the association of kinases to adaptors and scaffolds and the cellular colocalization of kinases and their substrates.

I this work we have generated PSSMs from the sequences targeted by 93 families of kinases and we have analyzed their statistical significance and performance. We have found negative correlations between the number of seed phosphorylation sites and *a*) the percent recall, *b*) the information content and *c*) the AUC-ROC of the PSSMs. Based on the IC we have estimated the statistical significance of the PSSMs. We have observed that statistical and non-statistically significant PSSMs show disgnificant differences in the number of seed phosphorylation sites and on their performance parameters (*i.e.*, the percernt recall and the AUC-ROC). Our results show the negative effect that the sequence degeneracy caused by the increase of the seed phosphorylation sites can impose on the performance and on the level of self-information of the PSSMs.

Starting from 22 statistically significant PSSMs, we have identified several SDRs that function as positive (or negative) elements for the substrate recognition by different kinases families. The SDRs identified among the different kinase families show high diversity in terms of the type of residue, the position relative to the phosphorylation site and the frequency among phosphorylated sequences available. Some kinase families are very specific towards particular SDRs, which occur in more than 80% of the sequences they target (*e.g.*, $AKT_{R\text{-}3}$, CDK_{P+1}, $MAPK_{P+1}$ and $PIKK_{Q+1}$). We have observed that multiple SDRs are generally identified in families for which the frequencies of the SDRs range approximately between 15% and 55% of the target sequences (*e.g.*, $CK1_{S\text{-}3}$, $CK2_{D\text{-}1}$, $GSK_{S\text{-}4}$, GSK_{P+1} and $PLK_{E\text{-}2}$). Our opinion is that in such cases cases, multiple SDRs may contribute cooperatively to the recognition of the phosphorylation site. We have also noted that the SDRs occur at low frequency (6.01% on average) among the complementary target sequences (*i.e.*, the phosphorylation sites corresponding to those kinase families that do not count with the given SDR). To our opinion, this suggests that an SDR contribute as a negative selection factors for non-cognate phosphorylation sites.

We have compiled a set of 191 proteins with known roles as adaptors or scaffolds and that associate to 55% of the human kinases, which account for 72.3% of all human kinase families. When compared to random proteins in the human interactome, this set of proteins was five times more likely to interact with a large fraction of the substrates of the human kinases to which they associate. To our opinion, these results suggest that the association to adaptors or scaffolds is a common mechanism among human kinases and also supports the concept of adaptors and scaffolds as mediators in the encounter of kinases with their cognate substrates.

In this work we devised a strategy for the identification of potential adaptors and scaffolds of human protein kinases. For 50% of the initial kinases in the analysis we identified a total of 279 potential adaptors/scaffolds. This set of proteins is enriched in functional terms and in domain families that suggest a tight link to protein-protein binding functions involved in cellular signalling events. We have also found that for 74.6% of the kinase–potential adaptor/scaffold associations identi-

fied, the adaptor/scaffold is annotated under cellular compartment terms found to be enriched among the set of substrates of the associated kinase. We consider that these results put forward a role for the potential adaptors/scaffolds in promoting the colocalization of the kinases and their sets of substrates.

Finally, we analyzed whether the association of different kinases to common adaptors/scaffolds, may relate with the *in vivo* substrate cross-specificity of that kinases. We have not found any case of two or more kinases that, having an adaptor or scaffold in common, also share a number of *in vivo* substrates larger than what would be expected by chance. To our opinion, these results suggest that the association of kinases to adaptors and/or scaffolds may play important roles in the localization of the enzymes with their set of cognate substrates and also in diminishing substrate cross-specificity *in vivo*.

5 Acknowledgements

MAAT is supported by a 'la Caixa'/IRB Barcelona International Ph.D. Programme Fellowship (01/09/FLC). MAAT would like to thank M. Duran-Frigola (IRB Barcelona), A. Stein and R.A. Pache (UCSF, USA) and A. Zanzoni (TAGC-U1090 Inserm, France) for helpful discussions.

References

1 P. Cohen, *Nature cell biology*, 2002, **4**, E127–30.
2 S. Arena, S. Benvenuti and a. Bardelli, *Cellular and molecular life sciences : CMLS*, 2005, **62**, 2092–9.
3 A. Forrest, T. Ravasi, D. Taylor, T. Huber, D. Hume and S. Grimmond, *Genome research*, 2003, **13**, 1443.
4 G. Manning, D. B. Whyte, R. Martinez, T. Hunter and S. Sudarsanam, *Science (New York, N.Y.)*, 2002, **298**, 1912–34.
5 A. Torkamani, G. Verkhivker and N. J. Schork, *Cancer letters*, 2009, **281**, 117–27.
6 P. Cohen, *Eur J Biochem.*, 2001, **268**, 5001–5010.
7 L. R. Pearce, D. Komander and D. R. Alessi, *Nature reviews. Molecular cell biology*, 2010, **11**, 9–22.
8 K. Deshmukh, K. Anamika and N. Srinivasan, *Progress in biophysics and molecular biology*, 2010, **102**, 1–15.
9 J. Alexander, D. Lim, B. a. Joughin, B. Hegemann, J. R. a. Hutchins, T. Ehrenberger, F. Ivins, F. Sessa, O. Hudecz, E. a. Nigg, A. M. Fry, A. Musacchio, P. T. Stukenberg, K. Mechtler, J.-M. Peters, S. J. Smerdon and M. B. Yaffe, *Science signaling*, 2011, **4**, ra42.
10 A. N. Kettenbach, D. K. Schweppe, B. K. Faherty, D. Pechenick, A. a. Pletnev and S. a. Gerber, *Science signaling*, 2011, **4**, rs5.
11 J. Ptacek, G. Devgan, G. Michaud, H. Zhu, X. Zhu, J. Fasolo, H. Guo, G. Jona, A. Breitkreutz, R. Sopko, R. R. McCartney, M. C. Schmidt, N. Rachidi, S.-J. Lee, A. S. Mah, L. Meng, M. J. R. Stark, D. F. Stern, C. De Virgilio, M. Tyers, B. Andrews, M. Gerstein, B. Schweitzer, P. F. Predki and M. Snyder, *Nature*, 2005, **438**, 679–84.
12 J. A. Ubersax and J. E. Ferrell, *Nature reviews. Molecular cell biology*, 2007, **8**, 530–41.
13 B. Kobe, T. Kampmann and J. K. Forwood, *Biochimica et biophysica acta*, 2005, **1754**, 200 – 209.
14 H. Zhu, J. F. Klemic, S. Chang, P. Bertone, a. Casamayor, K. G. Klemic, D. Smith, M. Gerstein, M. a. Reed and M. Snyder, *Nature genetics*, 2000, **26**, 283–9.
15 J. V. Olsen, B. Blagoev, F. Gnad, B. Macek, C. Kumar, P. Mortensen and M. Mann, *Cell*, 2006, **127**, 635–48.
16 J. Mok, P. M. Kim, H. Y. K. Lam, S. Piccirillo, X. Zhou, G. R. Jeschke, D. L. Sheridan, S. a. Parker, V. Desai, M. Jwa, E. Cameroni, H. Niu, M. Good, A. Remenyi, J.-L. N. Ma, Y.-J. Sheu, H. E. Sassi, R. Sopko, C. S. M. Chan, C. De Virgilio, N. M. Hollingsworth, W. a. Lim, D. F. Stern, B. Stillman, B. J. Andrews, M. B. Gerstein, M. Snyder and B. E. Turk, *Science signaling*, 2010, **3**, ra12.
17 B. Hegemann, J. R. a. Hutchins, O. Hudecz, M. Novatchkova, J. Rameseder, M. M. Sykora, S. Liu, M. Mazanek, P. Lenart, J.-K. Heriche, I. Poser, N. Kraut, a. a. Hyman, M. B. Yaffe, K. Mechtler and J.-M. Peters, *Science Signaling*, 2011, **4**, rs12–rs12.
18 A. Kreegipuu, N. Blom, S. Brunak and J. Ja, *FEBS letters*, 1998, **430**, 45–50.
19 R. H. Newman, J. Hu, H.-S. Rho, Z. Xie, C. Woodard, J. Neiswinger, C. Cooper, M. Shirley, H. M. Clark, S. Hu, W. Hwang, J. Seop Jeong, G. Wu, J. Lin, X. Gao, Q. Ni, R. Goel, S. Xia, H. Ji, K. N. Dalby, M. J. Birnbaum, P. a. Cole, S. Knapp, A. G. Ryazanov, D. J. Zack, S. Blackshaw, T. Pawson, A.-C. Gingras, S. Desiderio, A. Pandey, B. E. Turk, J. Zhang, H. Zhu and J. Qian, *Molecular Systems Biology*, 2013, **9**, 1–12.
20 G. Z. Hertz and G. D. Stormo, *Bioinformatics (Oxford, England)*, 1999, **15**, 563–77.
21 J. C. Obenauer, *Nucleic Acids Research*, 2003, **31**, 3635–3641.
22 N. F. W. Saunders, R. I. Brinkworth, T. Huber, B. E. Kemp and B. Kobe, *BMC Bioinformatics*, 2008, **11**, 1–11.
23 N. Blom, T. Sicheritz-Pontén, R. Gupta, S. Gammeltoft and S. r. Brunak, *Proteomics*, 2004, **4**, 1633–49.
24 M. C. Good, J. G. Zalatan and W. a. Lim, *Science*, 2011, **332**, 680–686.
25 C. D. White, M. D. Brown and D. B. Sacks, *FEBS letters*, 2009, **583**, 1817–24.
26 H. Zhang, A. Photiou, G. Arnhild, J. Stebbing and G. Giamas, *Cellular signalling*, 2012, **24**, 1173–1184.
27 A. S. Shaw and E. L. Filbert, *Nature reviews. Immunology*, 2009, **9**, 47–56.
28 A. Alexa, J. Varga and A. Reményi, *The FEBS journal*, 2010, **277**, 4376–82.
29 M. Colledge and J. D. Scott, *Trends in cell biology*, 1999, **9**, 216–21.
30 M. M. McKay, D. a. Ritt and D. K. Morrison, *Proceedings of the National Academy of Sciences of the United States of America*, 2009, **106**, 11022–7.
31 F. Ramírez and M. Albrecht, *Trends in cell biology*, 2010, **20**, 2–4.
32 D. F. Brennan, A. C. Dar, N. T. Hertz, W. C. H. Chao, A. L. Burlingame, K. M. Shokat and D. Barford, *Nature*, 2011, **472**, 366–9.
33 W. G. Cance, E. Kurenova, T. Marlowe and V. Golubovskaya, *Science Signaling*, 2013, **6**, pe10–pe10.
34 T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady and A. Pandey, *Nucleic acids research*, 2009, **37**, D767–72.
35 P. V. Hornbeck, I. Chabra, J. M. Kornhauser, E. Skrzypek and B. Zhang, *Proteomics*, 2004, **4**, 1551–61.
36 H. Dinkel, C. Chica, A. Via, C. M. Gould, L. J. Jensen, T. J. Gibson and F. Diella, *Nucleic acids research*, 2010, **39**, 261–267.
37 J. M. Claverie and S. Audic, *Computer applications in the biosciences : CABIOS*, 1996, **12**, 431–9.
38 G. D. Stormo, *Bioinformatics (Oxford, England)*, 2000, **16**, 16–23.
39 R. Mosca, A. Céol and P. Aloy, *Nature methods*, 2012, **10**, 47–53.
40 G. Badis, M. F. Berger, A. a. Philippakis, S. Talukder, A. R. Gehrke, S. a. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-F.

Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes and M. L. Bulyk, *Science (New York, N.Y.)*, 2009, **324**, 1720–3.
41 D. Matenia and E.-M. Mandelkow, *Trends in biochemical sciences*, 2009, **34**, 332–42.
42 A. Alonso, T. Zaidi, M. Novak, I. Grundke-Iqbal and K. Iqbal, *Proceedings of the National Academy of Sciences of the United States of America*, 2001, **98**, 6923–8.
43 H. Matsuzaki, A. Ichino, T. Hayashi, T. Yamamoto and U. Kikkawa, *Journal of biochemistry*, 2005, **138**, 485–91.
44 M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman and R. D. Finn, *Nucleic acids research*, 2012, **40**, D290–301.
45 M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, *Nature genetics*, 2000, **25**, 25–9.
46 A. Zeke, M. Lukács, W. a. Lim and A. Reményi, *Trends in cell biology*, 2009, **19**, 364–74.
47 S. Karthikeyan, T. Leung and J. A. A. Ladias, *The Journal of biological chemistry*, 2002, **277**, 18973–8.
48 K. Jiang, E. Pereira, M. Maxfield, B. Russell, D. M. Goudelock and Y. Sanchez, *The Journal of biological chemistry*, 2003, **278**, 25207–17.
49 G. A. Penman, L. Leung and I. S. Näthke, *Journal of cell science*, 2005, **118**, 4741–50.
50 C. Reichen, S. Hansen and A. Plückthun, *Journal of structural biology*, 2013, **In Press**, 0–0.
51 R. Roskoski, *Pharmacological research : the official journal of the Italian Pharmacological Society*, 2012, **66**, 105–43.
52 A. C. Lloyd, *Journal of Biology*, 2006, **5**, 13.