Aix-Marseille Université

CRCM, Inserm UMR 1068, CNRS UMR 7258

Centre de Recherche en Cancérologie de Marseille

Thèse de Doctorat

Bioinformatique, Biochimie Structurale et Génomique

2P2I$_{DB}$: UNE BASE DE DONNÉES DEDIÉE A LA DRUGGABILITÉ DES INTERACTIONS PROTÉINE-PROTÉINE

Présentée par **Raphaël BOURGEAS**

Sous la direction du **Dr. Philippe ROCHE**

Devant le jury composé de :

Pr. Pedro Coutinho - *Président du Jury*

Dr. Maria Miteva - *Rapporteur*

Dr. Françoise Ochsenbein - *Rapporteur*

Dr. Philippe Roche - *Directeur de thèse*

Soutenance le 20 décembre 2012

TABLE DES MATIERES

REMERCIEMENTS

Je tiens en premier lieu à remercier le ministère de la recherche ainsi que l'université Aix-Marseille dont les financement m'ont permis de réaliser ce doctorat.

Ensuite, bien sûr, merci énormément à mon directeur de thèse, le **Dr. Philippe Roche**, sans qui je n'aurais pu réaliser cette thèse. Merci pour ta disponiblité : tu es toujours ouvert pour discuter de science, ou plus simplement pour répondre à mes questions, auxquelles tu avais toujours une réponse simple. Merci aussi pour ton soutien et ton implication, tu m'as appris beaucoup.

Merci aussi au **Dr. Françoise Guerlesquin**, qui m'a acceuilli dans son unité lors de mon arrivée dans son laboratoire. **Interactions et Modulateurs de Réponses** fut une unité merveilleuse pour s'initier à la science.

Merci aux membres de notre équipe, « **Integrative Structural and Chemical Biology** » avec qui j'ai pu partager ces quatre dernières années. Plus particulièrement, merci au **Dr. Xavier Morelli**, pour tous ses conseils (scientifiques ou non), et d'avoir su créer un climat favorable à l'émancipation de la chémoinformatique dans son groupe et au **Dr. Véronique Hamon** pour m'avoir beaucoup appris en chémoinformatique. Merci à tous les étudiants qui se sont un jour arrétés – pour quelques semaines ou quelques années – au bureau 427, pour les agréables moments passés avec vous.

Je souhaite aussi remercier tous les enseignants chercheurs avec qui j'ai pu travailler lors de mon monitorat. Merci au **Pr. Céline Brochier**, pour m'avoir appris à structurer un cours et comment tenir une classe dissipée ! Merci au **Dr. Caroline Costedoat**, au **Dr. Emese Megliecz** et au **Dr. Benedicte Woerth** pour m'avoir appris ce qu'est l'enseignement.

Au cours de mon A.T.E.R., j'ai eu la joie de travailler avec le **Pr. Jacques Van Helden**, merci pour ton enthousiasme et pour m'avoir montré comment un professeur passioné est capable d'entrainer toute une classe dans sa passion. Merci aussi au **Dr. Jean-Pierre Duneau**, pour les connaissances sur l'organisation d'un programme universitaire que tu m'as transmis.

Je souhaite remercier également les membres d'**UPETEC** et particulièrement le **Dr. Jean-Pierre Mano**, pour avoir partagé avec moi ton savoir sur les systèmes émergents.

Et plus personellement…

Merci infiniment à Stéphane, pour son accueil chaleureux lorsque je n'étais que stagiaire, tous ses conseils et surtout pour son soutient moral ET concret. Merci à Adrien D. Luffy pour tous les bons moments, les fous-rires et les coups de mains.

Merci aussi à tous les anciens de la promo BBSG 2008 (Sylvain, Matt, guillaume, Fab, Titi…), et 2007 (John, Elo, Laura, Fab) pour avoir été là pendant ces 4 années, soit autour d'un d20 ou au 0'Bradys, mais quoi qu'il en soit, toujours heureux !

Merci une fois encore à Philippe pour le soutien, aux Steph (les trois !), et à Betty, pour toutes les rigolades et la bonne ambiance du bureau.

Enfin, un grand merci à mon père, qui m'a appris à être curieux et faire ce que je voulais faire, et à ma mère pour son soutien inconditionnel.
Merci à toute ma famille pour leur amour et leur approche de la vie ; qu'on se le dise, *on aurait pu tomber plus mal*. Merci Nanou pour tes yeux de lynx…
Merci surtout à toi, Lauranne, pour m'avoir tout simplement rendu heureux.

LISTE DES ABREVIATIONS

2P2I :	Protein-Protein Interaction Inhibition
ADME :	Absorption, Distribution, Métabolisme et Excrétion
ASA :	Accessible Surface Area
BEI :	Binding Efficiency Index
BAD :	Bcl-2 Associated Death promoter
BAK :	Bcl-2 homologous Antagonist Killer
Bcl :	B-Cell Lymphoma
BSA :	Buried Surface Area
FDA :	Food & Drug Administration
HDM2 :	Human Double Minute 2
hGH :	Human Growth Hormone
IL-2 :	Interleukin 2
IL-2Rα :	Interleukin 2 Receptor, subunit α
Kd :	constante de dissociation
LEI :	Ligand Efficiency Index
PDB :	Protein Data Bank
PPIM :	Protein-Protein Interaction Modulator
RCPG :	Récepteurs Couplés aux Protéines G
QSAR :	Quantitative Structure Activity Relationship
QSPR :	Quantitative Structure Property Relationship
PPI :	Protein-Protein Interaction
RMN :	Résonance Magnétique Nucléaire
SEI :	Surface Efficiency Index

AVANT-PROPOS

J'ai débuté ma thèse sous la direction du **Dr. Philippe Roche** en Octobre 2008, grâce à un financement du ministère de la recherche (allocation de thèse **MRT**). La première partie de ma thèse a consisté en l'étude structurale des paramètres physicochimiques qui gouvernent les complexes protéiques pour lesquels un inhibiteur orthostérique existe déjà. Pour cela, j'ai travaillé en collaboration avec **Marie Jeanne Basse** (AI, CRCM, Marseille), afin de répertorier tous les complexes caractérisés présents dans la Protein Data Bank (**PDB**) pour lesquels une structure tridimensionnelle est disponible à la fois pour le complexe protéique, le complexe protéine-inhibiteur, et la protéine cible seule, si elle est disponible. L'ensemble des familles de protéines ainsi trouvées ont été classées et intégrées dans une base de données : $2P2I_{DB}$, qui est disponible sur internet (http://2p2idb.cnrs-mrs.fr/). Les paramètres géométriques et physicochimiques des complexes protéiques ont été calculés et intégrés dans la base de données. A l'heure actuelle, $2P2I_{DB}$ comporte 14 complexes protéine-protéine, 16 protéines libres, 60 complexes protéine-ligand et 55 modulateurs. Elle est un outil extrêmement utile pour étudier la druggabilité des interactions protéine-protéine. La mise en place de cette base de données, ainsi que l'étude statistique des paramètres des interfaces protéine-protéine, ont donné lieu à une publication parue en 2010 dans PloS One (**Article 1**) et une seconde parue en 2012 dans Nucleic Acid Research (**Article 2**) lors de la mise à jour de la base de données et du site internet.

$2P2I_{DB}$ représente ainsi les fondations qui ont permis d'analyser les caractéristiques des inhibiteurs orthostériques d'interactions protéine-protéine. La seconde partie de mon travail de thèse a donc consisté à révéler et à étudier, en collaboration avec le **Dr. Véronique Hamon** (Postdoc, CRCM, Marseille), les paramètres physicochimiques qui sont propres aux inhibiteurs d'interactions protéine-protéine (appelés aussi PPIM pour Protein-Protein Interaction Modulators). L'étude de ces paramètres ainsi que la mise en place d'un algorithme permettant de focaliser une chimiothèque à l'aide de petites molécules enrichies en inhibiteurs d'interactions protéine-protéine ont fait l'objet d'une publication actuellement en révision dans le Journal of PloS Computational Biology (**Article 3**). L'étude des caractéristiques biophysiques des paramètres d'interaction entre les différents inhibiteurs d'interactions protéine-protéine présents dans $2P2I_{DB}$ et leurs partenaires nous a permis d'étudier l'efficacité atomique pour cette classe de molécules et de les comparer aux molécules médicaments

classiquement utilisées en pharmacopée. Ces travaux ainsi que « l'état de l'art » de la connaissance sur l'espace chimique représentant les inhibiteurs d'interactions protéine-protéine ont étés publiés dans une revue parue en 2011 dans *Current Opinion in Chemical Biology* (**Article 4**).

En parallèle de ces travaux principaux et puisque j'avais acquis des connaissances en dynamique moléculaire au cours de mes stages de Master 1&2, le **Dr. Roche** m'a proposé de m'impliquer dans l'étude de la dynamique de l'ouverture de la lipase pancréatique humaine en collaboration avec le **Dr Frédéric Carrière** (LEIPL, Marseille). Dans ce projet, initié par le **Dr. André Fournel** (BIP, Marseille), je me suis servi de la dynamique moléculaire afin d'apporter un éclairage aux résultats obtenus par résonance paramagnétique électronique. Ce projet n'est pas discuté dans ce manuscrit par soucis de logique scientifique, mais la publication relative à ce travail est présentée en Annexe 1 (**Article 5**).

LISTE DES PUBLICATIONS

Article 1. **Bourgeas R**, Basse M-J, Morelli X, Roche P: **Atomic Analysis of Protein-Protein Interfaces with Known Inhibitors: The 2P2I Database**. *PLoS ONE* 2010, 5:e9598.

Article 2. Basse MJ, Betzi S, **Bourgeas R**, Bouzidi S, Chetrit B, Hamon V, Morelli X, Roche P: 2P2Idb: **A Structural Database Dedicated to Orthosteric Modulation of Protein-Protein Interactions**. *Nucleic Acid Research* 41: *in press*.

Article 3. Hamon V, **Bourgeas R**, Ducrot P, Theret I, Xuereb L, Basse MJ, Brunel JM, Combes S, Morelli X, Roche P: 2P2IHUNTER: **A Tool for Filtering Orthosteric Protein-Protein Interaction Modulators via a Dedicated Support Vector Machine**. *En révision à Plos Computational Biology*.

Article 4. Morelli X, **Bourgeas R**, Roche P: **Chemical and structural lessons from recent successes in protein-protein interaction inhibition (2P2I).** *Curr Opin Chem Biol* 2011, 15:475-481.

Article non présenté dans ce manuscrit (Annexe 1) :

Article 5. Ranaldi S, Belle V, Woudstra M, **Bourgeas R**, Guigliarelli B, Roche P, Vezin H, Carriere F, Fournel A: **Amplitude of pancreatic lipase lid opening in solution and identification of spin label conformational sub-ensembles by combining CW and pulsed EPR spectroscopy and molecular dynamics.** *Biochemistry* 2010, 49:2140-2149.

INTRODUCTION

Préambule

I.1 Les protéines, cibles moléculaires thérapeutiques

Au sein de tout être vivant, les protéines remplissent des fonctions capitales pour assurer le bon fonctionnement de l'organisme. Ces fonctions, variées, peuvent être divisées en six groupes. Les **protéines de structure** constituent le cytosquelette, et permettent aux cellules de maintenir leur organisation interne. Les **protéines motrices** permettent aux cellules, et parfois aux organismes, de se déplacer. Les **protéines de signalisation** reçoivent les signaux extérieurs et s'assurent de la transmission de l'information à l'organisme. Les **protéines de transport** assurent le transfert des différentes molécules au sein et entre les cellules. Les **protéines enzymatiques** catalysent l'ensemble des réactions chimiques qui s'opèrent dans les cellules. Enfin, toutes ces activités sont modulées par les **protéines de régulation**, soit par interaction directe, soit par le contrôle de l'expression des gènes (Petsko and Ringe, 2008).

Toutes les fonctions de ces protéines peuvent, dans de très nombreux cas, être associées à un dysfonctionnement et donc jouer un rôle dans un phénotype (caractère anatomique, morphologique, moléculaire, physiologique, ou éthologique) anormal observé dans une maladie. L'idée de cibler certaines de ces protéines clés dans un but thérapeutique a donc vite germé avec l'accumulation des connaissances sur leurs modes d'action. En effet, bien que certaines protéines remplissent leur(s) fonction(s) de manière isolée, la plupart d'entre elles exercent leurs fonctions biologiques par l'intermédiaire d'interactions faisant intervenir différents types de reconnaissances entre molécules ou macromolécules biologiques. Ce sont ces interactions qu'il convient de moduler afin de tenter d'agir sur leur fonction et donc sur une maladie associée.

I.2 Le XXème siècle : l'ère des enzymes

Parmi toutes les interactions biologiques décrites, la plus étudiée est l'interaction entre une enzyme et son substrat. Catalysant toutes les réactions chimiques qui se produisent au sein des êtres vivants, les enzymes, leur interaction avec leur substrat et la réaction chimique

qui leur est associée, ont été très tôt décelés et décrits à partir de la découverte de l'amylase de l'orge par Anselme Payen et Jean-François Persoz en 1833. Les enzymes ont dès lors fait l'objet d'une intense activité de recherche académique avec en point culminant le prix Nobel de Chimie décerné à Arthur Harden et Hans von Euler-Chelpin en 1929 pour leurs travaux sur la fermentation des sucres et sur les enzymes qui y participent (Euler-Chelpin, 1930).

L'accumulation et l'amélioration des connaissances sur les enzymes, associées aux progrès de la chimie organique et de la chimie médicinale, ont initié le renouveau de l'industrie agroalimentaire puis de l'industrie pharmaceutique du XIXème siècle qui adopte un nouveau modèle pour la 2nde partie du XXème siècle : **le criblage**. L'idée est de tester l'effet d'une molécule chimique sur l'activité d'une protéine et de reproduire ce test avec un maximum de molécules afin d'identifier les médicaments de « demain ». Cette procédure d'abord très artisanale s'est rapidement automatisée et a permis de tester un nombre croissant de composés (100, 1000, 10 000 …) et aujourd'hui les firmes dominantes du secteur appelées « BIG PHARMA » sont capables de tester (de cribler) plusieurs millions de molécules associées dans ce que l'on appelle une **chimiothèque**.

Enzyme	Inhibiteur	Indication
	Grande voies métaboliques	
HMG-CoA réductase	statines	hypocholesterolémiant
xanthine-oxydase (Zyloric®)	allopurinol	anti-goutteux
topoisomérases,	irinotécan	étoposide anticancéreux
dihydrofolate-réductase	méthotrexate	anticancéreux
	Métabolisme des neuromédiateurs	
acétylcholinestérase	néostigmine	anti-Alzheimer
monoamine-oxydase	sélégiline	anti-Parkinson
cyclo-oxygénases	Paracétamol, aspirine	anti-inflammatoire, antalgique, antipyrétique
conversion de l'angiotensine	captopril	anti-hypertenseur
	Voies de signalisation	
phosphodiestérases	Théophylline, sildénafil	vasodilatateurs
tyrosine-kinases	dasatinib, pazopanib	anticancéreux
	Organismes pathogènes	
neuraminidases	oseltamivir	grippe
reverse transcriptase	zidovudine,	anti-VIH
protéases	amprénavir	anti-VIH
Bêta-lactamases	sulbactam	antibiotiques

Table 1 : Quelques exemples d'inhibiteurs majeurs d'enzymes utilisés ces dernières années.

Ce modèle a fait les beaux jours de l'économie mondiale du XXème siècle. Impulsée par les firmes anglo-saxonnes comme GlaxoSmithKline ou Pfizer, ces stratégies ont permis d'identifier de nombreux médicaments phares, les blockbusters pouvant parfois générer des dizaines de milliards de dollars de chiffre d'affaires par an (wikipedia, 2010), (Table 1).

I.3 Au-delà des cibles traditionnelles

L'industrie pharmaceutique a su capitaliser sur le modèle du criblage pour développer de nombreux inhibiteurs de cibles enzymatiques et également de quelques autres cibles non enzymatiques qui regroupent essentiellement des récepteurs (récepteurs nucléaires et membranaires, récepteurs couplés aux protéines G) et des canaux ioniques (Figure 1).

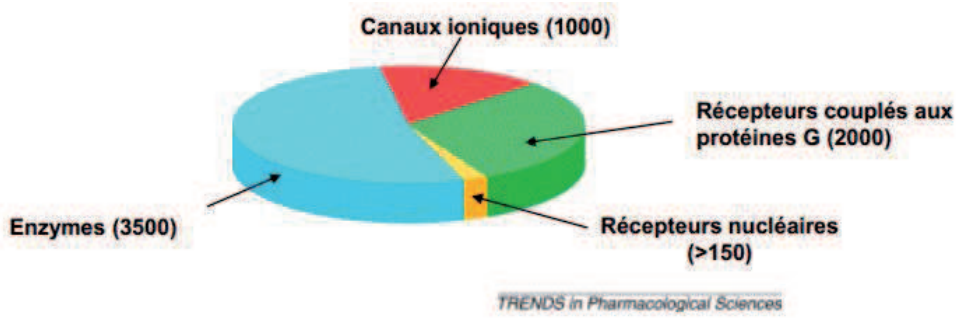


Figure 1 : Répartition des cibles potentielles des médicaments d'après Terstappen et Reggiani (2001)

Les récepteurs membranaires exprimés à la surface des cellules sont les principaux relais entre l'environnement extérieur et l'intérieur de la cellule. Ils représentent aujourd'hui les cibles de 30% des molécules approuvées par la FDA (Food & Drugs Administration) (Overington et al., 2006). Parmis ces récepteurs, la famille des récepteurs couplés aux protéines G (RCPG), présentant environ 800 membres, est particulièrement ciblée dans la recherche de nouveaux agents thérapeutiques encore aujourd'hui (Shoichet and Kobilka, 2012). Leur domaine d'application est particulièrement large, allant du traitement anti-VIH (Dorr et al., 2005) au traitement contre l'hyperparathyroïdie (Harrington and Fotsch, 2007).

Pourtant le début du XXIème siècle voit ce modèle s'effondrer. En effet, depuis 2005, l'argent investi dans la recherche de nouveaux médicaments par neuf des plus importantes entreprises pharmaceutiques (AstraZeneca, Bristol-Myers Squibb, Eli Lilly, GlaxoSmithKline, Merck, Novartis, Pfizer, Roche and Sanofi Aventis) est passé de 40 milliards de dollars en 2005 à 60 milliards de dollars en 2010. Pourtant, durant cette même

période, le nombre de nouvelles petites molécules approuvées par la FDA n'a cessé de diminuer. Avec une moyenne de 7 nouvelles molécules par an (soit moins d'une par entreprise), le nombre de molécules acceptées est passé de neuf en 2005 à seulement deux en 2010 (Bunnage, 2011).

Comme le montrent ces chiffres, l'industrie pharmaceutique paie un lourd tribut à sa dépendance à un nombre restreint de cibles. En effet, de récentes analyses suggèrent qu'il n'y a pas plus de 600 cibles thérapeutiques viables chez l'humain (Hopkins and Groom, 2002). En 2006, Overington *et al.* font le même constat et montrent que la pharmacopée occidentale (molécules acceptées par la FDA http://www.fda.gov/default.htm) n'agissait que sur 207 cibles différentes chez l'humain (Overington et al., 2006).

Il devient donc urgent de chercher au sein des cellules, les vecteurs thérapeutiques du XXIème siècle. En plus de l'interaction enzyme/substrat, de nombreux autres types d'interactions sont candidates pour remplir ce rôle telles que les liaisons protéine-sucre, protéine-lipide, protéine-acide nucléique, protéine-ligand et surtout **les interactions protéine-protéine (PPIs)** qui font l'objet de la suite de mon exposé.

I Les interactions protéine-protéine

En opposition au nombre restreint de cibles thérapeutiques identifiées viables chez l'humain, l'interactome humain coderait pour un nombre compris entre 130 000 (Venkatesan et al., 2009) et 650 000 (Stumpf et al., 2008) interactions protéine-protéine. En effet, parmi les milliers de protéines dans une cellule, si certaines peuvent agir seules, la plupart doivent opérer de concert avec d'autres protéines dans des complexes et souvent dans de vastes réseaux d'interaction essentiels à la réalisation de toutes les fonctions majeures de la biologie comme la réplication de l'ADN, la transcription, la traduction, l'épissage, la sécrétion, le contrôle du cycle cellulaire, la transduction du signal ou encore le métabolisme intermédiaire. Ainsi, même si la fonction biologique finale d'un processus biologique est portée par une enzyme, un nombre parfois très élevé d'interactions protéine-protéine sont mises en œuvre avant d'aboutir à l'activation, l'inactivation et la régulation fine de la réaction catalysée. Même si une petite fraction de ces interactions protéine-protéine est pertinente dans le cadre d'une utilisation thérapeuthique, la communauté scientifique dispose ici d'un réservoir important de nouvelles cibles qui ne demande qu'à être exploré.

I.1 Les interactions protéine-protéine : étudier et classifier

I.1.A Les réseaux d'interactions

Etudier les interactions protéine-protéine revient donc à identifier expérimentalement dans un contexte cellulaire ou dans un test *in vitro* si deux protéines sont capables d'interagir. L'étude de ces interactions a largement bénéficié des progrès technologiques développés pour les approches de protéomique (spectrométrie de masse, puces à protéines …) associés aux progrès dans la gestion des bases de données de connaissances biologiques. Combinées, ces méthodes ont permis de développer une nouvelle science appelée **interactomique**. L'objectif est d'identifier expérimentalement puis d'annoter à grande échelle les interactions puis les réseaux d'interactions protéine-protéine. Les résultats de ces recherches permettent de construire des diagrammes d'interaction sous forme de réseaux point à point, où chaque point représente une protéine (Figure 2 A). L'analyse spatiale de ces réseaux permet de repérer les « hub » ou nœuds fortement connectés (Figure 2 B) qui peuvent être associés à un processus biologique et/ou à une pathologie (Uetz et al., 2000).

A **B**

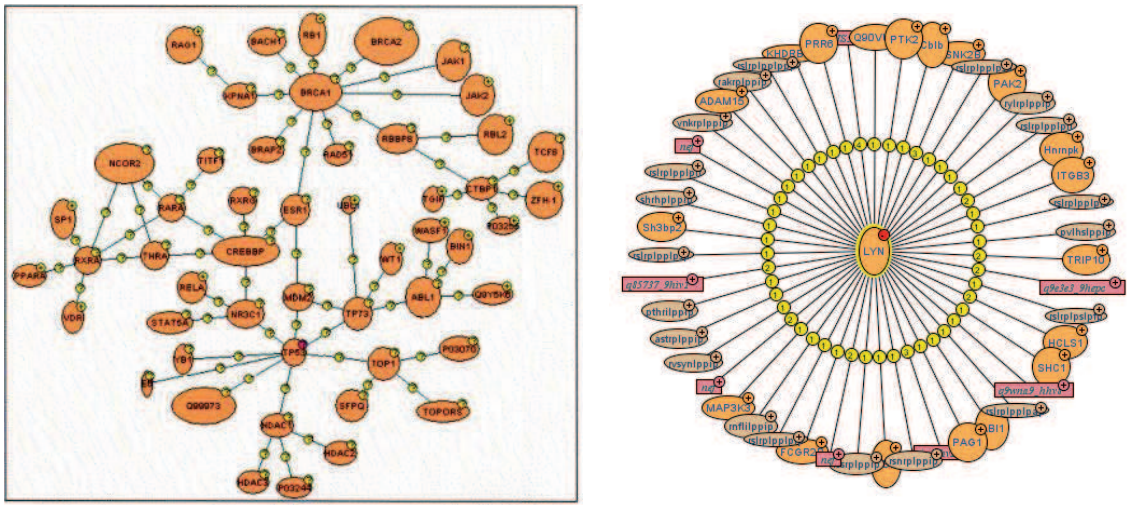


Figure 2 : Exemple de réseau d'interactions obtenues expérimentalement (MINT database) et représentation binaire centrée sur la protéine kinase humaine Lyn.

Il existe un grand nombre de bases de données regroupant les réseaux d'interactions protéine-protéine validés expérimentalements ou prédits. La plupart d'entre elles recensent les interactomes de plusieurs espèces tels BioGRID (Lehne and Schlitt, 2009) et HINT (Das and Yu, 2012), mais certaines sont focalisées sur une seule espèce, tel la « Human Protein Reference Database » (HPRD) (Prasad et al., 2009) et PIPs (McDowall et al., 2009) qui référencient uniquement l'interactome humain. Enfin, le « BIANA Interlog Prediction Server » (BIPS) recense non seulement les interactomes de plusieurs espèces, mais aussi les interactions inter-espèces (Garcia-Garcia et al., 2012).

En général, la combinaison des renseignements de ces diverses bases de données permettent une analyse exhaustive de l'interaction protéique. Pour faciliter cette recherche, des méta-serveurs rassemblant les informations de diverses bases de données ont été mis en place. **APID** : Agile Protein Interaction DataAnalyzer (Prieto and De Las Rivas, 2006) (http://bioinfow.dep.usal.es/apid), mis en ligne en 2006, rassemble par exemple les données de PPIs validées expérimentalement et contenues dans six bases de données (BIND, BioGRID, DIP, HPRD, IntAct et MINT). De même, STRING (Jensen et al., 2009) et CPDB (Kamburov et al., 2009), regroupent plusieurs bases de données.

Nom de la banque	Année	URL
DIP	2000	http://dip.doe-mbi.ucla.edu
STRING	2003	http://string.embl.de
IntAct	2004	http://www.ebi.ac.uk/intact
3did	2005	http://3did.irbbarcelona.org/
BOND / BIND	2005	http://bond.unleashedinformatics.com/
BioGRID	2006	http://www.thebiogrid.org
HPID	2006	http://wilab.inha.ac.kr/hpid/
MIPS	2005	http://mips.helmholtz-muenchen.de/proj/ppi/
MINT	2007	http://mint.bio.uniroma2.it/mint
HPRD	2009	http://www.hprd.org
PIPs	2009	http://www.compbio.dundee.ac.uk/www-pips
POINeT	2009	http://point.bioinformatics.tw/Welcome.do
BIANA	2010	http://sbi.imim.es/web/index.php/research/servers/biana
DOMINE	2011	http://domine.utdallas.edu/cgi-bin/Domine
Intermitobase	2011	http://mcube.nju.edu.cn/cgi-bin/intermitobase/home.pl
BIPS	2012	http://sbi.imim.es/web/index.php/research/servers/bips
Dr. PIAS 2.0	2012	http://www.drpias.net
DTome	2012	http://bioinfo.mc.vanderbilt.edu/DTome
HINT	2012	http://hint.yulab.org
IBIS	2012	http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi
IRView	2012	http://ir.hgc.jp/
UniHI	2012	http://www.unihi.org

Table 2 : Exemples de bases de données d'interactomes multi-espèces et humains.

La connaissance de ces réseaux peut permettre l'identification de la fonction d'une protéine de fonction inconnue peut être effectuée au sein de ce réseau. En effet, si la fonction d'au moins un des partenaires est connue, alors on pourra associer cette fonction et le chemin fonctionnel associé au travers de l'analyse du réseau intriqué de ces interactions. L'importance de ces interactions est telle dans la régulation des mécanismes biologiques, que des modules protéiques globulaires spécifiques ont été sélectionnés au cours de l'évolution pour assurer ces fonctions. Des modules protéiques tels que les domaines SH2, SH3, WW ou BH, dont une liste plus exhaustive est présentée Figure 3, sont spécifiquement inclus dans la séquence des protéines à réguler en amont ou en aval des domaines actifs/catalytiques.

Figure 3 : Liste de domaines protéiques globulaires utilisés pour les reconnaissances protéine-protéine. Liste hhtp//pawsonlab.mshri.on.ca

Les modules architecturaux permettent d'assurer le contact, la reconnaissance et donc la colocalisation de protéines devant se trouver proches spatialement pour assurer leurs fonctions. Les protéines kinases impliquées dans les réseaux de signalisation en sont des exemples parfaits. Les domaines SH3 de protéines kinase reconnaissent des motifs linéaires polyproline inclus dans d'autres protéines kinases afin de les phosphoryler et donc les activer dans un phénomène de cascade de réactions de phosphorylation (Shi et al., 2012).

L'étude des interactions entre protéines est donc primordiale pour assurer la compréhension de leurs fonctions à l'intérieur même de la cellule. Ces interactions entre protéines peuvent être stables ou transitoires. Elles peuvent également intervenir dans des complexes homodimèriques entre protéines identiques, ou au sein de complexes hétérodimériques entre protéines différentes. Ces différents types d'interactions ont conduit à proposer des classifications qui sont présentées dans le paragraphe suivant.

I.1.B Les différents types d'interactions protéine-protéine

En 2003, Nooren & Thornton ont proposé une classification des PPIs suivant trois propriétés différentes (Nooren and Thornton, 2003). Elles différencient les complexes protéiques en fonction de 1) leur partenaire, 2) du caractère obligatoire du complexe, ou enfin 3) suivant la durée de la formation du complexe.

I.1.B.a Définition biochimique

Complexes hétéro- et homo-oligomériques

La formation de complexes protéiques peut se faire soit avec des partenaires identiques (les homo-oligomères Figure 4 A) soit entre des protéines différentes (les hétéro-oligomères Figure 4 B). Parmi les hétéro-oligomères, on dénote deux catégories d'organisation des complexes : les hétérologues et les isologues (Goodsell and Olson, 2000). Une association isologue entre deux protéines implique la même surface d'interaction chez les deux partenaires (Figure 4 C), tandis que les associations hétérologues impliquent des surfaces d'interactions différentes (Figure 4 A). Chez les complexes isologues, on peut observer soit une symétrie de centre O entre les deux protéines, O étant un point se situant au centre de la zone d'interaction ; soit une symétrie orthogonale de plan P, P étant le plan des moindres carrés des atomes de l'interface . A l'inverse des complexes isologues qui ne peuvent former que des dimères, les hétérologues, utilisent des interfaces différentes, sans une symétrie cyclique (et donc fermée), et peuvent mener à des agrégations infinies.

Complexes obligatoires et non-obligatoires

On distingue deux types de complexes suivant qu'ils sont **obligatoires** ou **non obligatoires**. Lorsqu'il s'agit de complexes obligatoires, les partenaires ne sont généralement pas stables *in vivo* lorsqu'ils ne sont pas associés à leur partenaire. Cela implique aussi qu'ils sont souvent exprimés simultanément, et sont co-localisés. De tels complexes sont généralement nécessaires pour que les protéines soient fonctionnellement actives. A l'inverse, la plupart des complexes non-obligatoires sont formés à partir de monomères qui peuvent exister *in vivo* de manière stable. Ces monomères peuvent, comme pour les partenaires des complexes obligatoires, être co-localisés (ex : complexes participant à la signalisation intramoléculaire), mais ils peuvent aussi ne pas l'être (ex : complexes anticorps/antigènes). De précédentes études ont montré que la plupart des complexes obligatoires sont des homo-oligomères tandis que la plupart des complexes non-obligatoires sont des hétéro-oligomères. Cependant, certains complexes obligatoires sont des hétéro-oligomères, telle que la cathepsine D humaine (Figure 4 D), tandis que des homo-oligomères peuvent ne pas être obligatoires, tel que le dimère de la lysine du sperme (Figure 4 A), (Nooren and Thornton, 2003).

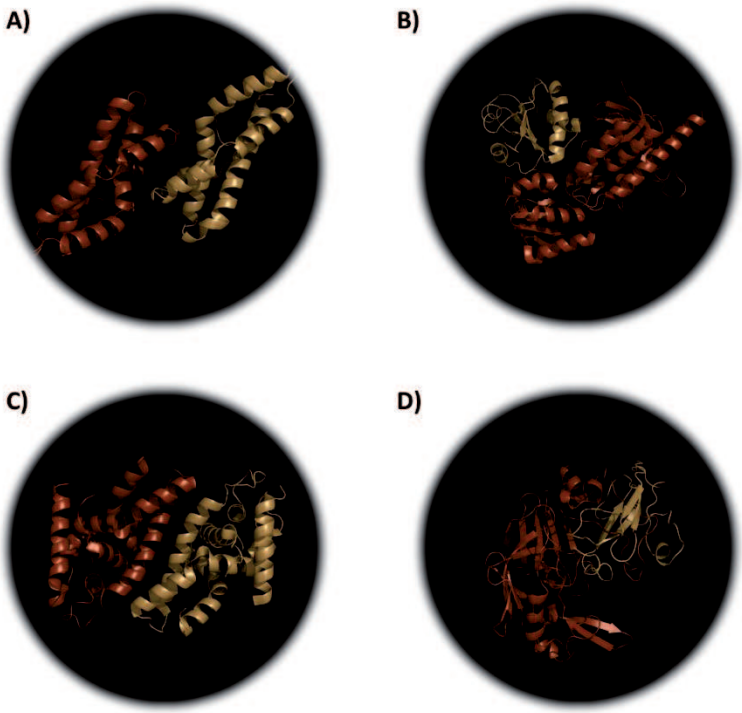


Figure 4 : Exemples de différents types d'interactions protéine-protéine tel que discutés dans le texte. A) Homodimère non-obligatoire : la lysine du sperme (PDB ID : 3LYN). B) Hétérodimère entre la thiorédoxine (jaune) et la thiorédoxine réductase (marron) (PDB ID 1F6M). C) Homodimère isologue du domaine Ribonucléase III de la dicer humaine (PDB ID : 2EB1). D) Hétérodimère obligatoire de la cathepsine D humaine (PDB ID : 1LYW).

Complexes transitoires et permanents

Les PPIs peuvent aussi être différenciées en fonction de leur durée de vie. Les interactions permanentes sont généralement extrêmement stables et n'existent que sous leur forme complexées. En revanche, les complexes transitoires s'associent et se dissocient régulièrement *in vivo*. On peut néanmoins distinguer, parmi ces derniers, les interactions transitoires faibles et les interactions transitoires fortes. Les interactions faibles mènent à des équilibres oligomériques dynamiques en solution, avec associations et dissociations des complexes continues. Les interactions fortes, quant à elles, nécessitent un levier afin de déplacer l'équilibre oligomérique. On peut prendre pour exemple la protéine G, hétéro-trimère (G_α, G_β, G_γ) qui est stable en présence de GDP, mais qui se dissocie en deux sous-unités G_α et $G_{\beta\gamma}$ lors de la liaison avec le GTP (Thaker et al., 2012). Généralement, les PPIs structurellement ou fonctionnellement obligatoires sont permanentes, tandis que les PPIs non-obligatoires peuvent être transitoires ou bien permanentes (Nooren and Thornton, 2003).

I.1.B.b Définition structurale

Les complexes protéiques peuvent être classifiés en fonction de la géométrie que présentent les protéines à la surface d'interactions. On considère qu'il existe deux classes structurales de complexes suivant le niveau de repliements présents à l'interface.

Complexes protéine-peptide

Les complexes protéine-peptide sont composés d'une protéine globulaire d'une part, et d'autre part, d'une protéine qui ne présente soit aucun repliement à l'interface (par exemple une protéine désordonnée), soit uniquement une structure secondaire (une hélice α ou un brin β). Cet élément de structure secondaire peut faire partie intégrante d'une protéine plus importante, mais qui ne fait pas partie de la zone d'interaction. Ces complexes sont ceux qui peuvent être le plus facilement inhibés, notamment grâce à des peptides ou des peptido-mimétiques. Les exemples parmi les plus étudiés de complexes protéine-peptide sont les complexes BclXL/Bak et MDM2/p53 tous deux impliqués dans l'apoptose et qui sont détaillés dans le chapitre III (Exemples et Succès).

Complexes protéine-protéine

Les complexes protéine-protéine proprement dits sont ceux composés par deux protéines globulaires en contact direct. A l'interface, chaque partenaire comporte plusieurs segments de résidus en interaction discontinue. Ils présentent une surface d'interaction la plupart du temps plus grande que pour les complexes protéine-peptide. Le complexe IL-2/IL-2R est l'un des complexes protéine-protéine les plus étudiés concernant le développement d'inhibiteurs et est détaillé dans le chapitre III (Exemples et Succès).

I.2 Moduler l'interaction entre deux partenaires protéiques

Malgré l'abondance de cibles, les PPIs ont été pendant longtemps mises à l'écart car leurs interfaces étaient considérées comme trop grandes et trop planes, et donc manquant de poches de liaisons pour des petites molécules (Mullard, 2012). En effet, les cibles enzymatiques traditionnelles présentent, la plupart du temps, une surface qui a évolué afin d'accepter des petites molécules endogènes (comme l'ATP dans le cas des kinases) qu'il est possible de mimer lors de la recherche de nouveaux agents thérapeutiques. En revanche, les surfaces protéiques, au niveau des interactions protéine-protéine, ont évolué afin de se lier avec un partenaire macromoléculaire et non une petite molécule hydrophobe comme c'est le

cas pour les enzymes. Bien qu'elles restent des cibles difficiles, les récents progrès en chimie médicinale ont permis de relancer l'intérêt pour les interactions protéine-protéine jusqu'à présent perçues comme 'non-druggable' (Bunnage, 2011; Mullard, 2012).

I.2.A Des dogmes remis en causes

Historiquement, les deux avancées majeures qui ont permis de rendre la modulation des PPIs plus facilement réalisable sont la découverte des hots-spots et le calcul du volume des poches à l'interface. En 1995, *Wells et Clackson* ont mis en évidence les « hot-spots », quelques résidus présents à l'interface et généralement bien conservés, représentant la majorité de l'énergie d'interaction du complexe protéique (Clackson et Wells, 1995). Ils ont montré qu'il suffit de rentrer en compétition non plus avec la totalité de l'interface, mais avec un sous-ensemble de résidus (généralement deux ou trois) situés au cœur de celle-ci. La première barrière que constituait la taille des PPIs ne représentait donc plus de problème. La seconde avancée majeure pour moduler les interactions protéine-protéine concerne le volume des poches à l'interface. En 2009, Fuller *et al.* ont quantifié le volume des poches présentes à l'interface des complexes protéine-protéine (Fuller et al., 2009). Même si celles-ci ne sont pas assez profondes pour recevoir une molécule, elles sont en revanche suffisamment nombreuses et proches les unes des autres pour qu'une molécule puisse se répartir entre elles. Tous ces arguments ont permis à la communauté scientifique de relancer les interfaces protéine-protéine comme cible thérapeutique de choix.

I.2.B Preuves de concept et réussites

Face au large réservoir de protéines inutilisées par l'industrie pharmaceutique et à la remise en question des dogmes sur les propriétés des surfaces de reconnaissance impliquées dans les interactions protéine-protéine, il a fallu répondre aux questions suivantes: **est-il possible d'inhiber les complexes protéine-protéine? Comment peut-on le faire? Peut-on concevoir des médicaments à visées thérapeutiques ?**

Ces questions ont largement été adressées depuis les 15 dernières années par un grand nombre de laboratoires qui se sont lancés dans des programmes ambitieux de recherche de modulateurs des quelques complexes protéine-protéine clés. La modulation des PPIs a ainsi progressé tout d'abord par l'utilisation d'anticorps et de peptides. Ces tout premiers modulateurs furent dérivés directement de ce qui était observé dans la nature, mais très

rapidement ces modèles ont permis de proposer des peptides mimétiques tels que les 'stappled-peptides' (Henchey et al., 2008; Walensky et al., 2004) ou les macrocycles inspirés des produits naturels (Driggers et al., 2008). Aujourd'hui, il est communément admis que des petites molécules sont également capables, en se fixant à la surface d'une protéine, de moduler la fixation d'un partenaire protéique.

Comme nous le décrirons plus en détail à la fin de l'introduction, parmi les récents succès, Sunesis Pharmaceuticals a développé une molécule pour le syndrome de Gougerot-Sjögren (SAR1118) qui est en phase III des essais cliniques depuis 2011. De même, le Navitoclax d'Abbott et l'Obatoclax de Teva sont tous deux en phase II en tant qu'agents anticancéreux. Enfin, le groupe Roche teste deux autres agents anticancéreux agissant sur le complexe p53/MDM2, et le composé RG7112 est actuellement en phase Ib des essais cliniques et RO553781 devraient entrer en phase clinique à la fin de l'année (Mullard, 2012).

I.2.C Nouveaux défis : Druggabilité et Chimiothèques

Bien que primordiaux, les progrès réalisés dans la science de l'inhibition des interactions protéine-protéine n'ont permis d'identifier qu'une infime partie des molécules pressenties utilisables en clinique. Il convient donc de comprendre la lenteur des progrès réalisés, ou néanmoins d'identifier les raisons de cet échec relatif. Deux paramètres importants ressortent déjà des études préliminaires. Le premier concerne le choix des protéines ciblées car indépendamment du rôle d'une protéine dans un processus biologique, il convient de définir si un complexe peut être dissocié et si une petite molécule peut être utilisée pour assurer cette fonction (étude dite de la « druggabilité », qui fait l'objet du chapitre II). Le deuxième est le choix des chimiothèques dans lesquelless les inhibiteurs des interactions protéine-protéine sont recherchés (ce point est discuté plus en détail dans le chapitre III).

Le choix initial qui conduit à choisir un complexe protéine-protéine en tant que cible thérapeutique est évidemment dépendant de son implication dans un processus biologique. Pourtant le faible recul que nous avons sur l'inhibition des interactions protéine-protéine impose la prudence. En effet, certains complexes protéine-protéine ne sont peut être pas adaptés à une modulation par une petite molécule, ou en tout cas le sont moins que d'autres. Il convient donc de définir à partir des données connues (complexes et inhibiteurs), si des caractéristiques favorables peuvent être identifiées au niveau des propriétés physicochimiques et géométriques des interfaces afin de classer et d'annoter le vaste espace chimique et

biologique que représentent les interactions protéine-protéine. C'est ce qui est appelé la « **druggabilité** » des PPIs. Au cours de ma thèse, j'ai analysé les paramètres physicochimiques et statistiques permettant de caractériser la druggabilité des interactions protéine/proteine en étudiant les complexes protéiques pour lesquels des inhibiteurs ont étés développés.

II La druggabilité des interfaces protéine-protéine

Comme nous l'avons vu dans le chapitre précédent, il est urgent d'élargir le spectre des cibles pour la recherche de nouvelles molécules. Les interactions protéine-protéine sont des cibles originales et innovantes lors de la mise en place de nouveaux programmes de recherche de médicaments. Or, elles sont généralement plus difficiles à étudier que les cibles traditionnelles. Pour éviter pertes de temps et d'argent, il est indispensable de pouvoir sélectionner *a priori* les interfaces protéiques que l'on peut plus facilement moduler. L'étude de la druggabilité des interactions protéine-protéine permet de quantifier la faisabilité de l'inhibition de ces interactions.

II.1 Qu'est-ce que la druggabilité ?

Le mot **druggabilité** est un néologisme parfois décrié en français. Il provient de la traduction littérale du terme anglosaxon « druggability » qui désigne la faculté pour une protéine d'interagir avec une molécule pouvant aboutir au développement d'un médicament (« drug »), (Egner and Hililig, 2008). Toutefois, cette définition est le plus souvent restreinte à la première partie de l'assertion, indépendamment des propriétés thérapeutiques de la molécule. C'est cette définition que nous adopterons dans le reste de ce manuscrit. Le concept de druggabilité a tout d'abord été développé pour les enzymes et est hautement associé à la notion de site de liaison de haute affinité. La recherche et la mise en évidence de ces sites de liaisons constitue une des voies pour « mesurer » la druggabilité d'une protéine. Cette définition a été étendue au cas des interactions protéine-protéine. Plusieurs définitions ont été proposées.

Si l'on s'intéresse à la maladie, la druggabilité d'une interaction protéine-protéine qualifie l'impact de la modulation de cette PPI sur cette maladie. Cette définition de la druggabilité est parfois nommée « targetabilité » (Schneider, 2004). Par ailleurs, la druggabilité d'une PPI est la vraisemblance que cette interaction puisse être sensible à une modulation fonctionnelle de la part d'un agent chimique. Cette définition est parfois appelée « inhibitabilité » (Jochim and Arora, 2010). Enfin, la druggabilité peut indiquer la vraisemblance que cette interaction puisse être inhibée par un agent thérapeutique (Fauman et al., 2011). L'objectif ultime de la druggabilité serait la mesure de l'impact de la modulation

d'une interaction protéine-protéine par une petite molécule sur un quelconque gain thérapeutique chez le patient. Cependant, cette approche est difficile à mettre en œuvre. Dans le cadre de cette thèse, la notion de druggabilité d'un complexe protéine-protéine représente la faculté de trouver une molécule pouvant empêcher la formation de ce complexe ou pouvant rentrer en compétition avec la formation du complexe. Il existe de nombreuses approches permettant d'évaluer la druggabilité des complexes protéine-protéine (Fry, 2011). Au cours de ma thèse, je me suis focalisé sur les paramètres structuraux de la druggabilité ; c'est-à-dire la mesure de l'impact des paramètres géométriques et chimiques de l'interface entre les deux protéines sur sa capacité à être modulée par une petite molécule organique. Afin de pouvoir qualifier la druggabilité des interactions protéine-protéine en des termes structuraux, je me suis, dans un premier temps, intéressé aux descripteurs permettant de caractériser une interface protéine-protéine.

II.2 Descripteurs des interfaces protéine-protéine

II.2.A « Hot spots »

Historiquement, c'est la découverte de quelques résidus cruciaux au niveau de l'interface appelés « hot spots » qui a permis de montrer la faisabilité d'inhiber les interactions protéine-protéine.

II.2.A.a Découverte et définition

Afin de connaitre la contribution de chacun des acides aminés de l'interface dans l'énergie d'interaction entre l'hormone de croissance humaine (hGH) et son récepteur (hGHbp), Tim Clackson et James A. Wells ont muté expérimentalement en alanine chacun des 33 résidus de hGH dont les chaines secondaires sont en contact avec hGHbp (Clackson and Wells, 1995). Dans un second temps, ils ont défini l'énergie libre d'interaction de chacun des mutants en mesurant le déplacement compétitif provoqué par un anticorps monoclonal sur le complexe hGHbp /hGH. Ce déplacement est mesuré par le biais d'hGH marqué à l'iode 125. Ils ont ainsi pu comparer ces énergies libres avec celles du complexe sauvage. Ils ont classé les différents résidus à l'interface en fonction de la perte d'énergie libre que la mutation entraîne lors de la formation du complexe. Ils ont montré que deux tryptophanes, situés au cœur de la zone d'interaction, représentent à eux seuls plus de 75% de l'énergie d'interaction. La région de l'interface qui contient les acides aminés contribuant fortement à l'énergie

d'interaction est appelée épitope fonctionnel tandis que les deux résidus les plus impliqués sont appelés les « Hot spots ». Wells et Clackson ont généralisé cette notion en définissant les « hot spots » comme étant les quelques acides aminés qui contribuent pour plus de 75 % de l'énergie d'interaction (Clackson and Wells, 1995).

La découverte des « hot spots » fut une avancée majeure pour la recherche d'inhibiteurs d'interactions protéine-protéine. En effet, plus de 75% de l'énergie d'interaction étant concentrée sur seulement quelques résidus, il est suffisant pour bloquer l'interaction de rentrer en compétition avec le partenaire protéique non plus sur la totalité de la surface d'interaction, mais uniquement sur quelques résidus de la zone d'interaction. De plus, la proximité spatiale des « hot spots » au cœur de la zone d'interaction, rend possible une compétition de la part d'une petite molécule.

II.2.A.bComposition des « Hot spots »

La surface d'interaction d'une protéine avec une autre peut être partitionnée en deux groupes d'acides aminés (Bahadur et al., 2007; Bahadur and Zacharias, 2008; Chakrabarti and Janin, 2002) qui constituent le « cœur » et la périphérie de l'interface. Les acides aminés du cœur ont été définis comme étant les acides aminés ayant au moins un atome perdant entièrement tout contact avec le solvant (c'est-à-dire avec une surface accessible au solvant égale à zéro) après formation du complexe protéique. Les acides aminés à la périphérie, quant à eux, n'ont aucun atome inaccessible au solvant après formation du complexe (Guharoy and Chakrabarti, 2009), mais en ont au moins un dont la surface accessible au solvant est restreinte. La différenciation entre ces deux populations est importante car il a été montré que la contribution des acides aminés à l'interface diffère suivant son appartenance à l'un des deux groupes. En effet, en 2009, Guharoy et Chakrabarti ont mesuré par « alanine scanning » le $\Delta\Delta G$ induit par la mutation des acides aminés de l'interface et en ont déduit des règles sur l'implication des acides aminés sur l'énergie d'interaction. Lorsque les acides aminés ne forment aucune liaison hydrogène intermoléculaires, les acides aminés qui font partie du cœur ont en moyenne une contribution de 26 cal/mol tandis que les résidus de la périphérie ont une contribution négligeable. En revanche, les résidus formant des liaisons hydrogène avec le partenaire protéique lors de la formation du complexe ; qu'ils fassent partie du cœur ou de la périphérie , contribuent en moyenne de 29 cal/mol à l'énergie d'interaction (Guharoy and Chakrabarti, 2009).

Les « hot spots » sont situés parmi les acides aminés du cœur de l'interface et leur composition en acides aminés est différente de celle du reste de la surface accessible au solvant des protéines. Bogan et Thorn ont calculé l'enrichissement de chaque acide aminé présent dans les « hot spots » de leur base de données par rapport au reste de la surface (Bogan and Thorn, 1998). Ne prenant en compte que les acides aminés dont la surface accessible au solvant est supérieure à 10 Å², ils ont construit un ensemble de 2325 résidus. Ils ont ensuite comparé la fréquence de chacun des acides aminés de leur ensemble avec la fréquence des acides aminés présents dans les « hot spots » de ces protéines. Ils ont ici définis les « hot spots » de leur jeu de données en étudiant par « alanin scanning » les différences d'énergies d'interaction, et en ne gardant que les résidus dont la contribution était supérieure à 2 kilocalories. Ils ont ainsi montré que plus de 45 % des résidus des « hot spots » sont constitués de seulement 3 acides aminés différents : l'arginine, le tryptophane et la tyrosine. Ils sont les seuls acides aminés à être plus de deux fois plus présents dans les « hot spots » que sur le reste de la surface (respectivement 2,47 ; 3,91 et 2,29). A l'inverse, certains acides aminés sont moins présents dans les « hot spots » que sur le reste de la surface, comme la méthionine, la sérine et la thréonine (Met : 0,54 ; Ser : 0,21 et Thr : 0,28). Enfin, certains sont totalement absents des « hot spots », comme la leucine et la valine (Leu : 0,01 et Val : 0). Lorsque l'on s'intéresse aux caractéristiques des acides aminés présents dans les « hot spots », on s'aperçoit qu'on n'observe pas seulement un seul type de résidus (chargé, aliphatique…). En fait, les trois acides aminés les plus fréquents dans les « hot spots » (Arg, Trp, et Tyr) incluent deux résidus hydrophobes et un acide aminé chargé positivement. On peut émettre l'hypothèse que les acides aminés favorisés dans les « hot spots » sont ceux qui sont capables de réaliser plusieurs types d'interactions favorables à l'interaction avec un partenaire protéique. A titre d'exemple, le tryptophane et la tyrosine peuvent créer une interaction aromatique (π-stacking), être donneur d'électron au sein d'une liaison hydrogène, et ont une grande surface hydrophobe. On peut enfin penser que la capacité à former des liaisons hydrogène de la tyrosine par l'intermédiaire de son groupement hydroxyle, explique pourquoi elle se retrouve trois fois plus fréquemment que la phénylalanine dans les « hot spots ».

Une autre étude menée en 2007 a montré des résultats sensiblement différents concernant les acides aminés les plus fréquents dans les « hot spots » (Ma and Nussinov, 2007). En effet, en comparant la proportion de chaque type d'acides aminés conservés à la surface des protéines avec celle des types d'acides aminés conservés dans la zone

d'interaction avec le partenaire protéique, ils ont montré que trois acides aminés, le tryptophane, la méthionine et la phénylalanine sont beaucoup plus représentés à l'interface (avec des proportions respectives de 2,02 ; 1,21 et 1,35). La présence de ces résidus et leur proximité peut être utilisée pour prédire les zones d'interactions avec un partenaire protéique lorsqu'aucune information structurale ou de mutagénèse n'est disponible.

II.2.A.c Détection des « hot spots »

Expérimentalement, la recherche de « hot spots » se fait par l'évaluation du changement de l'énergie libre du complexe protéine-protéine lorsque l'on mute les acides aminés de l'interface en alanine. C'est par cette méthode que les « hot spots » ont étés définis pour la première fois en 1995 (Clackson and Wells, 1995). Des mutations en alanine ont été réalisées sur des complexes protéiques et les informations inhérentes sont disponibles dans deux bases de données : BID, contenant les informations de plus de 1300 mutations disponibles pour 170 complexes protéiques (Fischer et al., 2003) et ASEdb, contenant les informations de 2915 mutations pour 91 complexes protéiques (Thorn and Bogan, 2001). Ces bases de données ne sont malheureusement plus mises à jour, mais les résultats sont toujours disponibles.

Une première approche informatique pour déterminer la présence de « hot spots » consiste à reproduire *in silico* les mutations des acides aminés de l'interface par des alanines sur la structure tridimensionnelle des protéines quand celle-ci est disponible et d'évaluer – et non plus mesurer expérimentalement – les variations de l'énergie libre théorique du complexe. Cette méthode, utilisant l' « alanine scanning » et aussi appelée méthode énergétique a été pour la première fois intégrée dans le logiciel FOLDEF en 2002 (Guerois et al., 2002) et a été mise à disposition de la communauté scientifique par le biais du serveur FoldX (Table 3). D'autres logiciels permettent de prédire les « hot spots » par alanine scanning (Kortemme and Baker, 2002; Krüger and Gohlke, 2010; Lise et al., 2009). Ces logiciels mutent *in silico* dans un premier temps les acides aminés de l'interface en alanine et calculent ensuite la différence d'énergie libre avec l'état natif, grâce à une équation prenant en compte la somme de la contribution de l'énergie de Van der Waals des atomes ; l'énergie de solvatation des groupes polaires et apolaires ; la différence d'énergie entre la formation d'une liaison hydrogène entre les deux protéines ou entre une protéine et le solvant ; et la contribution électrostatique des groupements chargés. Les poids associés à chacune des composantes des équations ont étés calculés empiriquement à partir d'ensembles

d'apprentissages, différents pour chaque logiciel (Table 3). Plus récemment, une méthode simplifiée a été développée en ne prenant en compte que la formation de liaisons hydrogène et le ΔASA de la chaine secondaire de l'acide aminé (à partir du C_β) (Guharoy and Chakrabarti, 2009). Parmi les serveurs utilisant ces algorithmes, les plus connus sont Robetta (Kim et al., 2004) et FoldX (Schymkowitz et al., 2005).

En plus des algorithmes énergétiques, il existe les algorithmes basés sur des paramètres structuraux, et des algorithmes basés sur l'évolution. Parmi les algorithmes utilisant des paramètres structuraux, KFC2 (Zhu and Mitchell, 2011) permet de prédire les « hot spots » grâce à 47 paramètres (dont l'ASA) et un algorithme de machine à support de vecteur. Les algorithmes utilisant l'évolution ne nécessitent pas d'information structurale sur la surface de la protéine, mais uniquement la structure primaire. Le serveur ISIS utilise la séquence d'acides aminés sans information sur la structure tertiaire de la protéine, ni sur le partenaire protéique (Ofran and Rost, 2007). Enfin, certains groupes ont construits leurs outils en combinant plusieurs types d'algorithmes (Guney et al., 2008; Metz et al., 2012; Tuncbag et al., 2009). C'est le cas du serveur $PCRP_i$-W qui utilise des informations fournies à la fois par les mutations, les paramètres géométriques et physicochimiques et aussi par l'évolution en acides aminés des interfaces (Segura and Fernandez-Fuentes, 2011; Segura Mora et al., 2010).

Nom du serveur	Année	URL
ASEdb	2000	http://nic.ucsf.edu/asedb
HotSprint & HotPoint	2007	http://prism.ccbb.ku.edu.tr/hotsprint
KFC2	2007	http://kfc.mitchell-lab.org
FTMAP	2009	http://ftmap.bu.edu
Hotspot Wizard	2009	http://loschmidt.chemi.muni.cz/hotspotwizard
$PCRP_i$, $PCRP_i$-W	2009	http://www.bioinsilico.org/PCRPi
CCRXP	2010	http://ccrxp.netasa.org/
DrugscorePPI	2010	http://cpclab.uni-duesseldorf.de/dsppi/
HSPred	2011	http://bioinf.cs.ucl.ac.uk/hspred
PRICE	2011	http://www.boseinst.ernet.in/resources/bioinfo/stag.html
HotSpot	2012	http://www.aporc.org/doc/wiki/HotSpot

Table 3 : Serveurs de prédiction des « hot spots » aux interfaces protéine-protéine.

La prédiction des « hot spots » permet d'identifier les acides aminés énergétiquement important. Cependant, un autre paramètre crucial pour l'étude de la druggabilité des PPIs est la présence de poches à l'interface.

II.2.B Poches à l'interface

II.2.B.a Définition

L'arrimage moléculaire d'une petite molécule au niveau de l'interface nécessite la présence d'une ou plusieurs poches capables de l'accueillir. La présence d'une ou de plusieurs poches au niveau de la zone d'interaction est donc primordiale pour estimer la druggabilité d'une interaction protéine-protéine. Elles sont d'autant plus importantes dans les liaisons protéine-ligand, puisqu'elles permettent d'augmenter la surface d'interaction entre la protéine et le ligand (Laskowski et al., 1996). Cette augmentation de surface permet un gain du nombre d'interactions et ainsi de stabilité du complexe ainsi formé.

D'après Fuller *et al* les poches situées aux interfaces des hétérodimères ne sont pas différentes de celles que l'on trouve sur le reste de la surface (Fuller et al., 2009). En effet, elles ont un volume moyen de 54 $Å^3$ tandis que celles trouvées sur la totalité de la surface ont un volume moyen de 55 $Å^3$. En revanche, leurs caractéristiques sont très différentes de celles des poches présentes au sein des sites catalytiques des enzymes. Les poches de la surface des enzymes sont en moyenne plus grandes que celles trouvées à la surface des hétérodimères (77 $Å^3$ et 55 $Å^3$ respectivement). Mais surtout, le volume des poches occupé par le partenaire naturel est bien plus élevé pour les poches enzymatiques que pour les interfaces protéine-protéine (260 $Å^3$ et 54 $Å^3$ respectivement). Le nombre de poches est aussi différent. Alors que les poches des sites catalytiques sont peu nombreuses (60 % des enzymes étudiées ont seulement une poche présente dans la zone d'interaction avec les substrats), on retrouve souvent plusieurs poches aux interfaces protéine-protéine (en moyenne 6±3) (Fuller et al., 2009). Le volume et le nombre de poches aux interfaces protéine-protéine varient aussi suivant la nature du complexe protéique. On peut ainsi différencier les hétérodimères et les homodimères. Dans une étude parue en 2008, Sonavane et Chakrabarti ont analysé le volume des poches pour un ensemble d'homodimères et d'hétérodimères. Le volume des poches aux interfaces d'homodimères est plus important que celui des poches présentes aux interfaces d'hétérocomplexes (324±498 $Å^3$ et 97±127 $Å^3$ respectivement). On retrouve également deux

fois plus de poches chez les homodimères : en moyenne 5±4,5 pour seulement 2,4±2,1 poches pour les hétérodimères (Sonavane and Chakrabarti, 2008).

Les valeurs trouvées par ces deux études ne semblent pas cohérentes entre elles (par exemple le volume moyen des poches aux interfaces d'hétérodimères est de 54 $Å^3$ pour l'étude menée par Fuller et de 97 $Å^3$ pour celle menée par Sonavane). Ces différences peuvent toutefois être expliquées par les méthodes de détection et de calcul du volume des poches. En effet, la détection et la mesure du volume des poches est un problème loin d'être trivial. En particulier, la détermination du contour d'une poche est subjective et des différences significatives peuvent être observées d'un logiciel à l'autre. Dans une étude comparative réalisée durant ma thèse entre 6 logiciels couramment utilisés, nous avons obtenu des coefficients de corrélations voisins de 0,2 en moyenne. Il est donc extrêmement difficile de comparer deux études entre elles si les méthodes utilisées pour le calcul du volume des poches sont différentes.

II.2.B.bDétection des poches

La détection de poches, contrairement à la détection des « hot spots », nécessite obligatoirement une connaissance structurale du récepteur. Les méthodes permettant de détecter ces poches peuvent être réparties en deux grandes catégories : les méthodes géométriques (par exemple : SURFNET (Laskowski, 1995), LIGSITE (Hendlich et al., 1997), PocketDepth (Kalidas and Chandra, 2008), PocketPicker (Weisel et al., 2007)) et PASS (Brady and Stouten, 2000) et les méthodes énergétiques basées soit sur l'utilisation des sondes, soit sur le calcul d'énergie (par exemple : GRID, Q-SiteFinder (Laurie and Jackson, 2005), CS_Map (Goodford, 1985), AutoLigand (Harris et al., 2008) et ICM-PocketFinder (An et al., 2005)) (Table 4).

La détection des poches à l'interface reste un challenge difficile à surmonter, en particulier à cause de la dynamique de l'interface. La détection des poches nécessite une structure tridimensionnelle de la protéine d'intérêt, or, il est connu de longue date que des changements conformationnels peuvent s'effectuer au niveau de l'interface lors de la formation d'un complexe rendant la détection de poches difficile lorsque la seule information structurale disponible est celle de la forme apo de la protéine. Ces changements conformationnels peuvent s'effectuer par ajustement mutuel des deux partenaires (Koshland, 1958) ou par sélection d'un conformère minoritaire de la forme libre. Les deux modèles ne

sont pas exclusifs mais peuvent coexister. Des poches peuvent apparaître de façon transitoire dans la forme libre. Par exemple, la forme apo de la protéine Bcl-X_L présente une interface plane dans la zone qui interagit avec BAK ; en effet, le serveur Q-SiteFinder (Laurie and Jackson, 2005) qui permet de trouver les poches à la surface d'une protéine, n'en trouve aucune dans la zone d'interaction de Bcl-X_L avec BAK. De la même façon, lorsque l'on essaye de trouver des poches sur la zone d'interaction de la forme complexée avec BAK, Q-SiteFinder ne trouve qu'une petite poche de 97 $Å^3$. En revanche lorsque l'on fait la même analyse sur la conformation de Bcl-X_L en interaction avec le ligand, Q-SiteFinder trouve une poche en plus de celle existant dans le complexe protéine-protéine. Cette poche a un volume de 179 $Å^3$ et permet la stabilisation du complexe protéine-ligand par l'intermédiaire de deux π-stacking entre la petite molécule et les résidus Phe 101 et Tyr 199 (PDB ID 2O2M).

Les poches peuvent donc être préformées (présentes dans la structure de la protéine libre) ou transitoires. Lorsqu'elles sont préformées, les logiciels cités ci-dessus permettent de les détecter. En revanche, lorsqu'elles ne sont pas présentes dans la structure de la protéine libre, il faut préalablement générer la ou les conformations permettant de les détecter (Durrant and McCammon, 2011). Récemment, des protocoles combinant génération de conformères et détection des poches ont permis de détecter des poches pour lesquelles l'information structurale n'était pas disponible. En utilisant des simulations de dynamique moléculaire de 10ns pour générer des conformères, Eyrisch et Helms ont ainsi pu mettre en évidences des poches à la surface de Bcl-X_L, IL-2 et MDM2 en utilisant l'algorithme PASS sur 4000 conformères issus de leurs simulations (Eyrisch and Helms, 2007). De même, plus récemment, la génération de 100 conformères grâce à une méthode basée sur les contraintes de distances (CONCOORD, (de Groot et al., 1997)) a permis de calculer la propension de chaque acide aminé de l'antitrypsine-α_1 à faire partis des résidus délimitant une poche (Patschull et al., 2012).

L'étude des poches est nécessaire lors de l'étude de la druggabilité des PPIs mais elle n'est pas suffisante. D'autres paramètres structuraux permettant de caractériser les interfaces, peuvent être pris en compte pour estimer la druggabilité.

Outil	Année	URL
SURFNET	1995	http://www.chem.ac.ru/Chemistry/Soft/SURFNET.en.html
CAST 2.0	1998	http://sts.bioengr.uic.edu/castp/
PASS	2000	http://www.ccl.net/cca/software/UNIX/pass/overview.shtml
CASTp	2003	http://sts.bioengr.uic.edu/castp/
DrugSite	2004	https://drugsite.msi.umn.edu/home
pvSOAR	2004	http://pvsoar.bioengr.uic.edu
Q-SiteFinder	2005	http://www.modelling.leeds.ac.uk/qsitefinder/
LIGSITE^{CSC}	2006	http://projects.biotec.tu-dresden.de/pocket/
PocketPicker	2007	http://gecco.org.chemie.uni-frankfurt.de/pocketpicker
CLIPPERS	2009	http://crystal.med.upenn.edu/software.html
Metapocket	2009	http://metapocket.eml.org
fpocket	2009	http://fpocket.sourceforge.net/
FTMAP	2009	http://ftmap.bu.edu/
ANCHOR	2010	http://structure.pitt.edu/anchor
POCASA	2010	http://altair.sci.hokudai.ac.jp/g6/Research/POCASA_e.html
DEPTH	2011	http://mspc.bii.a-star.edu.sg/tankp/run_depth.html
MDpocket	2011	http://fpocket.sourceforge.net
Paris	2011	http://cbio.ensmp.fr/paris/paris.html
POVME	2011	http://www2.nbcr.net/data/sw/hosted/POVME/
DoGSiteScorer	2012	http://dogsite.zbh.uni-hamburg.de

Table 4 : Outils de détection et de calcul du volume des cavités à la surface des protéines.

II.2.C Taille de l'interface

II.2.C.a Définition

La taille des interfaces peut être simplement calculée en dimension absolue (Å²), ou de manière plus correcte, en termes de variation de surface accessible au solvant (ΔASA) lors de la formation du complexe. L'algorithme qui permet de calculer cette surface, en faisant rouler une sphère d'un diamètre équivalent à la taille d'une molécule d'eau (généralement 1,4 Å) a été mis au point en 1971 (Lee and Richards, 1971), et modifié en 1994 (Harpaz et al., 1994) puis en 1995 (Gerstein et al., 1995). Le ΔASA, aussi appelé BSA pour Buried Surface Area,

est la mesure la plus couramment utilisée lorsque l'on veut connaître la taille d'une interface protéine-protéine et représente la surface qui devient inaccessible au solvant lors de la formation du complexe. Cette surface est la somme des S-ASA de chacune des sous-unités du complexe, moins la surface accessible du complexe formé (Bahadur and Zacharias, 2008).

$$BSA = SASA_A + SASA_B - SASA_{AB}$$

Où $SASA_{AB}$ est la surface accessible au solvant du complexe formé par les protéines A et B, et $SASA_A$ et $SASA_B$ sont les surfaces accessibles au solvant des protéines A et B respectivement. Certains auteurs divisent ce résultat par deux, afin d'obtenir une valeur représentant la surface accessible par protomère, mais il est aussi possible de calculer la contribution de chaque protéine à la surface d'interaction. On peut noter que la surface accessible au solvant est fortement corrélée avec le nombre d'atomes ou de résidus à l'interface (un atome ou un résidu est considéré à l'interface si son SASA diminue de plus de 0,1 Å² lors de la formation du complexe). En 2007, Bahadur et al. ont calculé la taille de toutes les interfaces de capsides présentes dans la Protein Data Bank (Bahadur et al., 2007). Ils ont ainsi pu mesurer le coefficient de corrélation avec le nombre d'atomes ou de résidus à l'interface. Ils ont trouvé que pour les 779 interfaces, un atome représentait environ 9,8 Å² du BSA (coefficient de corrélation R^2=0,997), tandis que l'on trouve un résidu pour 38 Å² (R^2=0,985) (Bahadur et al., 2007).

II.2.C.b Taille moyenne des interfaces

Les interfaces des complexes hétérodimériques sont assez grandes, avec une surface de 1 910 Å² en moyenne (Bahadur and Zacharias, 2008). On peut noter que le manque de complexes avec des petites interfaces (< 800Å²) indique soit qu'il n'existe pas –ou peu– d'interactions protéine-protéine de petite taille ; soit que les données sur les structures atomiques de ces complexes (par diffraction aux rayons X ou résonance magnétique nucléaire) sont incomplètes. En effet, la caractérisation des structures tridimensionnelles des protéines requiert des complexes stables, ce qui n'est pas nécessairement le cas des complexes dont l'interface est de petite taille. Ceci dénote la possibilité d'un biais sur l'étude de la taille des interfaces protéine-protéine.

II.2.D Forme de l'interface

II.2.D.a Circularité et excentricité

La taille de l'interface, en Å², doit toujours être nuancée par la forme de l'interface, du moins lors de la recherche d'inhibiteurs. En effet, pour une interaction dont la forme serait toute en longueur, cas typique d'une interaction qui fait intervenir un sillon entre deux hélices α sur l'un des partenaires (par exemple MDM2/p53), on utilisera de préférence soit des petites molécules linéaires (entre autres FRH pour l'interleukine 2, *PDB ID : 1PY2* (Thanos et al., 2003) voir § IV.1), soit des peptido-mimétiques d'hélice α. En revanche, pour des interfaces plus circulaires, on préférera des petites molécules non-linéaires, avec différents groupements hydrophobes, pouvant se positionner dans les différentes rugosités présentes dans la zone d'interaction. Pour quantifier la forme de l'interface, on utilise comme mesure la circularité (Figure 5). Cette mesure est définie comme le ratio de la plus petite et de la plus grande des longueurs des axes principaux du plan des moindres carrés des atomes de l'interface. La circularité étant un ratio de deux distances, elle n'a, par définition, pas d'unité et est comprise entre 0 (interface « linéaire ») et 1 (interface parfaitement circulaire).

La notion de circularité a plus tard été délaissée au profit de l'excentricité. Cette dernière, bien que proche de la circularité, est un concept plus général puisqu'elle s'applique à toute section conique. On la calcule d'après l'équation suivante : $\sqrt{1-\frac{b^2}{a^2}}$ dans laquelle a représente la moitié de la longueur du premier axe principal du plan des moindres carrés des atomes de l'interface, et b représente la moitié de la longueur du second axe principal de ce plan. Comme la circularité, l'excentricité n'a pas d'unité et elle varie entre 0 (interface circulaire) et 1 (interface « linéaire ») (Weissten).

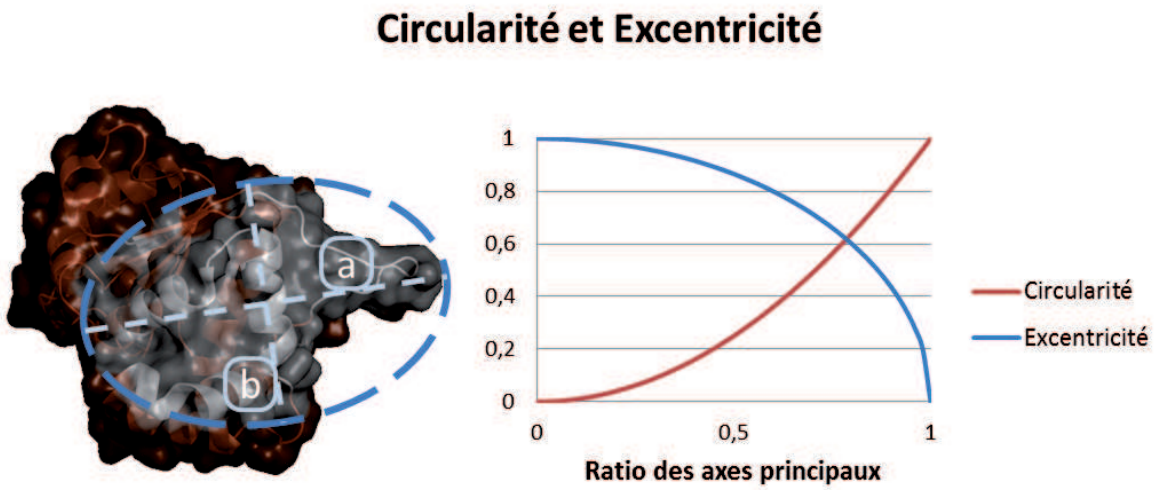


Figure 5 : La circularité et l'excentricité permettent toutes deux de quantifier le caractère circulaire du plan des moindres carrés de l'interface. Cependant, ces deux paramètres varient en sens inverse.

II.2.D.b Planéité de l'interface

L'étude de la taille et de la forme d'une interface protéine-protéine donne d'importantes informations sur l'interaction en elle-même (voir Résultats § I). Cependant, ces deux descripteurs de forme sont des descripteurs bidimensionnels. En effet, restant dans le plan des moindres carrés des atomes de l'interface, aucun d'eux ne donne d'information sur la topologie d'une interface. Or, sans même parler de poches proprement dites, les protéines présentent à leurs surfaces des reliefs qui peuvent -et doivent- être pris en compte lors de toute campagne de « drug-design ». La planéité permet de mesurer cette rugosité de la surface d'interaction entre deux protéines (Jones and Thornton, 1996). La planéité est établie en calculant la déviation de la valeur efficace (ou root mean square deviation) de chacun des atomes de l'interface par rapport au plan des moindres carrés de ces atomes. Cette mesure étant une mesure de distance, elle s'exprime en angström (Å). Plus la planéité est faible et plus l'interface est plane.

II.2.D.c Les interfaces protéine-protéine ne sont pas planes

Une des idées reçues est que les interfaces protéine-protéine ne présentent aucune rugosité permettant à une petite molécule de venir s'y nicher. Cette opinion est due au fait

que, pendant longtemps, les cibles privilégiées des études de « drug design » étaient des enzymes. Or les enzymes ont évolué pour interagir avec des petites molécules. En conséquence, elles possèdent des poches profondes permettant d'accueillir les substrats. De telles poches ne sont généralement pas présentes lors de la formation de complexes protéine-protéine, toutefois des poches de taille réduite, permanentes ou transitoires sont tout de même présentes au niveau de l'interface (Fuller et al., 2009). En effet, les caractérisations de structures de protéines avec un ligand dans la zone d'interaction et les études de dynamique moléculaire montrent que contrairement à l'idée couramment répandue, il existe certaines conformations qui présentent des poches de taille suffisante pour accueillir un ligand (Eyrisch and Helms, 2007, 2009) b (voir § II.2.B.b).

II.2.E Complémentarité de forme

Comme nous venons de le voir, les interfaces protéine-protéine ne sont pas lisses mais irrégulières. Ceci amène la question suivante : lorsque deux protéines s'assemblent, à quel point leurs surfaces sont-elles complémentaires ? En d'autres termes, y a-t-il des cavités qui se forment entre les deux protéines, au niveau de leur zone d'interaction, et si oui, quel volume représentent ces cavités ? Le volume inoccupé que représentent ces cavités entre les protéines à l'interface est appelé Gap Volume (Figure 6).

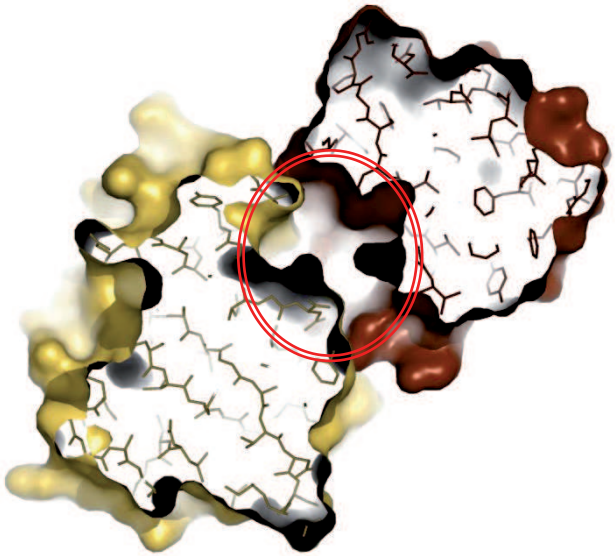


Figure 6 : Le Gap Volume entre deux protéines représente l'espace innoccupé entre deux protéines. Il mesure la complémentarité topologique entre les deux protéines à l'interface

Le premier à apporter un élément de réponse à cette question est Roman A. Laskowski, qui, en 1991, a mis au point un algorithme qui permet d'appréhender le volume de ces cavités (Gap Volume), et l'a intégré dans un logiciel gratuit pour les académiques : SURFNET (http://www.biochem.ucl.ac.uk/~roman/surfnet/surfnet.html). L'algorithme utilisé par SURFNET insère des sphères à l'interface entre des paires d'atomes des deux protéines. Les rayons de chacune des sphères est calculé de manière à ce qu'aucun atome d'une ou l'autre des protéines ne soit inclus dans la sphère. Si, à un moment ou à un autre, le rayon de la sphère vient à être inférieur à 1Å, cette sphère est mise de côté. En revanche, si la sphère est toujours pertinente après avoir vérifié la position des atomes les plus proches, son volume et ses coordonnées sont sauvegardées. Lorsque toutes les paires d'atomes ont été vérifiées, les volumes des sphères sauvegardées sont utilisés pour calculer le volume des cavités à l'interface (Laskowski, 1995). Cet algorithme est toujours celui utilisé pour calculer le Gap Volume entre deux protéines.

En 1995, Jones & Thornton ont montré, sur un jeu de 32 dimères non-homologues, que le volume des cavités à l'interface est corrélé avec la taille de cette interface (Jones and Thornton, 1995), (Figure 7). Afin de s'affranchir de la taille des interfaces, et de pouvoir comparer les volumes des cavités de complexes avec des interfaces de tailles différentes, elles ont défini le Gap Volume index (GV_i) qui correspond au GV normalisé par unité de surface :

$$GV_i = \frac{Gap\ Volume(\dot{A}^3)}{Buried\ Surface\ Area\ (\dot{A}^2)}$$

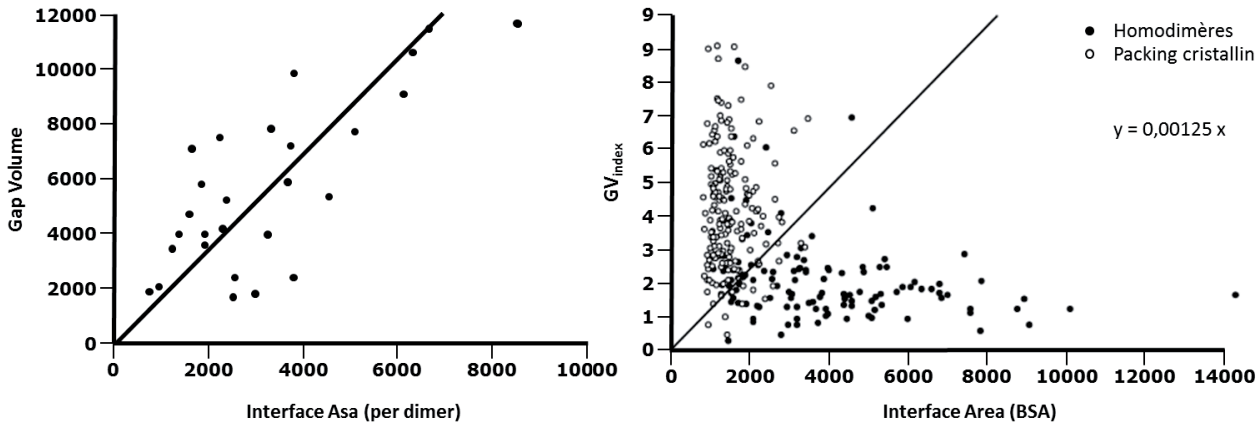


Figure 7 : à gauche : Gap volume à l'interface des homodimères en fonction de la surface accessible au solvant. A droite : Gap volume index des homodimères et du packing cristallin en fonction du ΔASA. Figure d'après Bahadur *et al*, 2008.

En utilisant cet index, Jones & Thornton ont montré que les complexes transitoires avaient des surfaces d'interactions moins complémentaires que les complexes permanents, ou que les complexes enzymes-substrats. En effet, alors que les GV_i des hétérodimères transitoires et des complexes anticorps-antigènes sont de 3,02±0,80 Å, les GV_i des homodimères et des complexes enzymes-substrats sont de 2,20±0,87 Å et celui des hétéro-complexes permanents sont de 1,47±1,34 Å.

Le GV_i a été par la suite utilisé en complément lors des études de planéité des interfaces protéiques, notamment en 2004 par Bahadur *et al*, sur un jeu de donnés plus conséquent puisque composé de 122 homodimères, 70 hétérodimères et 188 complexes dus aux contacts cristallin (Bahadur et al., 2004). Ils arrivent aux mêmes conclusions que Jones & Thornton, avec un GVi moyen de 4,4±1,9 Å pour les complexes dus aux contacts cristallins et un GVi moyen plus petit de moitié pour les homodimères (2,1±1,2 Å).

II.2.F Composition en acides aminés

L'étude de la composition en acides aminés aux interfaces protéine-protéine peut se révéler intéressante pour mieux comprendre les principes qui gouvernent la formation de complexes de deux –ou plusieurs– protéines.

Afin de pouvoir comparer la composition en acides aminés entre l'interface et le reste de la surface de la protéine, on peut calculer la proportion de divers acides aminés au niveau de la surface d'interaction. Ensuite, on la divise par la proportion de ce même acide aminé sur l'ensemble de la surface de la protéine. La propension de l'acide aminé AA_j à se trouver à l'interface est donnée par la formule suivante :

$$\text{Propension } (AA_j) = \left(\frac{\sum_{i=1}^{Ni} ASA_{AA_j(i)}}{\sum_{i=1}^{Ni} ASA_{(i)}}\right) / \left(\frac{\sum_{i=1}^{Ns} ASA_{AA_j(s)}}{\sum_{i=1}^{Ns} ASA_{(s)}}\right)$$ (Jones and Thornton, 1996).

Où $\sum_{i=1}^{Ni} ASA_{AA_j(i)}$ est la somme de l'ASA des acides aminés de type j à l'interface du monomère, $\sum_{i=1}^{Ni} ASA_{(i)}$ est la somme de l'ASA de tous les résidus à l'interface du monomère, $\sum_{i=1}^{Ns} ASA_{AA_j(s)}$ est la somme de l'ASA des acides aminés de type j à l'interface du monomère (y compris l'interface) et $\sum_{i=1}^{Ns} ASA_{(s)}$ est la somme de l'ASA de tous les résidus à la surface du monomère (y compris l'interface).

Cette façon de comparer les compositions en acides aminés a l'avantage de donner une seule valeur par acide aminé. Si cette valeur est supérieure à 1, cela signifie que le résidu en question est statistiquement plus représenté à l'interface que sur la protéine entière, et inversement s'il est inférieur à 1.

La première étude sur la propension des acides aminés à se situer au niveau des interfaces protéine-protéine a montré des différences significatives entre complexes homodimériques et hétérodimériques. En 1996, Jones & Thornton ont appliqué cette méthode sur leur jeu de données de 32 complexes et ont ainsi montré qu'à l'exception de la méthionine, la proportion des acides aminés hydrophobes entre la surface d'interaction avec le partenaire protéique et la surface totale de la protéine était plus élevée pour les complexes homodimériques qu'hétérodimériques. De plus, la faible proportion des acides aminés hydrophobes sur la surface d'interaction des complexes hétérodimériques est équilibrée par l'augmentation de la proportion en résidus polaires.

Par la suite, plusieurs autres équipes ont travaillé sur la composition en acides aminés des PPIs, notamment Bahadur et Zacharias qui ont pu réaliser une étude similaire, mais sur un nombre plus important de complexes protéine-protéine (122 homodimères et 70 hétérodimères) et ont obtenu des résultats plus précis (Bahadur and Zacharias, 2008). Les surfaces d'interactions des hétérodimères et homodimères sont enrichies en acides aminés aliphatiques (Leu, Val, Ile, Met) ainsi qu'en résidus aromatiques (His, Phe, Tyr, Trp). On peut noter que la tyrosine et le tryptophane sont deux des trois acides aminés retrouvés dans les « hot spots » les plus fréquents (voir chapitre II.2.A.b). Par ailleurs, ils sont pauvres en résidus chargés (Asp, Glu, Lys) autres que l'arginine (Bahadur et al., 2003; Lo Conte et al., 1999) qui est le troisième acide aminé fréquemment présent dans les « hot spots ».

Les protéines sont majoritairement composées de carbone, d'oxygène, d'azote et de soufre (ainsi que d'hydrogène) qui peuvent être subdivisés en deux groupes : les atomes polaire (N, O et S) et non polaire (C). En moyenne, 58% de la surface accessible au solvant des protéines est constituée d'atomes de carbone et 42% d'atomes polaires. Dans les zones d'interactions protéine-protéine, l'écart entre les atomes polaires et non polaires tend à s'accroître puisque 65% de la surface accessible au solvant est constituée d'atomes de carbone, et seulement 35% d'atomes polaires (Bahadur and Zacharias, 2008).

II.2.G Segments à l'interface et classification

Comme nous l'avons vu dans le chapitre précédent, il existe différents paramètres qui permettent de classer les interactions protéine-protéine en fonction de critères géométriques ou physicochimiques. Un autre paramètre, le nombre de segments présents à l'interface, permet de classer les différents complexes protéiques.

Un segment est défini comme un tronçon de protéine, qui commence par un résidu présent à l'interface et qui finit par un autre résidu à l'interface. La séparation entre deux segments est définie soit par un tronçon de plus de 6 résidus (Jones and Thornton, 1996) soit de plus de 5 résidus (Pal et al., 2007) consécutifs ne faisant pas partie de l'interface. Au cours de ma thèse, j'ai utilisé la définition de Jones. En 1996, Jones et Thornton ont utilisé 32 homodimères non homologues et 11 hétéro-oligomères dont 4 sont des complexes permanents et les 7 autres sont transitoires. On notera que ces 11 hétéro-oligomères ne sont ni des complexes enzyme/substrat, ni des complexes anticorps/antigène, puisqu'elles ont séparé ces interfaces particulières du reste du jeu de données. De leur côté, Pal et Bahadur ont utilisé 122 homodimères et 204 hétéro-oligomères (soit 2x204 = 408 polypeptides différents) et 273 complexes issus des contacts cristallins (103 homodimères isologues et 85 hétérodimères ou homodimères hétérologues, ce qui fait 103+2x85=273 polypeptides différents).

Il ressort de l'étude de Jones et Thornton que les homodimères possèdent en moyenne 5,2±2,6 et 5,0±2,5 segments à l'interface pour les homodimères et hétérodimères respectivement. De leur côté Pal et Bahadur aboutissent à des valeurs relativement identiques pour les hétéro-oligomères (4,7±2,9) et sensiblement plus élevées pour les homodimères : 6,0±3,6. Les contacts cristallins ont environ le même nombre de segments que les hétéro-oligomères, bien que leur répartition soit beaucoup plus centrée autour de la moyenne : 4,5±1,9. Les différences du nombre de segments à l'interface entre homodimères et hétérodimères étant très proches, ce paramètre ne permet pas de les dissocier. Afin de palier à ce problème, Pal et Bahadur ont normalisé les valeurs obtenues suivant une unité de mesure liée à la taille d'une PPI. Soit le nombre de segment pour 1000 Å² d'interface, soit le nombre de segment pour 100 résidus à l'interface. Ces deux ratios permettent de mieux différentier les homodimères des contacts cristallins qui sont moitiés moindres (6.3±2.2 segments contre 3.4±1.6 pour 1000 Å²). Les homodimères possèdent plus de segments qui font partie de cœur de l'interface que les autres types de complexes (1.7±1.6 contre 1.0±1.2 pour les hétéro-oligomères et 0.5±0.7 pour les contacts cristallins). En utilisant une classification linéaire

entre le nombre de segments par 1000 Å² à l'interface et le nombre de segments au cœur, Pal et Bahadur retrouvent 80 % des homodimères au-dessus de la ligne (Figure 8) et 91 % des contacts cristallins au-dessous.

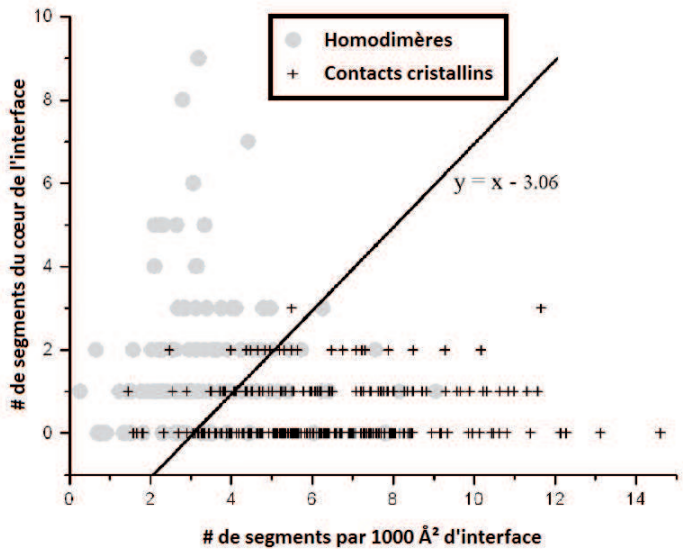


Figure 8 : En 2007, Pal et Bahadur ont développé une méthode structurale simple pour différencier les homodimères des simples contacts cristallins : une classification linéaire entre le nombre de segments par 1000 Å² d'interface et le nombre de segments du cœur de l'interface. Figure tirée de Pal et Bahadur, 2007.

Le nombre de segments à l'interface permet de différencier les hétérodimères, les homodimères et le « packing » cristallin, mais aussi les complexes entre deux protéines globulaires, des complexes impliquant, chez un partenaire au moins, un seul élément de structure secondaire. Les complexes protéine-protéine contiennent plus de segments à l'interface que les complexes protéine-peptide (voir Résultats § I). Cette différenciation est importante lorsque l'on étudie la druggabilité des interfaces protéine-protéine car ces deux types d'interaction possèdent des propriétés différentes. Par exemple, pour une interaction protéine-peptide, le partenaire peut être mimé par des peptides (Vlieghe et al., 2010), des peptoïdes (Lee et al., 2008), des peptidomimétiques (Ross et al., 2010), des « stappled peptides » (Bird et al., 2008) ou par des molécules organiques (Higueruelo et al., 2009).

Depuis 20 ans, les avancées de la génomique, de la protéomique et de l'informatique ont permis la création de bases de données qui répertorient les données expérimentales obtenues par diverses méthodes lors d'études de complexes protéiques *in vitro* et *in cellulo*. Certaines bases de données, accessibles à tous via des serveurs web, listent également des

interactions entre domaines protéiques, ou bien encore des interactions prédites informatiquement.

II.3 Outils bioinformatiques et bases de données

Nous avons vu dans le chapitre précédent les paramètres à considérer pour étudier la druggabilité des interfaces protéine-protéine d'un point de vue structural. Le nombre de données de plus en plus important nécessite le développement de bases de données qui répertorient les PPIs validées expérimentalement (présentées dans le chapitre I.1.A), et d'outils permettant d'analyser ces données.

L'ensemble des paramètres cités dans les chapitres précédents peuvent être utilisés pour estimer la druggabilité des interactions protéine-protéine. Afin de pouvoir calculer ces paramètres, un grand nombre de serveurs ont été développés ces dernières années. Les algorithmes utilisés par ces serveurs, peuvent, comme les algorithmes permettant de calculer le volume des poches, être subdivisés en deux catégories : les algorithmes géométriques et les algorithmes énergétiques (Schmidtke et al., 2010). Parmi les outils les plus populaires, la plupart sont accessibles via une interface web : PDBePISA (Krissinel and Henrick, 2007), InterProSurf (Negi et al., 2007), Dr. PIAS (Sugaya et al., 2012)et PROTORP (Reynolds et al., 2009) qui n'est cependant plus disponible. La plupart de ces bases de données décrivent les propriétés des interfaces des PPIs d'un point de vue structural : taille de l'interface, composition en acides aminés, charges à l'interface, etc… (Table 5).

Outil	Année	URL
SURFNET	1995	http://www.chem.ac.ru/Chemistry/Soft/SURFNET.en.html
ConSurf	2002	http://consurf.tau.ac.il/
PDBePISA	2005	http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html
PIBASE	2005	http://salilab.org/pibase/
Proface	2006	http://resources.boseinst.ernet.in/resources/bioinfo/interface
InterProSurf	2007	http://curie.utmb.edu/prosurf.html
PIC	2007	http://pic.mbu.iisc.ernet.in/
PPIDD	2008	http://ppidd.cgm.cnrs-gif.fr/ppidd/
Pisite	2009	http://pisite.hgc.jp/
Dr. PIAS 2.0	2012	http://www.drpias.net
PocketQuery	2012	http://pocketquery.csb.pitt.edu
ProtinDB	2012	http://protindb.cs.iastate.edu/index.py
PROTORP	2009	Indisponible

Table 5 : Liste de serveurs permettant d'analyser les interactions protéine-protéine .

III Chimiothèques dédiées

La première constatation qui peut être faite à propos des inhibiteurs des interactions protéine-protéine est qu'ils sont recherchés par des criblages de banques de petites molécules. Malheureusement, les banques actuellement criblées ont été développées spécifiquement depuis 50 ans dans le but d'identifier des inhibiteurs de poches enzymatiques. Elles contiennent énormément de molécules similaires car dévéloppées autour d'une même chimie cherchant à mimer les substrats enzymatiques les plus courants et notamment des analogues de nucléotides. S'il paraît évident que cribler de telles banques de molécules sur une enzyme donnera un grand nombre de molécules actives (« hits »), rien ne dit qu'elles soient adaptées au criblage d'interfaces protéine-protéine impliquées dans des interactions protéine-protéine. Les bases de données de petites molécules jusqu'à présent utilisées lors de la recherche de nouveaux médicaments sont donc biaisées et il est nécessaire de créer de nouvelles chimiothèques adaptées à ce nouvel espace chimique (voir § III.4).

Pour pouvoir créer une telle chimiothèque, nous avons étudié au cours de ma thèse les molécules connues pour inhiber les interactions protéine-protéine, afin de pouvoir en extraire des règles permettant de créer des chimiothèques enrichies en inhibiteurs de PPIs.

Les peptides sont très utilisés pour l'inhibition des interactions protéine-protéine en raison de leur facilité à mimer le partenaire naturel. Ils ont cependant un défaut majeur pour l'élaboration de nouveaux médicaments : leur vectorisation. En effet, la plupart de ces molécules a des difficultés à franchir les barrières biologiques et / ou a une demi-vie au sein de l'organisme beaucoup trop courte pour qu'elles puissent être actifs (Tanaka et al., 2011). En conséquence, un champ fortement dynamique a émergé dans la conception de petites molécules qui miment les structures secondaires et présentent des fonctionnalités de manière similaire (Ross et al., 2010). Les structures des molécules ainsi synthétisées sont donc fortement dépendantes des structures secondaires présentes à l'interface du complexe protéique. Une approche parallèle consiste à utiliser de petites molécules de synthèse, des produits naturels ou des fragments pour développer des modulateurs de PPIs. Cette approche qui est largement utilisée pour identifier des inhibiteurs d'une activité biologique de type enzymatique par exemple, et pouvant éventuellement conduire au développement de médicaments, est encore parfois controversée dans le cas des PPIs en raison des propriétés

structurales des interfaces protéine-protéine comme il a été décrit dans les chapitres précédents. En effet, les propriétés physicochimiques des interfaces des complexes protéiques étant très différentes de celles des interfaces enzyme-ligand (§ I.3.C), les molécules utilisées pour inhiber les complexes enzyme-ligand ne sont pas nécessairement adaptées à l'inhibition de complexes protéine-protéine. Pour définir les propriétés d'un ensemble de molécules et pour pouvoir comparer les molécules entre elles, on utilise généralement le concept d'espace chimique.

III.1 Univers et espaces chimiques

L'univers des molécules chimiques peut être subdivisé en **4 grandes catégories** suivant leur disponibilité (Hann and Oprea, 2004). **L'espace virtuel** regroupe l'ensemble des molécules qui peuvent en principe être créées. Ce nombre est grossièrement estimé à 10^{60} si on se limite aux molécules de bas poids moléculaire (MW≤500) (Bohacek et al., 1996). Seule une infime partie de cet espace chimique a actuellement été explorée dans le cadre de la recherche de nouveaux médicaments. **L'espace tangible** contient toutes les molécules qui peuvent être facilement synthétisées. Ce nombre de composés est compris entre 10^{20} et 10^{24} selon Peter Erl (Ertl, 2003). Ces chiffres ont été obtenus en analysant la base interne de composés commerciaux de Novartis en termes de substituants de moins de 13 atomes lourds. **L'espace Global** regroupe tous les composés réellement synthétisés. On peut estimer le nombre de ces composés à quelques millions ; il est toutefois très difficile de connaître de façon précise le nombre de molécules synthétisées dans le monde. D'une part, parce que beaucoup de molécules sont synthétisées par la recherche industrielle et sont donc confidentielles. D'autre part, un grand nombre de molécules synthétisées dans les laboratoires académiques ne sont pas répertoriées. L'absence d'information sur les composés académiques met en exergue l'utilité des projets regroupant ces composés dans des collections nationales, européennes ou même mondiales. En 2003, Marcel Hibert a initié un tel projet, dans le but de regrouper tous les composés présents dans les laboratoires académiques français : la Chimiothèque Nationale (http://chimiotheque-nationale.enscm.fr/). Aujourd'hui, elle comporte près de 50 000 produits de synthèse et 14 000 extraits d'origine naturelle. A contrario, l'ensemble des composés disponibles commercialement est regroupé dans la base de données Zinc (http://zinc.docking.org/) qui représente environ 21 millions de molécules

vendues par une centaine de fournisseurs. Différents sous-ensembles prédéfinis sont disponibles, par fournisseur, par propriétés physicochimiques des molécules, ou selon diverses propriétés (produits naturels, métabolites, médicaments, synthons …). Chaque catalogue a été filtré avec 54 règles pour sélectionner des composés pertinents dans le cadre de recherche de nouveaux médicaments (http://blaster.docking.org/filtering/rules_default.txt). **L'espace réel** correspond à tous les composés possédés par un même fournisseur ou une société pharmaceutique, ce nombre peut aller de quelque dizaine de milliers jusqu'à plusieurs millions (Schuffenhauer et al., 2004).

L'univers chimique couvert par les molécules physiques disponibles représente donc une infime partie de l'espace chimique théorique, ce qui peut expliquer en partie le faible nombre de nouvelles entités moléculaires introduites sur le marché ces dernières années (Bunnage, 2011; Overington et al., 2006). Toutefois, pour une famille de cibles donnée, il est possible de se focaliser sur une zone bien définie de cet espace. La notion d'espace chimique, ou pharmacologique est de plus en plus répandue, notamment pour la conception, la caractérisation ou la comparaison de collections de molécules (Dobson, 2004; Lipinski and Hopkins, 2004; Wong, 2012). Il est alors important de disposer d'outils chémoinformatiques pour pouvoir caractériser cet espace chimique et comparer les molécules entre elles.

III.2 Les outils chémoinformatiques

III.2.A Les descripteurs moléculaires

Les descripteurs moléculaires permettent de comparer et d'analyser les propriétés physicochimiques et structurales des molécules de façon rapide et efficace. Il existe un grand nombre de descripteurs à ce jour (par exemple la version 6 du logiciel Dragon propose actuellement 4 885 descripteurs moléculaires répartis en 29 classes). Ils peuvent correspondre à des propriétés simples et facilement interprétables comme le poids moléculaire ou le nombre d'atomes lourds ou à des propriétés beaucoup plus « complexes » comme les descripteurs WHIM qui permettent de transcrire des informations 3D pertinentes (taille, forme, symétrie et distribution atomique) et qui sont basés sur des indices statistiques calculés sur la projection des atomes selon les axes principaux des atomes des molécules. Certains descripteurs peuvent également être mesurés expérimentalement, comme le LogP qui représente le coefficient de partage octanol-eau. Toutefois, tous les paramètres doivent pouvoir être calculés ou prédits à

l'aide de logiciels chémoinformatiques dédiés (ADMET, DRAGON, MOE, VOLSURF, CHEMAXON, SYBYL). Les algorithmes permettant de calculer les descripteurs doivent être suffisamment rapides pour pouvoir être appliqués à un grand nombre de molécules. En général, ce temps de calcul est corrélé avec la quantité d'information inhérente au paramètre étudié. Par exemple, le poids moléculaire ou le nombre d'atomes de carbone sont extrêmement rapides à estimer mais apportent peu d'information sur les propriétés physicochimiques des molécules en question. En revanche, des descripteurs basés sur des propriétés quantiques de la molécule apportent des informations très précises des propriétés moléculaires mais ne sont généralement pas applicables en raison du temps de calcul prohibitif nécessaire pour les calculer. Il est communément admis de classer les descripteurs moléculaires en fonction de la dimensionnalité de la structure nécessaire pour leur calcul. On parlera donc de **descripteurs 1D**, **2D** et **3D** (Table 6).

Structure de la molécule	Informations	Exemple de descripteurs
1D	Formule brute Atomes présents	Masse moléculaire Présence/nombre d'un atome donné
2D	Enchaînement des atomes	CLog P Fingerprints
3D	Structure minimisée Conformations	Surfaces Volumes Pharmacophores

Table 6 : Exemples de descripteurs en fonction de la dimensionnalité de la structure de départ.

Les descripteurs de type 1D regroupent des descripteurs très rapides à calculer tels que le poids moléculaire, le nombre d'atomes d'un type particulier… Ils ne nécessitent que la formule brute de la molécule et permettent de comparer les molécules de façon très grossière.

Les descripteurs de type 2D nécessitent de disposer de la structure développée des molécules. Ils peuvent être subdivisés en différents groupes. Ils regroupent des caractéristiques telles que le nombre de rotules (ou liaisons rotatoires), le nombre de cycles, le nombre de donneurs ou d'accepteurs de liaison hydrogène… Un certain nombre de descripteurs 2D peuvent être estimés à l'aide de modèles prédictifs QSPR (Relation Quantitative Structure/Propriété) comme le CLogP ou ALogP, la solubilité.

Les descripteurs de type 3D permettent de prendre en compte des paramètres de forme qui sont d'une importance cruciale pour caractériser une interaction macromolécule-ligand. Toutefois, il convient de noter que les descripteurs 3D sont souvent mis de côté pour les calculs de diversité car ils nécessitent au préalable le calcul des structures 3D, ce qui pose deux gros problèmes : le temps de calcul et le choix de la ou des conformations bioactives. Les descripteurs 1D et 2D sont donc les plus utilisés.

Les très nombreux descripteurs moléculaires permettent de comparer des molécules et peuvent donc être utilisés pour définir un espace chimique relatif à un ensemble de composés ou encore pour corréler les propriétés de molécules à leur activité biologique sous forme de modèles QSAR. Ils sont également très utiles pour estimer la diversité d'un ensemble de composés donnés.

III.2.B Diversité moléculaire

Le principe de **similarité** stipule que des composés proches possèdent des activités biologiques similaires. Si l'on s'en réfère à ce principe, on peut accélérer et réduire le coût de campagnes de criblage en ne considérant que des molécules diverses. Par ailleurs, toujours selon ce principe, lorsqu'une ou plusieurs « touches » ont été identifiées, il est possible d'obtenir d'autres composés d'intérêt en explorant l'espace chimique au voisinage des touches initiales. Bien que le principe de similarité soit remis en cause dans de nombreux cas, la diversité moléculaire est très souvent utilisée lors de campagne de criblages.

Etant donné le grand nombre de descripteurs existants, une des premières difficultés est de choisir ceux à utiliser pour représenter la diversité au sein d'une collection de molécules. Une analyse en composantes principales permet généralement d'extraire les descripteurs les plus significatifs. Une pondération des descripteurs peut être apportée afin de prendre en compte leur importance. Une des approches largement répandue pour estimer la diversité moléculaire consiste à utiliser des **« empreintes moléculaires »** ou **fingerprints**. Les fingerprints encodent une structure moléculaire en une série de bits qui représentent la présence ou l'absence d'une sous-structure particulière dans la molécule. Comparer deux molécules revient alors à comparer leurs fingerprints, bits par bits. C'est le type de descripteurs le plus utilisé pour les études de similarité et de diversité. La raison est que ces descripteurs parviennent généralement à saisir un grand nombre d'informations de la molécule, mais surtout que ces informations sont stockées sous une forme très condensée. La

plupart des fingerprints sont codés par des chaînes allant de quelques dizaines à un millier de bits. Ce format est performant en termes d'espace disque (il occupe peu d'espace disque et il est donc facile de sauvegarder les fingerprints de bases de données de millions de molécules) et de performances (les opérations sur les chaînes de bits sont réalisées de manière très performante par les ordinateurs). Un des fingerprints les plus utilisés est le fingerprint MACCS. Il a été mis au point par MDL pour accélérer la recherche sous structurale dans les bases de données. Différents type de coefficients, pouvant être utilisés avec les fingerprints 2D, ont été décrits pour estimer la similarité (ou à contrario la dissimilarité) entre composés. Le plus couramment utilisé est le coefficient de Tanimoto (Willett, 2006).

III.2.C Efficacité de ligand

En 2004, Hopkins a introduit le concept **« d'efficacité de ligand »** (ligand efficiency ou LEI) en rapportant l'énergie d'interaction entre un ligand et son partenaire protéique au nombre d'atome lourds du ligand (Hopkins et al., 2004). Plus récemment, d'autres indices d'efficacité de ligands ont été proposés, tels que BEI et SEI qui représentent l'affinité divisée par le poids moléculaire et la surface accessible polarisable, respectivement (Abad-Zapatero and Metz, 2005). Ces indices sont notamment utilisés lors des phases d'optimisation des premières « touches » obtenues. La taille des molécules et leur polarité sont des critères essentiels pour le développement de composés à caractère thérapeutique. La combinaison des indices BEI et SEI permet de représenter l'espace chimico-biologique de grands ensembles de composés (Abad-Zapatero et al., 2010). Ces indices peuvent être utilisés pour déterminer la capacité de molécules à devenir de « bons candidats » médicaments. Ces indices présentent l'avantage, par rapport à des critères de type « règle des 5 » de Lipinski de donner des valeurs continues et non pas des cutoffs abrupts. L'analyse comparative des modulateurs d'interaction protéine-protéine a montré qu'ils sont en général légèrement moins efficaces que les molécules optimisées pour d'autres cibles. En effet, en 2007, Wells et McClendon ont estimé à 0,24 le LEI moyen de 13 inhibiteurs dérivant de 6 cibles PPI majeures (Wells and McClendon, 2007). Les auteurs en concluent que pour obtenir un modulateur avec une affinité de 10nM, il faudrait une molécule avec un poids moléculaire de 645 Da ; ce qui est largement au-dessus de la limite pour qu'un composé possède des propriétés pharmacocinétiques acceptables (ADME), selon les règles de « type Lipinski ». Deux années plus tard, le groupe de Sir Tom Blundel a mis en place la base de données de molécules chimiques, Timbal

(http://mordred.bioc.cam.ac.uk/timbal), regroupant l'ensemble des modulateurs d'interactions protéine-protéine connus à ce jour (Higueruelo et al., 2009). A partir des 76 molécules pour lesquelles les données d'affinité étaient disponibles et correspondant à 17 complexes protéine-protéine, ils ont obtenu une valeur moyenne de 0,27 ± 0,1 pour le LEI pour des composés possédant en moyenne 30 atomes (valeurs variant de 0,15 à 0,35). A titre de comparaison, le LEI obtenu pour des composés en phase d'optimisation pour d'autres types de cible varie de 0,32 à 0,43. Au cours de mon travail de thèse, nous avons calculé différents indices d'efficacité des modulateurs d'interaction protéine-protéine présents dans la base de données 2P2Idb et montré que, contrairement aux idées répandues, un grand nombre de ces composés sont dans un espace chimique qui permet d'envisager leur optimisation en molécules à visée thérapeutique (voir Résultats, chapitre III).

III.3 Espace chimique des modulateurs PPIs

Avec l'augmentation du nombre d'inhibiteurs d'interactions protéine-protéine, il est devenu possible de caractériser l'espace chimique représenté par ces composés afin de le comparer aux autres types de composés « drug-like » obtenus pour d'autres cibles biologiques. En 2004, Pagliaro *et al.* ont été les pionniers en réalisant une analyse en composantes principales sur 19 inhibiteurs d'interactions protéine-protéine (Pagliaro et al., 2004). Ils ont ainsi montré que seule la moitié des inhibiteurs étaient représentée dans l'espace chimique de diversité correspondant à 3 chimiothèques commerciales couramment utilisées (Chemical Diversity, 119 475 composés ; Maybridge, 59 223 composés et Asinex, 321 867 composés). Les 19 inhibiteurs pris en compte dans cette étude étaient issus de 12 cibles biologiques différentes. Cependant, seuls les inhibiteurs de 4 cibles Bak-BH3/Bcl-xL, MDM2/p53, NGF/p75 et LFA/ICAM-1 étaient représentés dans les 3 chimiothèques commerciales. En conséquence, ils ont été les premiers à suggérer de concevoir des **chimiothèques dédiées** aux interactions protéine-protéine pour améliorer les taux de succès dans la recherche de nouveaux modulateurs. En 2007, Neugebauer *et al.* ont pour la première fois combiné une approche chémoinformatique et des méthodes d'apprentissage pour extraire les caractéristiques des modulateurs d'interactions protéine-protéine, dans le but de pouvoir sélectionner ces composés parmi une collection de molécules (Neugebauer et al., 2007). Ils ont construit un jeu de données de 25 inhibiteurs issus de la littérature comme set positif d'entrainement et 1 137 composés de la FDA-approved (composés approuvés par la Food &

Drug Administration) comme set négatif. Ils ont ensuite construit un arbre de décision à partir des 3 descripteurs jugés les plus significatifs (un facteur de forme, un ensemble de distances interatomiques des atomes à la périphérie et représentatifs de la structure 3D, et le nombre de fonctions esters). Leur algorithme a démontré de bonnes performances sur le set d'entrainement en permettant de séparer le set négatif du set positif. Il a été appliqué sur un ensemble de 1 130 molécules issues de la base de données ZINC, sélectionnant 185 modulateurs potentiels (16%). Malheureusement, l'algorithme n'a pas été validé sur des données biologiques. En 2009, Higueruelo *et al.* ont développé une base de données dédiée aux modulateurs d'interactions protéine-protéine (Higueruelo et al., 2009). La première version de cette base de données comprenait 104 modulateurs dont 27 ont été cristallisés en présence de leur cible biologique. Les auteurs ont comparé les propriétés des molécules présentes dans TIMBAL avec différentes collections de composés : des composées en phase préclinique ou phase I à IV, des composés issus de chimiothèques commerciales diverses (Enamine, Asinex et Maybridge) et des ligands provenant de la PDB. Ils ont ainsi pu montrer qu'en moyenne les inhibiteurs d'interactions protéine-protéine sont plus « gros » en taille que des inhibiteurs d'enzymes, ou des ligands de récepteurs classiques. De plus, ils sont généralement plus hydrophobes, contiennent des cycles aromatiques et moins de donneurs et d'accepteurs de liaison hydrogène. D'un point de vue fonctionnalité chimique, les modulateurs PPI contiennent plus de groupement acide carboxylique et sulphonamides et moins de fonctions éther. La plupart de ces observations peut être facilement corrélée avec les propriétés géométriques très différentes des sites de liaisons par rapport aux sites enzymatiques, et notamment le nombre et la forme des poches comme il a été évoqué dans le § II.2.B. Une nouvelle version de la base de données TIMBAL recensant plus de 3 000 composés et 27 cibles protéine-protéine a été mise en ligne très récemment (http://mordred.bioc.cam.ac.uk/timbal). La maintenance est dorénavant effectuée automatiquement par des requêtes dans la base de données ChEMBL (https://www.ebi.ac.uk/chembl/).

Plus récemment, l'équipe de Bruno Villoutreix a également développé un algorithme basé sur des méthodes d'apprentissage dans le but d'enrichir les chimiothèques en modulateurs d'interactions protéine-protéine (Reynes et al., 2010; Sperandio et al., 2010). Leur jeu de données initial était composé de 145 modulateurs validés expérimentalement et 4 857 composés issus de la base de données DrugBank. Après élimination de la redondance et

des molécules indésirables de par leurs propriétés pharmacocinétiques, ils ont obtenu un jeu final d'entrainement pour leur modèle constitué de 66 modulateurs PPI pour le set positif et 557 molécules représentatives de l'espace chimique des composés « drug-like » pour le set négatif. Après une analyse comparative des deux jeux de données avec des descripteurs Dragon, ils ont identifié deux paramètres permettant de discriminer les deux ensembles de molécules : un facteur de forme (RDF070m) et un facteur lié au degré d'insaturation des molécules (UI). Deux arbres de décisions ont été construits à partir de ces deux descripteurs en utilisant des seuils de tolérance différents pour le UI (>3,95 et >4,13). Le premier modèle (UI>3,95) montre une forte sensibilité (81% contre 70% pour le second modèle). Par contre le second modèle (UI>4,13) conduit à une meilleure spécificité (80% au lieu de 70% pour le premier modèle). Ces modèles ont été validés sur 10 essais biologiques relatifs à des interactions protéine-protéine issus de la base de données PubChem (http://www.ncbi.nlm.nih.gov/pcassay) et un criblage *in vitro* sur le complexe p53/MDM2 réalisé sur la plateforme CDithem (http://www.cdithem.fr/) montrant des facteurs d'enrichissement allant de 1,4 à 5,4 pour les chimiothèques filtrées. Finalement, les collections de molécules de MayBridge (57 200 composés) et ChemBridge (50 000 composés) ont été filtrées sélectionnant respectivement 13 799 (24%) et 9 622 (19%) composés potentiellement inhibiteurs d'interactions protéine-protéine. L'algorithme a été mis à la disposition de la communauté scientifique sous le nom de PPI-HitProfiler (http://www.cdithem.fr/ppiHitProfiler.php) et peut être téléchargé pour une utilisation locale. Ce programme permet de filtrer une collection de composés à partir d'un fichier au format SDF, pour obtenir une chimiothèque enrichie en inhibiteurs d'interactions protéine-protéine.

Au cours de mon travail de thèse, j'ai également contribué au développement d'outils dans le but de définir le profil caractéristique des inhibiteurs d'interactions protéine-protéine. Ces travaux sont développés dans la partie Résultat (chapitre II).

Les études décrites ci-dessus montrent que l'utilisation d'outils chémoinformatiques et de méthodes d'apprentissage est particulièrement adaptée pour développer des banques de composés dédiées aux interactions protéine-protéine dans le but d'améliorer les taux de succès dans les campagnes de criblage. En effet, l'une des stratégies de prédilection pour identifier des modulateurs d'interaction protéine-protéine reste le criblage expérimental de molécules, à moyen et haut débit. Cette technique a fortement bénéficié des progrès

technologiques de robotisation et de miniaturisation des essais biologiques au cours des dernières années. Il est aujourd'hui possible de tester l'activité d'un grand nombre de molécules lors de ces campagnes de criblage. Ces campagnes de criblage, initiées dans l'industrie pharmaceutique dans les années 1980, se sont également développées plus récemment dans le milieu académique à une échelle généralement plus modeste en ce qui concerne le nombre de molécules testées. Lors de ces campagnes, le choix de la cible biologique revêt évidemment une importance primordiale mais il est également nécessaire de disposer de collections de molécules ou « chimiothèques » adaptées.

III.4 Les Chimiothèques dédiées PPI

Une chimiothèque est une librairie de composés chimiques pouvant être composée de quelques centaines à quelques millions de composés. On distingue les **chimiothèques virtuelles,** qui sont constituées de molécules stockées *in silico* sous format électronique (le plus souvent SDF), et **les chimiothèques physiques**, le plus souvent stockées sur plaques. Les collections de molécules proposées par les sociétés pharmaceutiques peuvent être chimiquement diverses (screening collections) ou être dédiées à un espace biologique particulier tel que les kinases, les protéases, les récepteurs nucléaires, les canaux ioniques, les récepteurs couplés aux protéines G (GPCR). Elles peuvent également être dédiées à un type de pathologie donné (anti-cancéreux, antiviraux, anti-microbiens …).

A la suite des récents succès dans la découverte d'inhibiteurs d'interactions protéine-protéine, plusieurs fournisseurs ont développé des chimiothèques dédiées à cet espace biologique. **La société Chemdiv** (http://eu.chemdiv.com/) propose une chimiothèque PPI de 123 000 composés obtenus par la combinaison de différentes approches, à partir de données de la littérature, par la recherche de nouveaux scaffolds, par la synthèse dirigée, la synthèse d'homologues de produits naturels. Ils ont notamment mis l'accent sur la forme des composés en utilisant le facteur sp3 qui rend compte de la tridimensionnalité des molécules (Lovering et al., 2009). Ces composés ont été subdivisés en 11 catégories en fonction des cibles biologiques ou des propriétés physicochimiques (Eccentric, Cyclic Ugi, mdm2 focused, Recognition elements, Hedgehog pathway focused, Nonpeptide Peptidomimetic, PDZ domain focused, Spiro, 3D Mimetic, Helix-Mimetics, Beyond Flatland). **La compagnie Life Chemicals** (http://www.lifechemicals.com/) propose une chimiothèque PPI d'environ 34 000

composés. Ces composés ont été obtenus en collectant dans un premier temps 10 031 molécules montrant une activité inhibitrice dans des essais PPI pour 7 cibles biologiques sur le site PubChem Bioassay (http://www.ncbi.nlm.nih.gov/pcassay). Cette première collection a été enrichie en recherchant des composés similaires dans la chimiothèque « référence » maison (coefficent de similarité Tanimoto de 90%). Au final, 34 041 composés ont été identifiés et incorporés dans la chimiothèque focalisée PPI. **La société Polyphor** (http://polyphor.com/) développe des modulateurs d'interactions protéine-protéine selon deux stratégies complémentaires MacroFinder® et PEMfinder® basées sur la synthèse de macrocycles dont le poids moléculaire varie de 400 à 2 000 Da. Les 2 approches ont été conçues pour cibler les interfaces protéine-protéine mais les molécules synthétisées au travers de la plateforme PEMfinder® ciblent essentiellement les complexes extracellulaires possédant une grande interface, tandis que les composés issus de la plateforme MacroFinder® ont la capacité de pénétrer à l'intérieur de la cellule et sont donc destinées à cibler des complexes intracellulaires. Ces chimiothèques ne sont pas disponibles à la vente.

IV Exemples et succès

De nombreux modulateurs d'interactions protéine-protéine ont été développés au cours de la dernière décennie soit par des laboratoires pharmaceutiques, soit par des laboratoires publics. Comme nous le verrons plus loin (voir Résultats, chapitre I), nous avons basé nos travaux sur l'analyse de données concernant des succès pour lesquels l'information structurale est disponible à la fois sur les cibles protéiques libres en solution mais aussi sur les complexes PPIs (cible-partenaire protéique) et cible-ligand. Nous avons ainsi obtenu une base de données regroupant quatorze complexes protéine-protéine pour lesquelles au moins un inhibiteur est connu. De nombreux inhibiteurs ont été découverts pour chacune de ces 14 familles et nous avons pu prendre un certain recul sur le mode d'action de ces composés au cours des cinq dernières années grâce à un effort spectaculaire de la communauté scientifique. La présentation de tous ces succès pouvant apparaître comme un catalogue qui n'apporterait que peu d'information au lecteur, j'ai choisi de ne présenter ici que 3 de ces succès historiques : un premier complexe protéine-protéine présentant plusieurs zones discontinues à l'interface (IL-2/IL-2Rα) et deux autres exemples bien connus également impliquant des complexes avec une zone continue (type protéine-peptide – Bcl/BAK – P53/MDM2).

IV.1 Le complexe IL-2/IL-2R$_{alpha}$

L'interleukine-2 a été parmi les premières cibles PPIs affectées par des petites molécules (Arkin et al., 2003; Braisted et al., 2003; Hyde et al., 2003; Raimundo et al., 2004; Thanos et al., 2006; Thanos et al., 2003; Waal et al., 2005), et reste l'un des rares exemples d'une petite molécule, imitant un épitope hautement discontinu. La richesse des données structurales recueillies au cours de l'exploration d'inhibiteurs de l'IL-2 a révélé une complexité surprenante de l'interface protéine-protéine. Celle-ci sert souvent désormais de modèle « d'école » permettant de guider le dépistage et la conception d'inhibiteurs.

Découvert il y a 30 ans (Taniguchi et al., 1983) pour son activité de facteur de croissance des lymphocytes T, la cytokine interleukine-2 joue un rôle clef dans le rejet de tissus lors de greffes et a donc un intérêt médical important.

Inactive lorsqu'elle est seule, elle devient effective après formation d'un complexe avec son récepteur : l'hétérotrimère IL-2R. Les trois sous-unités ne sont pas liées covalemment, et comprennent une chaine alpha (IL-2Rα), une chaine beta (IL-2Rβ, que l'on

retrouve aussi chez IL-15R) et la chaine commune des récepteurs de cytokines (IL-2Rγc). L'ensemble de ces trois domaines permet la création d'un complexe de très haute affinité avec IL-2 (K_d = 10^{-11} nM) via un taux d'association rapide (k = 10^7 $M^{-1}.s^{-1}$). La formation de ce complexe se déroule en trois temps. La première étape est la liaison de IL-2 avec la sous unité α du récepteur (Figure 9) ; ce complexe transitoire a une constante d'affinité d'environ 10^{-8} M. Cette première association permet ensuite à IL-2 d'être présentée à la sous unité β du récepteur, et enfin, la chaine commune aux récepteurs de cytokines (IL-2Rγ) vient stabiliser ce complexe quaternaire avec un taux de dissociation très lent (k' = $10^{-4}M^{-1}.s^{-1}$) (Johnson et al., 1994; Malek, 2008; Wang and Smith, 1987).

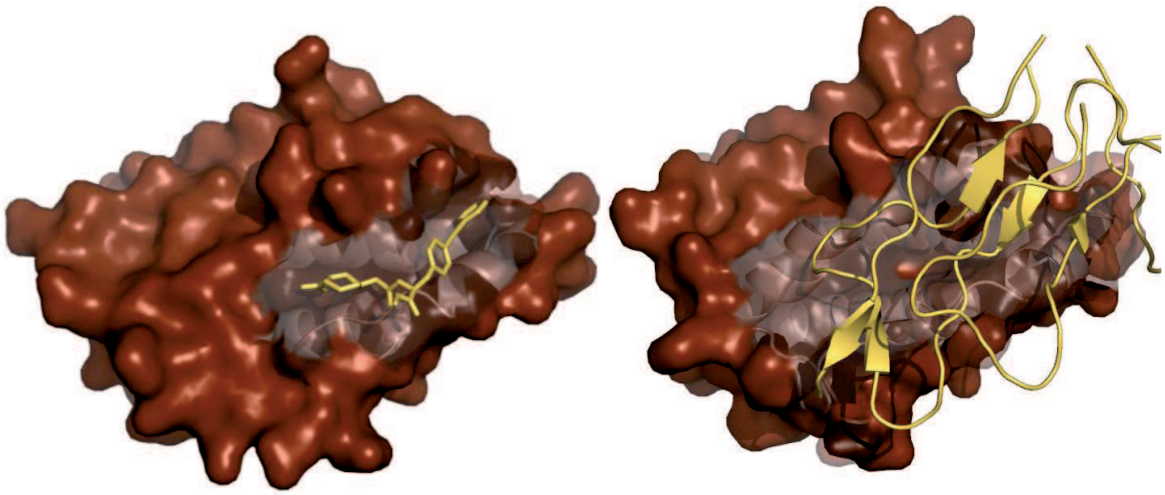


Figure 9 : L'interleukine-2 représentée en surface et en marron ainsi que le ligand 1M47 (à gauche) en jaune qui rentre en compétition avec le récepteur de l'interleukine-2 (en jaune, à droite).

La formation du complexe entre IL-2 et IL-2R engendre une voie de signalisation à travers les sous unités IL-2Rβ et IL-2Rγc grâce à l'association des tyrosines kinases Janus kinase 1 et 3 (JAK1 et JAK3) aux extremités cytoplasmiques des sous unités IL-2Rβ et IL-2Rγc. Cette association mène d'une part à la phosphorylation des tyrosines kinases JAK1 et JAK3 mais aussi à la phosphorylation de trois tyrosines clefs dans la partie cytoplasmique de IL-2Rβ. A partir de ces tyrosines, il existe trois voies de signalisations différentes (Gaffen, 2001; Nelson and Willerford, 1998). Deux de ces trois voies de signalisation, médiées par MAPK (mitogen-activated protein kinase) et PI3K (phosphoinositol 3 kinase), sont activées par la tyrosine la plus proche de la membrane cellulaire. Par ailleurs, la voie de signalisation STAT5 (signal transducer and activator of transcription 5) est activée par son association avec

les deux autres tyrosines d'IL-2Rβ. Ces voies de signalisation sont impliquées dans le cycle cellulaire, la croissance cellulaire et la différenciation (Cheng et al., 2011).

Des anticorps monoclonaux qui reconnaissent IL-2Rα et bloquent la liaison de l'IL-2 ont tout d'abord été mis au point et commercialisés sous le nom de basiliximab et le daclizumab, respectivement (Waldmann, 2003). Le succès de ces anticorps, appuyé par une étude clinique ayant permis de supprimer la réponse immunitaire associée au rejet de greffes d'organes (Lin et al., 2006), valide l'interaction IL-2/IL-2Rα comme cible thérapeutique. La cristallographie et diffraction des rayons X et la spectroscopie RMN ont alors été utilisées pour caractériser structurellement IL-2 (Fry, 2006), (Arkin et al., 2003).

L'entreprise pharmaceutique Roche a alors décidé de se lancer dans l'aventure consistant à découvrir de petits composés non peptidiques qui agiraient comme inhibiteurs de l'interaction en se liant à IL-2Rα. Ils ont préparé pour ce faire une série de dérivés acylphénylalanine conçus pour imiter l'Arg38 et la Phe42 de l'épitope de liaison fixant IL-2. Un composé s'est avéré présenter une valeur IC_{50} de ~ 45 uM. Ce composé 'touche' a été optimisé pour obtenir un composé présentant une valeur IC_{50} de 3 uM. C'est alors qu'ils ont fait la surprenante découverte (par RMN hétéronucléaire) que, bien que les dérivés acylphenyalanine aient été conçus pour lier IL-2Rα, le composé optimisé interagissait en fait avec IL-2 (Fry, 2006). Les structures 3D obtenues par cristallographie aux rayons X du complexe IL-2/IL-2Rα (par la société Sunesis Pharmaceuticals) ont permis de comprendre, à posteriori, l'erreur dans la stratégie de conception qui avait été employée. Elle a montré que l'idée d'essayer d'imiter Arg38 et Phe42 semblait raisonnable, mais que la poche de l'IL-2Rα était assez peu profonde. Elle a également montré que le composé optimisé qui a eu un succès (chanceux) mimait en fait IL-2Rα. L'optimisation du composé présentant un IC_{50} de 3 uM a permis de développer un inhibiteur non peptidique à 60 nM, SP4206 (Fry, 2011; Wilson and Arkin, 2011) (Figure 10). Cette étude de cas IL-2/IL-2Rα a permis de démontrer que le principe de mimétisme des principales interactions des chaînes latérales pourrait se présenter comme une stratégie efficace pour développer des inhibiteurs PPIs.

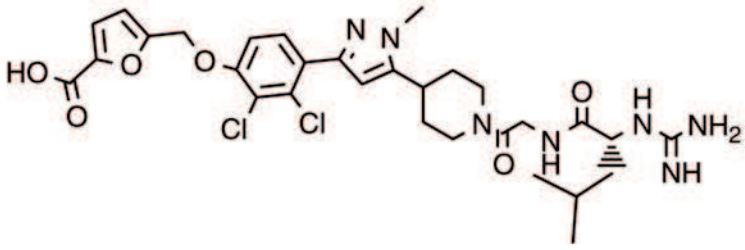
Compound number	Structure	IC_{50} (μM)
7 (SP4206)		0.06

Figure 10 : Représentation schématique du composé SP4206 (tiré de Wilson & Arkin 2011)

IV.2 Le complexe Bcl-X_L/BAK et le Navitoclax

Les membres de la famille Bcl (B-Cell Lymphoma) jouent un rôle important dans la régulation de la mort cellulaire programmée (Adams and Cory, 1998; Llambi and Green, 2011; Sattler et al., 1997). Par exemple, les protéines de la famille Bcl-2 (dont fait parti Bcl-X_L) jouent un rôle anti-apoptotique lorsqu'elles se lient à une portion d'hélice α de 16 résidus de la molécule pro-apoptotique BAK (Sattler et al., 1997), (Figure 11). De la même façon, elles peuvent se lier avec plusieurs protéines proapoptotiques telles que BAD (Bcl-2 Associated Death promoter), BAK (Bcl-2 Antagonist Killer), BAX (Bcl-2 Associated X protein), BID (BH3 Interacting-domain Death agonist), BOD (Bcl-2-related Ovarian Death gene) etc… (Petros et al., 2000).

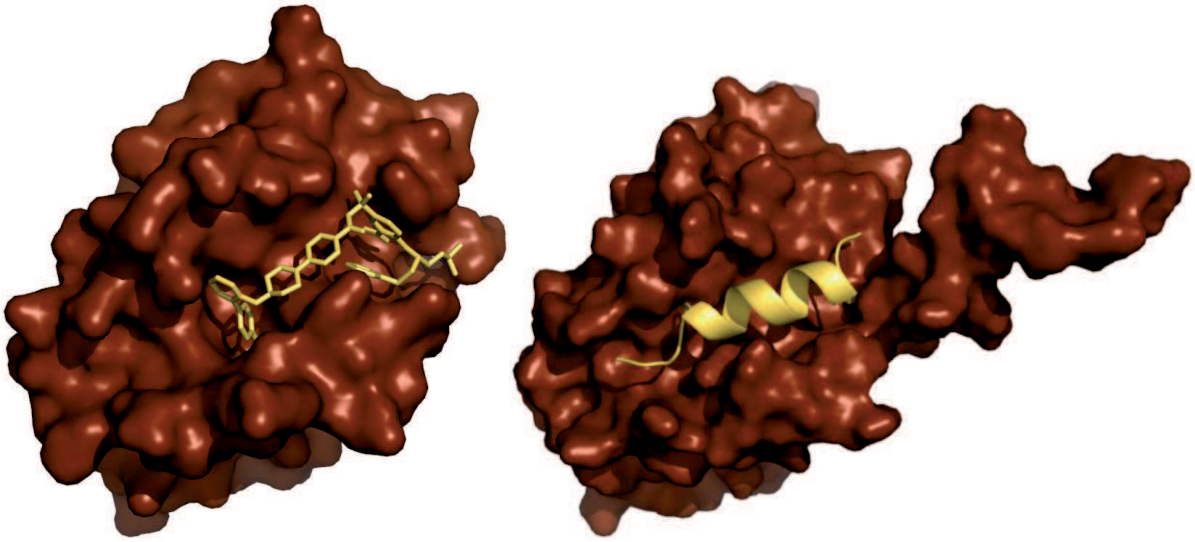


Figure 11 : Représentation du complexe entre Bcl-X_L (en marron) et le ligand N3C (en jaune à gauche) qui entre en compétition avec le peptide BAK (en jaune à droite).

L'importance de ces protéines comme cibles privilégiées dans la recherche contre le cancer a motivé la découverte de nombreuses molécules mimant les hélices α de ces protéines afin d'empêcher la formation des complexes (Bcl-2 ou Bcl-X_L avec un partenaire proapoptotique). Certaines de ces molécules atteignent même de hautes affinités avec Bcl-2 et Bcl-X_L (K_i entre 5 et 100 nM) (Sadowsky et al., 2007; Walensky et al., 2004; Yin et al., 2005). Récemment, les laboratoires Abbott ont découvert des composés organiques de très haute affinité qui se lient aussi bien à Bcl-2 qu'à Bcl-X_L, mais aussi Bcl-W, une autre protéine anti-apoptotique (Oltersdorf et al., 2005). La molécule la plus efficace parmi celles-ci, ABT-737, a un Ki de 0.6 nM et une masse moléculaire de 813 Da. Cette affinité est du même ordre de grandeur que l'affinité entre les deux partenaires protéiques, mais, du fait de la petite taille d'ABT-737 par rapport à l'hélice de BAK et BAD, son efficacité moléculaire est environ deux fois supérieure à celle de BAK et BAD (l'efficacité moléculaire ou « Ligand efficiency » représente la constante d'inhibition IC_{50} ou pIC_{50} rapportée à la surface moléculaire ou à la masse moléculaire de la molécule ; c'est un très bon indice de capacité ou efficacité moléculaire ; nous reviendrons en détail sur sa définition dans la partie III des résultats). La famille des composés inhibant ces interactions a été identifiée par « fragment based drug design » couplé à la résonance magnétique nucléaire (RMN), et ces composés ont ensuite été améliorés par chimie médicinale (Llambi and Green, 2011).

Des analyses de mutagénèse dirigées sur le peptide correspondant à la zone d'interaction de BAK ont permis d'identifier différents résidus jouant un rôle clef dans la formation du complexe avec Bcl- X_L : Val 74, Leu 78, Ile 81, Asp 83 et Ile 85. Or, même si le composé ABT-737 ne mime pas exactement les détails atomiques du peptide de BAK, il interagit directement avec ces résidus clefs, empêchant ainsi la formation du complexe. De plus ABT-737 piège Bcl- X_L dans une conformation différente de la forme apo, avec des poches plus profondes (Wells and McClendon, 2007).

L'optimisation d'ABT-737 a conduit à ABT-263 (Figure 12), plus généralement appelé Navitoclax, une petite molécule qui a une affinité avec Bcl-2 proche de celle d'ABT-737 mais qui a des propriétés pharmacocinétiques permettant une prise par voie orale. Cette molécule est en cours de test pour la phase 1 des essais cliniques en combinaison avec d'autres traitements, et en phase 2 en tant qu'agent seul.

Obatoclax (GX15-070, Figure 12) est une autre petite molécule découverte par la société Gemin X (et acquise ensuite par Céphalon). Cette molécule est active dans le

traitement des lymphomes, myelofibrose, mastocytose et des leucémies (elle est toujours en phase II en combinaison avec le sorafenib). Obatoclax a été conçue pour occuper une poche hydrophobe à l'intérieur de la poche de liaison de BH3 à Bcl-2. Cette molécule interfère spécifiquement dans les membranes mitochondriales externes qui ont été extraites de la cellule, et dans les membranes mitochondriales externes qui étaient présentes dans la cellule. Mcl-1 (un membre anti-apoptotoque de la famille Bcl-2) a été identifiée pour conférer une résistance à ABT-737 et au bortézomib (un inhibiteur du protéosome). Il a cependant été montré qu'Obatoclax pouvait surmonter cette résistance (Nguyen et al., 2007).

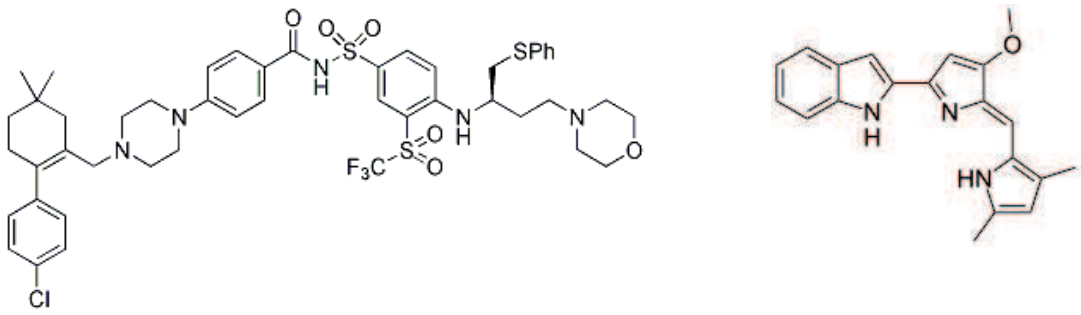


Figure 12 : Représentation schématique du composé Abt-263 (gauche) et GX15-070 (droite) les deux molécules actuellement en essai clinique de phase II.

IV.3 Le complexe HDM2/p53 et la Nutline-3

L'oncogène humain 'double mutant' code pour la protéine HDM2, qui est un régulateur négatif du facteur de transcription p53. Le facteur de transcription p53 est impliqué dans la régulation du cycle cellulaire, et agit donc en tant que suppresseur de tumeur. Son implication dans de nombreux cancers en fait un modèle très étudié. La surexpression de HDM2 a été observée dans de nombreuses tumeurs humaines, supprimant alors l'effet suppresseur de tumeur de p53 (Figure 13). L'inhibition de l'interaction p53-HDM2 peut rétablir l'effet de p53, et en tant que telle est donc une cible attrayante en cancérologie (Vassilev et al., 2004).

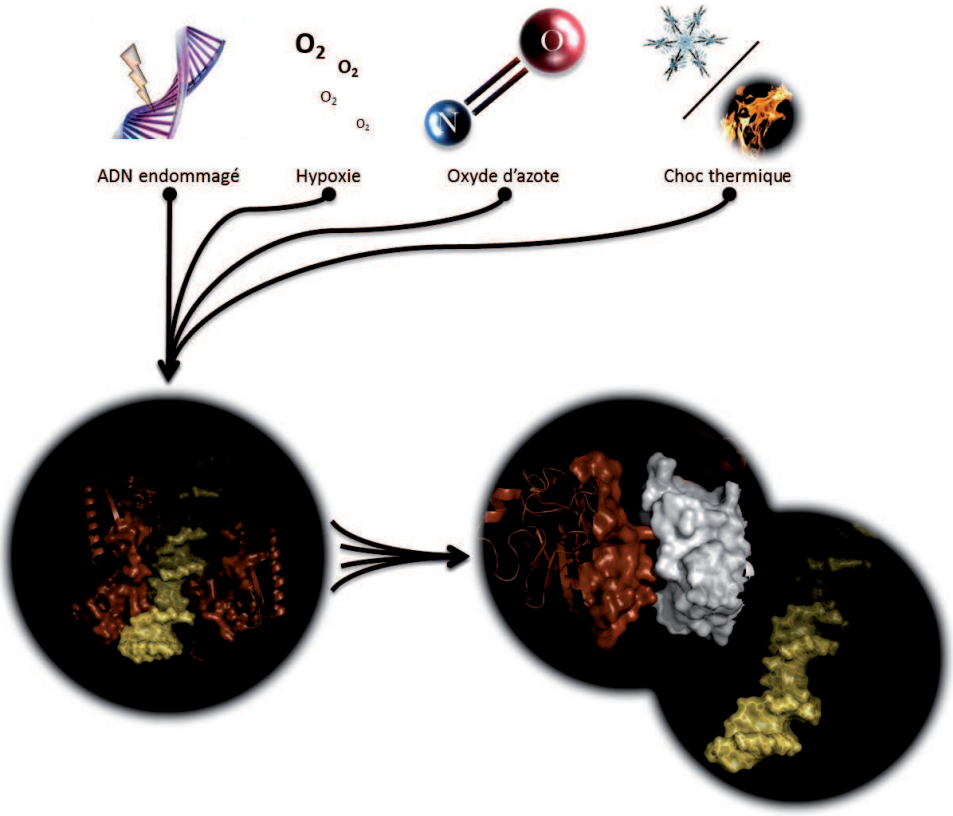


Figure 13 : L'activation, par différents facteurs, de la voie de signalisation p53 aboutit à l'interaction d'un homodimère de la protéine p53 (marron) avec l'ADN (jaune). Cet évènement est impliqué dans l'arrêt du cycle cellulaire, le vieillissement cellulaire, ou encore l'apoptose. La protéine HDM2 (blanc), quant à elle, régule cette interaction en entrant en concurrence avec l'ADN. Empêcher l'interaction entre HDM2 et p53 permettrait ainsi d'accélérer l'apoptose lors de cancers.

Dans cette voie de signalisation, la protéine HDM2 est devenue une cible pour la recherche contre le cancer, puisqu'elle réprime la transcription de p53, et donc empêche ainsi l'activation de la mort cellulaire. De nombreuses études portant sur des petites molécules se liant à HDM2 ont été amorcées. Une de ces études a été menée par F. Hoffmann-La Roche (New Jersey), et a permis de trouver une série de composés de type 'imidazoles tetra-substitués', qui ont été nommés nutlines. Après optimisation, il a été observé que la molécule la plus efficace de la série, la nutline-3 (Figure 14), pouvait rompre le complexe HDM2-p53 avec une IC_{50} de 90 nM. La nutline-3 a de plus montré une forte activation de l'activité de p53 *in vitro*, ainsi qu'une activité contre des tumeurs xéno-greffées *in vivo* (Vassilev et al., 2004). Une autre étude, menée chez Johnson & Johnson a permis de mettre en évidence une autre série de petites molécules (dérivées de benzodiazepinediones) (Grasberger et al., 2005) pouvant se lier à HDM2 et inhiber le complexe HDM2/p53. Après optimisation, ils ont

montré qu'une de ces molécules se lie à HDM2 avec un K_D de 67 nM (Parks et al., 2005). De plus, une étude a montré que l'une de ces benzodiazepinediones inhibait la prolifération de cellules tumorales *in vitro* avec un IC_{50} d'environ 10 µM, et a mis en évidence une synergie avec le médicament Doxorubicin sur des tumeurs chez la souris (Koblish et al., 2006).

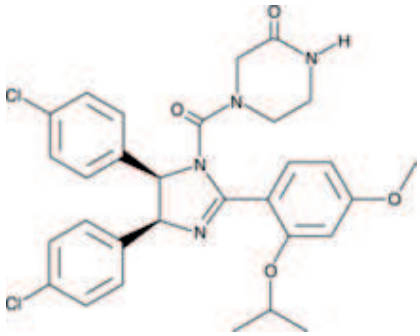


Figure 14 : Représentation schématique du composé nutline 3.

Différentes structures de la protéine HDM2 en complexe avec certaines de ces molécules ont été caractérisées, soit par Résonance Magnétique Nucléaire, soit par cristallographie aux rayons X (Fry et al., 2004; Grasberger et al., 2005; Vassilev et al., 2004). L'étude de ces structures permet de mettre en évidence que les molécules de chez Hoffmann-La Roche et celles de chez Johnson & Johnson se lient à la protéine dans la même région que l'hélice α de p53 qui interagit avec HDM2. Ils ont de plus tous un groupe aromatique ou aliphatique qui s'insère dans une poche à la surface de HDM2 et se lient aux trois résidus clefs de l'interaction avec p53 : Phe 19, Trp 23 et Leu 26.

De la même façon que pour les complexes IL-2/IL-2R et Bcl-X_L/BAK, les modulateurs du complexe HDM2/p53 piègent la zone d'interaction d'HDM2 avec p53 dans une conformation comprenant une poche plus grande que celles présentes lorsqu'elle est en complexe avec son partenaire biologique, montrant ainsi l'importance de la détection des poches dans les formes libres pour définir la druggabilité.

Dans le chapitre suivant, je présenter dans un premier temps les résultats obtenus sur l'analyse structurale des paramètres physicochimiques des interactions protéine-protéine. Je présente aussi les résultats obtenus sur les différences structurales entre les complexes protéiques pour lesquels un inhibiteur orthostérique existe déjà et l'ensemble des complexes protéiques dont la structure est caractérisée. Ces résultats ont abouti à la création de la base de données $2P2I_{DB}$.

Dans un second temps, j'expose mes travaux concernant la définition de l'espace chimique caractéristique des inhibiteurs d'interactions protéine-protéine. Je présente également le protocole informatique d'apprentissage que nous avons employé pour la conception de chimiothèques dédiées aux interactions protéine-protéine : $2P2I_{CHEM}$.

Enfin, je présente les indexes utilisés pour mesurer l'efficacité des ligands (LEI, BEI, SEI) ; je compare notamment la position des petites molécules présentes dans $2P2I_{DB}$ dans l'espace constitué par ces indexes avec les médicaments disponibles sur le marché.

RESULTATS

I La base de données structurale 2P2I$_{DB}$

Comme nous l'avons vu précédemment, les interfaces protéine-protéine sont des cibles thérapeutiques potentielles prometteuses ; d'autre part, la modulation de telles interactions est actuellement réalisable. Il n'en reste pas moins que la découverte d'un modulateur reste une expérience difficile à mener et coûteuse à réaliser.

Nous avons voulu au cours de ma thèse, caractériser dans un premier temps **les paramètres structuraux** des interfaces protéiques pour lesquelles des inhibiteurs (modulateurs) ont été identifiés. Par la suite nous nous sommes servis de ces résultats pour **caractériser l'espace chimique** des interactions protéine-protéine pour lesquelles un inhibiteur est connu. La caractérisation de cet espace chimique permettrait alors de sélectionner des cibles dans le même espace chimique pour lesquelles le développement d'inhibiteurs devrait être facilité.

- Construction de la base de données :

Le but de notre approche étant la caractérisation des paramètres structuraux qui orchestrent les interactions protéine-protéine, nous nous sommes intéressés uniquement aux familles de protéines pour lesquelles une structure tridimensionnelle à la fois du complexe protéine-protéine, mais aussi du complexe protéine-inhibiteur était disponible. L'ensemble de ces contraintes a permis, d'une part par une recherche dans la littérature, et d'autre part par une exploration exhaustive de la PDB, de créer un ensemble de 17 familles protéiques, et 56 petites molécules. L'existence de ces structures valide que les inhibiteurs présents dans la base de données soient tous orthostériques.

La base de données ainsi créée est constituée, pour chaque famille protéique, des structures tridimensionnelles du complexe protéine-protéine, d'un complexe protéine-ligand, pour chaque ligand trouvé se liant au niveau de la zone d'interaction du complexe protéique, et enfin, lorsqu'elles sont disponibles, des formes libres des protéines. L'ensemble des données présentes dans la base de données sont disponibles sur le site 2P2I$_{DB}$ (http://2p2idb.cnrs-mrs.fr/).

Ce travail a fait l'objet de ma première publication, publiée dans le journal PLoS ONE.

I.1 Article 1

Atomic Analysis of Protein-Protein Interfaces with Known Inhibitors: The inhibiteur d'interactions protéine-protéine Database.

Raphaël Bourgeas, Marie-Jeanne Basse, Xavier Morelli, Philippe Roche.

PLoS One. 2010 Mar 9;5(3):e9598.

Background: In the last decade, the inhibition of protein-protein interactions (PPIs) has emerged from both academic and private research as a new way to modulate the activity of proteins. Inhibitors of these original interactions are certainly the next generation of highly innovative drugs that will reach the market in the next decade. However, in silico design of such compounds still remains challenging.

Methodology/Principal Findings: Here we describe this particular PPI chemical space through the presentation of 2P2IDB, a hand-curated database dedicated to the structure of PPIs with known inhibitors. We have analyzed protein/protein and protein/inhibitor interfaces in terms of geometrical parameters, atom and residue properties, buried accessible surface area and other biophysical parameters. The interfaces found in 2P2IDB were then compared to those of representative datasets of heterodimeric complexes. We propose a new classification of PPIs with known inhibitors into two classes depending on the number of segments present at the interface and corresponding to either a single secondary structure element or to a more globular interacting domain. 2P2IDB complexes share global shape properties with standard transient heterodimer complexes, but their accessible surface areas are significantly smaller. No major conformational changes are seen between the different states of the proteins. The interfaces are more hydrophobic than general PPI's interfaces, with less charged residues and more non-polar atoms. Finally, fifty percent of the complexes in the 2P2IDB dataset possess more hydrogen bonds than typical protein-protein complexes. Potential areas of study for the future are proposed, which include a new classification system consisting of specific families and the identification of PPI targets with high druggability potential based on key descriptors of the interaction.

Conclusions: 2P2I database stores structural information about PPIs with known inhibitors and provides a useful tool for biologists to assess the potential druggability of their interfaces. The database can be accessed at http://2p2idb.cnrs-mrs.fr.

Ma contribution dans ce travail a consisté en premier lieu en la collecte des informations structurales nécessaires dans la PDB. J'ai ensuite utilisé le serveur PROTORP pour calculer les paramètres structuraux intéressants. J'ai analysés ces paramètres, et j'ai, par analyse statistique (t-test puis analyse en composante principale), mis en évidence les différents clusters d'interactions protéine-protéine présentés à la fin de la publication. J'ai de plus participé à la création du site internet en l'écriture des scripts Jmol qui permettent d'avoir des vues pré-calculées des complexes protéiques. Enfin, j'ai écrit le « materials & methods » et ai réalisé l'ensemble des tables et des figures excepté la première.

OPEN ACCESS Freely available online

PLoS one

Atomic Analysis of Protein-Protein Interfaces with Known Inhibitors: The 2P2I Database

Raphaël Bourgeas, Marie-Jeanne Basse, Xavier Morelli*, Philippe Roche*

Laboratoire Interactions et Modulateurs de Réponses (UPR3243), Centre National de la Recherche Scientifique (CNRS) & Aix-Marseille Universités, Institut de Microbiologie de la Méditerranée (IMM), Marseille, France

Abstract

Background: In the last decade, the inhibition of protein-protein interactions (PPIs) has emerged from both academic and private research as a new way to modulate the activity of proteins. Inhibitors of these original interactions are certainly the next generation of highly innovative drugs that will reach the market in the next decade. However, *in silico* design of such compounds still remains challenging.

Methodology/Principal Findings: Here we describe this particular PPI chemical space through the presentation of $2P2I_{DB}$, a hand-curated database dedicated to the structure of PPIs with known inhibitors. We have analyzed protein/protein and protein/inhibitor interfaces in terms of geometrical parameters, atom and residue properties, buried accessible surface area and other biophysical parameters. The interfaces found in $2P2I_{DB}$ were then compared to those of representative datasets of heterodimeric complexes. We propose a new classification of PPIs with known inhibitors into two classes depending on the number of segments present at the interface and corresponding to either a single secondary structure element or to a more globular interacting domain. $2P2I_{DB}$ complexes share global shape properties with standard transient heterodimer complexes, but their accessible surface areas are significantly smaller. No major conformational changes are seen between the different states of the proteins. The interfaces are more hydrophobic than general PPI's interfaces, with less charged residues and more non-polar atoms. Finally, fifty percent of the complexes in the $2P2I_{DB}$ dataset possess more hydrogen bonds than typical protein-protein complexes. Potential areas of study for the future are proposed, which include a new classification system consisting of specific families and the identification of PPI targets with high druggability potential based on key descriptors of the interaction.

Conclusions: 2P2I database stores structural information about PPIs with known inhibitors and provides a useful tool for biologists to assess the potential druggability of their interfaces. The database can be accessed at http://2p2idb.cnrs-mrs.fr.

Citation: Bourgeas R, Basse M-J, Morelli X, Roche P (2010) Atomic Analysis of Protein-Protein Interfaces with Known Inhibitors: The 2P2I Database. PLoS ONE 5(3): e9598. doi:10.1371/journal.pone.0009598

Editor: Narcis Fernandez-Fuentes, Leeds Institute of Molecular Medicine, United Kingdom

Received December 12, 2009; **Accepted** February 15, 2010; **Published** March 9, 2010

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: morelli@ifr88.cnrs-mrs.fr (XM); proche@ifr88.cnrs-mrs.fr (PR)

These authors contributed equally to this work.

Introduction

In the last decade, the inhibition of protein-protein interactions (PPIs) has emerged from both academic and private research as a new way to modulate the activity of proteins (for an in depth review see Roche and Morelli [1]). Based on this new focus, it is now more and more commonly accepted that protein-protein complexes are an important class of therapeutic targets [2]. PPIs can be involved in a network of complex interactions that play a central role in various cellular events. These interactions control processes involved in both normal and pathological pathways, which include signal transduction, cell adhesion, cellular proliferation, growth, differentiation, viral self-assembly, programmed cell death and cytoskeleton structure (for a review refer to [3]).

In parallel to this new field, large scale genomics and proteomics programs have permitted the identification of entire protein networks interactomes at the cellular level. These programs have led to major breakthroughs in understanding biological pathways, host-pathogen interactions and cancer development. With the growing tools of small molecules, the modulation of these networks of interactions represents a promising therapeutic strategy. Protein-protein interaction inhibitors (2P2Is) are certainly the next generation of highly innovative drugs that will reach the market in the next decade.

As a consequence of this enthusiasm, the exponential increase of published biomedical literature on PPIs and their inhibition has prompted the development of internet services and databases that help scientists to manage the available information. There is now a growing number of structural databases dedicated to protein-protein interactions [4–7]. A large variety of these PPIs databases depict protein-protein interactions at a structural level (for a summary of these available databases refer to [1]), but they focus only on this particular interface without taking into account the potential inhibitors related to one of the two partners. In a recent survey, Higueruelo *et al.* analyzed the atomic interactions and profile of small molecules disrupting PPIs in the TIMBAL database,

focusing on small molecules properties and comparing these results to drug-like databases [4]. Several other studies have also focused on subsets of small molecules that disrupt PPIs [5,6,7,8]. However, none of them have focused on both the protein-protein structural information available and the known inhibitors within the interface.

We describe here a chemical space, $2P2I_{DB}$, which is a hand-curated database dedicated to the structure of Protein-Protein complexes with known inhibitors thereby offering complementary information to these previous analyses ($2P2I_{DB}$ is available at http://2p2idb.cnrs-mrs.fr). We have analyzed the protein/protein and protein/inhibitor interfaces in terms of geometrical parameters, atom and residue properties, buried accessible surface area and other biophysical parameters, such as the protein-protein dissociation constant (Kd) of a complex. The interfaces found in $2P2I_{DB}$ were then compared to those of representative datasets of heterodimeric complexes from Bahadur and Zacharias [9] or from the ProtorP parameters (http://www.bioinformatics.sussex.ac.uk/protorp/ and [10].

The architecture present at the interface generally involves a globular interacting domain, a single secondary structure element (alpha-helix or beta strand) of a globular protein, or a short peptide. Complexes in $2P2I_{DB}$ present globally the same shape (planarity or eccentricity) than standard heterodimeric complexes, but their accessible surface areas are significantly smaller. More strikingly, no major conformational changes are observed between the different states of the proteins (bound to the biological partner, the equivalent free form and the form bound to the small molecule inhibitor). The interfaces are also more hydrophobic than general PPIs' interfaces, with less charged residues and more non-polar atoms. Moreover, fifty percent of the complexes in the $2P2I_{DB}$ dataset possess more hydrogen bonds than typical protein-protein complexes. A set of key descriptors were identified to distinguish between PPIs with known inhibitors and representative transient complexes in the protein databank. Transient protein-protein complexes are defined as protomers that, *in vivo*, can exist either on their own or in complex and also undergo an exchange between the free and complexed form [11].

A new classification based on these parameters is proposed with potential aims for the future to identify potential new druggable PPI targets.

Results and Discussion

Dataset Collection

As our goal was to define structural parameters that guide the development of PPI disruptors, we only considered those protein families for which a high resolution three dimensional structure was available for both the protein/protein and the protein/inhibitor complexes. Homodimers and covalently bound inhibitors were not taken into account due to their different behavior. When available, the best resolution structure of the unbound form of the proteins or a close homologue was included. The dataset was built through data mining from the literature and by exhaustive search of the Protein Data Bank (Figure 1 and Material & methods). The final dataset was compiled into a relational database ($2P2I_{DB}$) that was used to further analyze the general properties of protein/protein interfaces (PPIs) with a known inhibitor. The $2P2I_{DB}$ (http://2p2idb.cnrs-mrs.fr) contains a total of 17 protein/protein complexes corresponding to 14 families and 56 small molecule inhibitors bound to the corresponding target (Table 1 and figure 2).

There are a limited number of targets in the 2P2I database at this stage due to the structural prerequisites that were used. However, it is inevitable that high throughput structural genomic programs will generate a high level of data. In addition, the development of improved methodologies for the development of small molecule inhibitors will rapidly lead to the discovery and structural characterization of disruptors of new PPI families. These new targets and their corresponding ligands will be incorporated into the database as they appear in the literature and the Protein Data Bank (http://www.rcsb.org/).

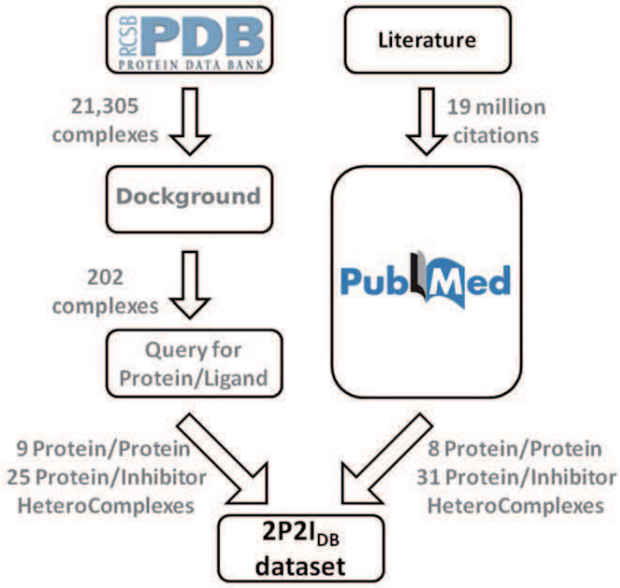


Figure 1. Flow chart indicating how the $2P2I_{DB}$ dataset was build from data mining. Two separate approaches were used to retrieve protein/protein complexes with known inhibitors for which structural information was available. The protein databank was search through the Dockground server [23] which led to 202 complexes that were filtered using an advanced query and manual inspection of the interface to give 9 protein/protein complexes and 25 protein/inhibitor complexes. Exhaustive search of the literature led to the discovery of 8 protein complexes corresponding to 31 protein/ligand complexes.
doi:10.1371/journal.pone.0009598.g001

To assess the characteristics of druggable PPIs, the general properties of the interfaces found in $2P2I_{DB}$ were compared to those of representative datasets of heterodimeric complexes retrieved from Bahadur and Zacharias [9] and from the ProtorP server [10].

Global and Local Rearrangements

PPI with known inhibitors do not undergo large conformational changes. The formation of heterodimeric complexes can lead to large rearrangements of the two protein partners [12]. To assess this point, we measured the root mean square deviation (rmsd) between the bound partners, the equivalent free forms and the form bound to a small molecule inhibitor for each complex family (Table 2 and Table S1). Strikingly, complexes stored in $2P2I_{DB}$ only underwent minor local adaptation during complex formation. The average rmsd (1.12 ± 0.4 Å) was not significantly different than the natural conformational dynamics of the free target protein and was in the same range as the resolution of the crystal structures. Some local rearrangements could be observed at the binding site; however, these rearrangements do not impair the possibility to design potent inhibitors with high affinity. The fact that there is no main rearrangement between the different forms in the $2P2I_{DB}$ dataset could mean that these types of complexes are easier to target. Other strategies with small molecule inhibitors binding at different

Table 1. Complex families in $2P2I_{DB}$.

Class	#	Family	Complex[a]	Number of Inhibitors[b]	Source[c]	Affinity[d] (nM)	Ref
I	1	$BclX_L$/Bak	1bxl	8	PubMed	340	[27]
I	2	MDM2/p53	1ycr 1ycq	3	PubMed	600	[28]
I	3	XIAP BIR3/CASPASE 9	1nw9	2	PubMed	20	[29]
I	4	XIAP BIR3/SMAC	1g73	5	PubMed	420	[29]
I	5	ZipA/FtsZ	1f47	4	PubMed	20,000	[30]
II	6	Chagasin/Papain	3e1z	1	PDB	0.036[e]	[31]
II	7	E2/E1	1tue	1	PubMed	na	[32]
II	8	FKBP12/TGFR	1b6c	17	PDB	na	[33]
II	9	IL-2/IL-2R	1z92	8	PubMed	10	[34]
II	10	MMP1/TIMP1	2j0t	1	PDB	0.40[e]	[35]
II	11	MMP3/TIMP1	1oo9	1	PDB	0.22[e]	[36]
II	12	Subtilisin/Eglin C	1cse 1r0r 1to2	1	PDB	0.029	[37–39]
II	13	Thrombin/Protein C inhibitor	3b9f	1	PDB	na	[40]
II	14	Trypsin/Trypsin inhibitor	2uuy	3	PDB	0.02	[41]

PPIs were subdivided into class I that correspond to protein/peptide interactions with less than six segments at the interface (families 1–5) and class II that represent more globular interacting domains with more segments (families 6–14).
[a]PDB code of protein/protein complexes.
[b]Number of inhibitors present in the database for a given protein/protein complex.
[c]Structures were retrieved through exhaustive search of the protein databank (PDB) or literature data mining (PubMed).
[d]Dissociation constant (K_D) of the protein/protein complexes are indicated in nanomolar when available.
[e]indicates Ki values.
doi:10.1371/journal.pone.0009598.t001

sites, such as an allosteric pocket away from the interface, would be necessary to disrupt PPIs with large conformational rearrangements. These classes of inhibitors are not present in the $2P2I_{DB}$, as we only kept those ligands that are present at the interface.

General Properties of the Interacting Partners

PPI with known inhibitors can be divided into two classes. We analyzed the general characteristics of the protein/protein interfaces (PPIs) using online servers [10] or local visualization programs. The 17 complexes could be divided into two classes according to the number of continuous segments at the interface (Tables 1 & 2). Class I contained six of the 17 PPIs (Table 1, families 1 to 5) and used a limited number of continuous segments for binding to the partner (average value 3.3 ± 1.4). Interestingly, small peptides are able to mimic the interacting partner for this class of complexes (Table 1, families 1 to 5 and references therein). The remaining eleven PPIs used a high number (average value of 8.4 ± 0.9) and corresponded to actual globular interacting domains (complex families 6–13 in Table 1).

A more detailed analysis revealed that complexes from class I contained a higher proportion of secondary structure elements at the interface. Four out of 6 complexes involved mainly an alpha helix at the interface, and the other two a beta strand (see supplementary material, Table S2). Class I PPIs involved a well ordered partner, which might be easier to mimic with small molecules. This later observation could account for the greater number of inhibitors developed for this type of interface [4]. Alternatively, PPIs from class II contained a higher proportion of nonstructured elements probably due to their larger size. However, it is noteworthy that the difference in nature of these two sets of PPIs does not seem to affect the size of the small molecule inhibitors because similar molecular weight ranges and averages were observed in our dataset for the two classes (data not shown).

When available, dissociation (K_D) or inhibitory (K_i) constants of the protein/protein complexes were compared (Table 1). On average, class I complexes corresponded to low affinity complexes in the micromolar range, whereas class II complexes revealed a higher affinity in the nano or sub-nanomolar range.

Subsequent analyses were performed on the two classes of complexes including analysis of the protein/protein and protein/inhibitor interfaces in terms of geometrical parameters, atom and residue properties, and buried surface area at the interface.

Geometry of the Interfaces

PPI with known inhibitors are smaller than standard heterodimers. The size of the interface was computed for each PPI by measuring the buried surface area between the protein/protein complexes and the unbound proteins. The average interface area of 685.2 ± 200 Å^2 (ranging linearly from 241 to 947 Å^2) was significantly smaller than the standard average values of approximately 1000 Å^2 observed for heterodimeric protein-protein complexes, as described in the literature by Bahadur and Zacharias [9] and on the ProtorP server [10]. The average values of 532 ± 198 Å^2 and 769 ± 150 Å^2 were observed for class I (families 1–5 in Table 1) and class II (family 6–13 in Table 1) complexes respectively. The average area of the interface is smaller when the interaction involves a short peptide segment. Moreover, this analysis also illustrate that interfaces with a known PPI inhibitor are slightly smaller than overall protein-protein complexes. However, the 17 interfaces analyzed cover a wide linear range in terms of size, which indicates that the size of the interface does not thwart the definition of a potential target.

PPI with known inhibitors share geometrical properties with heterodimers. The average planarity of interfaces in our dataset was 2.8 ± 0.4 Å, which is a value equivalent to that of overall heterodimeric complexes (2.7 ± 1.2 Å). Similarly, eccentricity (*i.e.*, the

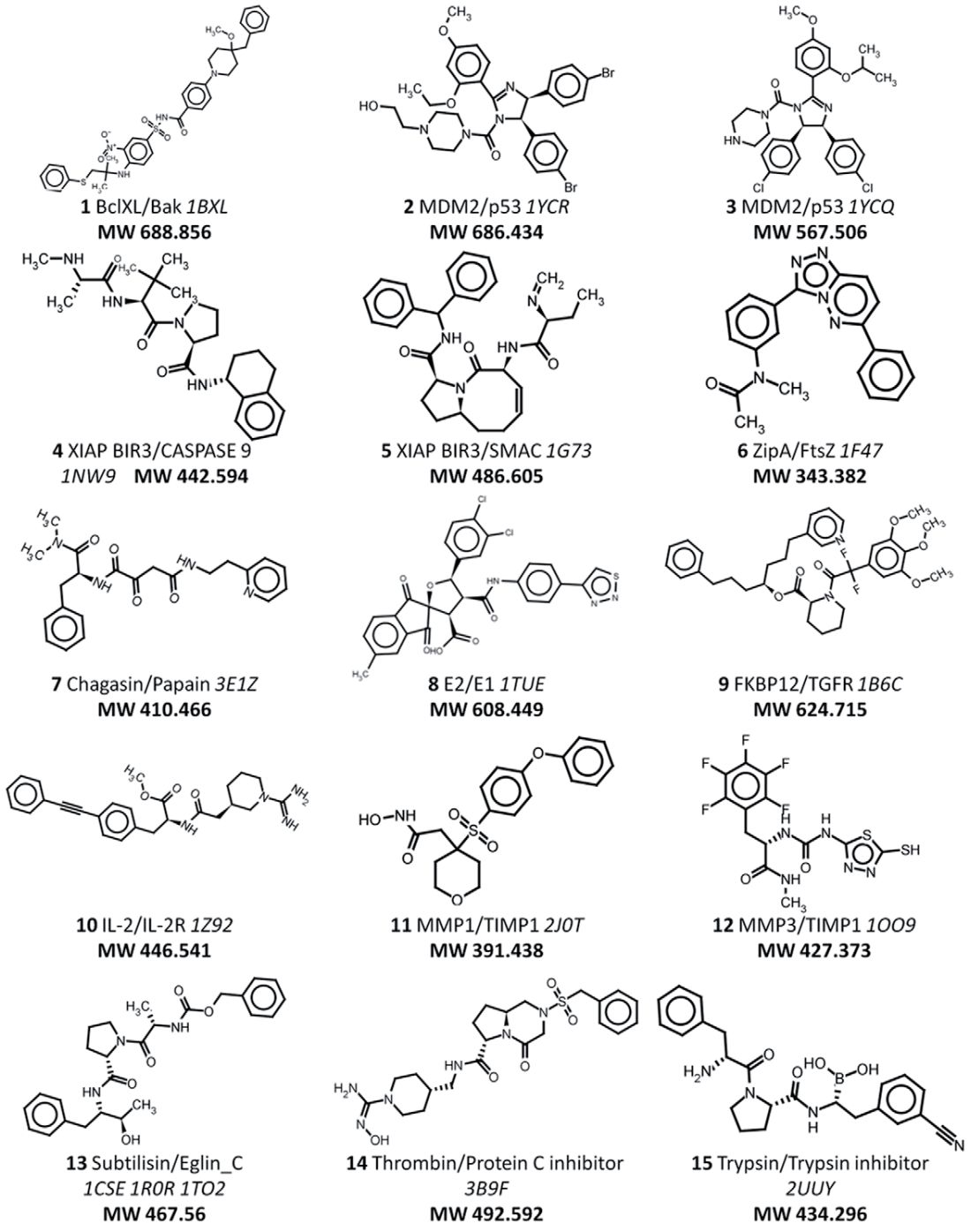


Figure 2. List of representative small molecule inhibitors for each protein/protein complex in $2P2I_{DB}$. Only the inhibitor used to define the subset of the interface at 4.5 Å around the ligand is shown. For each inhibitor, the name of the protein family, the PDB code of the complex and the molecular weight of the ligand are indicated.
doi:10.1371/journal.pone.0009598.g002

ratio of the lengths of the principal axes of the least-squares plane through the atoms in the interface) was not significantly distinguishable (0.72±0.12 vs. 0.70±0.12). Similar values were observed for the two classes defined above. The Gap volume index (GVi) provides a measure of the tightness of a protein-protein complex [13]. The average GVi values for our dataset and general protein-protein complexes are 2.8±1.1 and 2.8±1.4, respectively. On average, a tighter fit was observed for class I complexes (2.3±1.7), which could be due to their smaller size. However, a large variability was observed between complexes. The surface complementarity between the two partners was not correlated with the binding affinity, which could be accounted for by the entopic and desolvation terms of the binding energy. The chemical nature of the interface plays a more important role in defining the strength of the interaction.

Table 2. General interface parameters of PPIs in the $2P2I_{DB}$ dataset.

	PDB	Plan (Å)[a]	Ecc[b]	SecS[c]	GV (Å^3)[d]	GV_I (Å)[e]	Hb[f]	SB[g]	ASA (Å^2)[h]	H_ASA (%)[i]	RMSD (Å)[j]	Pockets (Å^3)[l]	Seg[m]
BclXL/Bak	1bxl	3,60	0,75	H/H	2892	1,75	0,05	0	825	45,6	2,3	97	5
MDM2/p53	1ycr	2,99	0,86	H/H	786	0,60	0,34	1	660	54,2	1,9	351	3
MDM2/p53	1ycq	2,14	0,62	H/H	1286	1,38	0,39	1	455	65,7	1,8	215	2
XIAP BIR3/CASPASE 9	1nw9	2,32	0,78	S/S	3567	1,79	0,78	0	241	100,0	2,4	244	2
XIAP BIR3/SMAC	1g73	2,18	0,73	S/S	3500	5,35	0,88	0	472	88,9	2,3	140	3
ZipA/FtsZ	1f47	2,76	0,75	S/H	3503	3,24	0,09	0	541	39,2	0,9	0	5
mean		***2,66***	***0,75***	-	***2589***	***2,35***	***0,42***	***0,33***	***532***	***65,6***	***1,9***	***174,5***	***3.3***
standard dev.		***0,57***	***0,08***	-	***1238***	***1,70***	***0,34***	***0,52***	***198***	***24,3***	***0,6***	***122,7***	***1.4***
Chagasin/papain	3e1z	3,00	0,89	C/C	4286	2,26	0,53	0	947	55,4	0,4	279	9
E2/E1	1tue	2,59	0,71	H/H	5042	2,86	0,55	3	946	32,6	1,4	202	7
FKBP12/TGFR	1b6c	2,82	0,42	S/H	5457	3,14	0,17	0	869	57,2	0,5	387	8
IL-2/IL-2R	1z92	2,40	0,86	H/C	4431	2,47	0,7	5	898	39,8	1,3	146	8
MMP1/TIMP1	2j0t	2,76	0,55	C/C	5380	4,08	0,83	0	660	48,8	0,5[k]	323	7
MMP3/TIMP1	1oo9	3,02	0,78	C/C	5157	2,76	0,24	0	936	39,7	1,2[k]	227	9
Subtilisin/Eglin C	1cse	2,67	0,65	C/S	3858	3,01	0,74	0	640	62,6	0,3	282	8
Subtilisin/Eglin C	1r0r	2,54	0,75	C/C	3763	2,99	0,78	0	630	66,3	0,3	230	9
Subtilisin/Eglin C	1to2	3,11	0,61	C/S	3277	2,25	0,74	1	728	66,4	0,3	275	8
Thrombin/Protein C inhibitor	3b9f	3,41	0,81	C/C	5538	4,33	0,72	2	639	38,0	0,6	350	9
Trypsin/trypsin inhibitor	2uuy	2,57	0,73	C/C	3500	3,12	1,09	1	562	70,1	0,7	154	10
mean		***2,81***	***0,71***	-	***4517***	***3,02***	***0,64***	***1,09***	***769***	***52,4***	***0,7***	***259,6***	***8.4***
standard dev.		***0.30***	***0.14***	-	***836***	***0,67***	***0,26***	***1,64***	***151***	***13,3***	***0,4***	***76,7***	***0.9***

[a]Planarity;
[b]eccentricity;
[c]secondary structure elements at the interface for the target and related partner;
[d]Gap volume;
[e]Gap volume index;
[f]Hydrogen bonds per 100 Å^2 of interface;
[g]Salt bridges;
[h]Accessible surface area buried at the interface of the protein/protein complex;
[i]Accessible surface area hidden by the inhibitor;
[j]Root mean square deviation (CA atoms) between unbound protein and complex;
[k]when unbound protein was not available, rmsd between protein/protein and protein/ligand complexes was computed;
[l]Total pocket volume at the interface;
[m]Number of interface residue segments. For each parameter the mean and standard deviation are presented.
doi:10.1371/journal.pone.0009598.t002

Overall, these results strongly suggest that PPIs with known inhibitors share similar shape properties than average transient protein-protein complexes.

PPIs with known inhibitors posses few pockets at the interface. PPIs and PPI inhibitor interactions use a greater number of small pockets than protein-ligand interactions [14]. The number and size of pockets at the interface for all the targets in $2P2I_{DB}$ were calculated using Q-SiteFinder [15]. Average values of 1.8 ± 0.7, 1.6 ± 0.7 and 1.7 ± 0.6 pockets corresponding to active volumes of 360 ± 244 Å^3, 303 ± 160 Å^3 and 247 ± 86 Å^3 were found for the target proteins in their free form, bound to the inhibitor and bound to their interacting partner, respectively. Similar values were observed for the two classes of PPIs. In their original paper, Fuller *et al.* reported 6 ± 3 pockets for PPIs with an average size of 54 Å^3. However, they calculated 99 pockets for each protein surface and only 10 could be visualized on the Q-SiteFinder web server. Because only the larger pockets could be visualized in that study, this limitation could account for the slightly smaller number of pockets observed in our dataset compared to those reported in the literature [14]. ZipA was not included in the general statistics because modulators of ZipA/FtsZ possess a remarkable way of binding and they do not penetrate the ZipA surface (supplementary material, Fig. S1). As a consequence, no pocket was found at the interface for this target.

We then analyzed the protein/protein and protein/inhibitor interfaces in terms of chemical properties.

Chemical Nature of the Interface

More hydrogen bonds at the interface. Hydrogen bonds play a key role in the specificity of the interaction between two proteins. The number of hydrogen bonds per 100 Å^2 of interface was estimated for the different complexes of the $2P2I_{DB}$ dataset. The average number of hydrogen bonds was comparable to that reported for protein-protein complexes (0.56 vs 0.52, [9]). On average, PPIs involving a peptide at the interface (class I) possessed less hydrogen bonds (0.42) than globular PPIs (class II, 0.64). However, large variations were observed for individual complexes; the average number of hydrogen bonds per 100 Å^2 varied from 0.05 to 1.1. Fifty percent of the complexes in the $2P2I_{DB}$ dataset possessed between 0.69 and 0.83 hydrogen bonds per 100 Å^2 of

interface indicating that the majority of 2P2Is possess more hydrogen bonds than typical protein-protein transient complexes (supplementary material, Fig. S2). Moreover, we analyzed the portion of the protein-protein interface that is occupied by the inhibitor in the protein/ligand form and found that for the majority of the PPIs, a large number of hydrogen bonds where located in that region.

Less salt bridges at the interface. PPIs with known inhibitors possess between zero and two salt-bridges, which are directly located in the region that is disrupted by the ligand (Table 2). The equivalent number for transient heterodimers could not be extracted from the literature, but it is very likely to be higher than the observed value in our dataset. The interaction between IL-2 and its receptor is an exception, as it contains five salt bridges with two of them corresponding to the binding site of the inhibitor. Interestingly, among the eight known inhibitors, six possess a guanidinium group that mimics a key arginine (Arg36) of the partner involved in one of the two salt bridges with a glutamic acid (Glu62) of the target.

Less charged residues at the interface. On average, PPIs with known inhibitors contain less charged residues at the interface (18.9±13.8%) than standard transient heterodimers (27.0±12.5%). This result indicates that PPIs with known inhibitors are more hydrophobic than typical heterodimers, which can be correlated to the observation that small molecule PPI modulators are hydrophobic [4]. However, a wide range of situations was observed with percentage values (from 0 to 46 percent), which confirms that PPIs with known inhibitors cover a diverse area of chemical space.

Descriptors of the Interface

We used a student's t-test to compare the main descriptors of PPIs in our dataset to standard heterodimers in an attempt to extract the most discriminating parameters that govern the druggability of a PPI. The number of interfacial segments, accessible surface area (ASA), Gap volume, hydrogen bonds and the percentage of charged residues were selected as the key parameters that typify PPIs with known inhibitors (Figure 3). The difference between the $2P2I_{DB}$ dataset and transient heterodimers for the four selected parameters corresponded to probabilities higher than 90% confidence according to the *t-distribution* table of significance.

PCA analysis. In addition to the five parameters defined above (ASA, Gap volume, number of segments at the interface, hydrogen bonds and percentage of charged residues), the combined pocket volumes at the interface were used in a principal component analysis (PCA) to separate the $2P2I_{DB}$ dataset into different families. Pockets volumes were not incorporated in the t-test study because statistics were not available to compare to the transient protein-protein dataset. However, they were used in the PCA analysis because they are known to play an important role in protein-protein specific interaction [14]. Out of the six parameters, four corresponded to geometric descriptors of the interface (ASA, Gap volume, number of interfacial segments and pocket volume) and two to the chemical nature of the interface (percentage of charged residues and hydrogen bonds). The influence of each descriptor is indicated by the direction and length of the corresponding arrows (Figure 4). All selected parameters have an important contribution

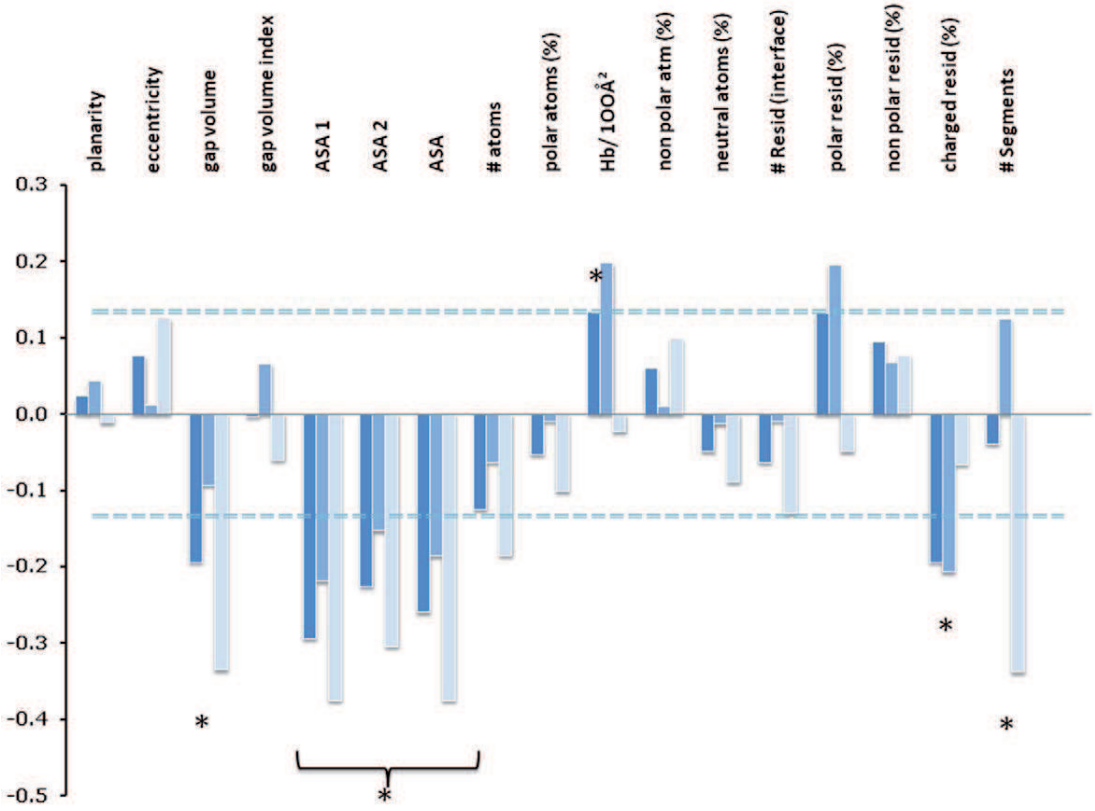


Figure 3. Student t-test allowing the selection of three main discriminating parameters for the PCA analysis. The t-test was calculated for the whole dataset and for each class (I and II) separately. Dotted lines indicate a threshold of confidence higher than 90% to differentiate $2P2I_{DB}$ complexes from transient heterodimers. On the basis of this analysis, ASA, Gap volume, Number of interfacial segments, hydrogen bonds and the percentage of charged residues were selected for the PCA analysis.
doi:10.1371/journal.pone.0009598.g003

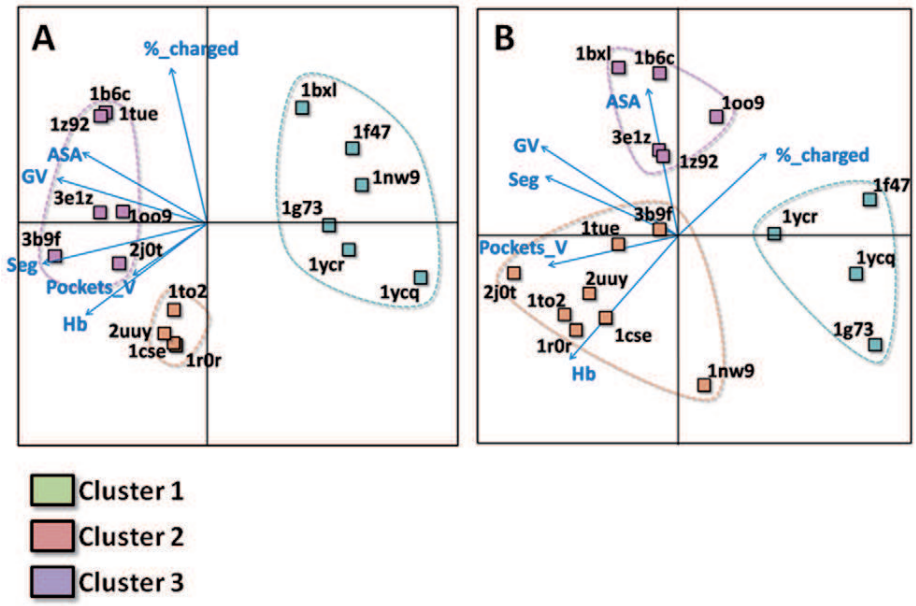


Figure 4. PCA analysis. Six key parameters were selected to perform the PCA analysis to separate the complexes into different groups: The five parameters based on the t-test and defined in figure 3 as well as total pocket volume at the interface. **A:** Analysis on the whole interface. Three different PPI clusters were defined. Cluster 1 (green) regrouped all complexes from class I corresponding to targets interacting with a peptide. Subtilisin and trypsin complexes defined cluster 2 (pink). Cluster 3 (purple) regrouped all other protein/protein complexes. **B:** Same analysis done on the part of the interface that is 4.5 Å around the ligand. The protein/protein complexes were in three slightly different clusters. Four out of six class I complexes were grouped together. Subtilisin and trypsin complexes remained very closely associated.
doi:10.1371/journal.pone.0009598.g004

to the clustering procedure (length of arrows) and they cover a large analytical space (direction of arrows). Three groups were found as a result of the clustering (Figure 4A). Interestingly, all six complexes that belong to class I were grouped together (cluster 1, green), which validates our choice of parameters for the PCA analysis. Subtilisin/Eglin C and trypsin/trypsin inhibitor formed a second group, mainly based on ASA and charged residue parameters (cluster 2, pink). The remaining seven complexes were grouped in a third cluster (cluster 3, purple). Parameters characteristic of weak and strong transient dimers as defined by Nooren and Thornton [11] were incorporated in a new PCA analysis. The parameters representative of the weak and strong dimers were grouped with cluster 1 and 3, respectively (data not shown). Therefore the PCA analysis with the six selected key parameters of the interface can be used to predict the type of interaction of the PPI, which can be of particular interest to define future targets for the discovery of new PPI inhibitors.

A similar PCA analysis was performed with parameters derived from the subset of the interface that is in direct contact with the inhibitor in the protein/ligand structure (Figure 4B). The same descriptors were used as mentioned above. Again, three clusters were identified; however, the separation between class I and class II was not as clear as for the whole interface. Two out of six complexes from class I behaved differently. In the case of the XIAP/Caspase complex (PDB code: 1nw9) the high number of hydrogen bonds and the absence of charged residues were responsible for its different behavior. In the $BclX_L$/Bak complex (PDB code: 1bxl), the high values for ASA and the number of segments and low hydrogen bond content led to its classification next to the FKBP12/TGFR complex (PDB code: 1b6c). Subtilisin/Eglin C and trypsin/trypsin, which formed a separated cluster when considering the whole interface, were grouped together when considering the part of the interface that interacts with the ligand.

2P2I Web Server Description

A web server was developed to facilitate the access to the data calculated for the different PPIs. It allows the user to search for specific complexes using different query procedures based on families, pdb codes, Uniprot numbers, cluster family or ligand properties (Figure 5A). For each query, a table is returned giving a list of PPIs matching the query. For each protein/protein complex, the cluster number, the family name, the pdb codes of the protein/protein complex, the unbound partners and the protein/inhibitor complex, the three letter code of the ligand and its molecular weight are given. Links to relevant databases and to literature are also provided. Finally, a data report containing the main analyses of PPIs can be accessed in different tabs.

An interactive menu allows the user to compare the properties of a given PPI to the $2P2I_{DB}$ dataset using the six key descriptors that were selected for the PCA analyses (Figure 5B). The parameters should be pre-calculated using online web servers (ProtorP [10], PDBSum [16] or Q-SiteFinder [15]) and entered on the $2P2I_{DB}$ website. Each parameter is compared to the equivalent parameter in the $2P2I_{DB}$ dataset and the number of parameters that are close or very close to $2P2I_{DB}$ parameters is returned. An indication of the type of PPI in terms of binding affinity (weak or strong) is also provided. Finally, individual scores for each parameter are calculated. Lower values are indicative of complexes with high potential to become druggable targets whereas higher scores are likely to correspond to poorly druggable complexes.

Conclusions

PPIs interfaces have long been considered to be poorly druggable because of their general properties. However, in the last few years, a growing number of PPI inhibitors have been discovered, and some of these inhibitors have even reached the

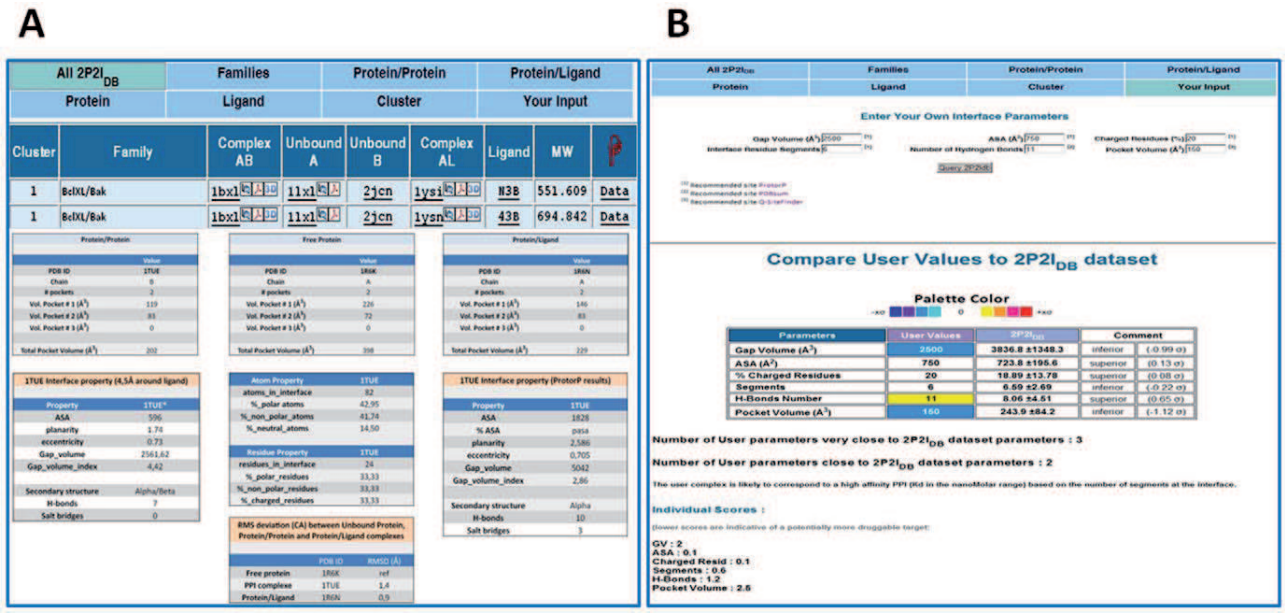


Figure 5. 2P2I$_{DB}$ Server (http://2p2idb.cnrs-mrs.fr). A: Query procedures to retrieve data stored in the database for each PPI. The database can be searched by family, by PDB codes of the protein/protein complexes, protein/ligand complexes, or unbound proteins, ligand properties or cluster number. A list of PDB codes or Uniprot numbers is returned as well as links to relevant databases. A detailed report of properties of each PPI is also provided. **B:** Interactive menu to compare user defined PPI properties with the 2P2I$_{DB}$ dataset. The user can enter pre-calculated values for the six key descriptors that were used in the PCA analyses. In return, these parameters are compared to the 2P2I$_{DB}$ dataset and an estimation of the druggability of the target is proposed.
doi:10.1371/journal.pone.0009598.g005

preclinical stage of investigation [8,17]. The number of currently available drug targets is very limited [18]. Due to their crucial role in various biological processes and in the dysfunction of cells, PPIs will probably become the next generation of successful therapeutics.

To this critical question *"What makes a PPI an attractive target for the discovery and development of small molecules?"* we can portray the good PPI candidates in a few points, which are summarized from existing published data, as follows: i) The target should be validated at least biologically (RNAi, knock-out gene etc…) or, even more importantly, clinically, and the inhibition of the PPI should not be associated with toxicity; ii) Key residues involved in the interaction and defining hot spots should be characterized through mutagenesis analyses, alanine scanning or by server prediction; iii) Structural information on the PPI complex and/or unbound forms should be available to accelerate the process; bridging water molecules should be taken into account when available and because the best conformation to target for PPI inhibition is not necessarily the one found in the complex, the natural conformational dynamics of the target should be investigated through molecular dynamic simulations [19], normal mode analysis [20], conformation ensemble predictions [21] or NMR [22]; iv) Three to five pockets should be available at the interface in the free form or in the simulated conformations of the target for a total volume of at least 250 Å^3 [14].

A recent study concluded that small molecule PPI modulators are larger (average molecular weight >400 Da), more hydrophobic (average alogP ~4), with more aromatic rings (~4 in average) and make fewer hydrogen bonds with the protein than average drugs [4]. However, a detailed description of the chemical and topological spaces of protein interfaces that can be disrupted by small drugs was not available. The present study leads to a better definition of a potentially successful PPI target.

We have gathered information available for 17 PPIs with known inhibitors whose three dimensional structural had been characterized. The interfaces were analyzed in terms of geometrical parameters, shape and chemical properties. The protein/protein complexes could be divided into two classes according to different parameters, such as the number of segments at the interface. Class I PPIs correspond to those that interact with peptide-like partners and show more secondary structure elements at the interface, whereas the class II group comprises more globular protein/protein complexes with more unstructured elements at the interface.

The different PPIs were further classified by PCA analysis using descriptors that were selected based on t-test evaluations and general analyses of the interfaces. Six interfacial parameters were selected corresponding to ASA, Gap volume, percentage of charged residues, number of segments, hydrogen bonds and total volume pockets. Three clusters were defined as a result of the PCA analysis; cluster 1 corresponded to class I PPIs, while class II PPIs were subdivided into two clusters.

Analysis of the part of the PPI that corresponds to the region directly in contact with the inhibitor led to similar results. However, minor differences could be observed, which suggests that parameters that define the druggability of a target are probably different from the parameters that define the chemical space of PPI inhibitors. The descriptors were selected for their ability to discriminate between whole PPIs with known inhibitors and transient dimers; additional parameters would have to be selected to predict the chemical space of the ligands that are likely to disrupt a given PPI.

The number of structurally characterized complexes with known inhibitors is small. Therefore, the chemical space of PPIs is not completely covered in the 2P2I$_{DB}$ because of the limited amount of data currently available. However, the definition of

what makes a PPI a potentially druggable target will become more and more reliable as the number of 3D structures increases.

The ZipA example, in which the ligand is lying on the protein surface in an unconventional way (supplementary material, Fig. S1), highlights the difficulty of defining general parameters for the druggability of PPI targets. However, we are at an early stage in the process of defining new relevant PPI targets and, as for HTS approaches, the goal is to be able to improve the selection of new PPI targets rather than define all the potential targets.

Further improvements will include incorporating other parameters such as available mutation data, known or predicted hot spots, dynamic behavior of the interface and the development of new databases dedicated to other types of PPI (such as disordered interacting segments and PPIs with large domain rearrangements, which will need specific descriptors).

The results of our study serve to expand current knowledge with new data and focuses at the interface of protein/protein complexes with prior structural knowledge. The proposed classification should lead to a better definition of potentially successful PPI targets and will accelerate the process of designing new PPI drugs. As successes in discovering PPI inhibitors accumulate, the parameters will be refined and the classification scheme updated.

The whole 2P2I dataset was organized as a relational database and can be accessed through a publicly available web server (http://2p2idb.cnrs-mrs.fr).

Materials and Methods

Dataset Collection

The hand curated dataset was constructed using two parallel approaches:

Search from the PDB. The entire PDB was searched using Dockground server [23] in order to collect protein-protein complexes. Crystal structures of heteromultimeric complexes with a resolution of 2.0 Å or lower were retrieved. Disordered protein and complexes with nucleic acid were discarded. A total of 202 heteromultimeric complexes were obtained.

The whole Protein Data Bank was then searched using an advanced query for free protein structures corresponding to each complex bound to small molecule inhibitors (supplementary material, Fig. S3). We then manually checked that the inhibitor was present at the interface between the two proteins and not covalently bounded to the protein. Nine protein/protein and 25 protein/ligand complexes were finally retrieved (Table 1, Source = PDB). When available, the unbound proteins were also included in the database.

Literature data mining. Eight protein/protein and 31 protein/ligand were retrieved by an exhaustive search of the literature (Table 1, Source = PubMed). IL-2/IL-2R and MDM2/p53 families were retrieved directly from the work of Pagliaro *et al.* [7]. Other complexes ($BclX_L$/Bak, E2/E1, XIAP BIR3/Caspase, XIAP BIR3/SMAC, ZipA/FtsZ) were not found in the PDB search either because the interacting partner was a peptide or because of the x-ray resolution (>2 Å).

The two lists were combined to form the final dataset, which can be downloaded at http://2p2idb.cnrs-mrs.fr/dataset/2P2Idataset.zip. The whole dataset is composed of 17 protein-protein complexes, 23 unbound proteins and 64 protein-inhibitor complexes. The PDB IDs are: 1bxl, 1lxl, 1ysi, 1ysn, 2o2m, 1ysg, 2o2n, 2o22 (BclXL/Bak); 1z1m, 1ttv (MDM2/p53); 1tfq, 1tft (XIAP_BIR3/CASPASE_9); 1f9x (XIAP_BIR3/SMAC); 1oo9 (MMP3/TIMP1) corresponds to solution NMR structures. All other PDB codes correspond to x-ray structures.

2P2I Database

The protein-protein interaction inhibition relational database was developed with MySqL. It stores information about the 17 PPIs described in this study. Scripts for interaction with the DB have been developed in PhP with the software MyAdmin.

Web Server

The $2P2I_{DB}$ database can be accessed through a web-based user interface (http://2p2idb.cnrs-mrs.fr). This platform allows users to query the database to get structural information about interfaces of stored complexes. The whole database can be searched by protein family, PDB codes of the free proteins, protein-protein or protein-ligand complexes, UniProt numbers, ligand three letter codes or by cluster number.

The user can also provide key parameters calculated from other web resources to compare the property of a given PPI to the 2P2I dataset.

Dataset Analysis

Root mean square deviations. The root mean square deviations (rmsd) between free and bound states of different proteins were computed over CA atoms using DaliLite server (http://www.ebi.ac.uk/Tools/dalilite/index.html) and are summarized into Table S1.

Geometrical parameters. Planarity, eccentricity, secondary structure in interface, Gap volume, Gap volume index, number of atoms in interface, % polar atoms in interface, % non polar atoms in interface, % neutral atoms in interface, number of residues in interface, % polar residues in interface and % non polar residues at the interfaces were calculated with the ProtorP server using default parameters (http://www.bioinformatics.sussex.ac.uk/protorp/). Continuous interface segment have been defined in the literature as a stretch of residues that starts and ends with interface residues and may contain intervening non-interface residues. While considering the length of the segment, only interface residues are counted [13,24].

Size of the interface. Accessible surface area (ASA) and percentage of charged residues were computed with the Naccess program with a probe radius of 1.4 Å. The size of the interface corresponded to the difference in ASA between the protein without its partner and in the complex.

Pocket size and volumes. Pockets at the interface were computed with the Q-SiteFinder server (http://www.bioinformatics.leeds.ac.uk/qsitefinder) on the protein-protein, protein-inhibitor complexes and the equivalent free proteins [14]. The homologous proteins were superimposed and only the pockets that were at least partly occupied by the inhibitor were retained.

Secondary structures. The percentage of secondary structure elements in the interface of the target protein and its partner were calculated with VMD using *in house* scripts. Four categories were defined for the overall class of the interface: H: Alpha Helix >30% and Beta strands <30%; S: Alpha Helix <30% and Beta strands >30%; HS: Alpha Helix >30% and Beta strands >30%; and Coil: Alpha Helix <30% and Beta strands <30%;

Hydrogen bonds and salt bridges at the interface. Hydrogen bonds were computed with the Pymol software (http://pymol.sourceforge.net/) using a 3.2 Å distance cutoff between the hydrogen atom and the acceptor atom and were checked manually.

A salt bridge was considered when an acidic residue (Asp or Glu) on one side of the interface and a basic residue (Arg, His or Lys) on the other side were less than 4.0 Å apart. Each putative

salt bridge was then validated manually using VMD (http://www.ks.uiuc.edu/Research/vmd/).

Analysis of a subset of the interface. A subset of the interface was defined by taking into account only atoms around 4.5 Å of the ligand in the protein-inhibitor complexes (see supplementary material Table S3).

Statistical Analyses and Clustering

T-test. The t-values were calculated as follows:

$$t = \frac{M_{2P2I} - M_{RCSB}}{\sqrt{(Var_{2P2I}/n_{2P2I}) + (Var_{RCSB}/n_{RCSB})}}$$

Where Var_{2P2I} and Var_{RCSB} are the variance of parameters in each group; M_{2P2I} and M_{RCSB} are means of these groups. The n_{2P2I} and n_{RCSB} are the total number of complexes in each group. Based on student's *t-distribution* table of significance (http://www.math.unb.ca/~knight/utility/t-table.htm), values higher than 1.34 correspond to probabilities of more than 90% confidence. On the one hand, if t-value is positive and greater than 1.34, then the mean of the studied parameter is significantly greater in the 2P2IDB dataset than in the RCSB transient dimers dataset at 90% or higher confidence level. On the other hand, if the t-value is negative and less than −1.34, then the mean of the studied parameter is significantly less in the $2P2I_{DB}$ dataset than in the RCSB transient dimers dataset.

PCA. Six parameters (ASA, Number of segments at the interface, Gap volume, pocket volume, hydrogen-bonds and the percent of charged residues in interface) were selected for the multivariate analysis performed according to the principal component analysis. Data were analyzed with the R software (http://www.R-project.org) and the ade4 package [25].

Clustering. Clustering of the PPIs into three groups was performed using the K-mean method [26].

Supporting Information

Figure S1 ZipA protein in complex with IQZ inhibitor (PDB code 1S1J). The IQZ inhibitor ZipA surface is shown as a stick representation. (Figure generated with pymol).
Found at: doi:10.1371/journal.pone.0009598.s001 (0.11 MB PDF)

Figure S2 Hydrogen bonds per 100 $Å^2$ of accessible surface area. The X-axis represents the number of hydrogen bonds per 100 $Å^2$. The Y-axis illustrates the number of complexes present in 2P2IDB having this number of hydrogen bonds (within ±0.1), i.e. y value at x = 0.6 indicates that there are 3 complexes having 0.3 to 0.5 hydrogen bonds per 100 $Å^2$.
Found at: doi:10.1371/journal.pone.0009598.s002 (0.09 MB PDF)

Figure S3 Advanced query search of the protein databank. This table lists the different parameters used to parse the protein databank to search for proteins bound to a small molecule inhibitor.
Found at: doi:10.1371/journal.pone.0009598.s003 (0.08 MB PDF)

Table S1 Root mean square deviations for the complexes in 2P2I database. The rms are computed between the unbound protein and its equivalent in the protein/protein complex; the unbound protein and its homologous in complex with the inhibitor; the protein in complex with its partner and the homologous in complex with the inhibitor. When several structures are compared, the average is shown. All RMSD were performed over CA atoms with the DaliLite web server.
Found at: doi:10.1371/journal.pone.0009598.s004 (0.15 MB PDF)

Table S2 Secondary structure at interface. This table lists secondary structures at interface (as defined in M&M) for each complex present in 2P2IDB. Information is detailed for both the target protein and its partner.
Found at: doi:10.1371/journal.pone.0009598.s005 (0.16 MB PDF)

Table S3 Geometrical and chemical parameters for the subset of the interface at 4.5 Å around the inhibitor. The parameters are detailed for each complex of 2P2IDB and mean and standard deviations are shown for class I, class II and the whole database.
Found at: doi:10.1371/journal.pone.0009598.s006 (0.16 MB PDF)

Author Contributions

Conceived and designed the experiments: XM PR. Performed the experiments: RB MJB. Analyzed the data: RB XM. Wrote the paper: RB XM PR.

References

1. Roche P, Xavier M (2010) Protein-Protein Interaction Inhibition (2P2I):Mixed Methodologies for the Acceleration of Lead Discovery. In: Miteva M, ed. In: Bentham. in press.
2. Patel S, Player MR (2008) Small-molecule inhibitors of the p53-HDM2 interaction for the treatment of cancer. Expert Opinion on Investigational Drugs 17: 1865–1882.
3. Toogood P (2002) Inhibition of protein-protein association by small molecules: approaches and progress. J Med Chem 45: 1543–1558.
4. Higueruelo AP, Schreyer A, Bickerton GRJ, Pitt WR, Groom CR, et al. (2009) Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. Chemical Biology & Drug Design 74: 457–467.
5. Fry DC (2008) Drug-like inhibitors of protein-protein interactions: a structural examination of effective protein mimicry. Current Protein & Peptide Science 9: 240–247.
6. Fry DC (2006) Protein-protein interactions as targets for small molecule drug discovery. Biopolymers 84: 535–552.
7. Pagliaro L, Felding J, Audouze K, Nielsen SJ, Terry RB, et al. (2004) Emerging classes of protein-protein interaction inhibitors and new tools for their development. Current Opinion in Chemical Biology 8: 442–449.
8. Wells JA, McClendon CL (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature 450: 1001–1009.
9. Bahadur RP, Zacharias M (2008) The interface of protein-protein complexes: analysis of contacts and prediction of interactions. Cellular and Molecular Life Sciences: CMLS 65: 1059–1072.
10. Reynolds C, Damerell D, Jones S (2009) ProtorP: a protein-protein interaction analysis server. Bioinformatics (Oxford, England) 25: 413–414.
11. Nooren I, Thornton J (2003) Structural characterisation and functional significance of transient protein-protein interactions. J Mol Biol 325: 991–1018.
12. Grünberg R, Leckner J, Nilges M (2004) Complementarity of structure ensembles in protein-protein binding. Structure (London, England: 1993) 12: 2125–2136.
13. Jones S, Thornton JM (1996) Principles of protein-protein interactions. Proceedings of the National Academy of Sciences of the United States of America 93: 13–20.
14. Fuller JC, Burgoyne NJ, Jackson RM (2009) Predicting druggable binding sites at the protein-protein interface. Drug Discovery Today 14: 155–161.
15. Laurie ATR, Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics (Oxford, England) 21: 1908–1916.
16. Laskowski R (2009) PDBsum new things. Nucleic Acids Res 37: D355–359.
17. Berg T (2008) Small-molecule inhibitors of protein-protein interactions. Current Opinion in Drug Discovery & Development 11: 666–674.
18. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nature Reviews Drug Discovery 5: 993–996.
19. Novak W, Wang H, Krilov G (2009) Role of protein flexibility in the design of Bcl-X(L) targeting agents: insight from molecular dynamics. Journal of Computer-Aided Molecular Design 23: 49–61.

20. Tama F, Brooks CL (2006) Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. Annual Review of Biophysics and Biomolecular Structure 35: 115–133.
21. Eyrisch S, Helms V (2009) What induces pocket openings on protein surface patches involved in protein-protein interactions? Journal of Computer-Aided Molecular Design 23: 73–86.
22. Lee GM, Craik CS (2009) Trapping moving targets with small molecules. Science (New York, NY) 324: 213–215.
23. Douguet D, Chen H-C, Tovchigrechko A, Vakser IA (2006) DOCKGROUND resource for studying protein-protein interfaces. Bioinformatics (Oxford, England) 22: 2612–2618.
24. Pal A, Chakrabarti P, Bahadur R, Rodier F, Janin J (2007) Peptide segments in protein-protein interfaces. J Biosci 32: 101–111.
25. Thioulouse J, Chessel D, Dolédec S, Olivier JM (1997) ADE-4: a multivariate analysis and graphical display software. Stat Comput. pp 75–83.
26. Hartigan JA, Wong MA (1979) Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society, Series C (Applied Statistics) 28: 100–108.
27. Muchmore SW, Sattler M, Liang H, Meadows RP, Harlan JE, et al. (1996) X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death. Nature 381: 335–341.
28. Kussie PH, Gorina S, Marechal V, Elenbaas B, Moreau J, et al. (1996) Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. Science (New York, NY) 274: 948–953.
29. Wu G, Chai J, Suber TL, Wu JW, Du C, et al. (2000) Structural basis of IAP recognition by Smac/DIABLO. Nature 408: 1008–1012.
30. Moy F, Glasfeld E, Mosyak L, Powers R (2000) Solution structure of ZipA, a crucial component of Escherichia coli cell division. Biochemistry 39: 9146–9156.
31. Redzynia I, Ljunggren A, Bujacz A, Abrahamson M, Jaskolski M, et al. (2009) Crystal structure of the parasite inhibitor chagasin in complex with papain allows identification of structural requirements for broad reactivity and specificity determinants for target proteases. FEBS J 276: 793–806.
32. Abbate EA, Berger JM, Botchan MR (2004) The X-ray structure of the papillomavirus helicase in complex with its molecular matchmaker E2. Genes & Development 18: 1981–1996.
33. Huse M, Chen Y, Massagué J, Kuriyan J (1999) Crystal structure of the cytoplasmic domain of the type I TGF beta receptor in complex with FKBP12. Cell 96: 425–436.
34. Rickert M, Wang X, Boulanger MJ, Goriatcheva N, Garcia KC (2005) The structure of interleukin-2 complexed with its alpha receptor. Science (New York, NY) 308: 1477–1480.
35. Iyer S, Wei S, Brew K, Acharya K (2007) Crystal structure of the catalytic domain of matrix metalloproteinase-1 in complex with the inhibitory domain of tissue inhibitor of metalloproteinase-1. J Biol Chem 282: 364–371.
36. Arumugam S, Van Doren S (2003) Global orientation of bound MMP-3 and N-TIMP-1 in solution via residual dipolar couplings. Biochemistry 42: 7950–7958.
37. Bode W, Papamokos E, Musil D (1987) The high-resolution X-ray crystal structure of the complex formed between subtilisin Carlsberg and eglin c, an elastase inhibitor from the leech Hirudo medicinalis. Structural analysis, subtilisin structure and interface geometry. Eur J Biochem 166: 673–692.
38. Horn J, Ramaswamy S, Murphy K (2003) Structure and energetics of protein-protein interactions: the role of conformational heterogeneity in OMTKY3 binding to serine proteases. J Mol Biol 331: 497–508.
39. Radisky E, Kwan G, Karen Lu C, Koshland DJ (2004) Binding, proteolytic, and crystallographic analyses of mutations at the protease-inhibitor interface of the subtilisin BPN'/chymotrypsin inhibitor 2 complex. Biochemistry 43: 13648–13656.
40. Li W, Adams T, Nangalia J, Esmon C, Huntington J (2008) Molecular basis of thrombin recognition by protein C inhibitor revealed by the 1.6-A structure of the heparin-bridged complex. Proc Natl Acad Sci U S A 105: 4661–4666.
41. Paesen G, Siebold C, Harlos K, Peacey M, Nuttall P, et al. (2007) A tick protein with a modified Kunitz fold inhibits human tryptase. J Mol Biol 368: 1172–1186.

I.2 Développements et mise à jour de la base de données $2P2I_{DB}$

L'analyse des différents paramètres structuraux, que nous avons comparés entre les complexes présents dans $2P2I_{DB}$ et l'ensemble des hétérodimères transitoires, nous a permis de mieux définir un espace chimique dans lequel les complexes présents dans la base de données forment un sous espace cohérent. La définition de cet espace pourrait permettre de choisir quelles cibles protéiques sont les plus semblables à celles déjà inhibées, et sont à leur tour les plus à même d'être inhibées facilement.

On peut notamment noter que les 17 complexes présents dans la base de données peuvent être séparés en deux classes : la classe I, comprenant 6 complexes, regroupe les interactions de type protéine-peptide, tandis que la classe II, comprenant 11 complexes, regroupe les complexes entre deux protéines globulaires.

Les complexes ont ensuite étés classés qualitativement grâce à une analyse en composante principale (ACP) à l'aide de 6 paramètres sélectionnés par test de Student que sont **l'ASA, le gap volume, le pourcentage de résidus chargés à l'interface, le nombre de segments, le nombre de liaisons hydrogènes et le volume total des poches présentes à l'interface**. L'ACP a classé les complexes en trois clusters, dont un correspond à la classe I, tandis que la classe II est subdivisée en deux clusters.

Après une analyse qualitative des paramètres régissant les interactions protéine-protéine, il est intéressant de réaliser une étude quantitative de ces paramètres. Une telle analyse permettrait de mesurer la druggabilité d'un complexe, et ainsi de diminuer les coûts et d'accélérer les campagnes de drug design. Une telle étude va être menée au sein du laboratoire, notamment par la mise en place d'une méthode d'apprentissage (telle que le SVM, les arbres de décisions ou encore les algorithmes bayésiens) qui permettra d'une part de définir l'espace conformationnel représentant les interactions protéine-protéine pour lesquelles un inhibiteur est connu, mais aussi de créer une fonction de score permettant d'attribuer un score à la druggabilité des interactions protéine-protéine.

Le nombre de structures tridimensionnelles de complexes protéiques avec un inhibiteur reste néanmoins assez restreint et peut entraîner pour l'instant un certain biais dans nos analyses. Cependant la sélectivité et la spécificité de modèles permettant de les étudier augmentera avec le nombre de structures connues. La base de données $2P2I_{DB}$ est

régulièrement mise à jour. Nous venons par exemple de passer de 40 molécules à 53 molécules depuis le début de l'année 2012.

Depuis sa première parution en 2010, la base de données $2P2I_{DB}$ a été entièrement mise à jour. D'une part de par son contenu, mais aussi par la mise en place de nouveaux outils développé au laboratoire et présentant un plus large éventail de réponses à qui s'interroge sur les interactions protéine-protéine. Aujourd'hui, $2P2I_{DB}$ contient 14 complexes protéine-protéine, 43 complexes protéine-ligand, 14 protéines libres et 53 inhibiteurs.

Les deux outils nouvellement mis en place depuis la parution de cet article sont $2P2I_{inspector}$ et $2P2I_{score}$. $2P2I_{inspector}$ permet une analyse en profondeur des paramètres d'une interaction protéine-protéine à partir de sa structure 3D. Un total de 58 descripteurs est calculé, incluant une grande variété de paramètres tels que le nombre de segments, des paramètres topologiques, les contacts à distance, les liaisons hydrogènes, la contribution des structures secondaires, les propriétés des atomes et des résidus et la composition atomique. Cet outil peut être utilisé soit pour calculer les paramètres d'une interface d'un complexe protéique présent dans la PDB avec un code 4 lettres, soit par téléchargement d'un fichier PDB personnel.

Le second outil, $2P2I_{score}$, est basé sur les six paramètres précédemment cités. Il permet de comparer chacun de ces six paramètres entre un complexe protéique du choix de l'utilisateur et l'espace conformationnel représentant les complexes protéiques pour lesquels un inhibiteur orthostérique existe. Pour chacun des paramètres, un score qualitatif est donné et un code couleur fournit la déviation de chaque paramètre comparé au même paramètre dans le cluster 2P2I équivalent.

Ce travail a fait l'objet d'une publication, publiée dans le journal Nucleic Acids Research.

I.3 Article 2

2P2I$_{DB}$: A Structural Database Dedicated to Orthosteric Modulation of Protein-Protein Interactions.

Basse, Marie Jeanne; Betzi, Stéphane; **Bourgeas, Raphael**; Bouzidi, Sofia; Chetrit, Bernard; Hamon, Véronique; Morelli, Xavier, Roche, Philippe.

Nucleic Acids Research (in press)

Protein-protein interactions are considered as one of the next generation of therapeutic targets. Specific tools thus need to be developed to tackle this challenging chemical space. In an effort to derive some common principles from recent successes, we have built 2P2Idb, a hand-curated structural database dedicated to protein-protein interactions with known orthosteric modulators. It includes all interactions for which both the protein-protein and protein-ligand complexes have been structurally characterized. A web server provides links to related sites of interest, binding affinity data, pre-calculated structural information about protein-protein interfaces and 3D interactive views through java applets. Comparison of interfaces in 2P2Idb to those of representative datasets of heterodimeric complexes has led to the identification of geometrical parameters and residue properties to assess the druggability of protein-protein complexes. A tool is proposed to calculate a series of biophysical and geometrical parameters that characterize protein–protein interfaces. A large range of descriptors are computed including, buried accessible surface area, gap volume, non-bonded contacts, hydrogen-bonds, atom and residue composition, number of segments and secondary structure contribution. All together the 2P2I database represents a structural source of information for scientists from academic institutions or pharmaceutical industries and is freely accessible at http://2p2idb.cnrsmrs.fr.

Ma contribution dans ce travail a consisté en la mise à jour des complexes présents dans 2P2I$_{DB}$ ainsi que dans la mise à jour des paramètres précalculés et présents dans la base de données.

D824–D827 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 30 November 2012*
doi:10.1093/nar/gks1002

2P2Idb: a structural database dedicated to orthosteric modulation of protein–protein interactions

Marie Jeanne Basse, Stéphane Betzi, Raphaël Bourgeas, Sofia Bouzidi, Bernard Chetrit, Véronique Hamon, Xavier Morelli* and Philippe Roche*

Laboratory of Integrative Structural and Chemical Biology (iSCB), Centre de Recherche en Cancérologie de Marseille (CRCM), CNRS UMR 7258, INSERM U 1068, Institut Paoli-Calmettes & Aix-Marseille Universités, 13009 Marseille, France

Received August 3, 2012; Revised September 18, 2012; Accepted October 1, 2012

ABSTRACT

Protein–protein interactions are considered as one of the next generation of therapeutic targets. Specific tools thus need to be developed to tackle this challenging chemical space. In an effort to derive some common principles from recent successes, we have built 2P2Idb (freely accessible at http://2p2idb.cnrs-mrs.fr), a hand-curated structural database dedicated to protein–protein interactions with known orthosteric modulators. It includes all interactions for which both the protein–protein and protein–ligand complexes have been structurally characterized. A web server provides links to related sites of interest, binding affinity data, pre-calculated structural information about protein–protein interfaces and 3D interactive views through java applets. Comparison of interfaces in 2P2Idb to those of representative datasets of heterodimeric complexes has led to the identification of geometrical parameters and residue properties to assess the druggability of protein–protein complexes. A tool is proposed to calculate a series of biophysical and geometrical parameters that characterize protein–protein interfaces. A large range of descriptors are computed including, buried accessible surface area, gap volume, non-bonded contacts, hydrogen-bonds, atom and residue composition, number of segments and secondary structure contribution. All together the 2P2I database represents a structural source of information for scientists from academic institutions or pharmaceutical industries.

INTRODUCTION

Protein–protein interactions (PPIs) represent a promising new class of attractive therapeutic targets, and the advancement in drug discovery efforts against PPIs has been recently referred as 'the unmined biology gold reserve' (1). However, PPIs are still considered as extremely difficult for targeting by small-molecules due to the structural characteristics of the interface, and specific strategies need to be undertaken to tackle this particularly challenging class of drug targets [for reviews see (2–5)]. Successes in drug discovery developments against PPI targets face two major issues, i.e. druggability assessment and adequacy of the chemical libraries used for screening. Over the last decade more and more orthosteric PPI modulators have been reported, and hundreds of small molecule inhibitors have now been developed for more than 40 PPI targets (4). Our goal is to use the structural knowledge from these success stories to derive some common principles to help future target selection and to accelerate the process of drug discovery in this field.

There are many structural databases dedicated to protein–protein complexes (6–14), to protein–ligand (15,16) or to small molecule inhibitors of PPIs (17–19). We have recently developed a hand-curated structural database (2P2Idb) by collecting information about protein–protein interfaces for which both the protein–protein and protein–inhibitor complexes have been structurally characterized, and we identified key descriptors of PPIs with known inhibitors (20). To our knowledge, 2P2Idb is the only structural database dedicated to orthosteric PPI modulators with structural information for protein–protein and protein–ligand complexes as well as for small molecule compounds. Although this database is relatively small at the moment, the hope is that, as it

*To whom correspondence should be addressed. Tel: +33 491164506; Fax: +33 491164540; Email: proche@imm.cnrs.fr
Correspondence may also be addressed to Morelli Xavier. Tel: +33 486977331; Fax: +33 486977499; Email: xavier.morelli@inserm.fr

grows, patterns will emerge for both protein–protein interfaces and small molecule inhibitors.

RESULTS

Presentation of 2P2Idb

2P2Idb is a relational database that was built through data mining from literature and by exhaustive search of the Protein Data Bank (20). To focus on orthosteric inhibitors, we have selected the cases for which both the protein–protein and protein–ligand complexes had been 3D-characterized (by X-ray or nuclear magnetic resonance) and for which the inhibitor is clearly competing at the interface. As of today, it contains 14 protein–protein complexes, 60 protein–inhibitor complexes, 16 free proteins and 55 small molecule modulators. The protein–protein complexes were subdivided into two classes corresponding to protein–peptide (cluster 1) and to globular protein–protein (cluster 2) complexes based on the number of segments at the interface. An interface segment is defined as a stretch of residues that starts and ends with interface residues and may contain intervening non-interface residues, but only in stretches of not more than four (21). The general interface properties are summarized for the two clusters in Table 1 showing that they differ notably. In particular, complexes from Cluster 1 can be disrupted with modified peptides such as staple peptides or with peptide mimetics whereas complexes that belong to Cluster 2 cannot. Furthermore, protein–protein complexes from Cluster 1 usually correspond to lower affinity complexes whereas those from Cluster 2 correspond to higher affinity complexes, on average. We have compared the general biophysical, biochemical and structural properties of the interfaces found in 2P2Idb with those of representative datasets of hetero and homodimers to establish a characteristic profile for 'druggable' protein–protein complexes (20 and Table 1). A web interface has been developed to facilitate access to pre-calculated data and to related websites.

Description of 2P2Idb web interface

Since its first release in 2010, the 2P2Idb website has been completely revisited by including a user friendly interface and more features. For each PPI family, clickable information can be found about protein–protein, protein–ligand complexes and free proteins as well as small molecule orthosteric modulators. Several links to relevant databases are provided such as published abstracts (PubMed), protein information (UniProt), 3D structures (PDBsum, PDBe), ligand properties (ChemSpider), protein–protein and protein–ligand binding affinities (PDBBind, BindingDB, ChEMBL or MOAD). A large number of pre-calculated interface parameters are accessible for each protein–protein complex. These interface descriptors include, total interface area, gap volume, percentage of charged residues, segments, non-bonded contacts, hydrogen bonds, salt bridges, disulfide bonds, secondary structure as well as atom and residue properties for each chain. The detailed list of non-bonded contacts, hydrogen bonds and salt bridges can be accessed through popup windows. Protein–protein and protein–ligand complexes can be interactively visualized through Jmol applets with customized menus and predefined representations. Furthermore, all protein structures in 2P2I database can be easily downloaded from our website (http://2p2idb.cnrs-mrs.fr/download.html) and analysed with external molecular visualization program viewers. In the downloaded files, 3D structures from the same family of complexes have been superimposed to the unbound form to facilitate user analysis and comparison. PDB structures can be downloaded by protein family or by complex type (protein–protein or protein–ligand).

2P2Iinspector: a protein–protein interface analysis tool

Several tools are available to analyse protein–protein interfaces. However, most of them are dedicated to the prediction of hotspots residues or binding pockets (22). Other servers provide structural and chemical information on protein–protein associations. Protein Interactions Calculator (PIC) is a server which computes contact information but does not calculate topological parameters such as gap volume and surface area (23). PISA is a tool for exploring macromolecular interfaces and surfaces (24). However, it is more dedicated to the prediction of probable quaternary structures from crystal contacts. In the first release of the 2P2I database, most interface parameters had been calculated through the ProtorP web

Table 1. The table provides 'means' and 'standard deviations' of several interface parameters calculated for the two classes of druggable complexes in 2P2Idb

Interface properties	$2P2I_{DB}$		Heterodimers		Homodimers	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
No. of complexes	7	7	189	336	331	1442
BASA (Å^2)	1384.7 ± 516.1	1793.3 ± 591.6	2149.2 ± 1017.6	2769.3 ± 1411.4	2307.2 ± 1503.1	3042.1 ± 1823.1
Gap volume (Å^3)	2282.8 ± 1351.5	5085.2 ± 2199.7	3906.7 ± 1745.2	6670.3 ± 3128.1	3780.8 ± 1775.1	6969.9 ± 3716.9
Non-bonded contacts	74.4 ± 27.5	94.4 ± 39.9	114.5 ± 57.2	151.5 ± 84.1	114.2 ± 80.4	164.6 ± 114.1
Total no. of segments	4.1 ± 1.1	8.1 ± 1.8	6.1 ± 2.7	10.8 ± 3.8	3.9 ± 1.1	10.7 ± 4.6
No. of hydrogen bonds	2.4 ± 1.3	3.3 ± 2.5	4.8 ± 3.7	6.7 ± 5.0	4.6 ± 5.4	7.2 ± 6.3
No. of salt bridges	0.6 ± 0.8	0.6 ± 0.8	1.8 ± 1.8	2.0 ± 1.8	1.4 ± 1.8	2.0 ± 2.5
No. of disulfide bonds	0.0 ± 0.0	0.0 ± 0.0	0.03 ± 0.2	0.03 ± 0.2	0.01 ± 0.1	0.01 ± 0.1
% Charged residues	20.9 ± 8.8	28.9 ± 11.5	28.7 ± 13.2	26.6 ± 11.7	26.6 ± 12.8	25.6 ± 11.5

Complexes from Cluster 1 correspond to protein–peptide complexes and can be disrupted with modified peptide or peptide mimetics. Complexes from Cluster 2 correspond to higher affinity complexes. Values for nonredundant representative datasets of hetero- and homo-dimeric complexes collected through the Dockground server are indicated as comparison for both classes.

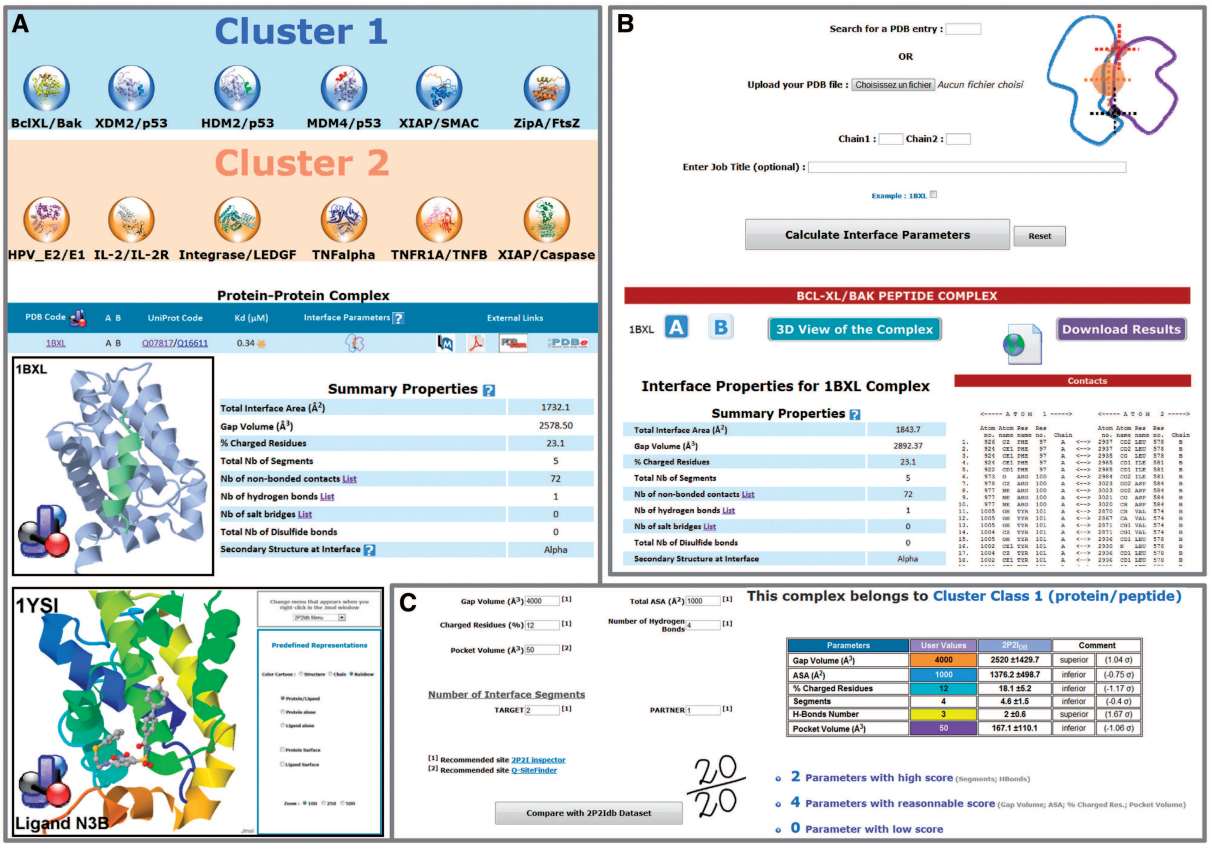


Figure 1. The 2P2I website and its main features. (**A**) 2P2Idb is a hand-curated database dedicated to the inhibition of protein–protein complexes with orthosteric modulators. It displays structural information about protein–protein, protein–ligand complexes and small molecule inhibitors. For each of the 14 families, sub-divided into two classes (protein–peptide and protein–protein), clickable html pages are provided with pre-calculated interface parameters, binding affinity data and links to related sites of interest (UniProt, PubMed, PDBsum, PDBe and ChemSpider). Protein–protein and protein–ligand complexes can be interactively visualized using Jmol applets and user-friendly menus. (**B**) 2P2Iinspector is a tool to analyse protein–protein interfaces in terms of geometric and physico-chemical descriptors. A total of 60 descriptors are computed including, buried accessible surface area, gap volume, non-bonded contacts, hydrogen-bonds, atom and residue composition, number of segments and secondary structure contribution. Users can analyze protein complexes from the PDB using standard four letter accession codes or upload their own files. (**C**) 2P2Iscore is a tool to assess the druggability of protein–protein interfaces. Comparison of protein–protein interfaces in 2P2Idb with standard heterodimers has allowed us to define six interface parameters to characterize protein–protein interfaces with a known modulator. Users are invited to compute five parameters using the 2P2Iinspector tool. The interfacial pocket volume should be calculated with Q-SiteFinder (http://www.modelling.leeds.ac.uk/qsitefinder). A color-coded table is provided to compare user defined parameters to those in 2P2Idb. A qualitative score is given for the six key parameters to assess the druggability of the interface. Detailed help documentation is available as PDF files for the different features.

server which has been discontinued and is no longer available (25). We have therefore developed our own, and enhanced, version of this tool by computing more interface parameters. 2P2Iinspector is a complete new tool that computes interaction properties from the 3D structure of protein–protein complexes. A total of 58 descriptors are now computed using in-house tcl scripts implemented in VMD (26) and SURFNET (27). These physical and chemical parameters include a large range of descriptors such as number of segments, buried accessible surface area, gap volume, non-bonded contacts, hydrogen-bonds, secondary structure contribution, atom and residue properties, and atomic composition. This new open access tool can be used to calculate interface parameters of protein complexes either from the PDB (using valid four-letter code) or by uploading a PDB file. The computed parameters can be accessed through the web interface for both chains of the protein–protein complex. Users can easily switch from the results of one chain to the other. Popup windows give easy access to the lists of non-bonded contacts, hydrogen bonds and salt bridges. The protein–protein complexes can be visualized interactively with a Jmol applet with the same functionalities described for the 2P2I database. Files are stored for 48 h before being deleted and during that period users can access their data via a direct unique link. Finally, results can be downloaded and then easily accessed locally as html files.

2P2Iscore: assessing the druggability of protein–protein interfaces

The difficulty of targeting PPIs emphasizes the importance of target selection. From a previous study, we defined six key interface parameters to characterize protein–protein complexes with a known modulator (20). We have used these descriptors to assess the druggability of protein–protein interfaces after the target has been assigned to a cluster type (protein–peptide or protein–protein) based on

the number of segments at the interface. A qualitative prediction is proposed, which is based on the standard deviation of each of the six interface parameters to the mean of the same parameter in the equivalent 2P2I cluster.

Users can compare parameters from their own protein–protein interface to a distribution of parameters from the 2P2Idb dataset. Five parameters can be easily computed on our website using 2P2Iinspector tool. We recommend using Q-SiteFinder to calculate the remaining descriptor, i.e. the interfacial pocket volume, because this server was used to estimate the size of pockets at the interface in 2P2Idb (20,28). A qualitative score is given for each parameter and a color coded table provides the deviation of each parameter compared to the mean of the same parameter in the equivalent 2P2I cluster.

CONCLUSIONS AND FUTURE DEVELOPMENTS

The 2P2I website (Figure 1) provides structural information about the modulation of PPIs with orthosteric inhibitors. A number of features give access to pre-calculated interface parameters and to related websites. A scoring function is available to qualitatively assess the druggability of protein–protein interfaces with prior 3D knowledge (2P2Iscore). A tool has been specifically developed to analyse protein–protein interfaces in terms of physico-chemical, topological or geometric features (2P2Iinspector). Future releases of the database will include new complexes and PPI modulators as they appear in the Protein Data Bank (29), new interface parameters (more particularly interfacial pockets) for the 2P2Iinspector tool and an automated version of 2P2Iscore with a quantitative scoring function.

We expect that this new version of the 2P2I database provides a useful source of information to characterize protein–protein interfaces and to design modulators of PPIs and is therefore of major interest for the scientific community.

ACKNOWLEDGEMENTS

The authors thank colleagues at the Cancer Research Center of Marseille for testing the webserver during construction.

FUNDING

Funding for open access charge: Agence Nationale de la Recherche [ANR-11-BS07-019-02].

Conflict of interest statement. None declared.

REFERENCES

1. Mullard,A. (2012) Protein–protein interaction inhibitors get into the groove. *Nat. Rev. Drug Discov.*, **11**, 173–175.
2. Wells,J.A. and McClendon,C.L. (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, **450**, 1001–1009.
3. Roche,P. and Morelli,X. (2010) In Miteva M (ed). In silico lead discovery. *Bentham Science Publishers, eBook*, pp.118–143.
4. Morelli,X., Bourgeas,R. and Roche,P. (2011) Chemical and structural lessons from recent successes in protein-protein interaction inhibition (2P2I). *Curr. Opin. Chem. Biol.*, **15**, 475–481.
5. Bienstock,R.J. (2012) Computational drug design targeting protein-protein interactions. *Curr. Pharm. Des.*, **18**, 1240–1254.
6. Davis,F.P. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.
7. Gong,S., Park,C., Choi,H., Ko,J., Jang,I., Lee,J., Bolser,D.M., Oh,D., Kim,D.-S. and Bhak,J. (2005) A protein domain interaction interface database: InterPare. *BMC Bioinformatics*, **6**, 207.
8. Teyra,J., Doms,A., Schroeder,M. and Pisabarro,M.T. (2006) SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces. *BMC Bioinformatics*, **7**, 104.
9. Winter,C., Henschel,A., Kim,W.K. and Schroeder,M. (2006) SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res.*, **34**, D310–D314.
10. Jefferson,E.R., Walsh,T.P., Roberts,T.J. and Barton,G.J. (2007) SNAPPI-DB: a database and API of Structures, iNterfaces and Alignments for Protein-Protein Interactions. *Nucleic Acids Res.*, **35**, D580–D589.
11. Kundrotas,P.J. and Alexov,E. (2007) PROTCOM: searchable database of protein complexes enhanced with domain-domain structures. *Nucleic Acids Res.*, **35**, D575–D579.
12. Günther,S., von Eichborn,J., May,P. and Preissner,R. (2009) JAIL: a structure-based interface library for macromolecules. *Nucleic Acids Res.*, **37**, D338–D341.
13. Stein,A., Panjkovich,A. and Aloy,P. (2009) 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res.*, **37**, D300–D304.
14. Bickerton,G.R., Higueruelo,A.P. and Blundell,T.L. (2011) Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the PICCOLO database. *BMC Bioinformatics*, **12**, 313.
15. Schüttelkopf,A.W. and van Aalten,D.M. (2004) PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 1355–1363.
16. Schreyer,A. and Blundell,T. (2009) CREDO: a protein-ligand interaction database for drug discovery. *Chem. Biol. Drug. Des.*, **73**, 157–167.
17. Higueruelo,A.P., Schreyer,A., Bickerton,G.R.J., Pitt,W.R., Groom,C.R. and Blundell,T.L. (2009) Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. *Chem. Biol. Drug Des.*, **74**, 457–467.
18. Reynes,C., Host,H., Camproux,A.C., Laconde,G., Leroux,F., Mazars,A., Deprez,B., Fahraeus,R., Villoutreix,B.O. and Sperandio,O. (2010) Designing focused chemical libraries enriched in protein-protein interaction inhibitors using machine-learning methods. *PLoS Comput. Biol.*, **6**, e1000695.
19. Sperandio,O., Reynès,C., Camproux,A. and Villoutreix,B. (2010) Rationalizing the chemical space of protein-protein interaction inhibitors. *Drug Discov. Today*, **15**, 220–229.
20. Bourgeas,R., Basse,M.-J., Morelli,X. and Roche,P. (2010) Atomic analysis of protein-protein interfaces with known inhibitors: the 2P2I database. *PLoS One*, **5**, e9598.
21. Pal,A., Chakrabarti,P., Bahadur,R., Rodier,F. and Janin,J. (2007) Peptide segments in protein-protein interfaces. *J. Biosci.*, **32**, 101–111.
22. Marabotti,A. and Milanesi,L. (2012) Finding inhibitors of protein-protein interactions (i-ppis): a support from bioinformatics. *World Res. J. Peptide Protein*, **1**, 9–20.
23. Tina,K.G., Bhadra,R. and Srinivasan,N. (2007) PIC: protein interactions calculator. *Nucleic Acids Res.*, **35**, W473–W476.
24. Krissinel,E. and Henrick,K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, **372**, 774–797.
25. Reynolds,C., Damerell,D. and Jones,S. (2009) ProtorP: a protein-protein interaction analysis server. *Bioinformatics*, **25**, 413–414.
26. Hsin,J., Arkhipov,A., Yin,Y., Stone,J.E. and Schulten,K. (2008) Using VMD: an introductory tutorial. *Curr. Protoc. Bioinformatics*, Chapter 5, Unit 5.7.
27. Laskowski,R. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph*, **13**, 323–330, 307–328.
28. Laurie,A. and Jackson,R. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, **21**, 1908–1916.
29. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

II Chimiothèques dédiées aux interactions protéine-protéine : 2P2I$_{CHEM}$

En 1995, *Wells et.al.* ont ouvert le champ de la modulation des interactions protéine-protéine en identifiant les hots-spots à l'interface de ces PPIs (Clackson and Wells, 1995). Depuis cette avancée majeure, un grand nombre d'études ont été menées sur les petites molécules permettant de moduler les interfaces protéine-protéine (Arkin and Whitty, 2009; Azmi and Mohammad, 2009; Blazer and Neubig, 2009; Fry, 2011; Koes and Camacho, 2011). Dans un même temps, plusieurs études se sont intéressées à la druggabilité des complexes protéine-protéine en se basant sur les propriétés de leurs interfaces (Fuller et al., 2009; Geppert et al., 2011; Kozakov et al., 2011; Sugaya and Furuya, 2011; Sugaya and Ikeda, 2009; Sugaya et al., 2007; Thangudu et al., 2012; Wanner et al., 2011). Cependant, malgré les progrès réalisés dans la recherche de nouveaux modulateurs de PPIs ces dernières années, il y a toujours un faible taux de succès lors des campagnes de criblages à haut débits. Ce faible taux d'inhibiteurs d'interactions protéine-protéine présent dans les chimiothèques commerciales révèle que ces dernières ne sont pas indiquées pour des campagnes de drug discovery sur des cibles protéine-protéine. L'inadéquation de ces chimiothèques commerciales met en valeur la nécessité de créer des chimiothèques dédiées à cet espace chimique précis que sont les interfaces protéine-protéine, pour accélérer les campagnes de criblages et en réduire le coût en augmentant le nombre de touches, tout en réduisant le nombre de molécules à tester. Pour parvenir à ce résultat, une des méthodologies est de créer un protocole de filtrage à partir de la plus grande chimiothèque possible, puis d'enlever les molécules qui ont peu de chance d'être des inhibiteurs d'interactions protéine-protéine tout en gardant le plus possible d'inhibiteurs potentiels. Plusieurs études ont étés menées en ce sens en utilisant les propriétés physicochimiques des inhibiteurs déjà connus (Higueruelo et al., 2009; Pagliaro et al., 2004; Wells and McClendon, 2007). Une conclusion consensus de ces études sur les caractéristiques des inhibiteurs d'interactions protéine-protéine est qu'ils sont en général plus grands, mais aussi plus hydrophobes. Ils forment par ailleurs plus d'interactions de type aromatique ou hydrophobe avec leur partenaire protéique.

Nous avons ici adopté cette méthodologie, et, afin de générer un profil physicochimique des inhibiteurs d'interactions protéine-protéine, nous avons analysé les 39 molécules présentes dans $2P2I_{DB}$. Afin de créer le protocole $2P2I_{HUNTER}$, nous avons généré un ensemble de modèles grâce aux machines à support de vecteurs (SVM) utilisant 11 descripteurs moléculaires. Les meilleurs ont été testés sur les deux seuls résultats de criblage à hauts débits représentatifs des interactions protéine-protéine présents sur la base de données PubChem (http://pubchem.ncbi.nlm.nih.gov/). Cette validation externe a montré la capacité de notre meilleur modèle à considérablement réduire la taille de la chimiothèque considérée, allant jusqu'à une réduction de 97%. Elle a de plus montré l'enrichissement en molécules actives du sous-ensemble sélectionné, allant jusqu'à une proportion 12 fois supérieure dans la chimiothèque filtrée par rapport à la chimiothèque initiale.

Ce modèle a enfin été appliqué à 25 chimiothèques commerciales (incluant les chimiothèques commerciales dédiées aux interactions protéine-protéine déjà disponible) et représentant un ensemble de 8 millions de petites molécules. En moyenne, seul trois pour cent des molécules furent retenues par le modèle, allant de 0,2% à 7,9% suivant les chimiothèques. Il y a cependant une chimiothèque d'environ 7000 composés : Eccentric de ChemDiv PPI, pour laquelle 70 % des molécules ont étés conservées. Il est intéressant de noter que cette chimiothèque, qui est composée de scaffolds inusités en chimie médicinale, coïncide en grande partie avec l'espace chimique défini par notre modèle SVM.

Le protocole inhibiteur d'interactions protéine-protéine $_{HUNTER}$ a fait l'objet d'un article, actuellement en cours de soumission dans le journal PloS Computational Biology.

Arkin, M.R., and Whitty, A. (2009). The road less traveled: modulating signal transduction enzymes by inhibiting their protein-protein interactions. Current Opinion in Chemical Biology *13*, 284-290 %U http://www.ncbi.nlm.nih.gov.gate281.inist.fr/pubmed/19553156.

Azmi, A.S., and Mohammad, R.M. (2009). Non-peptidic small molecule inhibitors against Bcl-2 for cancer therapy. Journal of Cellular Physiology *218*, 13-21 %U http://www.ncbi.nlm.nih.gov.gate11.inist.fr/pubmed/18767026.

Blazer, L.L., and Neubig, R.R. (2009). Small molecule protein-protein interaction inhibitors as CNS therapeutic agents: current progress and future hurdles. Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology *34*, 126-141 %U http://www.ncbi.nlm.nih.gov/pubmed/18800065.

Clackson, T., and Wells, J.A. (1995). A hot spot of binding energy in a hormone-receptor interface. Science (New York, NY) *267*, 383-386 %U http://www.ncbi.nlm.nih.gov.gate381.inist.fr/pubmed/7529940.

Fry, D. (2011). Small-Molecule Inhibitors of Protein-Protein Interactions (Nutley, New Jersey).

Fuller, J.C., Burgoyne, N.J., and Jackson, R.M. (2009). Predicting druggable binding sites at the protein-protein interface. Drug Discovery Today *14*, 155-161.

Geppert, T., Hoy, B., Wessler, S., and Schneider, G. (2011). Context-based identification of protein-protein interfaces and "hot-spot" residues. Chem Biol *18*, 344-353.

Higueruelo, A.P., Schreyer, A., Bickerton, G.R.J., Pitt, W.R., Groom, C.R., and Blundell, T.L. (2009). Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. Chemical Biology & Drug Design *74*, 457-467 %U http://www.ncbi.nlm.nih.gov/pubmed/19811506.

Koes, D.R., and Camacho, C.J. (2011). Small-Molecule Inhibitor Starting Points Learned From Protein-Protein Interaction Inhibitor Structure. Bioinformatics.

Kozakov, D., Hall, D.R., Chuang, G.Y., Cencic, R., Brenke, R., Grove, L.E., Beglov, D., Pelletier, J., Whitty, A., and Vajda, S. (2011). Structural conservation of druggable hot spots in protein-protein interfaces. Proc Natl Acad Sci U S A *108*, 13528-13533.

Pagliaro, L., Felding, J., Audouze, K., Nielsen, S.J., Terry, R.B., Krog-Jensen, C., and Butcher, S. (2004). Emerging classes of protein-protein interaction inhibitors and new tools for their development. Current Opinion in Chemical Biology *8*, 442-449.

Sugaya, N., and Furuya, T. (2011). Dr. PIAS: an integrative system for assessing the druggability of protein-protein interactions. BMC Bioinformatics *12*, 50.

Sugaya, N., and Ikeda, K. (2009). Assessing the druggability of protein-protein interactions by a supervised machine-learning method. BMC Bioinformatics *10*, 263 %U http://www.ncbi.nlm.nih.gov.gate261.inist.fr/pubmed/19703312.

Sugaya, N., Ikeda, K., Tashiro, T., Takeda, S., Otomo, J., Ishida, Y., Shiratori, A., Toyoda, A., Noguchi, H., Takeda, T.*, et al.* (2007). An integrative in silico approach for discovering candidates for drug-targetable protein-protein interactions in interactome data. BMC Pharmacology *7*, 10 %U http://www.ncbi.nlm.nih.gov.gate11.inist.fr/pubmed/17705877.

Thangudu, R.R., Bryant, S.H., Panchenko, A.R., and Madej, T. (2012). Modulating protein-protein interactions with small molecules: the importance of binding hotspots. J Mol Biol *415*, 443-453.

Wanner, J., Fry, D.C., Peng, Z., and Roberts, J. (2011). Druggability assessment of protein-protein interfaces. Future Med Chem *3*, 2021-2038.

Wells, J., and McClendon, C. (2007). Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature *450*, 1001-1009.

II.1 Article 3

2P2IHUNTER: A Tool for Filtering Orthosteric Protein-Protein Interaction Modulators via a Dedicated Support Vector Machine.

Véronique Hamon, **Raphael Bourgeas**, Pierre Ducrot, Isabelle Théret, Laura Xuereb, Marie Jeanne Basse, Jean Michel Brunel, Sebastien Combes, Xavier Morelli and Philippe Roche.

Article soumis pour publication.

Abstract

To get back on stage with approval of new molecular entities (NMEs), R&D pharmaceuticals necessitate tackling two main hurdles: i) target profiling optimization and ii) exploration of new chemical spaces. In the crowded population of 'druggable' candidates relevant for human diseases, protein-protein interactions (PPIs) and their inhibitors (protein-protein interaction inhibitors, 2P2Is) have witnessed obvious progress in recent years. However, the chemical space associated to this 'highly hanging fruit' is still clearly under debates. The purpose of the present research is to present a general profile for pharmacologically and structurally validated orthosteric PPI inhibitors identified so far and to subsequently propose an optimized protocol to design focused PPI chemical libraries using support vector machines approaches. The filtering protocol has been validated using external datasets from PubChem bioassay and results from *in house* screening campaigns. To characterize the applicability domain of our protocol, this $2P2I_{HUNTER}$ algorithm has been applied to the main chemical providers commercially available to pharmaceutical industries and academic researchers representing more than 8 million compounds. The algorithm keeps in average 0.2 to 7.9% of the commercial libraries demonstrating the low enrichment in 'PPI-like inhibitors' in these available databases. In a very surprising contrast, the $2P2I_{HUNTER}$ filtration tool selects more than 70 percent of the 7,076 compounds from the ChemDiv eccentric PPI library that corresponds to new and unusual scaffolds in medicinal chemistry. We believe that the resulting chemical space herein identified will provide to the scientific community a concrete support to search for PPI inhibitors during HTS campaigns.

Despite high levels of R&D investment, the number of new approved drugs on the market still has not found a sufficient level. This is partly due to the lack of innovative targets and appropriate corresponding chemical libraries. Over the last ten years protein-protein interactions have shown their potential as new therapeutic targets. The human 'interactome' is estimated to contain between ~130,000 and ~650,000 protein-protein interactions (PPIs); many of them are related to diseases and therefore represent a tremendous reservoir of putative original targets. However, the design of PPI modulators still remains challenging and the average success rate of high throughput screening campaigns stays critically low for these targets. One of the reasons is the lack of chemical libraries dedicated to this particular chemical space. In this research article we present an original tool ($2P2I_{HUNTER}$) to design focused PPI chemical libraries. $2P2I_{HUNTER}$ has been applied to a set of 25 commercial libraries corresponding to the major providers and representing more than 8 million compounds. Among these compound collections a PPI dedicated library from ChemDiv showed a great superposition with the chemical space defined by our SVM model. This new tool should lead to better success rates in screening campaigns dedicated to the modulation of protein-protein complexes.

Ma contribution dans ce travail a consisté en premier lieu en l'analyse des petites molécules présentes dans $2P2I_{DB}$ et qui a mené à la mise en place de la « Rule of 4 ». J'ai de plus aidé lors de la réalisation du modèle SVM, bien que je n'en sois pas le développeur. Lors de la validation du modèle, j'ai effectué les analyses de PubChem qui ont permis de sélectionner les cribles qui ont servis comme validation au modèle.

Introduction

Protein-protein interactions (PPIs) play a central role in many cellular processes ranging from signal transduction to cell adhesion through cell proliferation, growth, differentiation, viral self-assembly, programmed cell death (see reviews (Morelli et al., 2011; Roche and Morelli, 2010; Wells and McClendon, 2007)). They are involved in numerous pathways including diseases, different stages of cancer development and in host-pathogen interactions. Therefore, modulation of these networks of transient interactions represents a promising therapeutic strategy. For these reasons, PPIs are becoming more and more accepted as potential drug targets despite the longtime assumption that they were poorly druggable (Mullard, 2012). Inhibitors of these original interactions are certainly the next generation of highly innovative drugs that will reach the market in the next decade. However, the *in silico* design of such compounds still remains challenging. (Berg, 2008; Morelli et al., 2011; Patel and Player, 2008; Roche and Morelli, 2010; Vu and Vassilev, 2011; Wilson, 2009; Wilson and Arkin, 2011).

The identification of hot spots at the interface of PPIs (Wells, 1995) has given a rationale for the possible disruption of protein-protein complexes with small molecules. Since then, there have been an increasing number of studies reporting small molecules disrupting protein–protein interactions (Arkin and Whitty, 2009; Azmi et al., 2011; Blazer and Neubig, 2009; Dudkina and Lindsley, 2007; Fry, 2011; Koes and Camacho, 2011; Vu and Vassilev, 2011). Several studies on interface properties have addressed the druggability of protein-protein complexes (Bourgeas et al., 2010; Fuller et al., 2009; Geppert et al., 2011; Kozakov et al., 2011; Sugaya and Furuya, 2011; Sugaya and Ikeda, 2009; Sugaya et al., 2007; Thangudu et al., 2012; Wanner et al., 2011). However, despite the progress in protein-protein interaction drug discovery during the last decade, there is still limited success in HTS campaigns (the success rate to find hit compound is generally very low) suggesting that most chemical libraries available today are not appropriate for screening protein-protein interaction targets (Fry, 2006). This poor suitability of commercial libraries emphasizes the necessity to design targeted chemical libraries dedicated to this particular chemical space (PPI targets) to accelerate and reduce the cost of screening campaigns by enhancing the number of hits while reducing the number of compound tested. This would certainly help bringing pharmaceutical companies back on track (Bunnage, 2011). One way to achieve this goal is to design filtering algorithm to remove compounds that are unlikely to disrupt PPI interfaces from large

chemical libraries while preserving a large number of potential disruptors in the selected subset. Several studies have focused on the chemical properties of known PPI inhibitors (Higueruelo et al., 2009; Pagliaro et al., 2004; Wells and McClendon, 2007). A general profile has been defined for these PPI inhibitors by compiling a collection of known PPI inhibitors and comparing them to other drugs. As a general profile, the authors evidenced that PPI inhibitors compared to other small molecule–protein complexes are generally larger and more hydrophobic, they tend to form fewer hydrogen bonds and to present more aromatic and hydrophobic interactions at the protein-ligand interface.

Decision tree methods have also been used to design PPI-inhibitor focused libraries (Neugebauer et al., 2007; Reynes et al., 2010; Sperandio et al., 2010). However, these studies focused on a set of validated drug-like PPI inhibitors regardless of their mode of inhibition. Small molecule PPI Inhibitors can be classified as orthosteric or allosteric modulators depending upon their mode of interaction (Buchwald, 2010). The former compete directly with hot spots at the interface (Wells, 1995) while the later can bind a cavity away from the interface, usually preventing conformational changes necessary for binding to the protein partner. In addition, in vivo, small molecules can also prevent the formation of a protein-protein complex through non-direct mechanisms. To target specifically PPI inhibitors directly interfering with the interface of protein-protein complexes, we have focused on those cases where the 3D structures of both the protein-protein and protein-ligand complexes have been characterized. This resulted in the freely accessible $2P2I_{DB}$ structural database (http://2p2idb.cnrs-mrs.fr) (Bourgeas et al., 2010).

In the present research article, we have analyzed the properties of the 39 unique small molecule inhibitors found in $2P2I_{DB}$ to define a general profile of orthosteric inhibitors. We propose an original protocol, $2P2I_{HUNTER}$, to filter out general chemical libraries using machine-learning approach. Models were built using support vector machine (SVM) with 11 standard Dragon molecular descriptors. The best models were tested externally on the only two representative PPI bioassays from the publically available PubChem bioassay database for biological activities of small molecules (http://pubchem.ncbi.nlm.nih.gov/). This external blind validation showed the ability of the SVM model to considerably reduce the size of the filtered chemical library up to 97% of compounds elimination as well as to enhance the proportion of active compounds by a factor up to 12. Finally, the $2P2I_{HUNTER}$ protocol was applied to 25 commercial libraries including three PPI targeted libraries, representing a total

of more than 8 million compounds. On average three percent of the molecules were selected by the SVM tool ranging from 0.2 to 7.9 percent for the 25 commercial libraries showing that it is highly effective to filter large chemical libraries. Interestingly, more than 70 percent of the compounds from a subset of the ChemDiv PPI library were selected showing that this library of new and unusual scaffolds in medicinal chemistry overlaps with the chemical space defined by the SVM tool.

Results and Discussion

Orthosteric PPI inhibitors follow a general 'Rule of four'

We have recently developed the 2P2I structural database (http://2p2idb.cnrs-mrs.fr) that gathers all known protein-protein and protein-inhibitor complexes for which both 3D structures are available (Bourgeas et al., 2010). There are 39 small molecule orthosteric inhibitors present in the 2P2I database. These PPI modulators correspond to 12 PPI targets that cover most SCOP (Structural Classification of Proteins) and CATH (Class Architecture Topology Homologous) database fold classes and corresponding to various topological spaces – primarily helix-based domains, beta-strand domains, mixed folding (helix/beta strand) and loop-binding groove domains (Morelli et al., 2011). PPI inhibitors have properties that distinguish them from other medicinal chemistry compounds as has been suggested by several studies (Higueruelo et al., 2009; Morelli et al., 2011; Pagliaro et al., 2004). The detailed analysis of the protein-protein interaction inhibitors in $2P2I_{DB}$ revealed key features that make them different than standard drugs. Average values for the molecular weight (540±126 $g.mol^{-1}$, thus MW >= 400 $g.mol^{-1}$), ALogP (3.9±2.1, thus ALogP >= 4), number of rings (4.5±1.0, thus #Rings >= 4) and number of hydrogen bond acceptors (6.6±2.0, thus #HBA >= 4) define the generic profile of a PPI inhibitor compound that could be further derived into a more specific inhibitor. The general profile based on a distribution of compounds led us to propose a 'rule-of-four' as a guideline to characterize this particular chemical space (Morelli et al., 2011). We have analyzed the percentage of molecules in the training dataset that follow the different molecular properties of the Ro4 as defined above (Table S1). We found that ninety one percent of the small molecule compounds in the 2P2I database had a MW over 400 $g.mol^{-1}$. Ninety one percent of the compounds had at least 4 rings and eighty seven percent had more than four hydrogen bond acceptors. Forty-four percent of the compounds had a LogP value greater than 4. Overall, eighty one percent of the small molecule inhibitors in 2P2I database obey Ro4 (allowing 1 violation). These overall properties raised the question about the capacity of these inhibitors to be used as drugs in medicinal chemistry.

PPI inhibitors as therapeutic agents

Molecular weight and lipophilicity of drugs are usually increased during lead optimization. As a consequence, HTS chemical libraries designed to search for hits generally contain low molecular weight compounds that have been filtered using Lipinski's 'Rule of Five' (Ro5) (Keller et al., 2006). PPI inhibitors are sometimes considered as an exception to the Ro5, mainly due to high MW and hydrophobicity of the compounds (Higueruelo et al., 2009; Pagliaro et al., 2004). However, more than half of the active compounds in $2P2I_{DB}$ are Ro5 compliant showing that they could possess ADME properties that are not incompatible with further developments as oral drugs (Table S1). Furthermore, some of the PPI small-molecule inhibitors, such as Nutline-3 analogs or Abbott navitoclax (former ABT-263) have progressed to early phase clinical trials as anticancer agents. Interestingly, the potent and orally bioavailable Bcl-2 family inhibitor ABT-263 was improved from the non-orally available predecessor ABT-737 by increasing the molecular weight (Tse et al., 2008). Similarly, ligand efficiency analyses have allowed us to show that PPI inhibitors could present pharmacokinetics profiles acceptable for oral dosing (Morelli et al., 2011). However, pharmaceutical companies interested in the development of protein–protein interaction inhibitors should also consider parallel technologies, such as nanoparticle drug delivery systems for encapsulation, stabilization and delivery of the potential drugs (Amstad and Reimhult, 2012; Morelli et al., 2011).

Structural diversity of PPI inhibitors

To guarantee structural diversity of the compounds, the 39 small molecules in $2P2I_{DB}$ were clustered on the basis of Tanimoto similarity criterion of 0.8 with the OptiSim algorithm implemented in the Tripos package leading to 32 non-redundant molecules that were used as the positive set in our learning approach (see 2D representation of molecules in supplementary material Figure S2). Selection of the decoy molecules (presumed to be inactive against PPI targets) is more complicated due to the impossibility to test the compounds against all known PPI targets. The human 'interactome' is estimated to number between ~130,000 [26] and ~650,000 [27] protein-protein interactions, and to be definitively validated as a non-PPI inhibitor, a compound would have to be tested against a significant number of these protein-protein complexes. The aim of this work is to develop a filtration tool able to increase the hit rate in a screening campaign (*i.e.* to accelerate and reduce the cost of hit

finding in screening campaigns). This can be achieved either by selecting the maximum number of true PPI inhibitors in a screening library or by removing the maximum number of non-PPI inhibitors or a combination of both. We selected a library of small drug-like molecules as decoy based on the ability to be separated from the population of the positive dataset for some molecular key descriptors that are known to characterize PPI inhibitors and particularly the predefined 'Ro4' rule (Higueruelo et al., 2009; Pagliaro et al., 2004). We compared the molecular properties of the validated PPI inhibitors and several chemical libraries and selected the NCI Diversity set II chemical library as decoy (1364 compounds). As for the positive dataset, compounds in the decoy were clustered using Tanimoto similarity comparisons of the UNITY fingerprints to remove redundancy. Compounds with poor PK properties and those for which it was not possible to calculate all molecular descriptors were also removed from this dataset. Less than seven percent of the compounds in this decoy were compliant with the 'Ro4' rule (Table S1). Among the four parameters, MW is the most discriminating between the two datasets. High molecular weight compounds (MW>=500) are usually removed from chemical libraries using Lipinsky 'rule-of-five' which results in very low success rates in screening campaigns against PPI targets. This is the case for the decoy, as indicated by the very high number of molecules following 'Ro5' (Table S1). The final training dataset selected for the SVM approach contains 32 active (PPI inhibitors) and 1018 inactive (non-PPI inhibitors) compounds. The chemical diversity of the positive dataset and decoy was assessed using a large set of 2D BCUT descriptors (Figure S1 in supplementary material). As expected, this analysis showed the lack of redundancy within each dataset. In addition, compounds from the positive dataset cover most of the chemical space defined by the decoy.

2P2I$_{HUNTER}$: a filtering protocol to design PPI libraries

To gain further insights into the selection of PPI inhibitors, support vector machine (SVM) models were built and optimized based on the positive dataset and decoy described above. We used classification SVM with a linear kernel and 5-fold internal cross-validation. After comparison of the spread of values and of various molecular properties between the positive dataset and decoy, we performed t-test to select standard molecular descriptors that allowed at least partial separation of the positive dataset and decoy. The final 11 molecular descriptors selected are listed in Table S2 of the supplementary material. SVM models were

optimized using Receiver Operating Characteristic (ROC) curve value to consider the imbalance ratio of active/inactive (1/31) (Li et al., 2009). This choice which represents a good compromise between specificity and sensitivity was also guided by the characteristics of the desired filtration tool.

The whole dataset was randomly split into five folds, with four folds used for training and the other fold for testing. This process was repeated 30 times and the final average performance was calculated. The best selected model resulted in 99% accuracy, 99% specificity and 62% sensitivity (Table 1). To validate the robustness of the model the PPI classification property (active or not) of the 32 PPI inhibitors from the positive dataset were randomly reordered (y scrambling). The average ROC AUC obtained over 20 successive Y-scrambling runs fell down from 0.96 with the actual activity labels to about 0.6 (the random threshold value being 0.5).

	Cost	Sigma	Accuracy	Sensitivity	Specificity	AUC of ROC
optimized SVM model	**1.5**	**0.02**	**0.99**	**0.62**	**0.99**	**0.98**

Table 1: Parameters of best SVM model and overall internal performances on the training dataset composed of 32 PPI inhibitors from the 2P2I database as a positive set and 1018 compounds as decoy.

PCA analyses of the training dataset

Separation of the positive and decoy compounds in the training dataset using the 11 molecular descriptors was assessed using PCA analyses (Figure 1A). Contrary to the BCUT

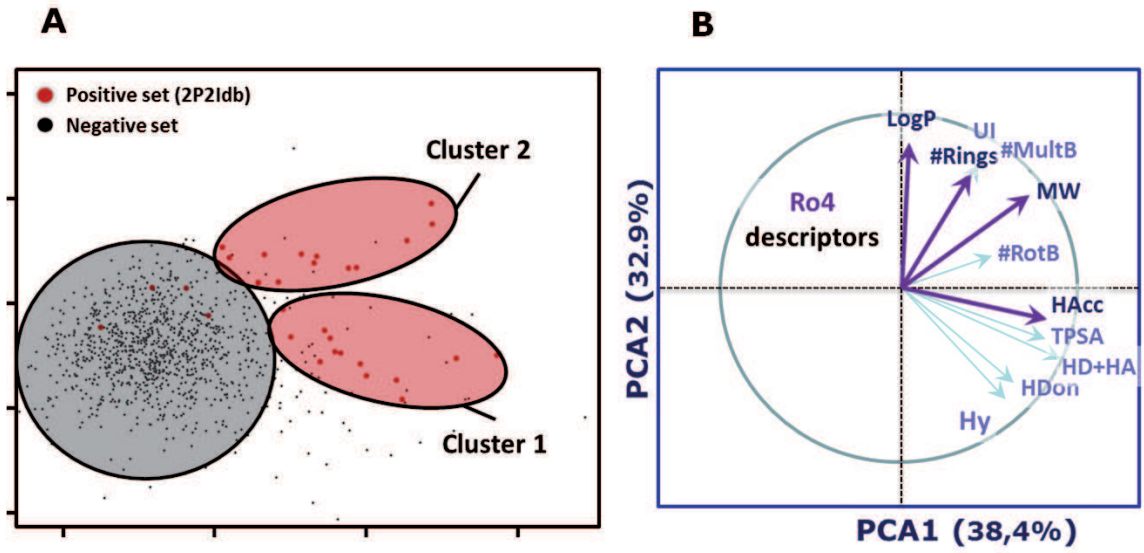


Figure 1 : 2D PCA (A) and variable contributions (B) for the training dataset. (A) The 32 PPI inhibitors from $2P2I_{DB}$ (positive set) and the 1018 compounds from Diversity are indicated as red and black spheres respectively. Most compounds from the decoy are clustered in the center left of the representation (black ellipse). Four PPI inhibitors are located in a region overlapping with the decoy. The remaining 28 compounds are split into two equally distributed clusters (red ellipses). (B) Distribution and contribution of the 11 Dragon molecular descriptors in the first two principal components. The first two principal axes represent more 71% of the data variance. Hydrogen bonds (Acceptor and Donor), TPSA and molecular weight play a major contribution in the first principal component whereas ALogP, the level of insaturation and the number of rings are the major contributors of the second principal component. Ro4 descriptors are highlighted; interestingly, HAcc and MW are among the major contributors in the first principal component and ALogP and the number of rings are important in the second dimension.

metrics, the use of the 11 DRAGON descriptors allows a good separation between true PPI modulators and decoy molecules. The first two components of the PCA account for more than 70% of the whole data variance. The most significant descriptors in the first PCA dimension correspond to the number of hydrogen bond donor and acceptors (Figure 1B). Topological

polar surface area and molecular weight also have important contributions. The second dimension of the PCA analysis is dominated by ALogP, the level of unsaturation and the number of rings. Altogether, these principal components represent the four parameters that define our 'Ro4' (Molecular weight, ALogP, number of rings and Hydrogen bond acceptors all contribute significantly to the first or second principal component as shown in Figure 1B).

Most compounds from the decoy are clustered together (black ellipse in Figure 1). The 32 PPI inhibitors selected as positive dataset occupy three cluster regions in the PCA representation. Four PPI compounds (IQZ, 976, CL3 and WAC) are located in the center of the region occupied by the decoy. They correspond to low molecular weight compounds (average MW=335.05±68.45 $g.mol^{-1}$). It is noteworthy to mention that three of these compounds (IQZ, 976 and CL3) do not follow the Ro4, mainly due to their low MW (IQZ:MW=240; CL3:MW=343; 976:MW=353). The fourth compound (WAC) only barely obeys Ro4 (MW=403). Three of these small molecules (IQZ, CL3 and WAC) correspond to ZipA/FtsZ inhibitors that have been described to bind on the protein surface in an unconventional way [17,26] and exhibit low binding affinity constants compared to other PPI inhibitors (kd 10-100μM). Compound 976 was found to disrupt the Integrase/LEDGF complex (IC50 1.4 μM). Compound 723, the other Integrase/LEDGF complex inhibitor (IC50 12.2 μM) had been removed from the positive dataset due to its chemical similarity with 976. When represented in the PCA analysis, this compound was also found in the middle of the decoy.

The remaining 28 PPI compounds are split between two equal size regions (red ellipses in Figure 1) that show little overlap with the decoy. The 14 compounds from cluster 1 (bottom red ellipse in Figure 1) correspond almost exclusively to inhibitors of the XIAP and IL-2 families. They are characterized by lower average values for ALogP (2.1±1.0) and molecular weight (525.3±77.2 $g.mol^{-1}$) and more hydrogen bond donors (5.0±1.7). All these compounds follow the Ro4. The 14 compounds from the second cluster (top red ellipse in Figure 1) are disruptors of families $Bcl\text{-}_{XL}$, MDM2, MDM4, HPVE2/E1, TNF and TNFR1A. They are characterized by higher average values for ALogP (5.7±1.3), slightly higher molecular weight (597.7±84.1 $g.mol^{-1}$) and less hydrogen bond donors (1.4±0.8). Only two of these compounds did not follow the Ro4.

External validation of the SVM model: Bioassay selection

The freely accessible PubChem BioAssay database (http://pubchem.ncbi.nlm.nih.gov) contains comprehensive information of small molecules and their biological activities (Li et al., 2009; Wang et al., 2010; Xie, 2010). It contains experimental descriptions and biological test results for more than 600,000 bioassays and therefore brings great opportunities for academic researchers in the fields of chemical biology, medicinal chemistry and chemoinformatics. We used an advanced query to retrieve relevant protein-protein bioassays to evaluate our SVM models with external data (Figure S3). We first selected the series of bioassays grouped by AID summaries corresponding to protein-protein interaction using a keyword search. This step led to 28 different bioassays (Figure S3). Among these, we selected hits for which both primary and secondary (dose response) assays were available (17 bioassays). At this stage bioassays were inspected manually and only those truly corresponding to protein-protein interactions were selected (5 bioassays). To make sure that the inactivity was not due to membrane barrier crossing, cytotoxicity effects or metabolic conversion of the compounds in the cell, only *in vitro* bioassays were selected (3 bioassays). The SVM model is a generalistic model therefore we did not take into account bioassays dedicated to the development of target-specific inhibitors. For example, several bioassays correspond to the development of selective inhibitors of myeloid cell leukemia sequence 1 (MCL1). In these series of bioassays the authors have selected compounds that are not active on other Bcl-2 proteins such as Bcl-$_{XL}$. Since our model was trained with modulators of the Bcl-$_{XL}$ family it is not expected to perform well for these types of bioassays. Overall two sets of bioassays were selected (AID 1645 and 1683, Figure S3 and Table S3). AID 1496 and 1438 correspond to the primary and secondary HTS identification of compounds inhibiting the binding between the RUNX1 Runt domain and the Core-binding factor, beta subunit. AID 1531 (primary screen) and 1892/1897 (secondary screens) correspond to HTS Assay for modulators of MEK Kinase PB1 Domain interactions via MEKK5.

Blind validation of 2P2I$_{HUNTER}$ with selected bioassays

The ability to enhance hit rates in screening campaigns using Ro4 and the best SVM model as filtering tools was tested on the two high-throughput screening compound bioassays selected from PubChem. Overall performances for the different models are summarized in Table 2. Confusion matrices, ROC curves and overall performances in terms of accuracy,

sensitivity and specificity are shown in Figure 2. To quantify the performance of the SVM approach, we used the enrichment factor metric which compares the performance of the model to what would be expected if the compounds were randomly selected. The enrichment factor is given by the ratio between the hit rate obtained with the whole chemical library and the hit rate for the filtrated library. On the basis of this metric, we observe a significant improvement of the HTS performances for both bioassays

AID	Tested	Active	Hit Rate[a] (%)	Model	Selected[b]	TP[c]	Hit Rate[d] (%)	EF[e]	%selected[f]
1531	289,475	97	0.033	Ro4	58,388	53	0.09	2.7	20
				SVM	7,621	27	0.35	10.6	3
				Ro4+SVM	5,145	22	0.43	12.8	2
1496	215,676	45	0.021	Ro4	44,009	24	0.055	2.6	20
				SVM	4,962	4	0.081	3.9	3
				Ro4+SVM	3,573	4	0.112	5.4	2

Table 2: Overall performances of Ro4 and SVM models for the two available PPI bioassays. [a]Experimental hit rate. [b]Number of molecules selected by the model at a probability threshold of 0.5. [c]Number of True positive compounds selected. [d]Hit rate of the model. [e]Enrichment factor. [f]Percentage of molecules selected from the initial library.

With both external validation datasets, the application of the 'Ro4' allows to select about 20% of the compounds from the starting chemical library with a significantly improved hit rate (enrichment factor around 3). Therefore, 'Ro4' constitutes a simple and very fast method to design chemical libraries enriched in PPI inhibitors. The SVM model showed a high accuracy and specificity for both validation datasets with values of 0.98 and 0.97 for AID 1496 and 1531 respectively compared to 0.8 for the Ro4 filtering (Figure 2). Although these high values are probably related to the relative imbalanced ratio of active and inactive compounds in the training dataset, they clearly indicate that the SVM model can be used as a highly efficient tool to remove non-PPI compounds from large chemical libraries. The true positive rate as measured by the sensitivity is lower for the SVM with values of 8.9% and 27.8% for AID 1496 and 1531 respectively. Ro4 exhibits a higher sensitivity than the SVM

model, which shows that it is able to retrieve more true active compounds in absolute numbers however; the SVM model is more stringent.

A

SVM				
	Actives		Non actives	
	TP	FN	TN	FP
AID1496	4	41	210,637	4,958
AID1531	27	70	281,748	7,594

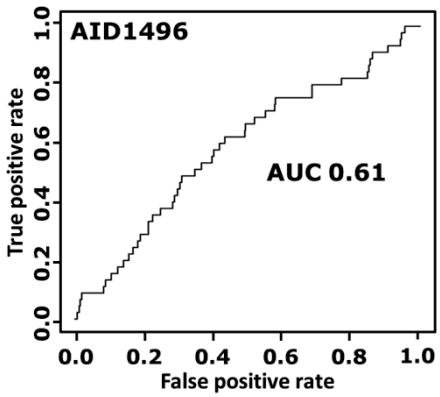
	accuracy	sensitivity	specificity
AID1496	0.98	0.09	0.98
AID1531	0.97	0.28	0.97

B

Ro4				
	Actives		Non actives	
	TP	FN	TN	FP
AID1496	24	21	171,642	43,989
AID1531	53	44	231,043	58,335

	accuracy	sensitivity	specificity
AID1496	0.80	0.53	0.80
AID1531	0.80	0.55	0.80

C



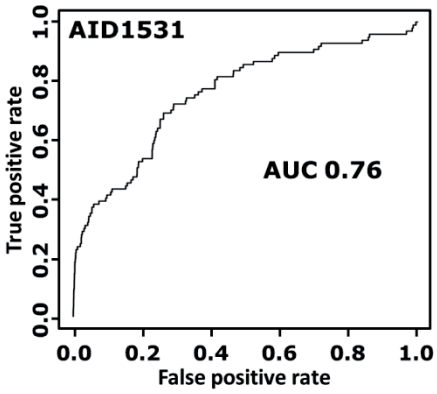


Figure 2 : Confusion matrices, performances and ROC curves of the SVM model for the two external validation datasets. A: Repartition of the molecules according to the SVM model prediction (TP True positive, FN False Negative, TN, True negative, FP False positive). B: Overall performances of the Ro4 model in terms of accuracy, sensitivity and specificity. ROC curves and Area Under the Curve (AUC) are shown for AID 1496 (C) and 1531 (D).

As mentioned before, this model is able to eliminate a large number of non-PPI compounds therefore leading to an increase in the hit rate up to 10. Interestingly, most true positive compounds selected by the SVM model follow the 'Ro4' (82% in the case of AID1531 and

100% in the case of AID1496). This prompted us to combine the two approaches to refine the filtering process. We applied the Ro4 filtering tool to the compounds selected by the SVM model. The results in Table 2 showed a significant increase in the hit rate compared to the SVM model due to a smaller number of selected compounds and almost the same number of true positive PPI inhibitors predicted. Molecules selected by the SVM model for both validation bioassays overlap the region defined by the positive dataset as shown by PCA analysis (Figure S4). Most false negative molecules are located in the middle of the decoy. However, a series of 12 highly related compounds is found in a totally different chemical space (insert Figure S4B). These Ro4 compliant inhibitors are characterized by molecular properties that make them invisible by our SVM model (MW=1063.1±42.0 $g.mol^{-1}$; #Rings=8.1±1.2; HAcc=17.8±0.9; Number of rotatable bonds 27.3±0.9).

AID1531 led to better overall enrichment factors indicating that active compounds in this bioassay are closer to the 2P2I chemical space as can be seen on the representation of the two principal components (Figure S4). X-ray structures were available for the unbound Runx1/Runt domain (PDB code 1EAN) and for the core binding factor beta (PDB code 2JHB) corresponding to the protein targets in bioassay AID1496. However, no structural information was available for the protein targets in AID1531 or for the protein-protein complexes. Therefore, we were not able to compare MEKKK2/MEKK5 and RuntX/CBFb complexes to those present in $2P2I_{DB}$ to determine whether or not the difference between the two bioassays was related to the structure of the protein-protein interfaces.

Chemical diversity of selected molecules

Structural diversity of the selected molecules with the SVM models was checked (using a Tanimoto index of 0.8) and showed that on average they do not share the same scaffold. In the case of AID1531, the 27 true positive compounds corresponded to 23 different classes after the similarity clustering whereas all 4 true positive compounds in AID1496 were dissimilar indicating that this SVM protocol did not retrieve a series of compounds from the same family.

Blind validation with in house PPI bioassays

Overall performance of $2P2I_{HUNTER}$ was also assessed using results from *in house* PPI bioassays. Prior to the present study, a high-throughput screen against a PPI target had been performed by the industrial partner to identify PPI modulators using FRET assay in the primary screening (*to be confirmed by IdRS*) and IC50 (*to be confirmed by IdRS*) for the secondary dose response confirmatory screening which led to a low 0.004% success rate (Table 3A). Application of the SVM protocol to the whole collection of compounds used in the HTS campaign selected 7.2% of the molecules corresponding to the filtrated library (Table 3).

A

Assay	Hit Rate[a] (%)	SVM[b] (%)	Hit Rate SVM[c] (%)	EF
HTS	0.004	7.2	0.025	7.0

B

Assay	PPI modulators (%)
Protein/Peptide (Primary screen)	1.9
Protein/protein (Primary screen)	0.7
Protein/Peptide (Dose response)	4.1
Protein/Protein (Dose Response)	1.9

A significant improvement was observed in the hit rate for this resulting filtrated chemical library (0.025%) corresponding to an enrichment factor of 7. It is especially noteworthy that fifty percent of the active compounds found in the HTS campaign were present in the filtrated library. In a parallel approach, results from a large series of *in house* protein-protein and protein-peptide related primary or dose response screenings were pooled together which allowed us to search for PPI modulators in the filtrated library (Table 3B). We

Table 3: Overall performances of SVM model for *in house* bioassays. A: High Throughput Screening of PPI target. [a]Experimental hit rate. [b]Number of molecules selected by the model at a probability threshold of 0.5. [c]Hit rate of the model. [d]Enrichment factor. **B:** Cumulative percentage of true PPI modulators found on a set of *in house* PPI targets.

particularly looked for evidence of activity in PPI modulation bioassays for each compound in the filtrated library. We were not able to calculate enrichment factors as the bioassays were not conducted on the whole non-filtered collection of compounds; however we could estimate the percentage of true PPI modulators in the filtrated library (Table 3B) which ranged from 0.7 to 4.1%. Not all compounds in the filtrated library have been tested in the various bioassays so the percentages are underestimated. Higher percentage of predicted PPI modulators were observed for protein-peptide related bioassays which could indicate that the SVM protocol performs better for these families of targets.

Applicability domain of 2P2I$_{HUNTER}$ SVM filtration tool

Our SVM filtration protocol, 2P2I$_{HUNTER}$ was applied on a set of 25 commercial screening libraries of compounds. The complete list of libraries corresponding to more than 8 million compounds is shown in Table S4. The percentage of compounds selected by the SVM model varies from 0.18 to 7.9 representing a rate of more than 40 times (data not shown). This result clearly indicates that all libraries do not share the same chemical space and therefore should lead to various success rates in drug discovery for PPI targets. Except for the PPI library from ChemDiv, most of the available commercial libraries have been filtered with Lipinski-like rules (MW $\leq$ 500, clogP $\leq$ 5, tPSA $\leq$ 100, rotatable bonds $\leq$ 8, hydrogen bond acceptors $\leq$ 10 and hydrogen bond donors $\leq$ 5) as indicated by the high percentage of Ro5 compliant compounds within them (Table S4). Interestingly, the commercial library with the lowest percentage of selected compounds by our SVM model (ASDI global collection) also corresponds to the library with the highest percentage of Ro5 compliant molecules (99.6%) and in general commercial libraries with the lowest percentage of selected compounds (<1.5%) all correspond to libraries with very high percentages of Ro5 compliant molecules (>92%).

The 123,000 compounds from ChemDivPPI have been subdivided by the provider into 11 subsets based on their molecular properties (Table 4). Compared to all other chemical libraries tested, the percentage of selected compounds is extremely high for the eccentric subset (~73%) and still quite high for Cyclic Ugi and mdm2 focused (~15% and 7% respectively). The selected compounds from the eccentric subset cover almost the entire PPI chemical space described by our SVM model (Figure 3B). This subset contains about 7,000

compounds that were grouped together based on their original scaffolds outside the field of (hetero)aromatic systems and their low toxicity (Marson, 2011). Compounds from the Cyclic Ugi subset overlap with 50% of PPI inhibitors in 2P2I database (Figure 3C). Compounds selected from the mdm2 focused subset cover a smaller chemical space that encompass most mdm2/p53 inhibitors found in $2P2I_{DB}$ (Figure 3D). Thus, the 2P2I chemical space herein identified, is able to select new and/or unusual aromatic scaffolds that exhibit low toxicity and lie outside the field of extensively used benzenoid and heteroaromatic ring systems (Marson, 2011). Some of these systems are present in agents with therapeutic potential and others are found in agents already in clinical use.

Compounds Series	% selected compounds (SVM)
Total unique compounds (123,001)	6.9
Eccentric	71.6
Cyclic Ugi	14.6
mdm2 focused	7.4
Recognition elements	2.0
Hedgehog pathway focused	1.9
Nonpeptide Peptidomimetic	2.3
PDZ domain focused	2.5
Spiro	1.2
3D Mimetic	1.1
Shape (Helix) Mimetics	0.8
Beyond Flatland	0.1

Table 4 : Percentage of compounds selected by the SVM model in the different subsets of the ChemDiv PPI library.

Figure 3

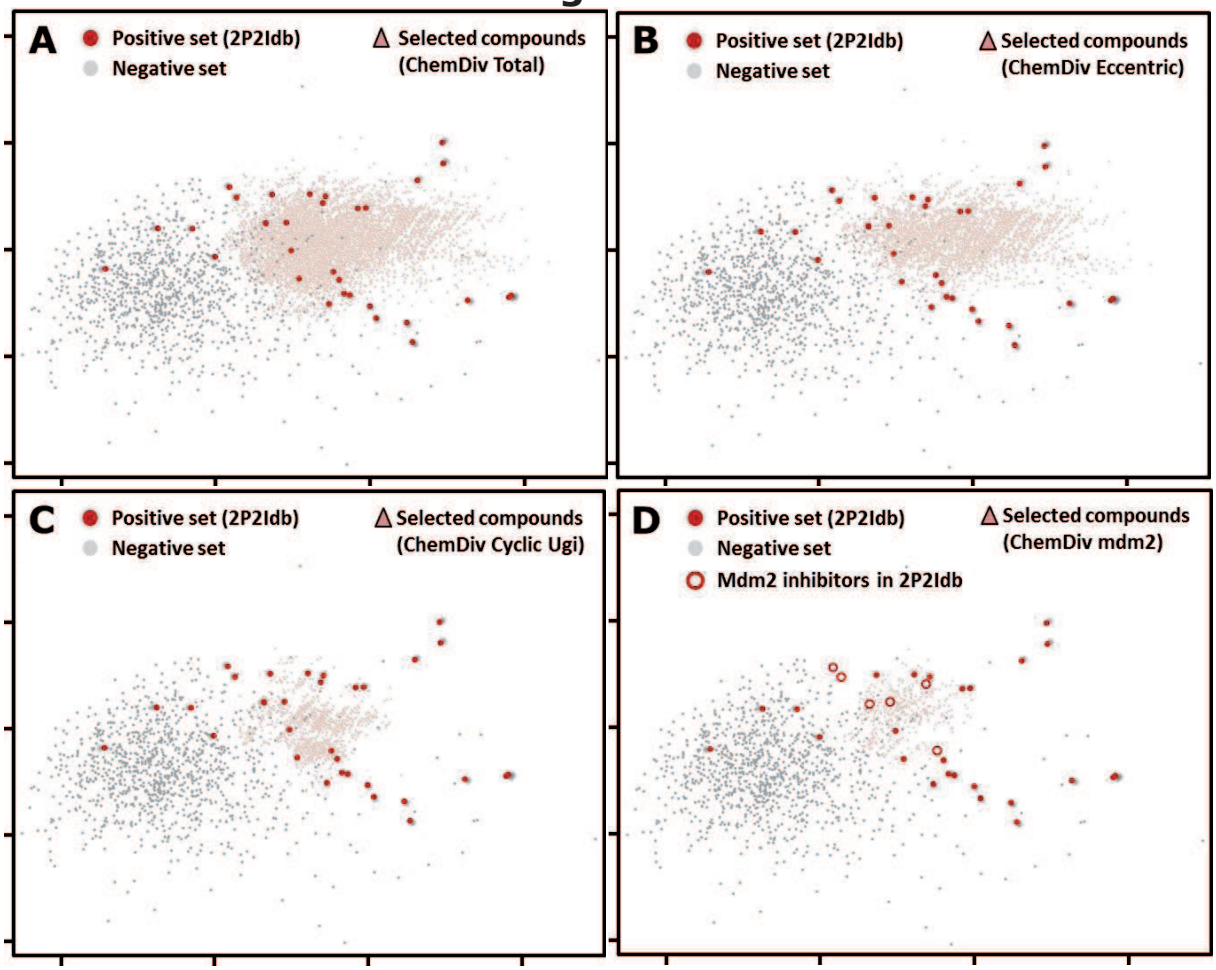


Figure 3 : 2D representation of ChemDiv subsets in the first two principal components of the training set chemical space. (A) Total set of 123,001 compounds, (B) Eccentric, (C) Cyclic Ugi and (D) mdm2 focused subsets. Mdm2 inhibitors found in the 2P2I database are indicated as unfilled red circles.

Conclusion

In the post-genomic era identification of complete networks of protein-protein interactions within a cell has open the way to major breakthroughs in understanding biological pathways, host–pathogen interactions and cancer development and has led to the identification of novel therapeutic drug targets (Lievens et al., 2010). Protein-protein interactions (PPIs) represent promising new class of attractive therapeutic targets and the advancement in drug discovery efforts against PPIs has been recently referred as "the unmined biology gold reserve" [4]. However, this class of targets is still considered as extremely difficult for targeting by small-molecules due to the structural characteristics of the interface and specific strategies needs to be undertaken to tackle this particularly challenging class of drug targets. Successes in drug discovery developments against PPI targets face two major issues, druggability assessment (or target selection) and adequacy of the chemical libraries used for screening. Our strategy to address these questions was to focus on orthosteric PPI modulators therefore we only considered PPIs for which structural information was available to validate that the inhibitor was interacting at the interface. We have recently developed the hand curated structural 2P2I database by collecting information about protein-protein interfaces for which both the protein-protein and protein-inhibitor complexes have been structurally characterized and we identified key descriptors of protein-protein interactions with a known inhibitor (Bourgeas et al., 2010). Although this database is rather small at the moment, the hope is that as it grows patterns will emerge for both protein-protein interfaces and small molecule inhibitors.

Here we report the analysis and characterization of the chemical space of PPI modulators present in the 2P2I database (http://2p2idb.cnrs-mrs.fr) in order to provide tools to build focused chemical libraries dedicated to PPI targets. We propose a "Rule-of-Four" to define the generic profile of PPI modulators (MW >=400 Da, ALogP >=4, number of rings >=4 and number of hydrogen bond acceptors >=4) [48] which is not in complete contrast with the well-known Lipinski's "Rule-of-Five" (Lipinski et al., 2001). We have developed a general protocol to filter chemical libraries using SVM approaches. This $2P2I_{HUNTER}$ filtration tool has been validated using external bioassays from PubChem and *in house* screening results and showed its potential to significantly increase success rates leading to enrichment factors of up to 12.8. The protocol has been applied to a set of 25 commercial libraries corresponding to the major providers and in all cases it showed its ability to dramatically decrease the size of the

resulting focused library (on average 97 percent of the compounds were removed from the original library). However, more than 70 percent of the 7,076 compounds from the ChemDiv eccentric PPI library were selected as putative PPI modulators by the $2P2I_{HUNTER}$ filtration tool.

The $2P2I_{HUNTER}$ protocol is currently used to build *in house* chemical libraries that will be tested on druggable PPI targets selected based on interface structural properties (Bourgeas et al., 2010). $2P2I_{HUNTER}$ represents a useful tool for both academics and pharmaceutical companies to expand the scope of chemical libraries dedicated to protein-protein targets and to enhance hit rates in high throughput experiments thus reducing the cost of screening campaigns.

Material and Methods

Training Dataset collection and preparation

Ligand preparation

A standard ligand preparation protocol was applied using Chemaxon tools. Briefly, molecules were first checked for errors in valence, coordination, aromaticity and covalently bound counter ions. Molecules were then standardized as follows: the largest fragment was kept, explicit hydrogens were removed, molecules were dearomatized, aromatized and neutralized, explicit hydrogens were added and finally the structures were cleaned in 2D. Major species at physiological pH 7.4 were then determined using cxcalc module.

Molecular Descriptors

The 10 molecular descriptors used for machine learning process (molecular weight; number of multiple bonds; number of rings; number of rotatable bonds; number of donor atoms for H-bonds; number of acceptor atoms for H-bonds; unsaturation count; hydrophilic factor; topological polar surface area and ALogP) were computed with DRAGON version 6 (http://www.talete.mi.it) on the charged compounds prepared with ChemAxon as described above. The 11th descriptor corresponded to the sum of hydrogen bond donors and acceptors. Molecules that generated errors in descriptors computation were discarded.

BCUT Descriptors

The DiverseSolution module from Tripos was used to compute BCUT based metrics. The program performs a partitioning of the BCUT chemical space and determines which molecules from the studied set occupy each mapped cell. Cells coordinates were retrieved and submitted to a Principal Components analysis.

Bioassay selection

Bioassays were retrieved from the PubChem Bioassay website that provides searchable descriptions of more than 600,000 bioassays (http://www.ncbi.nlm.nih.gov/pcassay). The detailed query is shown in supplementary material Figure S3. Briefly, the keyword "protein protein" was search in AID corresponding to bioassay type summary. Then bioassays with both a confirmatory and primary screenings were selected. Finally, bioassays were grouped by type summary. The 17 remaining bioassays were checked manually. Bioassays corresponding to enzymatic assays or cell-based were discarded leaving only 3 AID.

SVM model construction and validation

Support Vector Machine algorithm implemented in the statistical software package R was used for model training (Cortes and Vapnik, 1995). Original data were mapped through a kernel function on to a higher dimensional space where the two sets are more easily separable with a linear classifier. The Caret library in R environment was used to perform SVM modeling. The positive or negative activity of our training set was typed as a factor variable to specify the classification mode of the problem. The Radial Gaussian Basis (RGB) method was chosen as kernel function. The training is performed through a 5-fold cross-validation procedure repeated 30 times with a random selection of the training and test sets at each time. The final optimal model was selected according to the average best performance of the ROC AUC (Area Under Curve) statistical metric. The model training in RGB method consists in tuning two hyperparameters cost C and scaling function sigma along a grid of candidate values. The best final model were found with C=1.5 and sigma=0.02.

The enrichment factor which corresponds to the ratio of hit rates after and before filtration with the SVM model is given by the following formula:

$$EF = \frac{Hit_{selected}/N_{selected}}{Hit_{Total}/N_{total}} = \frac{Hit\ Rate_{selected}}{Hit\ Rate_{Total}}$$

Where EF is the enrichment factor obtained with the SVM filtered chemical library, $Hit_{selected}$ is the number of active compounds in the selected library, Hit_{total} is the total number of active molecules in the library, $N_{selected}$ is the number of compounds selected by

the SVM model, and N_{total} is the total number of compounds in the original (non-filtered) library.

Y-Scrambling

In order to assess the reliability of the model, activity labels of the training set were randomly re-ordered by keeping positive/negative ratio. The two parameters cost-sigma remained unchanged and the same training configuration was applied, 5-fold cross-validation repeated 30 times.

PCA analyses

Principal Component Analysis (PCA) was performed using the library FactoMineR in R package (Josse, 2008). All the 11 molecular descriptors were used to derive the principle components that defined the 2P2I chemical space, calculated on the whole training dataset. For the PCA of the external validation datasets and commercial libraries, properties of molecules were projected onto the first two principal components of the training set.

References

1. Roche P, Morelli X (2010) Protein-Protein Interaction Inhibition (2P2I): Mixed Methodologies for the Acceleration of Lead Discovery. In: Miteva M, editor. *In silico* lead discovery: Bentham. pp. 118-143.
2. Morelli X, Bourgeas R, Roche P (2011) Chemical and structural lessons from recent successes in protein-protein interaction inhibition (2P2I). Curr Opin Chem Biol 15: 475-481.
3. Wells J, McClendon C (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature 450: 1001-1009.
4. Mullard A (2012) Protein–protein interaction inhibitors get into the groove. Nature Reviews Drug Discovery 11: 173-175.
5. Wilson AJ (2009) Inhibition of protein-protein interactions using designed molecules. Chem Soc Rev 38: 3289-3300.
6. Wilson CG, Arkin MR (2011) Small-Molecule Inhibitors of IL-2/IL-2R: Lessons Learned and Applied. Curr Top Microbiol Immunol 348: 25-59.
7. Berg T (2008) Small-molecule inhibitors of protein-protein interactions. Current Opinion in Drug Discovery & Development 11: 666-674.
8. Patel S, Player MR (2008) Small-molecule inhibitors of the p53-HDM2 interaction for the treatment of cancer. Expert Opinion on Investigational Drugs 17: 1865-1882.
9. Vu BT, Vassilev L (2011) Small-Molecule Inhibitors of the p53-MDM2 Interaction. Curr Top Microbiol Immunol 348: 151-172.
10. Wells JA (1995) Structural and functional epitopes in the growth hormone receptor complex. Bio/Technology (Nature Publishing Company) 13: 647-651.
11. Dudkina AS, Lindsley CW (2007) Small molecule protein-protein inhibitors for the p53-MDM2 interaction. Current Topics in Medicinal Chemistry 7: 952-960.
12. Fry D (2011) Small-Molecule Inhibitors of Protein-Protein Interactions; Vassilev L, Fry D, editors. Nutley, New Jersey: Springer. 184 p.
13. Arkin MR, Whitty A (2009) The road less traveled: modulating signal transduction enzymes by inhibiting their protein-protein interactions. Current Opinion in Chemical Biology 13: 284-290.
14. Koes DR, Camacho CJ (2011) Small-Molecule Inhibitor Starting Points Learned From Protein-Protein Interaction Inhibitor Structure. Bioinformatics.
15. Azmi AS, Wang Z, Philip PA, Mohammad RM, Sarkar FH (2011) Emerging Bcl-2 inhibitors for the treatment of cancer. Expert Opin Emerg Drugs 16: 59-70.
16. Blazer LL, Neubig RR (2009) Small molecule protein-protein interaction inhibitors as CNS therapeutic agents: current progress and future hurdles. Neuropsychopharmacology. United States. pp. 126-141.
17. Fuller JC, Burgoyne NJ, Jackson RM (2009) Predicting druggable binding sites at the protein-protein interface. Drug Discovery Today 14: 155-161.

18. Bourgeas R, Basse M-J, Morelli X, Roche P (2010) Atomic Analysis of Protein-Protein Interfaces with Known Inhibitors: The 2P2I Database. PLoS ONE 5: e9598.

19. Wanner J, Fry DC, Peng Z, Roberts J (2011) Druggability assessment of protein-protein interfaces. Future Med Chem 3: 2021-2038.

20. Sugaya N, Ikeda K, Tashiro T, Takeda S, Otomo J, et al. (2007) An integrative in silico approach for discovering candidates for drug-targetable protein-protein interactions in interactome data. BMC Pharmacology 7: 10.

21. Sugaya N, Ikeda K (2009) Assessing the druggability of protein-protein interactions by a supervised machine-learning method. BMC Bioinformatics 10: 263.

22. Sugaya N, Furuya T (2011) Dr. PIAS: an integrative system for assessing the druggability of protein-protein interactions. BMC Bioinformatics 12: 50.

23. Kozakov D, Hall DR, Chuang GY, Cencic R, Brenke R, et al. (2011) Structural conservation of druggable hot spots in protein-protein interfaces. Proc Natl Acad Sci U S A 108: 13528-13533.

24. Thangudu RR, Bryant SH, Panchenko AR, Madej T (2012) Modulating protein-protein interactions with small molecules: the importance of binding hotspots. J Mol Biol 415: 443-453.

25. Geppert T, Hoy B, Wessler S, Schneider G (2011) Context-based identification of protein-protein interfaces and "hot-spot" residues. Chem Biol 18: 344-353.

26. Fry DC (2006) Protein-protein interactions as targets for small molecule drug discovery. Biopolymers 84: 535-552.

27. Bunnage ME (2011) Getting pharmaceutical R&D back on target. Nat Chem Biol 7: 335-339.

28. Pagliaro L, Felding J, Audouze K, Nielsen SJ, Terry RB, et al. (2004) Emerging classes of protein-protein interaction inhibitors and new tools for their development. Current Opinion in Chemical Biology 8: 442-449.

29. Higueruelo AP, Schreyer A, Bickerton GRJ, Pitt WR, Groom CR, et al. (2009) Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. Chemical Biology & Drug Design 74: 457-467.

30. Neugebauer A, Hartmann RW, Klein CD (2007) Prediction of protein-protein interaction inhibitors by chemoinformatics and machine learning methods. Journal of Medicinal Chemistry 50: 4665-4668.

31. Reynes C, Host H, Camproux AC, Laconde G, Leroux F, et al. (2010) Designing focused chemical libraries enriched in protein-protein interaction inhibitors using machine-learning methods. PLoS Comput Biol 6: e1000695.

32. Sperandio O, Reynès C, Camproux A, Villoutreix B (2010) Rationalizing the chemical space of protein-protein interaction inhibitors. Drug Discov Today 15: 220-229.

33. Buchwald P (2010) Small-molecule protein-protein interaction inhibitors: therapeutic potential in light of molecular size, chemical space, and ligand binding efficiency considerations. IUBMB Life 62: 724-731.

34. Keller TH, Pichota A, Yin Z (2006) A practical view of 'druggability'. Curr Opin Chem Biol 10: 357-361.

35. Tse C, Shoemaker AR, Adickes J, Anderson MG, Chen J, et al. (2008) ABT-263: a potent and orally bioavailable Bcl-2 family inhibitor. Cancer Res 68: 3421-3428.

36. Amstad E, Reimhult E (2012) Nanoparticle actuated hollow drug delivery vehicles. Nanomedicine (Lond) 7: 145-164.

37. Li Q, Wang Y, Bryant SH (2009) A novel method for mining highly imbalanced high-throughput screening data in PubChem. Bioinformatics 25: 3310-3316.

38. Wang Y, Bolton E, Dracheva S, Karapetyan K, Shoemaker BA, et al. (2010) An overview of the PubChem BioAssay resource. Nucleic Acids Res 38: D255-266.

39. Xie XQ (2010) Exploiting PubChem for Virtual Screening. Expert Opin Drug Discov 5: 1205-1220.

40. Marson CM (2011) New and unusual scaffolds in medicinal chemistry. Chem Soc Rev 40: 5514-5533.

41. Lievens S, Eyckerman S, Lemmens I, Tavernier J (2010) Large-scale protein interactome mapping: strategies and opportunities. Expert Rev Proteomics 7: 679-690.

42. Lipinski C, Lombardo F, Dominy B, Feeney P (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46: 3-26.

43. Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20: 273-297.

44. Josse J (2008) FactoMineR: An R Package for Multivariate Analysis. Journal Of Statistical Software 25: 1-18.

Supplementary Material

Table S1

	MW	nRing	HbAc	LogP	Ro4[a]	Ro5[b]
[a]PPI inhibitors	91	91	87	44	81	59
SVM decoy	5	16	63	19	<7	99

Table S1: Percentage of compounds that satisfy the different parameters of Ro4 in $2P2I_{DB}$ and in the decoy selected for the SVM. [a]Percentage of compounds that follow Ro4. [b]Percentage of compounds that follow Ro5.

Table S2

DRAGON Molecular descriptor	Description
ALOGP	Ghose-Crippen octanol-water partition coefficient
Hy	Hydrophilic factor
MW	Molecular weight
nBM	Number of multiple bonds
nCIC	Number of rings
nHAcc	Number of acceptor atoms for H-bonds (N,O,F)
nHDon	Number of donor atoms for H-bonds (N and O)
nHDHA	Sum of acceptors and donors (nHAcc+ nHDHA)
RBN	Number of rotatable bonds
TPSA	Topological polar surface area using N,O,S,P polar contributions
Uc	Unsaturation count

Table S2 : List of Dragon molecular descriptors used in the SVM model

Table S3

PPI	AID	Assay
Core-binding factor, beta subunit isoform 1/Runt-related transcription factor 1 (AID 1645)	1496	Primary Screen (FRET)
	1438	Dose Response Confirmation
MEKKK2/MEKK5 (AID 1683)	1531	Primary Screen (FRET)
	1892	Single Concentration Confirmation Screen
	1897	Dose Response Confirmation

Table S3: Description of the two high-throughput screening PPI bioassays selected from Pubchem for external validation of the SVM model.

Table S4

Library	Collection	# Compounds	%Ro5	website
ASDI	Global Collection	16 656	99.6	http://www.frontierssi.com/
ASDI	HTS	65 119	90.3	http://www.frontierssi.com/
Asinex	Merged Libraries	436 012	91.2	http://www.asinex.com/
AsisChem	Screening Libraries	234 509	86.7	http://www.asischem.com/
ChemBridge	EXPRESS-Pick Collection	503 803	94.9	http://www.chembridge.com
ChemDiv	Discovery Chemistry collection	668 052	84.2	http://www.chemdiv.com
ChemDiv	**PPI**	35 829	27.4	http://www.chemdiv.com
DrugBank	FDA Approved Drugs	1 412	84.7	http://www.epa.gov/ncct/dsstox/
Enamine	**PPI**	363	93.4	http://www.enamine.net
IBScreen	Screening Collection	441 574	89.3	http://www.ibscreen.com/
IBScreen	Natural Compounds	47 229	84.1	http://www.ibscreen.com/
Innovapharm	synthetic organic compounds	316 334	85.7	http://www.innovapharm.com.ua/
LaboTest	OnStock	106 743	91.1	http://www.labotest.com/
LifeChemicals	**PPI**	31 143	90.1	http://www.lifechemicals.com/
MayBridge	Screening Collection	55 717	91.9	http://www.maybridge.com/
MayBridge	HitFinder	14 400	93.9	http://www.maybridge.com/
NCI	Screening compounds	377 581	95.4	http://dtp.nci.nih.gov/
Otava	Tangible compounds library	487 428	98.1	http://www.otavachemicals.com/
Pharmeks	Main	259 523	83.2	http://www.pharmeks.com/
Princeton Biomolecular	Express Stock	794 160	88.2	http://www.princetonbio.com/
Specs	Screening Compounds	203 434	85.9	http://www.specs.net/
TimTech	ActiMol Collection (HTS)	97 721	93.5	http://www.timtec.net/
Uorsy	Screening Compounds	1 643 662	73.0	http://www.ukrorgsynth.com/
Vitas	HTS Compounds	1 101 503	87.7	http://www.vitasmlab.com/
Zinc	Natural Products	225 118	84.0	http://zinc.docking.org/

Table S4 : List of commercial libraries assessed for their ability to screen PPI targets.

Figure S1

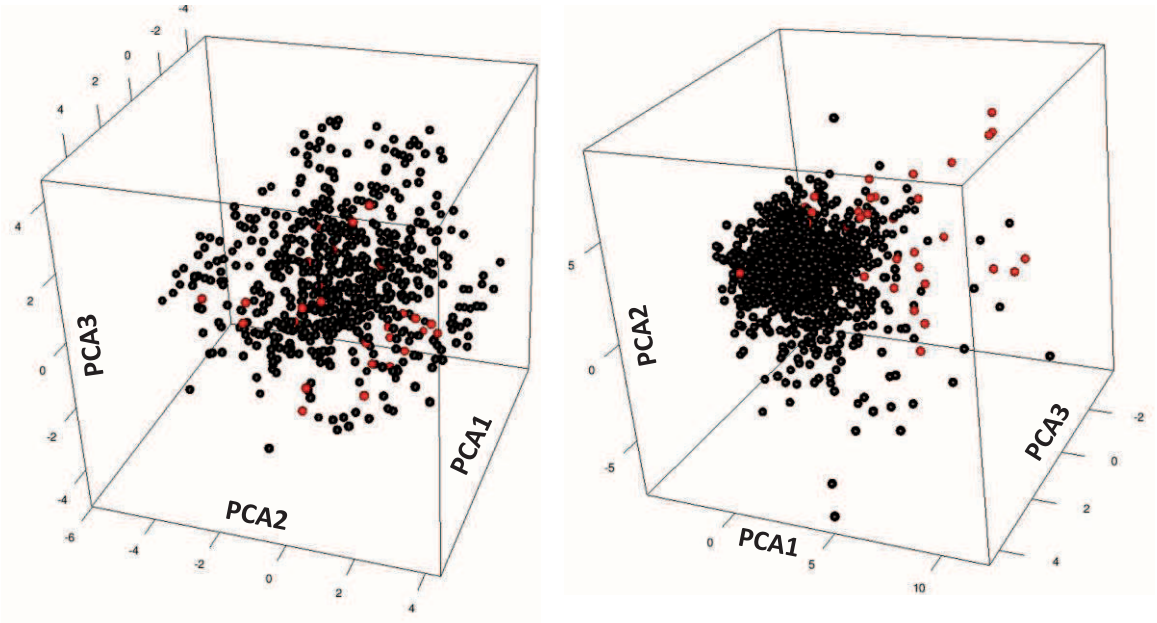


Figure S1 : Representation of the chemical diversity of the positive dataset and decoy. A: A set of 2D BCUT descriptors were generated for the positive dataset and decoy. The first three principal components defined by these BCUT descriptors that illustrate the diversity of the compounds have been selected and used to represent the chemical space of the training dataset. Compounds are clustered based on their chemical properties calculated with the BCUT descriptors. Cluster of molecules are shown as red and black spheres for the positive dataset and decoy respectively. The average number of molecules in each cluster is 1.03 and 1.05 for the positive dataset and decoy respectively. **B:** Projection of the same training dataset according to the first three principal components calculated with the 11 Dragon descriptors selected for the SVM models.

Figure S2

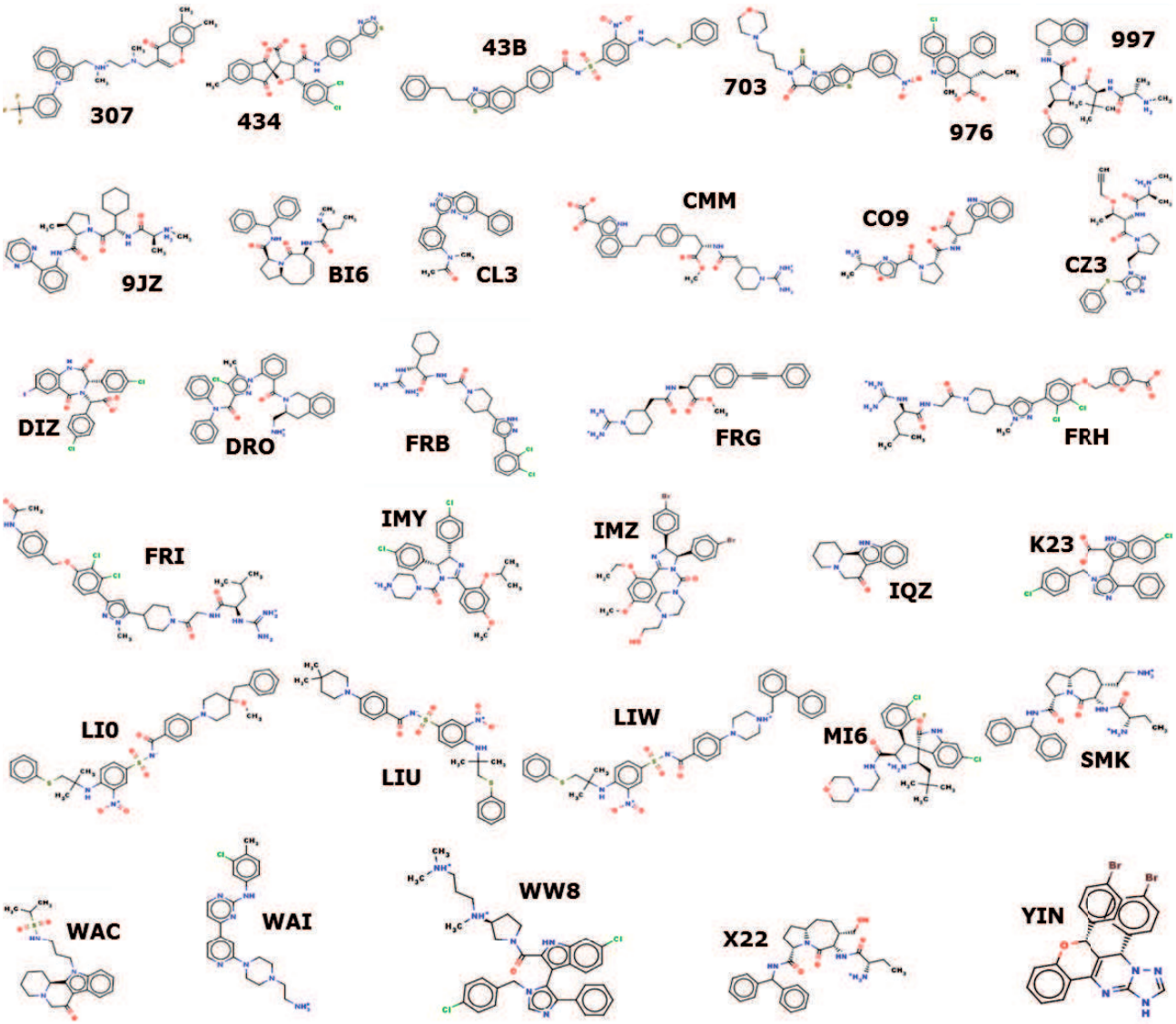


Figure S2: 2D representation of the 32 PPI modulators used as a positive set in the SVM approach. To guarantee structural diversity of the compounds, the 39 small molecules present in 2P2I database were clustered on the basis of Tanimoto similarity criterion of 0.8 with the OptiSim algorithm implemented in the Tripos package leading to 32 non-redundant molecules that were used as the positive set in the SVM protocol.

Figure S3

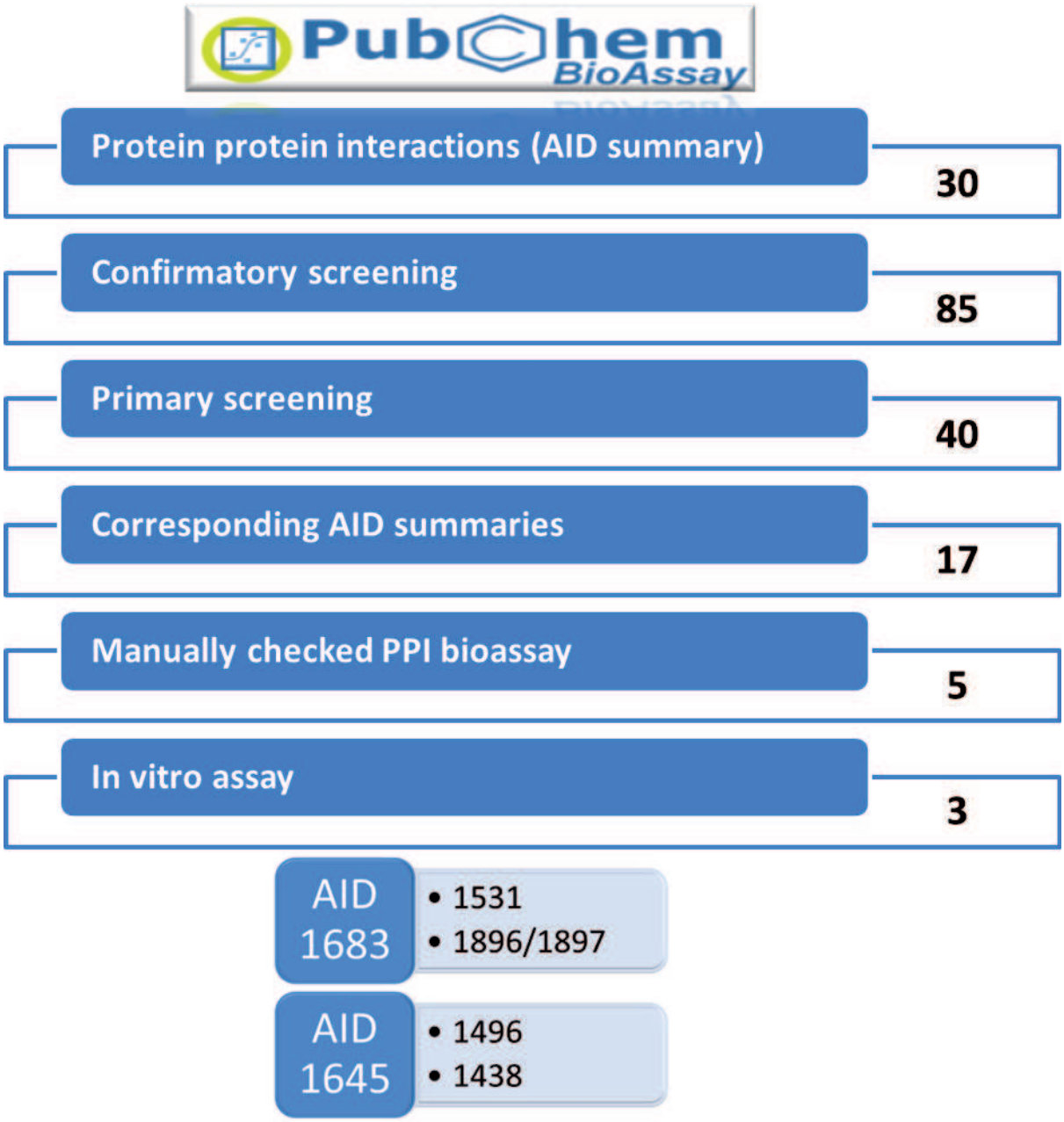
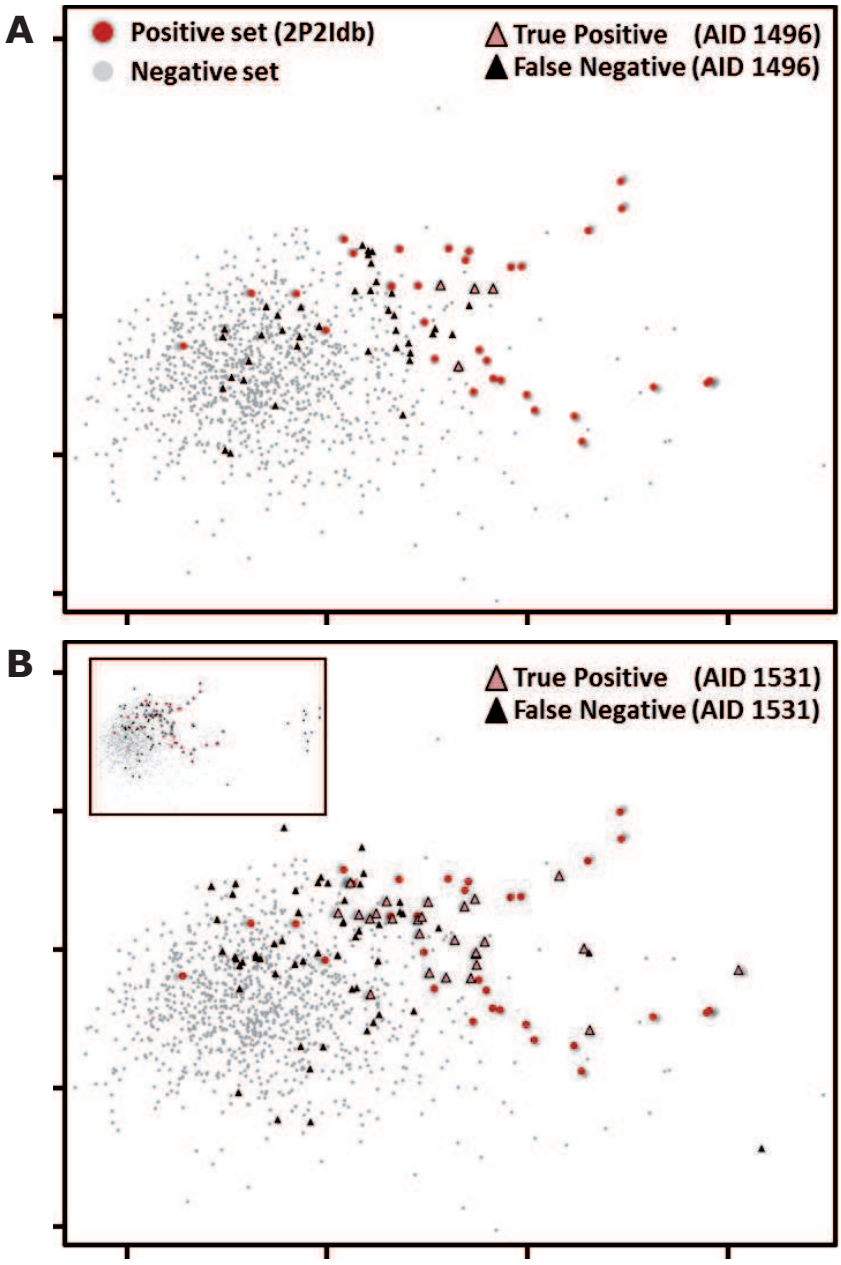


Figure S3: Flowchart of the external validation bioassay selection from Pubchem bioassay. PPI related bioassays for which both a primary and confirmatory screenings were available have been selected through an advanced query. Manual inspection of the 17 series of bioassays revealed that only five corresponded to protein-protein disruption. Cell-based bioassays were discarded to avoid inactive molecules due to cell penetration and only take into account the potentiality of the small molecules to act as PPI modulators.

Figure S4



gure S4: 2D PCA plot of training dataset and external validation dataset. The training set is shown as red and grey spheres for the positive and decoy compounds respectively. Active compounds from AID1496 (top) and AID1531 (bottom) are shown as triangles and divided into true positives (light red) and false negatives (black). The insert in the bottom panel (B) shows the entire 2D chemical space for the active compounds of AID1531.

II.2 Conclusion et perspectives

Nous avons, grâce à l'étude de l'espace chimique des petites molécules présentes dans la base de données $2P2I_{DB}$, créé un protocole permettant de filtrer *in silico* n'importe quelle chimiothèque afin de les enrichir en molécules à fort potentiel inhibiteur d'interactions protéine-protéine (Figure 15). Ce filtre a été validé en utilisant des résultats d'études de criblage expérimentaux à haut débit disponibles sur PUBCHEM (http://pubchem.ncbi.nlm.nih.gov/). L'efficacité de ce filtre a été évaluée en calculant l'enrichissement en petites molécules inhibant une interface protéine-protéine de la chimiothèque après application du filtre. Ainsi, nous avons pu, pour le complexe formé par le domaine Runt de RUNX1 et CBFb, créer un sous-ensemble de la chimiothèque criblée, qui ne représente que 2 % de la chimiothèque initiale, et comprenant un pourcentage de molécules actives 5,4 fois plus important (facteur d'enrichissement de 5,4). De même, pour le complexe formé entre MEK5 et MEK Kinase 2, nous avons créé un sous ensemble de la chimiothèque criblée, ne comprenant toujours que 2 % des molécules, et avec un facteur d'enrichissement de 12,8. Il est à noter que le facteur d'enrichissement théorique maximal est dans ce cas de 50.

Ce protocole est composé de deux filtres : le premier, appelé « Règle des 4» (Ro4), correspond au profil générique des inhibiteurs d'interactions protéine-protéine (MW ≥ 400 Da, AlogP ≥ 4, nombre de cycles ≥ 4 et accepteurs d'hydrogènes ≥ 4). Après application de cette Ro4 sur une chimiothèque donnée, un second filtre, créé grâce à une machine à support de vecteurs, permet d'éliminer encore plus de molécules à faible potentiel pour la modulation des interfaces protéine-protéine.

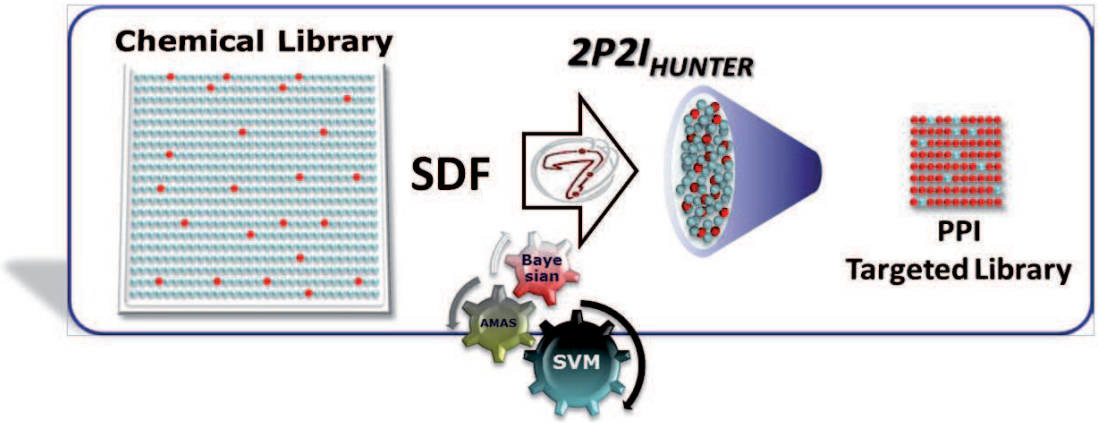


Figure 15 : Schéma de l'approche développée au laboratoire pour la création de chimiothèques focalisées sur les interactions protéine-protéine.

Nous avons appliqué ce protocole à 25 des plus grandes chimiothèques commerciales aujourd'hui sur le marché représentant environ 8,3 millions de composés (Figure 16). Nous avons ainsi créé une chimiothèque dédiée aux interfaces protéine-protéine, $2P2I_{REF}$, composée de 143 218 composés. Les composés répondant à des critères importants en chimie médicinale et présentant des structures privilégiées (identifiées comme structures de base dans de nombreuses molécules utilisées en clinique) ont été sélectionnés prioritairement pour former une version orientée de cette base de données, $2P2I_{PRIVILEGED}$, de 51 476 composés. Les capacités de criblage en milieu académique ne permettent pas d'envisager le criblage d'un grand nombre de cibles sur un grand nombre de composés. Pour pallier à ce problème financier et humain, des chimiothèques de diversité ont été créées dans tous les grands instituts de criblage pour être utilisées en criblage primaire. Les touches pouvant alors être élargies par recherche de similarité dans la chimiothèque principale. Une chimiothèque de diversité a donc été calculée à l'aide de fingerprints moléculaires 2D, $2P2I_{DIV}$, représentant 8 217 composés.

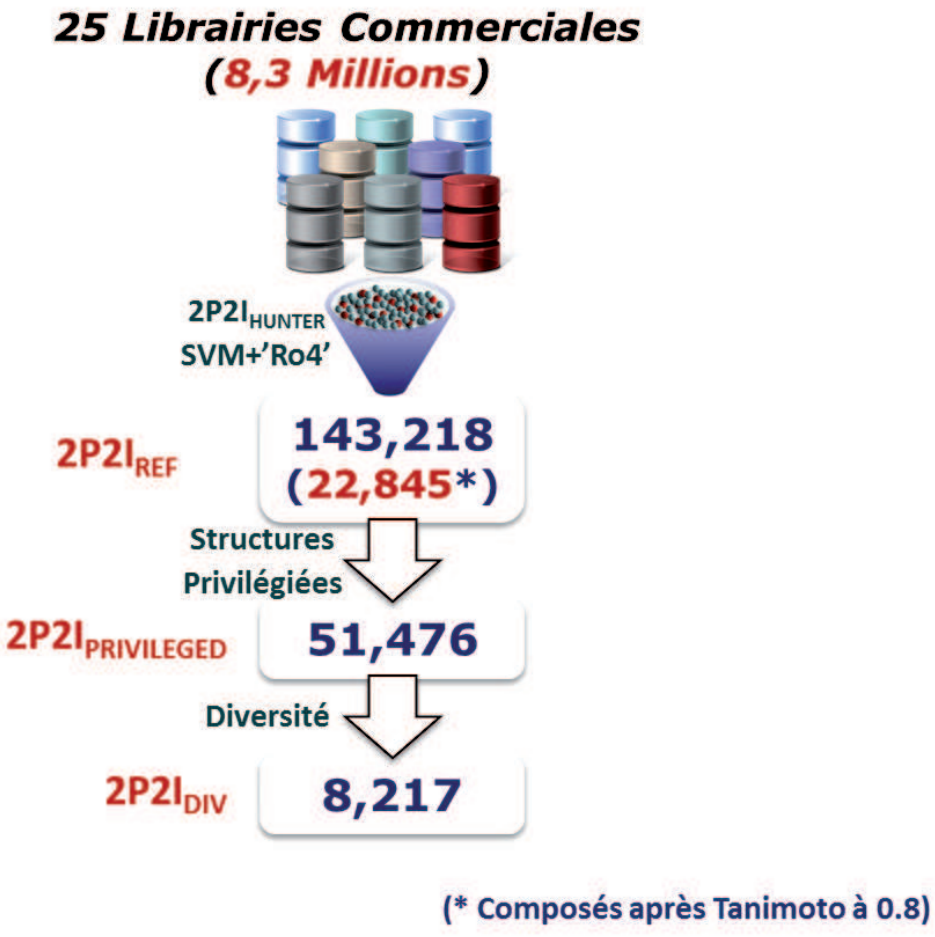


Figure 16 : Logigramme de construction des chimiothèques $2P2I_{REF,}$ $2P2I_{PRIVILEGED}$ et $2P2I_{DIV}$.

Bien que je n'ai pas pu participer à la création finale de ces chimiothèques ni à la rédaction de l'article correspondant (Hamon *et al*, soumis pour publication au journal « MedChemCom »), il me paraît important de les présenter ici puisqu'elles représentent l'aboutissement de mon travail de thèse.

Enfin, un projet de criblage utilisant la chimiothèque finale $2P2I_{DB}$ par des méthodes biochimiques et biophysiques (HTRF et BRET essentiellement) a vu le jour au laboratoire et une collaboration a commencé avec les sociétés CISBIO, Hybrigenics et ValiRX. Ce projet a pour but de cibler une dizaine de complexes importants en cancérologie et en virologie (projet « MATRIX »). Les résultats obtenus au cours de ce projet permettront de confirmer la pertinence de l'utilisation de cette chimiothèque dédiée lors de nouvelles campagnes de drug design orientées sur la modulation des interactions protéine-protéine, et sera mis à disposition de la communauté scientifique.

III Espaces chimiques et mesures d'efficacité des ligands

III.1 Modulateurs d'interactions protéine-protéine et index d'efficacités

La revue suivante, publiée en 2011, présente « l'état de l'art » de la connaissance sur l'espace chimique représentant les inhibiteurs d'interactions protéine-protéine, avec un accent particulier sur un sous-ensemble de cet espace, celui représentant les inhibiteurs présents dans la base de données $2P2I_{DB}$. De plus, des analyses de l'efficacité moléculaire (« ligand efficiency ») des molécules présentes dans $2P2I_{DB}$ nous ont permis d'établir un rapport entre la taille et la polarité et de discuter sur le développement des nano-systèmes de distribution des médicaments en cours de développement dans la plupart des sociétés pharmaceutiques.

Lors de la dernière décennie, un grand nombre d'inhibiteurs d'interactions protéine-protéine sont apparus. Plusieurs centaines de petites molécules ont été signalées ciblant plus de 40 protéines différentes. Les zones d'interactions sur ces protéines représentent la majorité des bases de données SCOP (Structural Classification Of Proteins) et CATH (Class Architecture Topology Homologous), correspondant aux espaces topologiques recensés à ce jour – essentiellement hélice α, feuillet β et une structure α/β mixte. L'ensemble de ces topologies sont retrouvées parmi les inhibiteurs d'interactions protéine-protéine orthostériques présents dans $2P2I_{DB}$.

Nous avons montré que la « Règle des 4 » (Ro4) est un outil rapide et efficace afin d'accélérer les campagnes de drug design ciblant les interactions protéine-protéine en filtrant la chimiothèque testée. Or, on peut noter que le profil de molécule représenté par la Ro4 rentre en contradiction, tout au moins partiellement, avec la « Règle des 5 » décrite par le Dr. Christopher Lipinski, et pourrait donc invalider les 2P2Is comme potentiels agents thérapeutiques. Cependant, une campagne de drug discovery doit optimiser un grand nombre de variables. Abad-Zapatero et Metz ont présenté de nouveaux indices chimiques comme indicateurs pour la recherche de nouveaux médicaments. Nous avons calculé ces paramètres pour les molécules présentes dans $2P2I_{DB}$ et avons ainsi pu les comparer avec les 92 médicaments discutés par Abad-Zapatero et Metz (Abad-Zapatero and Metz, 2005).

III.2 Article 4

Chemical and structural lessons from recent successes in protein–protein interaction inhibition (2P2I)

Xavier Morelli, **Raphaël Bourgeas** and Philippe Roche.

Current Opinion in Chemical Biology. 2011 Aug;15(4):475-81. Review.

Worldwide research efforts have driven recent pharmaceutical successes, and consequently, the emerging role of Protein– Protein Interactions (PPIs) as drug targets has finally been widely embraced by the scientific community. Inhibitors of these Protein–Protein Interactions (2P2Is or i-PPIs) are likely to represent the next generation of highly innovative drugs that will reach the market over the next decade. This review describes up-to-date knowledge on this particular chemical space, with a specific emphasis on a subset of this ensemble. We also address current structural knowledge regarding both protein–protein and protein–inhibitor complexes, that is, the inhibiteur d'interactions protéine-protéine database. Finally, ligand efficiency analyses permit us to relate potency to size and polarity and to discuss the need to codevelop nanoparticle drug delivery systems.

Ma contribution dans ce travail a consisté en premier lieu en le calcul et l'analyse des différents index d'efficacité (LEI, BEI et SEI) ainsi que la réalisation de l'ensemble des figures présentes dans la revue.

Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Chemical Biology

Chemical and structural lessons from recent successes in protein–protein interaction inhibition (2P2I)

Xavier Morelli, Raphaël Bourgeas and Philippe Roche

Worldwide research efforts have driven recent pharmaceutical successes, and consequently, the emerging role of Protein–Protein Interactions (PPIs) as drug targets has finally been widely embraced by the scientific community. Inhibitors of these Protein–Protein Interactions (2P2Is or i-PPIs) are likely to represent the next generation of highly innovative drugs that will reach the market over the next decade. This review describes up-to-date knowledge on this particular chemical space, with a specific emphasis on a subset of this ensemble. We also address current structural knowledge regarding both protein–protein and protein–inhibitor complexes, that is, the 2P2I database. Finally, ligand efficiency analyses permit us to relate potency to size and polarity and to discuss the need to co-develop nanoparticle drug delivery systems.

Address
Interactions et Modulateurs de Réponses (UPR3243), Centre National de la Recherche Scientifique (CNRS) & Université de Provence; 31 Chemin Joseph Aiguier, 13402 Marseille cedex 20, France

Corresponding author: Roche, Philippe (philippe.roche@ifr88.cnrs-mrs.fr)

Current Opinion in Chemical Biology 2011, **15**:475–481

This review comes from a themed issue on
Next Generation Therapeutics
Edited by Alex Matter and Thomas H. Keller

Available online 20th June 2011

1367-5931/$ – see front matter

DOI 10.1016/j.cbpa.2011.05.024

Introduction

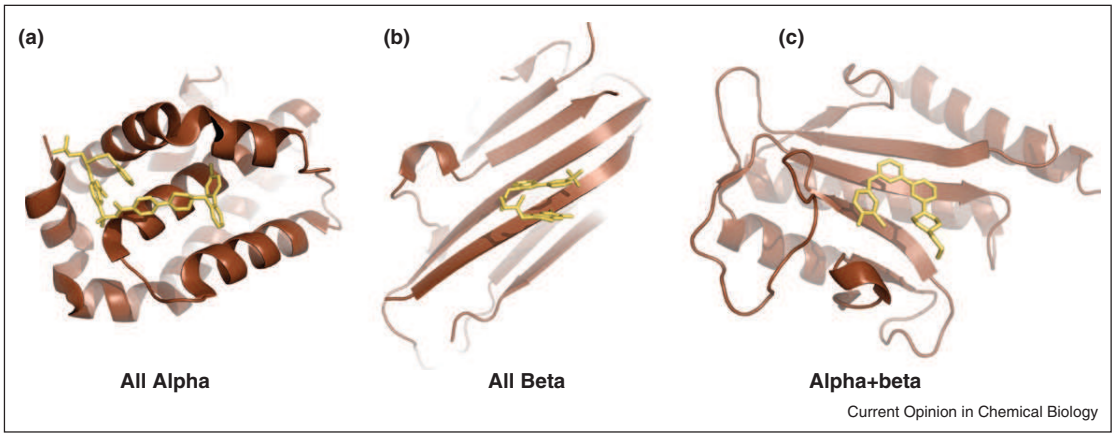
Most drugs currently on the market are competitive inhibitors of G-protein-coupled receptors, nuclear receptors, ion channels and enzymatic targets [1••]. However, there remains a significant unmet clinical need in many diseases for other types of drug intervention – a predicament that is likely to worsen owing to the emergence of serious clinical complications with current therapeutic drugs [2]. Large-scale genomics and proteomics programs have identified complete networks of protein interactions within a cell (called the interactome) and have led to major breakthroughs in understanding biological pathways, host–pathogen interactions and cancer development [3]. Although the size of the human interactome is still a matter of debate, it has been estimated to consist of between 130,000 [4] and 650,000 interactions [5] and therefore represents a vast source of novel targets for the development of new therapeutic drugs. In parallel to these discoveries, recent successes in the Inhibition of Protein–Protein Interactions with small molecules (2P2I or i-PPIs) have emerged from both academic and private research as a new way to modulate the activity of proteins and generate new drugs against this tremendous reservoir of potential targets [6••,7,8•,9–19,20•,21–23]. Certainly, tools to identify potential drugs currently exist. However, questions remain in more general areas, such as the predictability of the chemical space dedicated to these interaction interfaces or the druggability of each potential new PPI target identified in a biological program. Because there is currently no established structure-based general approach for the discovery of PPI modulators, we and others have recently analyzed data from the literature, as well as from patent databases, and have presented chemotypes for potential PPI inhibitors [24,25,26•,27–29].

In this review, we survey recent successes in the discovery of PPI inhibitors for which 3D structures of PPI interfaces in both the free form and bound to the inhibitor are known. A general analysis of the molecular descriptors of these small molecule inhibitors led us to propose a 'Rule-of-Four' to describe the chemical space covered by these compounds. Finally, we compare recent metrics related to size, activity and polarity, such as ligand efficiency (LE), binding efficiency index (BEI) and surface efficiency index (SEI) [30•] for known drugs that target this specific ensemble and discuss necessary developments in the future of drug discovery that should permit these highly innovative drugs to reach the market in the near future.

Each representative PPI class has specific inhibitors

In the past decade, an increasing number of success stories have appeared regarding the development of PPI inhibitors. Hundreds of small molecule compounds have been reported to target more than 40 different proteins, covering most SCOP (Structural Classification of Proteins) and CATH (Class Architecture Topology Homologous) database fold classes and corresponding to various topological spaces – primarily helix-based domains, beta-strand domains, mixed folding (helix + -beta strand) and loop-binding groove domains [28]. Some of these drugs can be very effective as PPI disruptors, and a few, such as ABT-263, have reached pre-clinical or clinical trials [31–36]. However, the mode of action and common properties of these inhibitors remain unclear.

Figure 1



Schematic representation of 3 protein–ligand complexes indicative of protein scaffolds in the 2P2I database. $2P2I_{DB}$ (http://2p2idb.cnrs-mrs.fr) contains protein–protein and protein–ligand complexes with known 3D structures from different CATH or SCOP classes. Crystal structure of targets with small molecules inhibitors are shown as example. **(a)** Bcl-X_L in complex with ABT-737 (all alpha, PDB code 2YXJ). **(b)** TNF-alpha with ligand 307 (all beta, PDB code 2AZ5). **(c)** ZipA with ligand WAI (alpha + beta, PDB Code 1Y2F).

Several groups have tried to rationalize the chemical space of PPI inhibitors using machine learning strategies and a set of molecular descriptors [7,24,25,26•,27–29]. These studies have highlighted the fact that most available commercial libraries are not fully adequate for targeting PPIs for therapeutic purposes. PPI inhibitors can be subdivided into two main classes: orthosteric and allosteric. Orthosteric inhibitors directly interfere with the interface of the protein complex, while allosteric inhibitors bind away from the interface and cause or prevent conformational changes that preclude formation of the complex [37]. To obtain a clear picture of what makes a good PPI inhibitor at the interface, we developed the structural 2P2I database (http://2p2idb.cnrs-mrs.fr), which contains information on all the families for which structural knowledge was available for both the protein–protein and protein–ligand complexes [27]. Complexes in $2P2I_{DB}$ cover the three main classes of CATH domain classifications (Figure 1).

Presentation of the 2P2I database

The newly updated version (v1.0) of the 2P2I database contains a total of 12 protein–protein complexes and 39 non-redundant small molecule inhibitors bound to their corresponding targets (Table 1 and Figure 2). We have proposed a classification of these interactions into two classes, depending on the number of segments at the interface [27]. Class I corresponds to targets that use a limited number of continuous segments for binding to their partner (average value of 3.29 ± 1.25); this class contains 7 of the 12 PPIs, corresponding to 5 target families (Table 1, lines 1–7). The remaining 5 PPIs representing Class II use a higher number of continuous segments (average value of 8.20 ± 1.30) and correspond to more globular interacting domains (Table 1, lines 8–12). Detailed analysis of the properties of the interfaces found in $2P2I_{DB}$ revealed that 6 interface parameters (gap volume, ASA, number of hydrogen bonds, % of charged residues, number of segments and pocket volume) can be used to characterize druggable PPIs [27].

Chemico-biological space of the 2P2I dataset

Toward a 'Rule-of-Four'

As discussed by Fry and Vassilev in their recent book [38•], the concept of what constitutes a drug-like molecule has evolved recently, particularly in the context of protein–protein modulators. The traditional profile of an organic compound with a molecular weight (MW) in the 200–500 Da range, as defined by Lipinski, has been expanded to include compounds of significantly higher molecular weight [38•]. A statistical analysis of the 39 inhibitors present in $2P2I_{DB}$ enabled us to calculate the general characteristics of this particular chemical space. Average values for the molecular weight (547 ± 154 Da, thus MW > 400 Da), ALogP (3.99 ± 2.37, thus ALogP > 4), number of rings (4.44 ± 1.02, thus #Rings > 4) and number of hydrogen bond acceptors (6.62 ± 2.60, thus #HBA > 4) define the generic profile of a PPI inhibitor compound that could be further derived into a more specific inhibitor. These chemical 'Rules-of-Four' properties could be used to filter '*in house*' databases and accelerate the process of hit identification by lowering both cost and time. Interestingly, the values that constitute the 'Rule-of-Four' contradict medicinal chemistry's 'Rule-of-Five' dogma and would thus invalidate 2P2I compounds as potential drugs. However, successful drug discovery requires the optimization of a large number of variables in two different domains: chemical and biological. Abad-Zapatero and Metz have defined original indices as guideposts for drug discovery and effective mapping of

Table 1

Families of protein–protein and protein–ligand complexes in the freely available 2P2I database (http://2p2idb.cnrs-mrs.fr) and binding efficiency indices of the ligands. Complexes have been divided into two classes: Class I encompasses Bcl-2/Bcl-X_L-BAK/BAD, HDM2/XDM2/HDM4-p53, XIAP BIR3-Caspase 9, XIAP BIR3-SMAC and ZIPA-FtsZ complexes (top 7 rows). Class II encompasses HPVE2-E1, IL2-IL2R, Integrase-LEDGFp75, TNFalpha-TNFRc1 and TNFR1A-TNFB complexes (bottom 5 rows). The P_{DB} codes of the protein–protein complexes (AB), unbound proteins (A) or protein–inhibitors complexes (AL) are given. At least one PPI inhibitor is indicated for each PPI target family. The ligand three-letter codes (as defined in the PDB files) of 24 representative small molecule inhibitors out of 39 in 2P2IDB are shown as well as binding constants. When possible, binding efficiency (BEI) and surface efficiency (SEI) indexes are calculated as defined in Figure 3. Where a direct binding dissociation constant was not available, K_i or IC_{50} values were used instead

Family	AB	K_d[a] (μM)	A	AL	Ligand[b]	MW[c]	Kd[d] (μM)	PSA/100 Å^2	BEI	SEI
Bcl/Bak	1BXL	0.34	1R2D	1YSI	1YSI	N3B	552	0.036[f]	1.37	13.48
					2O22	LIU	597	0.067[f]	1.58	12.01
					2YXJ	N3C[e]	813	0.001[f]	1.47	11.07
HDM2/p53	1YCR	0.6	1Z1M	1RV1	IMZ	686	0.14[g]	0.67	9.99	10.22
				1T4E	DIZ	581	0.08	0.67	12.21	10.59
				3JZK	YIN	536	1.23[g]	0.40	11.05	14.80
XDM2/p53	1YCQ	n/a	1Z1M	1TTV	IMY	567	0.16[g]	0.66	11.97	10.3
HDM4/p53	3DAB	0.21	3JZO	3LBJ	WW8	630	n/a	0.60	n/a	n/a
Xiap/Caspase 9	1NW9	75	1F9X	1TFQ	998	443	0.012	0.91	17.88	8.70
				1TFT	997	535	0.005	1.00	15.52	8.30
				2JK7	BI6	487	0.067[f]	0.91	14.73	7.88
				2OPY	CO9	439	30	1.55	10.28	2.92
				3CM2	X23	492	0.34[f]	1.31	13.15	4.94
				3G76	CZ3	969	0.23[g]	2.44	6.78	2.72
				3HL5	9JZ	507	34[f]	1.16	8.81	3.85
Xiap/Smac	1G73	n/a	2VSL	2JK7	BI6	487	0.067[f]	0.91	14.73	7.88
				2OPY	CO9	439	30	1.55	10.28	2.92
				3CM2	X23	492	0.34[f]	1.31	13.15	4.94
ZipA/FtsZ	1F47	21.6	1F46	1Y2F	WAI	424	12	0.52	11.61	9.46
				1Y2G	CL3	343	83	0.63	11.90	6.48
HPV E2/E1	1TUE	0.06	1QQH	1R6N	434	608	0.04	1.44	12.17	5.14
IL-2/IL-2R	1Z92	0.01	1M47	1M48	FRG	447	8.2	0.65	11.38	7.82
				1PW6	FRB	534	6[g]	0.77	9.78	6.78
				1PY2	FRH	663	0.06[g]	1.26	10.89	5.73
Integrase/LEDGFp75	2B4J	0.01	2ITG	3LPT	723	314	n/a	0.47	n/a	n/a
TNF-α/TNFRc1	1TNF	n/a	2E7A	2AZ5	307	548	22[g]	0.38	8.50	12.25
TNFR1A/TNFB	1TNR	n/a	1EXT	1FT4	703	457	0.27[g]	1.44	14.37	4.56

n/a: not available.
[a] Protein–protein dissociation constants.
[b] Ligand 3-Letter code as defined in the PDB chemical component dictionary.
[c] Ligand MW in g mol^{-1}.
[d] Protein–ligand binding constants (k_d, k_i or IC_{50}).
[e] Compound ABT-737.
[f] K_i.
[g] IC_{50}.

chemico-biological space (CBS) using the concept of an atlas-like representation [30•]. We have calculated these parameters for our 2P2I subset and present a comparable analysis.

Binding efficiency index (BEI) and surface efficiency index (SEI) of the 2P2I dataset

The use of ligand efficiency as a numerical metric in lead assessment was introduced a few years ago as the ratio between the free energy of binding ($\Delta G = -RT \ln K_d$) at 300 K and the number of non-hydrogen atoms in the compound [39]. Variations and extensions of this concept continue to appear in the literature and are gaining a much wider acceptance among medicinal chemists [40–42]. The binding efficiency index (BEI, defined as binding affinity divided by MW) is a simpler index, that provides an easy and effective ranking of compounds leading to drugs [30•,42]. Although it has always been considered to be of great importance, the concept of ligand efficiency related to compound polarity has not generally been directly used in lead optimization. A specific definition of this concept (the surface efficiency

Figure 2

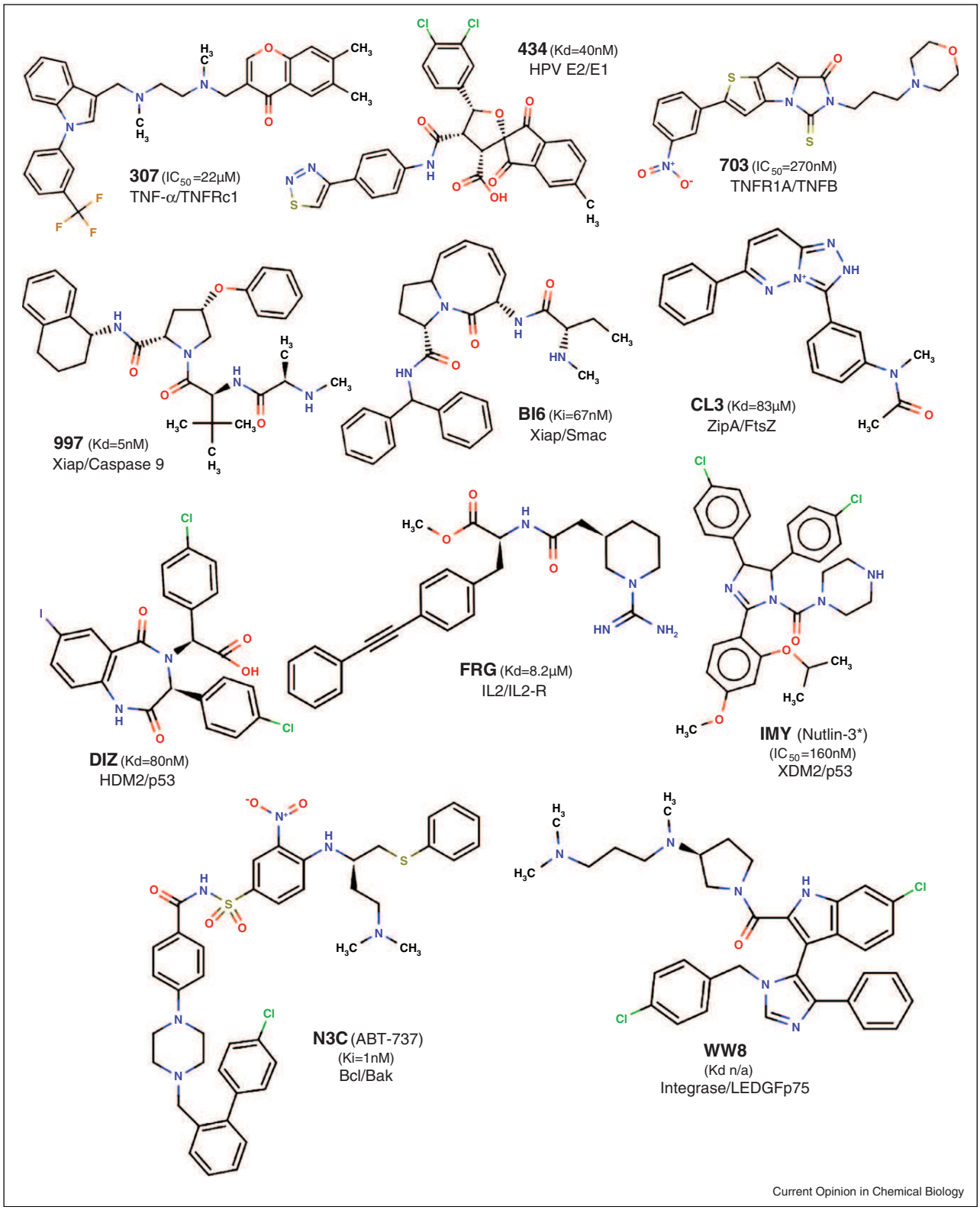
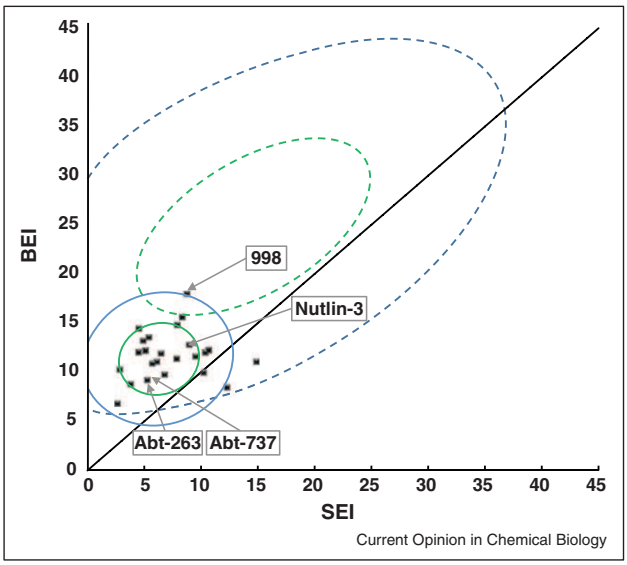

307 (IC50=22μM)
TNF-α/TNFRc1
434 (Kd=40nM)
HPV E2/E1
703 (IC50=270nM)
TNFR1A/TNFB
997 (Kd=5nM)
Xiap/Caspase 9
BI6 (Ki=67nM)
Xiap/Smac
CL3 (Kd=83μM)
ZipA/FtsZ
DIZ (Kd=80nM)
HDM2/p53
FRG (Kd=8.2μM)
IL2/IL2-R
IMY (Nutlin-3*)
(IC50=160nM)
XDM2/p53
N3C (ABT-737)
(Ki=1nM)
Bcl/Bak
WW8
(Kd n/a)
Integrase/LEDGFp75
Current Opinion in Chemical Biology

Figure 3



Mapping of the surface-binding and binding efficiency indexes for PPI inhibitors and a sample of marketed drugs. SEI–BEI values for 22 PPI inhibitors from $2P2I_{DB}$ (as indicated in Table 1) with ellipsoids including 50% (plain green line) and 95% (plain blue line) of the PPI inhibitors. Compounds ABT-263 (Bcl-2/Bcl-X_L target) and nutlin-3 (HDM2 target) in preclinical or clinical trials and discussed in the text are also indicated (ABT-263, MW = 975, PSA/100 $Å^2$ = 1.70, K_d < 1 nM, BEI > 9.2, SEI > 5.3; Nutlin-3, MW = 581, PSA/100 $Å^2$ = 0.83, K_d = 40 nM, BEI = 12.73, SEI = 8.91). As a comparison SEI–BEI values for the 92 examples of marketed drugs have been plotted together with the ellipsoids including 50% (dashed green line), 95% (dashed blue line) of the sample. The vast majority of marketed oral drugs map near and above the diagonal line, reflecting a reasonable optimization of both BEI (MW) and SEI (PSA) [30•]. BEI = (pK_i or pK_d or pIC_{50}/MW) and (SEI = pK_i or pK_d or pIC_{50}/PSA) as defined by Abad-Zapatero [30•].

index or SEI) has been defined using the affinity of the ligand (pK_i, pk_d or pIC_{50}) divided by the molecular Polar Surface Area (PSA) of the ligand scaled to $100A^2$ (Table 1 and Figure 3).

A mapping in the SEI–BEI plane of the location of 92 examples of marketed drugs discussed by Abad-Zapatero and Metz from Abbott Laboratories [30•] compared with the 2P2I dataset is presented in Figure 3. The centroid of the distribution for the marketed drugs has mean values of 25.8 ± 7.9 (BEI) and 14.5 ± 8.7 (SEI). It was originally noted that "*the majority of marketed oral drugs map near and above the diagonal line, reflecting a reasonable optimization of both BEI (MW) and SEI (PSA)*". Our 2P2I dataset presents a centroid distribution with mean values of 11.7 ± 2.4 (BEI) and 7.2 ± 3.0 (SEI), corresponding to '*sub-optimal series that could not get optimized*' in the Abbott paper. As an example of this category, the well-known nutlin-3 [20•,43–46] falls into this 'poor chemico-biological space plane' with a BEI value of 12.73 and a SEI value of 8.91. Although the majority of the 2P2I inhibitors cluster far below the optimum values of BEI and SEI, a few examples reach higher values, such as the peptide mimetic 998, a protein–protein inhibitor of the XIAP/Caspase 9 complex (presenting a BEI value of 17.88 and a SEI value of 8.70) (Figure 3). Another notable example is the B-Cell lymphoma 2 (Bcl-2/X_L) inhibitor, Navitoclax (ABT-263), which is a small molecule BH3 mimetic that inhibits Bcl-2 and is in phase II clinical trials for cancer treatment [31,35,47]. This compound also presents poor BEI (9.2) and SEI (5.3) values. However, Navitoclax (MW of 974 Da) presents a pharmacokinetics profile acceptable for oral dosing. Moreover, in a preliminary report of a Phase IIa trial of Navitoclax, 39 patients were treated with this drug with a median time of study of 49 days. Only adverse events were reported from this preliminary study, including thrombocytopenia in 29% of patients and diarrhea in 43% of patients [47]. These two specific examples (nutlin-3 and Navitoclax) demonstrate that 'poor' chemico-biological space candidates can present promising therapeutic values and that such compounds should not be rejected for their 'low probability to be developed as drugs'. It is important to point out that the 'Rule-of-Five' or BEI/SEI indices were meant to identify the chemical space of compounds with 'appropriate' pharmacokinetics or ADME behavior for oral delivery (better absorption, better bioavailability). They cannot be expected to predict high specific interactions with proteins. However, these low values in the BEI–SEI plane demonstrate the urgent need for pharmaceutical companies interested in the development of protein–protein interaction inhibitors to think differently and develop parallel technologies, such as nanoparticle drug delivery systems (NPDDS). These changes in thought and development tactics must occur if pharmaceutical companies want to address 2P2I Achilles' heels, such as molecular weight (BEI) and hydrophobicity (SEI) indexes, and enable this class of compounds to become the next generation of highly innovative drugs to reach the market within the next few decades.

Pitfalls and future developments

Nanoparticle drug delivery systems (NPDDS) and their uptake in biological systems at cellular levels are certainly the best approach for 2P2I candidates to become successful in the next decade. Recent breakthroughs in this highly dynamic field have demonstrated that nanocarriers offer a promising approach to obtain desirable delivery properties by altering the biopharmaceutic and pharmacokinetic properties of the molecules [48]. Various nanotechnology

(Figure 2 Legend) Representative set of 11 PPI inhibitors found in the 2P2I database and corresponding to the different target families described in Table 1. For each compound the PDB three letter code is indicated together with the binding constant (k_d, k_i or IC_{50}) when available and the target family. *Compound IMY is very closely related to nutlin-3.

platforms are under investigation as potential vehicles for encapsulation, stabilization and delivery of potential drugs, such as liposomal nanocarrier development [49], micro (nano) emulsions [50], mesoporous silicon-based methods [51,52], dendrimers [53,54] and chemically modified viral nanoparticles [55•]. More work is required to understand and assess the interaction of proteins with small-molecule ligands for direct applications of nanoparticle drug delivery systems in the clinic. However, the application of nanotechnology to drug delivery is widely expected to create novel therapeutics capable of changing the landscape of the pharmaceutical and biotechnology industries. Further, this technology is particularly promising in the field of protein–protein inhibition.

Concluding remarks

Most PPI inhibitors described thus far do not fit with the prevalent 'Rule-of-Five' drug profile, as defined by Lipinski and others. Instead, they present chemical properties shifted toward higher molecular weight, increased hydrophobicity and a higher unsaturation index and ring complexity than common drugs (what we have proposed for protein–protein inhibitors as a 'Rule-of-Four'). Moreover, compound libraries used to perform high throughput drug discovery have generally been filtered using a 'Rule-of-Five' that eliminates a large number of potential PPI disruptors. To develop innovative drugs targeting PPIs in the near future, we will certainly need to move away from the current paradigm of what makes a chemical compound a potential drug with pharmacological and therapeutic properties for this class of targets.

Acknowledgements

We would like to thank Dr P. Nuno Palma (Department of Research and Development, BIAL, Portugal) for helpful comments and for critical reading of the manuscript. RB is supported by a MENRT grant from the French ministry of research and education.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Overington JP, Al-Lazikani B, Hopkins AL: **How many drug targets are there?** *Nat Rev Drug Discov* 2006, **5**:993-996.
•• Comprehensive survey of the literature proposing a consensus number of drug targets for all classes of approved therapeutic drugs. Relevant statistics such as patterns in drug targets, average affinity of the drug for its target or the rate of developing new drugs are discussed.

2. Fosslien E: **Cardiovascular complications of non-steroidal anti-inflammatory drugs**. *Ann Clin Lab Sci* 2005, **35**:347-385.

3. Lievens S, Eyckerman S, Lemmens I, Tavernier J: **Large-scale protein interactome mapping: strategies and opportunities**. *Expert Rev Proteomics* 2010, **7**:679-690.

4. Venkatesan K, Rual J, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh K *et al.*: **An empirical framework for binary interactome mapping**. *Nat Methods* 2009, **6**:83-90.

5. Stumpf M, Thorne T, de Silva E, Stewart R, An H, Lappe M, Wiuf C: **Estimating the size of the human interactome**. *Proc Natl Acad Sci USA* 2008, **105**:6959-6964.

6. Wells J, McClendon C: **Reaching for high-hanging fruit in drug discovery at protein-protein interfaces**. *Nature* 2007, **450**:1001-1009.
•• Review of success stories in the development of small molecules inhibitors of protein–protein interfaces for which the 3D structures of the protein–protein and protein–ligand complexes are known. The authors describe how a protein interacts with its natural protein (or peptide) partner and with a synthetic small molecule at the atomic level focusing on the notion of 'hotspots' on the contact surfaces.

7. Fry DC: **Drug-like inhibitors of protein-protein interactions: a structural examination of effective protein mimicry**. *Curr Protein Peptide Sci* 2008, **9**:240-247.

8. Lessene G, Czabotar PE, Colman PM: **BCL-2 family antagonists for cancer therapy**. *Nat Rev Drug Discov* 2008, **7**:989-1000.
• Review of preclinical and clinical data on several compounds that have been described inhibiting the interaction between Bcl-2 family members and their natural ligand, a helical peptide sequence known as the BH3 domain. Recommendation of four criteria that define potential antagonists of the Bcl-2 protein family.

9. Azmi AS, Mohammad RM: **Non-peptidic small molecule inhibitors against Bcl-2 for cancer therapy**. *J Cell Physiol* 2009, **218**:13-21.

10. Cossu F, Mastrangelo E, Milani M, Sorrentino G, Lecis D, Delia D, Manzoni L, Seneci P, Scolastico C, Bolognesi M: **Designing Smac-mimetics as antagonists of XIAP, cIAP1, and cIAP2**. *Biochem Biophys Res Commun* 2009, **378**:162-167.

11. Crisostomo FR, Feng Y, Zhu X, Welsh K, An J, Reed JC, Huang Z: **Design and synthesis of a simplified inhibitor for XIAP-BIR3 domain**. *Bioorg Med Chem Lett* 2009, **19**:6413-6418.

12. Christ F, Voet A, Marchand A, Nicolet S, Desimmie BA, Marchand D, Bardiot D, Van der Veken NJ, Van Remoortel B, Strelkov SV *et al.*: **Rational design of small-molecule inhibitors of the LEDGF/p75-integrase interaction and HIV replication**. *Nat Chem Biol* 2010, **6**:442-448.

13. De Luca L, Ferro S, Gitto R, Barreca ML, Agnello S, Christ F, Debyser Z, Chimirri A: **Small molecules targeting the interaction between HIV-1 integrase and LEDGF/p75 cofactor**. *Bioorg Med Chem* 2010, **18**:7515-7521.

14. Porter JR, Helmers MR, Wang P, Furman JL, Joy ST, Arora PS, Ghosh I: **Profiling small molecule inhibitors against helix-receptor interactions: the Bcl-2 family inhibitor BH3I-1 potently inhibits p53/hDM2**. *Chem Commun (Camb)* 2010, **46**:8020-8022.

15. Azmi AS, Wang Z, Philip PA, Mohammad RM, Sarkar FH: **Emerging Bcl-2 inhibitors for the treatment of cancer**. *Expert Opin Emerg Drugs* 2011, **16**:59-70.

16. De Luca L, De Grazia S, Ferro S, Gitto R, Christ F, Debyser Z, Chimirri A: **HIV-1 integrase strand-transfer inhibitors: design, synthesis and molecular modeling investigation**. *Eur J Med Chem* 2011, **46**:756-764.

17. Popowicz GM, Dömling A, Holak TA: **The structure-based design of Mdm2/Mdmx-p53 inhibitors gets serious**. *Angew Chem Int Ed Engl* 2011, **50**:2680-2688.

18. Rinaldi M, Tintori C, Franchi L, Vignaroli G, Innitzer A, Massa S, Esté JA, Gonzalo E, Christ F, Debyser Z *et al.*: **A versatile and practical synthesis toward the development of novel HIV-1 integrase inhibitors**. *ChemMedChem* 2011, **6**:343-352.

19. Voet A, Callewaert L, Ulens T, Vanderkelen L, Vanherreweghe JM, Michiels CW, De Maeyer M: **Structure based discovery of small molecule suppressors targeting bacterial lysozyme inhibitors**. *Biochem Biophys Res Commun* 2011, **19**:19.

20. Vu BT, Vassilev L: **Small-molecule inhibitors of the p53-MDM2 interaction**. *Curr Top Microbiol Immunol* 2011, **348**:151-172.
• Review of a class of potent and selective small-molecule antagonists of the p53-MDM2 interaction that emerged as potential drugs and led to actual preclinical and clinical studies.

21. Wang S: **Design of small-molecule Smac mimetics as IAP antagonists**. *Curr Top Microbiol Immunol* 2011, **348**:89-113.

22. White PW, Faucher AM, Goudreau N: **Small molecule inhibitors of the human papillomavirus e1-e2 interaction**. *Curr Top Microbiol Immunol* 2011, **348**:61-88.

23. Wilson CG, Arkin MR: **Small-molecule inhibitors of IL-2/IL-2R: lessons learned and applied**. *Curr Top Microbiol Immunol* 2011, **348**:25-59.

24. Pagliaro L, Felding J, Audouze K, Nielsen SJ, Terry RB, Krog-Jensen C, Butcher S: **Emerging classes of protein-protein interaction inhibitors and new tools for their development**. *Curr Opin Chem Biol* 2004, **8**:442-449.

25. Neugebauer A, Hartmann RW, Klein CD: **Prediction of protein-protein interaction inhibitors by chemoinformatics and machine learning methods**. *J Med Chem* 2007, **50**:4665-4668.

26. • Higueruelo AP, Schreyer A, Bickerton GRJ, Pitt WR, Groom CR, Blundell TL: **Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database**. *Chem Biol Drug Des* 2009, **74**:457-467.
Description of the first hand-curated database dedicated to small molecule inhibitors of protein–protein interactions. The general molecular properties of the protein–protein modulators are analyzed as well as potency compared to medicinal chemistry hit or lead.

27. Bourgeas R, Basse M-J, Morelli X, Roche P: **Atomic analysis of protein-protein interfaces with known inhibitors: the 2P2I database**. *PLoS ONE* 2010, **5**:e9598.

28. Reynes C, Host H, Camproux AC, Laconde G, Leroux F, Mazars A, Deprez B, Fahraeus R, Villoutreix BO, Sperandio O: **Designing focused chemical libraries enriched in protein-protein interaction inhibitors using machine-learning methods**. *PLoS Comput Biol* 2010, **6**:e1000695.

29. Sperandio O, Reynès C, Camproux A, Villoutreix B: **Rationalizing the chemical space of protein-protein interaction inhibitors**. *Drug Discov Today* 2010, **15**:220-229.

30. • Abad-Zapatero C, Metz JT: **Ligand efficiency indices as guideposts for drug discovery**. *Drug Discov Today* 2005, **10**:464-469.
Presentation of an effective atlas-like representation of chemico-biological space (CBS) based on a series of cartesian planes mapping the ligands with the corresponding targets connected by an affinity parameter. This framework facilitates navigation in the multidimensional drug discovery space using map-like representations based on pairs of combined variables related to the efficiency of the ligands per Dalton and per unit of polar surface area.

31. Tse C, Shoemaker AR, Adickes J, Anderson MG, Chen J, Jin S, Johnson EF, Marsh KC, Mitten MJ, Nimmer P *et al.*: **ABT-263: a potent and orally bioavailable Bcl-2 family inhibitor**. *Cancer Res* 2008, **68**:3421-3428.

32. Ackler S, Mitten MJ, Foster K, Oleksijew A, Refici M, Tahir SK, Xiao Y, Tse C, Frost DJ, Fesik SW *et al.*: **The Bcl-2 inhibitor ABT-263 enhances the response of multiple chemotherapeutic regimens in hematologic tumors in vivo**. *Cancer Chemother Pharmacol* 2010, **66**:869-880.

33. Tahir SK, Wass J, Joseph MK, Devanarayan V, Hessler P, Zhang H, Elmore SW, Kroeger PE, Tse C, Rosenberg SH *et al.*: **Identification of expression signatures predictive of sensitivity to the Bcl-2 family member inhibitor ABT-263 in small cell lung carcinoma and leukemia/lymphoma cell lines**. *Mol Cancer Ther* 2010, **9**:545-557.

34. Vogler M, Furdas SD, Jung M, Kuwana T, Dyer MJ, Cohen GM: **Diminished sensitivity of chronic lymphocytic leukemia cells to ABT-737 and ABT-263 due to albumin binding in blood**. *Clin Cancer Res* 2010, **16**:4217-4225.

35. Gandhi L, Camidge DR, Ribeiro de Oliveira M, Bonomi P, Gandara D, Khaira D, Hann CL, McKeegan EM, Litvinovich E, Hemken PM *et al.*: **Phase I study of Navitoclax (ABT-263), a novel Bcl-2 family inhibitor, in patients with small-cell lung cancer and other solid tumors**. *J Clin Oncol* 2011, **29**:909-916.

36. Sakuma Y, Tsunezumi J, Nakamura Y, Yoshihara M, Matsukuma S, Koizume S, Miyagi Y: **ABT-263, a Bcl-2 inhibitor, enhances the susceptibility of lung adenocarcinoma cells treated with Src inhibitors to anoikis**. *Oncol Rep* 2011, **25**:661-667.

37. Buchwald P: **Small-molecule protein-protein interaction inhibitors: therapeutic potential in light of molecular size, chemical space, and ligand binding efficiency considerations**. *IUBMB Life* 2010, **62**:724-731.

38. • Fry D: In *Small-Molecule Inhibitors of Protein-Protein Interactions.* Edited by Vassilev L, Fry D. Nutley, New Jersey: Springer; 2011.
Success stories from a number of leaders in the challenging field of protein–protein interaction inhibition. These researchers describe their unique approaches, and share experiences, results, thoughts, and opinions. Numerous lessons to be taken away.

39. Hopkins AL, Groom CR, Alex A: **Ligand efficiency: a useful metric for lead selection**. *Drug Discov Today* 2004, **9**:430-431.

40. Bembenek SD, Tounge BA, Reynolds CH: **Ligand efficiency and fragment-based drug discovery**. *Drug Discov Today* 2009, **14**:278-283.

41. Nissink JW: **Simple size-independent measure of ligand efficiency**. *J Chem Inf Model* 2009, **49**:1617-1622.

42. Abad-Zapatero C, Perišić O, Wass J, Bento AP, Overington J, Al-Lazikani B, Johnson ME: **Ligand efficiency indices for an effective mapping of chemico-biological space: the concept of an atlas-like representation**. *Drug Discov Today* 2010, **15**:804-811.

43. Vassilev LT, Vu BT, Graves B, Carvajal D, Podlaski F, Filipovic Z, Kong N, Kammlott U, Lukacs C, Klein C *et al.*: **In vivo activation of the p53 pathway by small-molecule antagonists of MDM2**. *Science (New York, NY)* 2004, **303**:844-848.

44. Beretta GL, Gatti L, Benedetti V, Perego P, Zunino F: **Small molecules targeting p53 to improve antitumor therapy**. *Mini Rev Med Chem* 2008, **8**:856-868.

45. Endo S, Yamato K, Hirai S, Moriwaki T, Fukuda K, Suzuki H, Abei M, Nakagawa I, Hyodo I: **Potent in vitro and in vivo antitumor effects of MDM2 inhibitor nutlin-3 in gastric cancer cells**. *Cancer Sci* 2011, **102**:605-613.

46. Sonnemann J, Palani CD, Wittig S, Becker S, Eichhorn F, Voigt A, Beck JF: **Anticancer effects of the p53 activator nutlin-3 in Ewing's sarcoma cells**. *Eur J Cancer* 2011, **47**:1432-1441.

47. Rudin CM, Oliveira MR, Garon EB, Bonomi P, Camidge DR, Nolan C, Busman T, Krivoshik A, Humerickhouse R, Gandhi L: **A phase IIa study of ABT-263 in patients with relapsed small-cell lung cancer (SCLC)**. *J Clin Oncol* 2010, **28**: (suppl; abstr 7046).

48. Prokop A, Davidson JM: **Nanovehicular intracellular delivery systems**. *J Pharm Sci* 2008, **97**:3518-3590.

49. Drummond DC, Noble CO, Hayes ME, Park JW, Kirpotin DB: **Pharmacokinetics and in vivo drug release rates in liposomal nanocarrier development**. *J Pharm Sci* 2008, **97**:4696-4740.

50. Gupta S, Moulik SP: **Biocompatible microemulsions and their prospective uses in drug delivery**. *J Pharm Sci* 2008, **97**:22-45.

51. Salonen J, Kaukonen AM, Hirvonen J, Lehto VP: **Mesoporous silicon in drug delivery applications**. *J Pharm Sci* 2008, **97**:632-653.

52. Santos HA, Bimbo LM, Lehto VP, Airaksinen AJ, Salonen J, Hirvonen J: **Multifunctional porous silicon for therapeutic drug delivery and imaging**. *Curr Drug Discov Technol* 2011. [in press].

53. Cheng Y, Xu Z, Ma M, Xu T: **Dendrimers as drug carriers: applications in different routes of drug administration**. *J Pharm Sci* 2008, **97**:123-143.

54. Cheng Y, Zhao L, Li Y, Xu T: **Design of biocompatible dendrimers for cancer diagnosis and therapy: current status and future perspectives**. *Chem Soc Rev* 2011, **40**:2673-2703.

55. • Koudelka KJ, Manchester M: **Chemically modified viruses: principles and applications**. *Curr Opin Chem Biol* 2010, **14**:810-817.
Use of viruses as natural nanomaterials/substrates for chemical modification, materials development, and therapeutic design. Presentation of recent advances in chemical strategies for modifying viruses, and the applications of these technologies.

CONCLUSION GENERALE

Le travail effectué tout au long de ma thèse a été majoritairement orienté sur une approche structurale de l'inhibition des interactions protéine-protéine. Afin d'augmenter l'efficacité de notre approche sur ce sujet, j'ai travaillé sur l'étude des paramètres structuraux des complexes protéiques ; mais également sur les petites molécules permettant de moduler de tels complexes et susceptibles de devenir des agents thérapeutiques. Cette méthodologie basée sur deux approches différentes de la problématique a permis de répondre aux deux questions que l'on se posait, à savoir :

I. Quels sont les paramètres structuraux qui permettent de définir si un complexe protéique est adapté pour la recherche de nouveaux agents thérapeutiques ?

II. Est-il possible de définir des chimiothèques dédiées à cet espace biologique que constituent les interactions protéine-protéine ?

Le travail effectué au laboratoire au cours de ma thèse a permis de répondre à ces deux questions. Nous avons mis au point un protocole permettant à la fois de sélectionner parmis les PPIs celles qui seront les cibles thérapeutiques de demain, mais aussi de créer des chimiothèques de petites molécules enrichies en PPIMs.

A la suite de ce travail, un protocole de sélection de cibles, et de recherche de PPIMs peut être proposé (Figure 17 **A**). Notre approche étant structurale, la première étape consiste à recherche s'il existe une structure de la protéine sous sa forme libre. Si aucune structure n'est disponible, il sera possible de réaliser un modèle si une structure d'une protéine homologue est disponible. Notre approche, basée sur l'étude des paramètres physicochimiques régissant l'interaction entre la protéine cible et son partenaire, nécessite de connaitre la zone d'interaction. Pour cela, la structure tridimensionnelle du complexe protéique permet de la définir avec certitude. Cependant, si la structure du complexe protéique n'est pas disponible, certaines méthodes *in silico* peuvent être utilisées pour estimer cette zone d'interaction. En particulier, la recherche des « hot-spots », que l'on retrouve à l'interface et qui régissent l'essentiel de l'énergie d'interaction, peut être facilitée par les outils informatiques disponibles. Cette approche peut être appuyée par la recherche des poches présentes sur la

surface de la protéine, puisqu'elles sont plus concentrées et surtout plus volumineuses dans la zone d'interaction que sur le reste de la surface de la protéine. Enfin, ces approches *in silico* peuvent être complétées par des campagnes de mutation *in vitro* en alanine afin de déterminer quels acides aminés sont nécessaires à la reconnaissance entre les deux partenaires. Une fois la zone d'interface entre les deux protéines bien définie, la comparaison avec les complexes présents dans la base de données $2P2I_{DB}$ permet de juger la pertinence d'une campagne de drug design sur cette cible.

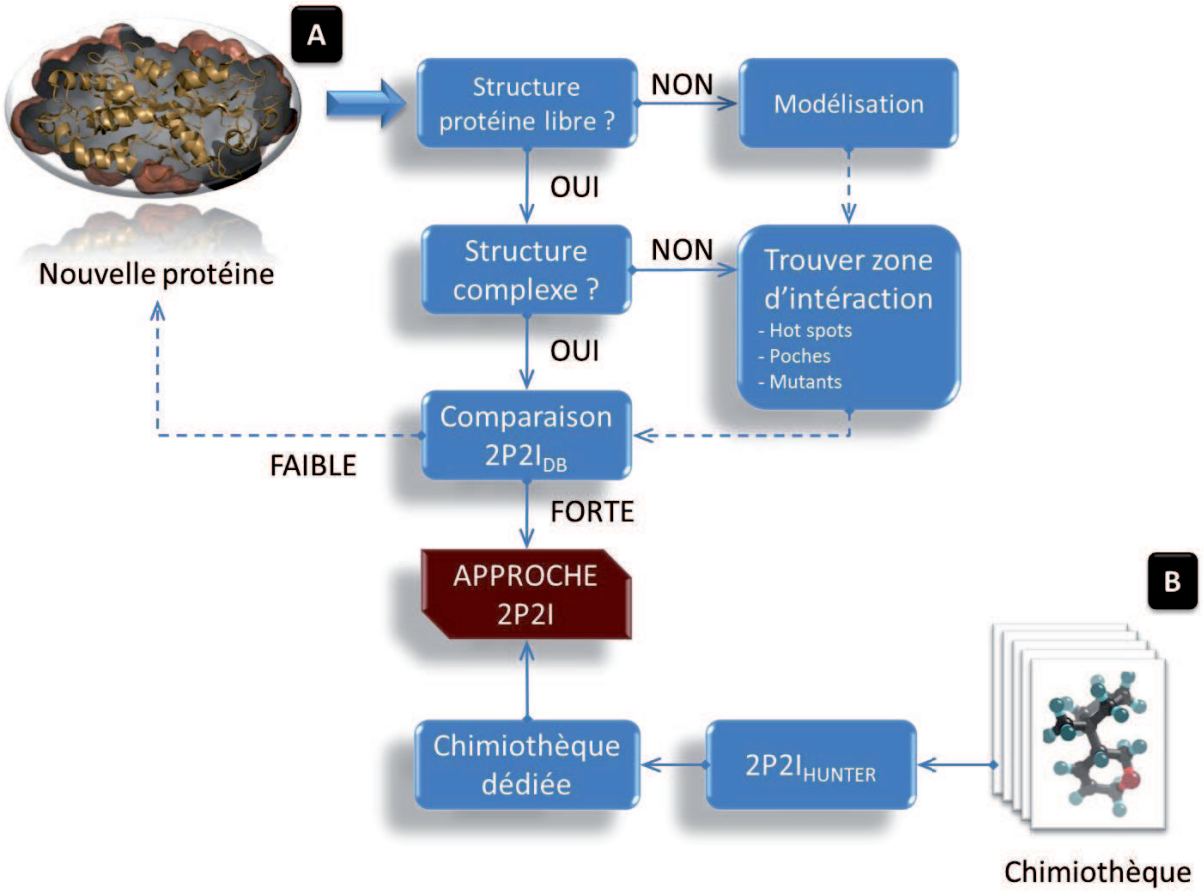


Figure 17 : Flowchart de l'approche de l'inhibition des interactions protéine-protéine développée au laboratoire.

L'ensemble de l'approche que nous avons mise au point à vue de définir la druggabilité d'un complexe protéine-protéine, repose aujourd'hui sur une étude qualitative des interfaces présentes dans $2P2I_{DB}$ (http://2p2idb.cnrs-mrs.fr/). La prochaine étape de ce projet consiste à mettre en place d'une approche quantitative par le biais d'une méthode d'apprentissage informatique. Cela permettra de définir l'espace conformationnel des interfaces protéine-

protéine pour lesquelles un inhibiteur orthostérique existe déjà, et d'accroitre la sensibilité et la spécificité de notre prédiction de la druggabilité.

L'étude atomique de l'ensemble des inhibiteurs orthostériques d'intéractions protéine-protéine nous a permis de définir l'espace chimique les caractérisant. Tous présents dans $2P2I_{DB}$, l'ensemble de ces petites molécules a servi à la mise en place d'une approche qualitative, ainsi que d'une approche quantitative permettant de définir l'espace chimique des PPIMs et de proposer un protocole par filtreage des larges collections de molécules (Figure 17 **B**).

Le protocole mis en place au sein du groupe : $2P2I_{HUNTER}$, per met de filtrer une chimiothèque afin de créer un sous-ensemble de molécules dédiées aux PPIs. $2P2I_{HUNTER}$ est composé de deux étapes, une qualitative et une quantitative. L'approche qualitative, appelée « **Règle des 4** » est un simple filtre permettant d'éliminer rapidement un grand nombre de molécules ne présentant pas les caractéristiques de modulateurs d'interactions protéine-protéine. Avec une forte sélectivité, et une spécificité plus modérée, cette Ro4 permet de filtrer en moyenne les chimiothèques commerciales à 60 %. L'approche quantitative, quant à elle, est basée sur une **machine à support de vecteurs**. Ayant appris des inhibiteurs présents dans $2P2I_{DB}$, cette approche permet, grâce à l'étude de onze paramètres (et non plus les 4 de la Ro4) de définir l'espace chimique auquel appartiennent ces molécules. En comparant les molécules des chimiothèques commerciales avec cet espace chimique et en ne gardant uniquement celles qui font partie de cet espace, nous filtrons en moyenne 96 % des composés.

La suite de ce travail consiste à valider le protocole au laboratoire dans un projet de criblage utilisant la chimiothèque finale composée de 1683 molécules par des méthodes biochimiques et biophysiques (HTRF et BRET essentiellement). Des collaborations ont été initialisées avec les sociétés CISBIO, Hybrigenics, ValiRX et les laboratoires du CRCM. Ce projet a pour but de cibler une dizaine de complexes importants en cancérologie et en virologie (projet « MATRIX »). Les résultats de ce projet, permettront de confirmer la pertinence de l'utilisation de cette chimiothèque dédiée lors de nouvelles campagnes de drug design orientées sur la modulation des interactions protéine-protéine et sera mis à disposition de la communauté scientifique.

REFERENCES BIBLIOGRAPHIQUES

Abad-Zapatero, C., and Metz, J.T. (2005). Ligand efficiency indices as guideposts for drug discovery. Drug Discov Today *10*, 464-469.

Abad-Zapatero, C., Perišić, O., Wass, J., Bento, A.P., Overington, J., Al-Lazikani, B., and Johnson, M.E. (2010). Ligand efficiency indices for an effective mapping of chemico-biological space: the concept of an atlas-like representation. Drug Discov Today *15*, 804-811.

Adams, J.M., and Cory, S. (1998). The Bcl-2 protein family: arbiters of cell survival. Science *281*, 1322-1326.

Amstad, E., and Reimhult, E. (2012). Nanoparticle actuated hollow drug delivery vehicles. Nanomedicine (Lond) *7*, 145-164.

An, J., Totrov, M., and Abagyan, R. (2005). Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics *4*, 752-761.

Arkin, M.R., Randal, M., DeLano, W.L., Hyde, J., Luong, T.N., Oslob, J.D., Raphael, D.R., Taylor, L., Wang, J., McDowell, R.S.*, et al.* (2003). Binding of small molecules to an adaptive protein-protein interface. Proceedings of the National Academy of Sciences of the United States of America *100*, 1603-1608 %U http://www.ncbi.nlm.nih.gov/pubmed/12582206.

Arkin, M.R., and Whitty, A. (2009). The road less traveled: modulating signal transduction enzymes by inhibiting their protein-protein interactions. Current Opinion in Chemical Biology *13*, 284-290 %U http://www.ncbi.nlm.nih.gov.gate281.inist.fr/pubmed/19553156.

Azmi, A.S., and Mohammad, R.M. (2009). Non-peptidic small molecule inhibitors against Bcl-2 for cancer therapy. Journal of Cellular Physiology *218*, 13-21 %U http://www.ncbi.nlm.nih.gov.gate11.inist.fr/pubmed/18767026.

Azmi, A.S., Wang, Z., Philip, P.A., Mohammad, R.M., and Sarkar, F.H. (2011). Emerging Bcl-2 inhibitors for the treatment of cancer. Expert Opin Emerg Drugs *16*, 59-70.

Bahadur, R.P., Chakrabarti, P., Rodier, F., and Janin, J. (2003). Dissecting subunit interfaces in homodimeric proteins. Proteins *53*, 708-719 %U http://www.ncbi.nlm.nih.gov/pubmed/14579361.

Bahadur, R.P., Chakrabarti, P., Rodier, F., and Janin, J. (2004). A dissection of specific and non-specific protein-protein interfaces. Journal of Molecular Biology *336*, 943-955.

Bahadur, R.P., Rodier, F., and Janin, J. (2007). A dissection of the protein-protein interfaces in icosahedral virus capsids. J Mol Biol *367*, 574-590.

Bahadur, R.P., and Zacharias, M. (2008). The interface of protein-protein complexes: analysis of contacts and prediction of interactions. Cellular and Molecular Life Sciences: CMLS *65*, 1059-1072.

Berg, T. (2008). Small-molecule inhibitors of protein-protein interactions. Current Opinion in Drug Discovery & Development *11*, 666-674.

Bird, G.H., Bernal, F., Pitter, K., and Walensky, L.D. (2008). Synthesis and biophysical characterization of stabilized alpha-helices of BCL-2 domains. Methods Enzymol *446*, 369-386.

Blazer, L.L., and Neubig, R.R. (2009). Small molecule protein-protein interaction inhibitors as CNS therapeutic agents: current progress and future hurdles. Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology *34*, 126-141 %U http://www.ncbi.nlm.nih.gov/pubmed/18800065.

Bogan, A.A., and Thorn, K.S. (1998). Anatomy of hot spots in protein interfaces. Journal of Molecular Biology *280*, 1-9 %U http://www.ncbi.nlm.nih.gov.gate1.inist.fr/pubmed/9653027.

Bohacek, R.S., McMartin, C., and Guida, W.C. (1996). The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev *16*, 3-50.

Bourgeas, R., Basse, M.-J., Morelli, X., and Roche, P. (2010). Atomic Analysis of Protein-Protein Interfaces with Known Inhibitors: The 2P2I Database. PLoS ONE *5*, e9598.

Brady, G.P., and Stouten, P.F. (2000). Fast prediction and visualization of protein binding pockets with PASS. Journal of Computer-Aided Molecular Design *14*, 383-401 %U http://www.ncbi.nlm.nih.gov/pubmed/10815774.

Braisted, A.C., Oslob, J.D., Delano, W.L., Hyde, J., McDowell, R.S., Waal, N., Yu, C., Arkin, M.R., and Raimundo, B.C. (2003). Discovery of a potent small molecule IL-2 inhibitor through fragment assembly. Journal of the American Chemical Society *125*, 3714-3715 %U http://www.ncbi.nlm.nih.gov/pubmed/12656598.

Buchwald, P. (2010). Small-molecule protein-protein interaction inhibitors: therapeutic potential in light of molecular size, chemical space, and ligand binding efficiency considerations. IUBMB Life *62*, 724-731.

Bunnage, M.E. (2011). Getting pharmaceutical R&D back on target. Nat Chem Biol *7*, 335-339.

Chakrabarti, P., and Janin, J. (2002). Dissecting protein-protein recognition sites. Proteins *47*, 334-343 %U http://www.ncbi.nlm.nih.gov/pubmed/11948787.

Cheng, G., Yu, A., and Malek, T.R. (2011). T-cell tolerance and the multi-functional role of IL-2R signaling in T-regulatory cells. Immunol Rev *241*, 63-76.

Clackson, T., and Wells, J.A. (1995). A hot spot of binding energy in a hormone-receptor interface. Science (New York, NY) *267*, 383-386.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. Machine Learning *20*, 273-297.

Das, J., and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol *6*, 92.

de Groot, B.L., van Aalten, D.M., Scheek, R.M., Amadei, A., Vriend, G., and Berendsen, H.J. (1997). Prediction of protein conformational freedom from distance constraints. Proteins *29*, 240-251.

Dobson, C.M. (2004). Chemical space and biology. Nature *432*, 824-828.

Dorr, P., Westby, M., Dobbs, S., Griffin, P., Irvine, B., Macartney, M., Mori, J., Rickett, G., Smith-Burchnell, C., Napier, C*., et al.* (2005). Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-

spectrum anti-human immunodeficiency virus type 1 activity. Antimicrob Agents Chemother *49*, 4721-4732.

Driggers, E.M., Hale, S.P., Lee, J., and Terrett, N.K. (2008). The exploration of macrocycles for drug discovery--an underexploited structural class. Nat Rev Drug Discov *7*, 608-624.

Dudkina, A.S., and Lindsley, C.W. (2007). Small molecule protein-protein inhibitors for the p53-MDM2 interaction. Current Topics in Medicinal Chemistry *7*, 952-960.

Durrant, J.D., and McCammon, J.A. (2011). Molecular dynamics simulations and drug discovery. BMC Biol *9*, 71.

Egner, U., and Hililig, R.C. (2008). A structural biology view of target drugability. Expert opinion on drug discovery *3*, 10.

Ertl, P. (2003). Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. J Chem Inf Comput Sci *43*, 374-380.

Euler-Chelpin, H.v. (1930). Nobel Lecture. Nobel Lectures, Chemistry *1922-1941*.

Eyrisch, S., and Helms, V. (2007). Transient pockets on protein surfaces involved in protein-protein interaction. Journal of Medicinal Chemistry *50*, 3457-3464 %U http://www.ncbi.nlm.nih.gov.gate3451.inist.fr/pubmed/17602601.

Eyrisch, S., and Helms, V. (2009). What induces pocket openings on protein surface patches involved in protein-protein interactions? Journal of Computer-Aided Molecular Design *23*, 73-86 %U http://www.ncbi.nlm.nih.gov.gate71.inist.fr/pubmed/18777159.

Fauman, E.B., Rai, B.K., and Huang, E.S. (2011). Structure-based druggability assessment--identifying suitable targets for small molecule therapeutics. Curr Opin Chem Biol *15*, 463-468.

Fischer, T.B., Arunachalam, K.V., Bailey, D., Mangual, V., Bakhru, S., Russo, R., Huang, D., Paczkowski, M., Lalchandani, V., Ramachandra, C.*, et al.* (2003). The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. Bioinformatics (Oxford, England) *19*, 1453-1454 %U http://www.ncbi.nlm.nih.gov/pubmed/12874065.

Fry, D. (2011). Small-Molecule Inhibitors of Protein-Protein Interactions (Nutley, New Jersey).

Fry, D.C. (2006). Protein-protein interactions as targets for small molecule drug discovery. Biopolymers *84*, 535-552.

Fry, D.C., Emerson, S.D., Palme, S., Vu, B.T., Liu, C.-M., and Podlaski, F. (2004). NMR structure of a complex between MDM2 and a small molecule inhibitor. Journal of Biomolecular NMR *30*, 163-173 %U http://www.ncbi.nlm.nih.gov/pubmed/15557803.

Fuller, J.C., Burgoyne, N.J., and Jackson, R.M. (2009). Predicting druggable binding sites at the protein-protein interface. Drug Discovery Today *14*, 155-161.

Gaffen, S.L. (2001). Signaling domains of the interleukin 2 receptor. Cytokine *14*, 63-77.

Garcia-Garcia, J., Schleker, S., Klein-Seetharaman, J., and Oliva, B. (2012). BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. Nucleic Acids Res *40*, W147-151.

Geppert, T., Hoy, B., Wessler, S., and Schneider, G. (2011). Context-based identification of protein-protein interfaces and "hot-spot" residues. Chem Biol *18*, 344-353.

Gerstein, M., Tsai, J., and Levitt, M. (1995). The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. J Mol Biol *249*, 955-966.

Goodford, P.J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. Journal of Medicinal Chemistry *28*, 849-857 %U http://www.ncbi.nlm.nih.gov/pubmed/3892003.

Goodsell, D.S., and Olson, A.J. (2000). Structural symmetry and protein function. Annu Rev Biophys Biomol Struct *29*, 105-153.

Grasberger, B.L., Lu, T., Schubert, C., Parks, D.J., Carver, T.E., Koblish, H.K., Cummings, M.D., LaFrance, L.V., Milkiewicz, K.L., Calvo, R.R., *et al.* (2005). Discovery and cocrystal structure of benzodiazepinedione HDM2 antagonists that activate p53 in cells. Journal of Medicinal Chemistry *48*, 909-912 %U http://www.ncbi.nlm.nih.gov/pubmed/15715460.

Guerois, R., Nielsen, J.E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol *320*, 369-387.

Guharoy, M., and Chakrabarti, P. (2009). Empirical estimation of the energetic contribution of individual interface residues in structures of protein-protein complexes. J Comput Aided Mol Des *23*, 645-654.

Guney, E., Tuncbag, N., Keskin, O., and Gursoy, A. (2008). HotSprint: database of computational hot spots in protein interfaces. Nucleic Acids Research *36*, D662-666 %U http://www.ncbi.nlm.nih.gov.gate661.inist.fr/pubmed/17959648.

Hann, M.M., and Oprea, T.I. (2004). Pursuing the leadlikeness concept in pharmaceutical research. Curr Opin Chem Biol *8*, 255-263.

Harpaz, Y., Gerstein, M., and Chothia, C. (1994). Volume changes on protein folding. Structure *2*, 641-649.

Harrington, P.E., and Fotsch, C. (2007). Calcium sensing receptor activators: calcimimetics. Curr Med Chem *14*, 3027-3034.

Harris, R., Olson, A.J., and Goodsell, D.S. (2008). Automated prediction of ligand-binding sites in proteins. Proteins *70*, 1506-1517.

Henchey, L.K., Jochim, A.L., and Arora, P.S. (2008). Contemporary strategies for the stabilization of peptides in the alpha-helical conformation. Curr Opin Chem Biol *12*, 692-697.

Hendlich, M., Rippmann, F., and Barnickel, G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model *15*, 359-363, 389.

Higueruelo, A.P., Schreyer, A., Bickerton, G.R.J., Pitt, W.R., Groom, C.R., and Blundell, T.L. (2009). Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. Chemical Biology & Drug Design *74*, 457-467.

Hopkins, A.L., and Groom, C.R. (2002). The druggable genome. Nat Rev Drug Discov *1*, 727-730.

Hopkins, A.L., Groom, C.R., and Alex, A. (2004). Ligand efficiency: a useful metric for lead selection. Drug Discov Today *9*, 430-431.

Hyde, J., Braisted, A.C., Randal, M., and Arkin, M.R. (2003). Discovery and characterization of cooperative ligand binding in the adaptive region of interleukin-2. Biochemistry *42*, 6475-6483 %U http://www.ncbi.nlm.nih.gov/pubmed/12767230.

Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M.*, et al.* (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Research *37*, D412-416 %U http://www.ncbi.nlm.nih.gov/pubmed/18940858.

Jochim, A.L., and Arora, P.S. (2010). Systematic analysis of helical protein interfaces reveals targets for synthetic inhibitors. ACS Chem Biol *5*, 919-923.

Johnson, K., Choi, Y., Wu, Z., Ciardelli, T., Granzow, R., Whalen, C., Sana, T., Pardee, G., Smith, K., and Creasey, A. (1994). Soluble IL-2 receptor beta and gamma subunits: ligand binding and cooperativity. Eur Cytokine Netw *5*, 23-34.

Jones, S., and Thornton, J.M. (1995). Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol *63*, 31-65.

Jones, S., and Thornton, J.M. (1996). Principles of protein-protein interactions. Proceedings of the National Academy of Sciences of the United States of America *93*, 13-20.

Josse, J. (2008). FactoMineR: An R Package for Multivariate Analysis. Journal Of Statistical Software *25*, 1-18.

Kalidas, Y., and Chandra, N. (2008). PocketDepth: a new depth based algorithm for identification of ligand binding sites in proteins. J Struct Biol *161*, 31-42.

Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2009). ConsensusPathDB--a database for integrating human functional interaction networks. Nucleic Acids Res *37*, D623-628.

Keller, T.H., Pichota, A., and Yin, Z. (2006). A practical view of 'druggability'. Curr Opin Chem Biol *10*, 357-361.

Kim, D.E., Chivian, D., and Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res *32*, W526-531.

Koblish, H.K., Zhao, S., Franks, C.F., Donatelli, R.R., Tominovich, R.M., LaFrance, L.V., Leonard, K.A., Gushue, J.M., Parks, D.J., Calvo, R.R.*, et al.* (2006). Benzodiazepinedione inhibitors of the Hdm2:p53 complex suppress human tumor cell proliferation in vitro and sensitize tumors to doxorubicin in vivo. Mol Cancer Ther *5*, 160-169.

Koes, D.R., and Camacho, C.J. (2011). Small-Molecule Inhibitor Starting Points Learned From Protein-Protein Interaction Inhibitor Structure. Bioinformatics.

Kortemme, T., and Baker, D. (2002). A simple physical model for binding energy hot spots in protein-protein complexes. Proceedings of the National Academy of Sciences of the United States of America *99*, 14116-14121 %U http://www.ncbi.nlm.nih.gov.gate14111.inist.fr/pubmed/12381794.

Koshland, D.E. (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis. Proc Natl Acad Sci U S A *44*, 98-104.

Kozakov, D., Hall, D.R., Chuang, G.Y., Cencic, R., Brenke, R., Grove, L.E., Beglov, D., Pelletier, J., Whitty, A., and Vajda, S. (2011). Structural conservation of druggable hot spots in protein-protein interfaces. Proc Natl Acad Sci U S A *108*, 13528-13533.

Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. J Mol Biol *372*, 774-797.

Krüger, D.M., and Gohlke, H. (2010). DrugScorePPI webserver: fast and accurate in silico alanine scanning for scoring protein-protein interactions. Nucleic Acids Res *38*, W480-486.

Laskowski, R.A. (1995). SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. Journal of Molecular Graphics *13*, 323-330 %U http://www.sciencedirect.com/science/article/B326VNC-323Y450FT-328/322/c300fbb327a755d336d619e328b835f393bb318.

Laskowski, R.A., Luscombe, N.M., Swindells, M.B., and Thornton, J.M. (1996). Protein clefts in molecular recognition and function. Protein Sci *5*, 2438-2452.

Laurie, A.T.R., and Jackson, R.M. (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics (Oxford, England) *21*, 1908-1916.

Lee, B., and Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. J Mol Biol *55*, 379-400.

Lee, B.C., Chu, T.K., Dill, K.A., and Zuckermann, R.N. (2008). Biomimetic nanostructures: creating a high-affinity zinc-binding site in a folded nonbiological polymer. J Am Chem Soc *130*, 8847-8855.

Lehne, B., and Schlitt, T. (2009). Protein-protein interaction databases: keeping up with growing interactomes. Human Genomics *3*, 291-297.

Li, Q., Wang, Y., and Bryant, S.H. (2009). A novel method for mining highly imbalanced high-throughput screening data in PubChem. Bioinformatics *25*, 3310-3316.

Lievens, S., Eyckerman, S., Lemmens, I., and Tavernier, J. (2010). Large-scale protein interactome mapping: strategies and opportunities. Expert Rev Proteomics *7*, 679-690.

Lin, M., Ming, A., and Zhao, M. (2006). Two-dose basiliximab compared with two-dose daclizumab in renal transplantation: a clinical study. Clin Transplant *20*, 325-329.

Lipinski, C., and Hopkins, A. (2004). Navigating chemical space for biology and medicine. Nature *432*, 855-861.

Lipinski, C., Lombardo, F., Dominy, B., and Feeney, P. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev *46*, 3-26.

Lise, S., Archambeau, C., Pontil, M., and Jones, D. (2009). Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. BMC Bioinformatics *10*, 365 %U http://www.ncbi.nlm.nih.gov/pubmed/19878545.

Llambi, F., and Green, D.R. (2011). Apoptosis and oncogenesis: give and take in the BCL-2 family. Curr Opin Genet Dev *21*, 12-20.

Lo Conte, L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. Journal of Molecular Biology *285*, 2177-2198 %U http://www.ncbi.nlm.nih.gov/pubmed/9925793.

Lovering, F., Bikker, J., and Humblet, C. (2009). Escape from flatland: increasing saturation as an approach to improving clinical success. J Med Chem *52*, 6752-6756.

Ma, B., and Nussinov, R. (2007). Trp/Met/Phe hot spots in protein-protein interactions: potential targets in drug design. Current Topics in Medicinal Chemistry *7*, 999-1005 %U http://www.ncbi.nlm.nih.gov.gate1001.inist.fr/pubmed/17508933.

Malek, T.R. (2008). The biology of interleukin-2. Annu Rev Immunol *26*, 453-479.

Marson, C.M. (2011). New and unusual scaffolds in medicinal chemistry. Chem Soc Rev *40*, 5514-5533.

McDowall, M.D., Scott, M.S., and Barton, G.J. (2009). PIPs: human protein-protein interaction prediction database. Nucleic Acids Res *37*, D651-656.

Metz, A., Pfleger, C., Kopitz, H., Pfeiffer-Marek, S., Baringhaus, K.H., and Gohlke, H. (2012). Hot spots and transient pockets: predicting the determinants of small-molecule binding to a protein-protein interface. J Chem Inf Model *52*, 120-133.

Morelli, X., Bourgeas, R., and Roche, P. (2011). Chemical and structural lessons from recent successes in protein-protein interaction inhibition (2P2I). Curr Opin Chem Biol *15*, 475-481.

Mullard, A. (2012). Protein–protein interaction inhibitors get into the groove. Nature Reviews Drug Discovery *11*, 173-175.

Negi, S.S., Schein, C.H., Oezguen, N., Power, T.D., and Braun, W. (2007). InterProSurf: a web server for predicting interacting sites on protein surfaces. Bioinformatics (Oxford, England) *23*, 3397-3399 %U http://www.ncbi.nlm.nih.gov.gate3391.inist.fr/pubmed/17933856.

Nelson, B.H., and Willerford, D.M. (1998). Biology of the interleukin-2 receptor. Adv Immunol *70*, 1-81.

Neugebauer, A., Hartmann, R.W., and Klein, C.D. (2007). Prediction of protein-protein interaction inhibitors by chemoinformatics and machine learning methods. Journal of Medicinal Chemistry *50*, 4665-4668.

Nguyen, M., Marcellus, R.C., Roulston, A., Watson, M., Serfass, L., Murthy Madiraju, S.R., Goulet, D., Viallet, J., Bélec, L., Billot, X.*, et al.* (2007). Small molecule obatoclax (GX15-070) antagonizes MCL-1 and overcomes MCL-1-mediated resistance to apoptosis. Proc Natl Acad Sci U S A *104*, 19512-19517.

Nooren, I.M., and Thornton, J.M. (2003). Diversity of protein-protein interactions. EMBO J *22*, 3486-3492.

Ofran, Y., and Rost, B. (2007). Protein-protein interaction hotspots carved into sequences. PLoS Comput Biol *3*, e119.

Oltersdorf, T., Elmore, S.W., Shoemaker, A.R., Armstrong, R.C., Augeri, D.J., Belli, B.A., Bruncko, M., Deckwerth, T.L., Dinges, J., Hajduk, P.J.*, et al.* (2005). An inhibitor of Bcl-2 family proteins induces regression of solid tumours. Nature *435*, 677-681 %U http://www.ncbi.nlm.nih.gov.gate671.inist.fr/pubmed/15902208.

Overington, J.P., Al-Lazikani, B., and Hopkins, A.L. (2006). How many drug targets are there? Nat Rev Drug Discov *5*, 993-996.

Pagliaro, L., Felding, J., Audouze, K., Nielsen, S.J., Terry, R.B., Krog-Jensen, C., and Butcher, S. (2004). Emerging classes of protein-protein interaction inhibitors and new tools for their development. Current Opinion in Chemical Biology *8*, 442-449.

Pal, A., Chakrabarti, P., Bahadur, R., Rodier, F., and Janin, J. (2007). Peptide segments in protein-protein interfaces. J Biosci *32*, 101-111.

Parks, D.J., Lafrance, L.V., Calvo, R.R., Milkiewicz, K.L., Gupta, V., Lattanze, J., Ramachandren, K., Carver, T.E., Petrella, E.C., Cummings, M.D.*, et al.* (2005). 1,4-Benzodiazepine-2,5-diones as small molecule antagonists of the HDM2-p53 interaction: discovery and SAR. Bioorganic & Medicinal Chemistry Letters *15*, 765-770 %U http://www.ncbi.nlm.nih.gov/pubmed/15664854.

Patel, S., and Player, M.R. (2008). Small-molecule inhibitors of the p53-HDM2 interaction for the treatment of cancer. Expert Opinion on Investigational Drugs *17*, 1865-1882.

Patschull, A.O., Gooptu, B., Ashford, P., Daviter, T., and Nobeli, I. (2012). In silico assessment of potential druggable pockets on the surface of α1-antitrypsin conformers. PLoS One *7*, e36612.

Petros, A.M., Nettesheim, D.G., Wang, Y., Olejniczak, E.T., Meadows, R.P., Mack, J., Swift, K., Matayoshi, E.D., Zhang, H., Thompson, C.B.*, et al.* (2000). Rationale for Bcl-xL/Bad peptide complex formation from structure, mutagenesis, and biophysical studies. Protein Sci *9*, 2528-2534.

Petsko, G.A., and Ringe, D. (2008). Structure et fonction des protéines (Bruxelles).

Prasad, T.S., Kandasamy, K., and Pandey, A. (2009). Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. Methods Mol Biol *577*, 67-79.

Prieto, C., and De Las Rivas, J. (2006). APID: Agile Protein Interaction DataAnalyzer. Nucleic Acids Research *34*, W298-302 %U http://www.ncbi.nlm.nih.gov.gate291.inist.fr/pubmed/16845013.

Raimundo, B.C., Oslob, J.D., Braisted, A.C., Hyde, J., McDowell, R.S., Randal, M., Waal, N.D., Wilkinson, J., Yu, C.H., and Arkin, M.R. (2004). Integrating fragment assembly and biophysical methods in the chemical advancement of small-molecule antagonists of IL-2: an approach for inhibiting protein-protein interactions. Journal of Medicinal Chemistry *47*, 3111-3130 %U http://www.ncbi.nlm.nih.gov/pubmed/15163192.

Reynes, C., Host, H., Camproux, A.C., Laconde, G., Leroux, F., Mazars, A., Deprez, B., Fahraeus, R., Villoutreix, B.O., and Sperandio, O. (2010). Designing focused chemical libraries enriched in protein-protein interaction inhibitors using machine-learning methods. PLoS Comput Biol *6*, e1000695.

Reynolds, C., Damerell, D., and Jones, S. (2009). ProtorP: a protein-protein interaction analysis server. Bioinformatics (Oxford, England) *25*, 413-414.

Roche, P., and Morelli, X. (2010). Protein-Protein Interaction Inhibition (2P2I): Mixed Methodologies for the Acceleration of Lead Discovery. In *In silico* lead discovery, M. Miteva, ed. (Bentham), pp. 118-143.

Ross, N.T., Katt, W.P., and Hamilton, A.D. (2010). Synthetic mimetics of protein secondary structure domains. Philos Transact A Math Phys Eng Sci *368*, 989-1008.

Sadowsky, J.D., Murray, J.K., Tomita, Y., and Gellman, S.H. (2007). Exploration of backbone space in foldamers containing alpha- and beta-amino acid residues: developing protease-resistant oligomers that bind tightly to the BH3-recognition cleft of Bcl-xL. Chembiochem *8*, 903-916.

Sattler, M., Liang, H., Nettesheim, D., Meadows, R.P., Harlan, J.E., Eberstadt, M., Yoon, H.S., Shuker, S.B., Chang, B.S., Minn, A.J., *et al.* (1997). Structure of Bcl-xL-Bak peptide complex: recognition between regulators of apoptosis. Science *275*, 983-986.

Schmidtke, P., Souaille, C., Estienne, F., Baurin, N., and Kroemer, R.T. (2010). Large-scale comparison of four binding site detection algorithms. J Chem Inf Model *50*, 2191-2200.

Schneider, M. (2004). A rational approach to maximize success rate in target discovery. Arch Pharm (Weinheim) *337*, 625-633.

Schuffenhauer, A., Popov, M., Schopfer, U., Acklin, P., Stanek, J., and Jacoby, E. (2004). Molecular diversity management strategies for building and enhancement of diverse and focused lead discovery compound screening collections. Comb Chem High Throughput Screen *7*, 771-781.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. Nucleic Acids Res *33*, W382-388.

Segura, J., and Fernandez-Fuentes, N. (2011). PCRPi-DB: a database of computationally annotated hot spots in protein interfaces. Nucleic Acids Res *39*, D755-760.

Segura Mora, J., Assi, S.A., and Fernandez-Fuentes, N. (2010). Presaging critical residues in protein interfaces-web server (PCRPi-W): a web server to chart hot spots in protein interfaces. PLoS One *5*, e12352.

Shi, X., Betzi, S., Lugari, A., Opi, S., Restouin, A., Parrot, I., Martinez, J., Zimmermann, P., Lecine, P., Huang, M., *et al.* (2012). Structural recognition mechanisms between human Src homology domain 3 (SH3) and ALG-2-interacting protein X (Alix). FEBS Lett *586*, 1759-1764.

Shoichet, B.K., and Kobilka, B.K. (2012). Structure-based drug screening for G-protein-coupled receptors. Trends Pharmacol Sci *33*, 268-272.

Sonavane, S., and Chakrabarti, P. (2008). Cavities and atomic packing in protein structures and interfaces. PLoS Computational Biology *4*, e1000188 %U http://www.ncbi.nlm.nih.gov.gate1000181.inist.fr/pubmed/19005575.

Sperandio, O., Reynès, C., Camproux, A., and Villoutreix, B. (2010). Rationalizing the chemical space of protein-protein interaction inhibitors. Drug Discov Today *15*, 220-229.

Stumpf, M.P., Thorne, T., de Silva, E., Stewart, R., An, H.J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. Proc Natl Acad Sci U S A *105*, 6959-6964.

Sugaya, N., and Furuya, T. (2011). Dr. PIAS: an integrative system for assessing the druggability of protein-protein interactions. BMC Bioinformatics *12*, 50.

Sugaya, N., and Ikeda, K. (2009). Assessing the druggability of protein-protein interactions by a supervised machine-learning method. BMC Bioinformatics *10*, 263 %U http://www.ncbi.nlm.nih.gov.gate261.inist.fr/pubmed/19703312.

Sugaya, N., Ikeda, K., Tashiro, T., Takeda, S., Otomo, J., Ishida, Y., Shiratori, A., Toyoda, A., Noguchi, H., Takeda, T., *et al.* (2007). An integrative in silico approach for discovering candidates for drug-targetable protein-protein interactions in interactome data. BMC Pharmacology *7*, 10 %U http://www.ncbi.nlm.nih.gov.gate11.inist.fr/pubmed/17705877.

Sugaya, N., Kanai, S., and Furuya, T. (2012). Dr. PIAS 2.0: an update of a database of predicted druggable protein-protein interactions. Database (Oxford) *2012*, bas034.

Tanaka, K., Kanazawa, T., Sugawara, K., Horiuchi, S., Takashima, Y., and Okada, H. (2011). A cytoplasm-sensitive peptide vector cross-linked with dynein light chain association sequence (DLCAS) enhances gene expression. Int J Pharm *419*, 231-234.

Taniguchi, T., Matsui, H., Fujita, T., Takaoka, C., Kashima, N., Yoshimoto, R., and Hamuro, J. (1983). Structure and expression of a cloned cDNA for human interleukin-2. Nature *302*, 305-310.

Thaker, T.M., Kaya, A.I., Preininger, A.M., Hamm, H.E., and Iverson, T.M. (2012). Allosteric mechanisms of G protein-Coupled Receptor signaling: a structural perspective. Methods Mol Biol *796*, 133-174.

Thangudu, R.R., Bryant, S.H., Panchenko, A.R., and Madej, T. (2012). Modulating protein-protein interactions with small molecules: the importance of binding hotspots. J Mol Biol *415*, 443-453.

Thanos, C.D., DeLano, W.L., and Wells, J.A. (2006). Hot-spot mimicry of a cytokine receptor by a small molecule. Proceedings of the National Academy of Sciences of the United States of America *103*, 15422-15427 %U http://www.ncbi.nlm.nih.gov/pubmed/17032757.

Thanos, C.D., Randal, M., and Wells, J.A. (2003). Potent small-molecule binding to a dynamic hot spot on IL-2. Journal of the American Chemical Society *125*, 15280-15281 %U http://www.ncbi.nlm.nih.gov/pubmed/14664558.

Thorn, K.S., and Bogan, A.A. (2001). ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics (Oxford, England) *17*, 284-285 %U http://www.ncbi.nlm.nih.gov/pubmed/11294795.

Tse, C., Shoemaker, A.R., Adickes, J., Anderson, M.G., Chen, J., Jin, S., Johnson, E.F., Marsh, K.C., Mitten, M.J., Nimmer, P.*, et al.* (2008). ABT-263: a potent and orally bioavailable Bcl-2 family inhibitor. Cancer Res *68*, 3421-3428.

Tuncbag, N., Gursoy, A., and Keskin, O. (2009). Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. Bioinformatics (Oxford, England) *25*, 1513-1520 %U http://www.ncbi.nlm.nih.gov.gate1511.inist.fr/pubmed/19357097.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P.*, et al.* (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature *403*, 623-627.

Vassilev, L.T., Vu, B.T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., Kong, N., Kammlott, U., Lukacs, C., Klein, C.*, et al.* (2004). In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. Science (New York, NY) *303*, 844-848 %U http://www.ncbi.nlm.nih.gov/pubmed/14704432.

Venkatesan, K., Rual, J.F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.I.*, et al.* (2009). An empirical framework for binary interactome mapping. Nat Methods *6*, 83-90.

Vlieghe, P., Lisowski, V., Martinez, J., and Khrestchatisky, M. (2010). Synthetic therapeutic peptides: science and market. Drug Discov Today *15*, 40-56.

Vu, B.T., and Vassilev, L. (2011). Small-Molecule Inhibitors of the p53-MDM2 Interaction. Curr Top Microbiol Immunol *348*, 151-172.

Waal, N.D., Yang, W., Oslob, J.D., Arkin, M.R., Hyde, J., Lu, W., McDowell, R.S., Yu, C.H., and Raimundo, B.C. (2005). Identification of nonpeptidic small-molecule inhibitors of interleukin-2. Bioorganic & Medicinal Chemistry Letters *15*, 983-987 %U http://www.ncbi.nlm.nih.gov/pubmed/15686897.

Waldmann, H. (2003). The new immunosuppression: just kill the T cell. Nat Med *9*, 1259-1260.

Walensky, L.D., Kung, A.L., Escher, I., Malia, T.J., Barbuto, S., Wright, R.D., Wagner, G., Verdine, G.L., and Korsmeyer, S.J. (2004). Activation of apoptosis in vivo by a hydrocarbon-stapled BH3 helix. Science (New York, NY) *305*, 1466-1470 %U http://www.ncbi.nlm.nih.gov/pubmed/15353804.

Wang, H.M., and Smith, K.A. (1987). The interleukin 2 receptor. Functional consequences of its bimolecular structure. J Exp Med *166*, 1055-1069.

Wang, Y., Bolton, E., Dracheva, S., Karapetyan, K., Shoemaker, B.A., Suzek, T.O., Wang, J., Xiao, J., Zhang, J., and Bryant, S.H. (2010). An overview of the PubChem BioAssay resource. Nucleic Acids Res *38*, D255-266.

Wanner, J., Fry, D.C., Peng, Z., and Roberts, J. (2011). Druggability assessment of protein-protein interfaces. Future Med Chem *3*, 2021-2038.

Weisel, M., Proschak, E., and Schneider, G. (2007). PocketPicker: analysis of ligand binding-sites with shape descriptors. Chemistry Central Journal *1*, 7 %U http://www.ncbi.nlm.nih.gov/pubmed/17880740.

Weissten. http://mathworld.wolfram.com/Eccentricity.html.

Wells, J.A. (1995). Structural and functional epitopes in the growth hormone receptor complex. Bio/Technology (Nature Publishing Company) *13*, 647-651.

Wells, J.A., and McClendon, C.L. (2007). Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature *450*, 1001-1009.

wikipedia (2010). List of pharmaceutical companies, pp. http://en.wikipedia.org/wiki/List_of_pharmaceutical_companies.

Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. Drug Discov Today *11*, 1046-1053.

Wilson, A.J. (2009). Inhibition of protein-protein interactions using designed molecules. Chem Soc Rev *38*, 3289-3300.

Wilson, C.G., and Arkin, M.R. (2011). Small-Molecule Inhibitors of IL-2/IL-2R: Lessons Learned and Applied. Curr Top Microbiol Immunol *348*, 25-59.

Wong, Y.S. (2012). Exploring chemical space: recent advances in chemistry. Methods Mol Biol *800*, 11-23.

Xie, X.Q. (2010). Exploiting PubChem for Virtual Screening. Expert Opin Drug Discov *5*, 1205-1220.

Yin, H., Lee, G.-i., Park, H.S., Payne, G.A., Rodriguez, J.M., Sebti, S.M., and Hamilton, A.D. (2005). Terphenyl-based helical mimetics that disrupt the p53/HDM2 interaction. Angewandte Chemie (International Ed in English) *44*, 2704-2707 %U http://www.ncbi.nlm.nih.gov.gate2701.inist.fr/pubmed/15765497.

Zhu, X., and Mitchell, J.C. (2011). KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. Proteins *79*, 2671-2683.

ANNEXE

Amplitude of Pancreatic Lipase Lid Opening in Solution and Identification of Spin Label Conformational Subensembles by Combining Continuous Wave and Pulsed EPR Spectroscopy and Molecular Dynamics.

Sebastien Ranaldi, Valérie Belle, Mireille Woudstra, **Raphael Bourgeas**, Bruno Guigliarelli, Philippe Roche, Hervé Vezin, Fredéric Carrière and André Fournel.

Biochemistry. 2010 Mar 16;49(10):2140-9.

The opening of the lid that controls the access to the active site of human pancreatic lipase (HPL) was measured from the magnetic interaction between two spin labels grafted on this enzyme. One spin label was introduced at a rigid position in HPL where an accessible cysteine residue (C181) naturally occurs. A second spin label was covalently bound to the mobile lid after introducing a cysteine residue at position 249 by site-directed mutagenesis. Double electron-electron resonance (DEER) experiments allowed the estimation of a distance of 19 (2A° between the spin labels when bilabeled HPL was alone in a frozen solution, i.e., with the lid in the closed conformation. A magnetic interaction was however detected by continuous wave EPR experiments, suggesting that a fraction of bilabeled HPL contained spin labels separated by a shorter distance. These results could be interpreted by the presence of two conformational subensembles for the spin label lateral chain at position 249 when the lid was closed. The existence of these conformational subensembles was revealed by molecular dynamics experiments and confirmed by the simulation of the EPR spectrum.When the lid opening was induced by the addition of bile salts and colipase, a larger distance of 43(2A° between the two spin labels was estimated from DEER experiments. The distances measured between the spin labels grafted at positions 181 and 249 were in good agreement with those estimated from the known X-ray structures of HPL in the closed and open conformations, but for the first time, the amplitude of the lid opening was measured in solution or in a frozen solution in the presence of amphiphiles.

Ma contribution dans ce travail a consisté en l'analyse des dynamiques moléculaires réalisées sur la Lipase Pancréatique Humaine. Notamment, l'étude des angles Chi1 et Chi2 du « spin label » en position 249. J'ai aussi réalisé la figure 5 représentant la distribution de ces angles dihèdres.

2140 *Biochemistry* **2010,** *49,* 2140–2149
DOI: 10.1021/bi901918f

BIOCHEMISTRY
including biophysical chemistry & molecular biology
Article

Amplitude of Pancreatic Lipase Lid Opening in Solution and Identification of Spin Label Conformational Subensembles by Combining Continuous Wave and Pulsed EPR Spectroscopy and Molecular Dynamics[†]

Sebastien Ranaldi,[‡] Valérie Belle,[‡] Mireille Woudstra,[‡] Raphael Bourgeas,[§] Bruno Guigliarelli,[‡] Philippe Roche,[§] Hervé Vezin,[⊥] Frédéric Carrière,*[,∥] and André Fournel*[,‡]

[‡]*CNRS Laboratoire de Bioénergétique et Ingénierie des Protéines, UPR 9036,* [§]*CNRS Laboratoire Interactions et Modulateurs de Réponses, FRE3083,* [∥]*CNRS Laboratoire d'Enzymologie Interfaciale et de Physiologie de la Lipolyse, UPR 9025, Institut de Microbiologie de la Méditerranée, Aix-Marseille Universités, Marseille, France, and* [⊥]*CNRS Laboratoire de Chimie Organique et Macromoléculaire, UMR 8009, Villeneuve d'Ascq, France*

Received November 8, 2009; Revised Manuscript Received February 3, 2010

ABSTRACT: The opening of the lid that controls the access to the active site of human pancreatic lipase (HPL) was measured from the magnetic interaction between two spin labels grafted on this enzyme. One spin label was introduced at a rigid position in HPL where an accessible cysteine residue (C181) naturally occurs. A second spin label was covalently bound to the mobile lid after introducing a cysteine residue at position 249 by site-directed mutagenesis. Double electron−electron resonance (DEER) experiments allowed the estimation of a distance of 19 ± 2 Å between the spin labels when bilabeled HPL was alone in a frozen solution, i.e., with the lid in the closed conformation. A magnetic interaction was however detected by continuous wave EPR experiments, suggesting that a fraction of bilabeled HPL contained spin labels separated by a shorter distance. These results could be interpreted by the presence of two conformational subensembles for the spin label lateral chain at position 249 when the lid was closed. The existence of these conformational subensembles was revealed by molecular dynamics experiments and confirmed by the simulation of the EPR spectrum. When the lid opening was induced by the addition of bile salts and colipase, a larger distance of 43 ± 2 Å between the two spin labels was estimated from DEER experiments. The distances measured between the spin labels grafted at positions 181 and 249 were in good agreement with those estimated from the known X-ray structures of HPL in the closed and open conformations, but for the first time, the amplitude of the lid opening was measured in solution or in a frozen solution in the presence of amphiphiles.

Lipases (triacylglycerol hydrolase, EC 3.1.1.3) are key enzymes in major physiological processes such as fat digestion and lipoprotein metabolism (*1*, *2*). These enzymes are also used in many industrial processes (biotransformation of oils and fats, synthesis of structured triacylglycerols, enantioselective reactions in organic synthesis) and products such as detergents for cleaning fat stains (*3*, *4*). Understanding their mechanism of action for then improving it is therefore a major challenge in biotechnology. One particular structural feature of lipases is the so-called "lid" that controls the access to the active site and the amphiphilic properties of these enzymes that are highly soluble in water but act at the surface of oil droplets (*5*). When the lid is "closed", the active site is not accessible to solvent, and the enzyme mainly presents a hydrophilic surface (*6*−*8*). When the lid is "open", the active site becomes accessible and functional, and a large hydrophobic surface is revealed at the surface of the protein surrounding the active site (*9*−*11*). This large hydrophobic surface becomes part of the active site, but it is also involved in the interaction of the lipase with the lipid−water interface. All of these structural characteristics have been revealed by X-ray crystallography since the last 20 years, the open conformations of lipases being often obtained by crystallization of lipase−inhibitor complexes and cocrystallization with detergents (*9*−*15*). Although these findings have been major breakthroughs in lipase studies, the structural behavior of lipases in solution, in the presence of detergent or organic solvents, or when they bind at the interface of two liquid phases remains still largely unknown. The use of NMR spectroscopy is not largely developed due to the high molecular masses of many lipases, and only a few lipases like cutinase (*16*, *17*) and *Pseudomonas mendocina* lipase (*18*) have been structurally characterized in solution by NMR.

The use of site-directed spin labeling (SDSL)[1] coupled to electron paramagnetic resonance spectroscopy (EPR) was however recently introduced for studying the conformational changes of the lid in human pancreatic lipase (HPL). By grafting a nitroxide spin label (MTSL) on the HPL lid at position 249, it was possible to identify specific EPR spectra of the spin label corresponding to the closed and open lid, and these spectra were later used for monitoring the HPL lid opening in solution (*19*). In particular, this study showed how an increasing concentration of

[†]The Ph.D. research of S.R. was supported by a grant from the French Ministery of Research and Education.

*To whom correspondence should be addressed. A.F.: tel, (33) 4 91 16 41 34; fax, (33) 4 91 71 58 57; e-mail, fournel@ifr88.cnrs-mrs.fr. F.C.: tel, (33) 4 91 16 41 34; fax, (33) 4 91 71 58 57; e-mail, carriere@ifr88.cnrs-mrs.fr.

[1]Abbreviations: CW, continuous wave; DEER, double electron−electron resonance; EPR, electron paramagnetic resonance spectroscopy; HPL, human pancreatic lipase; MD, molecular dynamics; MTSL, (1-oxy-2,2,5,5-tetramethyl-Δ^3-pyrroline-3-methyl)methanethiosulfonate; SDSL, site-directed spin labeling.

 Published on Web 02/05/2010

bile salts above their critical micellar concentration promotes the opening of the lid and how colipase, the specific HPL cofactor, plays a role in stabilizing the open conformation of the lid. An important finding was that the lid opening process was reversible as observed by decreasing the bile salt concentration under the micellar concentration. More recently, the structural changes induced in HPL by lowering the pH were investigated using a combined approach involving SDSL-EPR and Fourier transform infrared (ATR-FTIR) spectroscopy (*20*). A reversible opening of the lid was observed when the pH decreased from 6.5 to 3.0, giving an EPR spectrum similar to the one observed in the presence of bile salts and colipase. Below pH 3.0, ATR-FTIR measurements indicated that HPL had lost most of its secondary structure. In parallel, EPR studies were more precise in revealing a local unfolding in the vicinity of the active site, whereas the lid was found to resist to the global unfolding and to keep a stable structure. SDSL-EPR therefore appeared as a useful technique for studying both the mechanism of action and the stability of HPL.

The work reported in the present report is a further step forward in studying the HPL lid opening process by EPR. Using a double spin labeling strategy and pulsed EPR spectroscopy, the amplitude of the lid opening was estimated from double electron−electron resonance (DEER). Molecular dynamics led to the identification of two conformational subensembles for the spin label grafted on the lid, and the complete analysis of the magnetic interaction between this spin label and another one grafted at a rigid part of HPL confirmed this finding.

MATERIALS AND METHODS

Production of Recombinant HPL and HPL Mutant. All of the procedures used here have been previously described in detail in Belle et al. (*19*). The cDNA encoding HPL was previously obtained from human placenta mRNA using PCR methods (*21*). A 1411 bp *Bam*HI DNA fragment containing the entire HPL coding region was subcloned into the pGAPZB *Pichia pastoris* transfer vector (Invitrogen) downstream of the GAP constitutive promoter for further expression of HPL the yeast *P. pastoris*. Two HPL mutants (D249C and C181Y-D249C) were constructed using the PCR overlap extension technique as previously described, and they were also produced in *P. pastoris* after inserting their DNA into the pGAPZB vector. The wild-type *P. pastoris* strain X-33 was transformed by electroporation using linearized pGAPZB vectors containing either HPL or HPL mutant DNA. Cell cultures were then performed in 1 L Erlenmeyer flasks containing 200 mL of YPD medium without any zeocin, and the cell growth was stopped after 40 h in order to limit the proteolysis of the recombinant HPL (or mutant) secreted into the culture medium.

For the purification of HPL and HPL mutants, 2 L of yeast culture medium was collected, and the pure proteins were obtained after a single cation-exchange chromatography step on S-Sepharose gel (Pharmacia). The purified lipases were characterized by performing SDS−PAGE, N-terminal sequencing, and MALDI-TOF mass spectrometry analysis. Protein concentration in all samples used for EPR experiments was determined by amino acid composition analysis, and the enzyme specific activity was estimated from activity measurement performed with the pHstat technique and using tributyrin as substrate (*19*, *21*).

Spin Labeling Procedure. This procedure was described in previous reports (*19*, *20*). Since it was observed that free cysteines were oxidized in the recombinant lipase recovered from *Pichia* culture medium, the recombinant proteins were initially reduced with DTT prior to the spin labeling reaction with (1-oxy-2,2,5,5-tetramethyl-Δ^3-pyrroline-3-methyl)methanethiosulfonate (MTSL; Toronto Research Chemicals Inc., Toronto, Canada) for 1 h in ice. Upon performing the labeling and EPR measurements with the D249C HPL mutant for double spin labeling experiments, it appeared however that disulfide bridges might be slighty cleaved by DTT, and the resulting free cysteine might be also labeled with MTSL as suggested by a large background absorption revealed by EPR spectroscopy. In this case, the labeling procedure was therefore modified: the treatment by DTT was suppressed, and spin labeling with MTSL was performed for a longer period (4 h) at 4 °C. Four consecutive additions of MTSL at a 10 to 1 molar ratio versus HPL were required. When a single addition of MTSL at a 40 to 1 molar ratio was used, HPL was found to precipitate. All of the samples of HPL mutants treated with MTSL were checked by EPR spectroscopy, and those giving an EPR spectral shape typical of a labeled protein were pooled and concentrated at around 4 mg of HPL/mL (80 μM).

EPR Data Collection by CW EPR Spectroscopy. The spin-labeled HPL samples were injected into a quartz capillary tube with a useful volume of about 20 μL for both room and cryogenic temperature experiments, and the spin-labeled enzyme concentration ranged from 40 to 80 μM. Spectra were recorded at room temperature (296 K) on an ESP 300E Bruker spectrometer equipped with an ELEXSYS Super High Sensitivity resonator operating at 9.9 GHz. EPR spectra of the same samples were recorded at 150 K with an ELEXSYS E500 Bruker spectrometer fitted with an Oxford Instruments ESR 900 helium flow cryostat.

For room temperature experiment, the microwave power was set to 10 mW, and the magnetic field modulation frequency and amplitude were 100 kHz and 0.1 mT, respectively. At 150 K, the microwave power was set to 0.1 mW to avoid saturation of the signal, and the magnetic field modulation frequency and amplitude were 100 kHz and 0.4 mT, respectively.

EPR Data Collection by Pulsed EPR Spectroscopy. DEER experiments were achieved with a Bruker ELEXSYS E580 X band spectrometer using the standard MD5 dielectric resonator and equipped with an Oxford helium temperature regulation unit. All of the spectra were recorded at 70 ± 5 K. These experiments were performed using the four-pulse DEER sequence $(\pi/2)\nu_1-\tau_1-(\pi)\nu_1-\tau-(\pi)\nu_2-(\tau_1+\tau_2)-\tau-(\pi)\nu_1-\tau_2-$echo (*22*). The pump pulse (ν_2) length was set to 12 ns and applied at the maximum of the nitroxide spectrum corresponding to the $m_I = 0$ transition line, and its amplitude was optimized at the maximum of echo inversion. The observer pulses (ν_1) $\pi/2$ and π length were set respectively to 12 and 24 ns and positioned at a 72 MHz higher frequency corresponding to the transition $m_I = +1$. Signal processing was achieved using the DeerAnalysis2008 software package under Matlab (*23*). The signal was corrected by subtracting the unmodulated background echo decay by using a homogeneous three-dimensional spin distribution. The Tikhonov regularization was applied to the corrected dipolar evolution data set to obtain the distance distributions (*24*).

Simulation of the Room Temperature CW EPR Spectra. The EPRSIM-C software program used to simulate the EPR spectra was kindly provided by Dr. J. Strancar (University of Ljubljana, Slovenia). This program is based on the so-called motional-restricted fast-motion approximation and is described in detail in refs *25* and *26*. Five parameters were used to simulate the EPR spectra: an effective rotational correlation time τ, two angles, θ_0 and φ_0, which describe respectively the amplitude and the anisotropy of the spin label rotational motion within a cone,

a residual width (w), and a scalar parameter p_A which allows to adjust the value of the principal values of the hyperfine tensor describing the polarity of the environment of the probe. For a given polarity of the milieu, the shape of the EPR spectrum of a spin label grafted on a protein is governed by the partial averaging of its hyperfine and g tensors. This partial averaging is described by the values of the parameters τ, θ_0, and φ_0, the two last parameters being normalized by $\Omega = (\theta_0\varphi_0)/(\pi/2)^2$, representing the free rotational space which varies from zero (totally restricted movement) to 1 (unrestricted movement).

Simulation of the CW EPR Spectra at 150 K. The method used to simulate the EPR spectrum of a frozen solution of noninteracting nitroxide spin labels was previously described in Morin et al. (*27*). This home-built program was based on the diagonalization of the spin Hamiltonian describing the spin system (Zeeman and hyperfine interaction), the determination of the resonant magnetic field and the associated transition probabilities, and the evaluation of the line width of each transition line. For the present study, we added to the model the calculation of spin−spin interaction resulting from first-order dipolar interaction. The simulation was performed to estimate how the EPR spectrum of a frozen solution of spin-labeled HPL could be broadened by a fraction of strongly interacting spin labels (around 25%) compared to the EPR spectrum of a frozen solution containing 100% of noninteracting labels and not to precisely determine the distance between the closed spin labels. The calculated broadening was only arising from the first-order dipolar interaction, even when the spin labels were closed, which was a rough approximation since, in this later case, spin exchange and second-order dipolar interactions should be taken into account.

Molecular Dynamic Simulations. Molecular dynamics simulations were performed with the CHARMM19 extended atom force field (*28*). Spin label parameters and topology files were taken from a study by Dr. Piotr Fajer reported in ref *29*. Initial atomic coordinates of HPL in its closed conformation were retrieved from the Protein Data Bank (PDB ID 1N8S, HPL−colipase complex). All atoms corresponding to the colipase (chain C) were discarded. Serine residue 30B (numbering introduced by Winkler et al. (*7*)) was renumbered 31 and residues 31− 404 were renumbered accordingly (32−405). Residue Asp249 (Winkler's numbering) was replaced by a cysteine residue covalently linked to the MTSL nitroxide spin label using VMD and the psfgen software package (http://www.ks.uiuc.edu/Research/vmd/plugins/psfgen/). Missing coordinates were added with the hbuild procedure in CHARMM. A Monte Carlo search on the spin label torsion angles was used to define reasonable starting conformations for the spin label as described previously (*30*). The three lowest energy conformers were used as initial conformations. In addition, the three major χ_1,χ_2 rotamers observed in T4 lysozyme were used as initial structures (*31*) (Table 1). The structures were heated at 300 K (10 ps), equilibrated (75 ps), and subjected to free molecular dynamics (5 ns). Atoms beyond 30 Å of atom N1 of the spin label were fixed during the simulation. Molecular dynamics simulations were performed using a cutoff of 15 Å for nonbonded interaction. All bonds between hydrogens and heavy atoms were constrained with the SHAKE algorithm (*32*). An integration step of 1 fs was used. For each conformer, four independent trajectories were performed using different initial velocities.

Analyses were done with VMD (http://www.ks.uiuc.edu/Research/vmd/). Statistical analyses were performed with the R package.

Table 1: Initial χ_1 and χ_2 Dihedral Angles Used in MD Simulations

rotamer	χ_1	χ_2
1	305	120
2	60	230
3	292	53
t,p	180	60
t,m	180	270
m,m	280	310

RESULTS

Production and Biochemical Characterization of the Wild-Type (wt) HPL and Mutants. In addition to wt-HPL, two HPL mutants (D249C and C181Y-D249C) were produced in the yeast *P. pastoris* and further purified as previously reported in ref *19*. The wt-HPL was used for introducing a single spin label at position 181 where a free cysteine residue is naturally present and accessible to solvent. The HPL C181Y-D249C mutant was used for introducing a single spin label at position 249 within the HPL lid, after the substitution of C181 by a tyrosine residue. The HPL D249C mutant contained two accessible free cysteine residues at positions 181 and 249 for double spin labeling experiments. As previously done for wt-HPL and the HPL C181Y-D249C mutant (*19*), a biochemical characterization of the HPL D249C mutant was performed before the EPR experiments. N-Terminal sequencing confirmed that the signal peptide of this HPL mutant was correctly processed in the yeast and that no additional proteolytic cleavage occurred. MALDI-TOF analysis revealed a molecular mass of 51754 ± 36 Da similar to the mass previously measured for the glycosylated wt-HPL polypeptide. The specific activity on tributyrin of the spin-labeled HPL D249C mutant was found to be 6466 ± 1078 units/mg in the presence of bile salts (4 mM NaTDC) and colipase. This enzyme activity was similar to those previously recorded with HPL and the HPL C181Y-D249C mutant (*19*).

HPL Spin Labeling and Yield. The yield of protein spin labeling was estimated from the double integration of the CW EPR spectra of the labeled HPL recorded under nonsaturating conditions and at room temperature and compared with that given by a 3-carboxyproxyl sample of known concentration. Typical values obtained for the labeling yield were 70% spin label per protein molecule for a single spin label grafted at either position 181 (wt-HPL) or 249 (C181Y-D249C HPL mutant) in HPL and 140% when HPL was simultaneously labeled on both sites using the HPL D249C mutant. Assuming that each site was labeled with the same yield (70%), the proportion of bi-, mono-, and nonlabeled enzymes was estimated to be 49% (product of the labeling yield for both sites, 0.7×0.7), 42% (twice the product of the labeling yield for the first site and the nonlabeling proportion for the second site, $2 \times 0.7 \times 0.3$), and 9% (product of the nonlabeling proportions for both sites, 0.3×0.3), respectively. Since the total HPL concentration was 80 μM, the respective concentrations of bi-, mono-, and nonlabeled HPL were estimated to be 39, 34, and 7 μM.

Observation of the Magnetic Interaction between the Two Spin Labels Grafted at Positions 181 and 249 (Lid) of HPL. *(A) Pulsed EPR Spectroscopy*. DEER experiments were performed with double spin-labeled HPL in order to measure the variations in the distance distributions during the lid opening process. Taking into account the low concentration of the bilabeled HPL (39 μM), DEER experiments data sets were recorded for 12 h at 70 K.

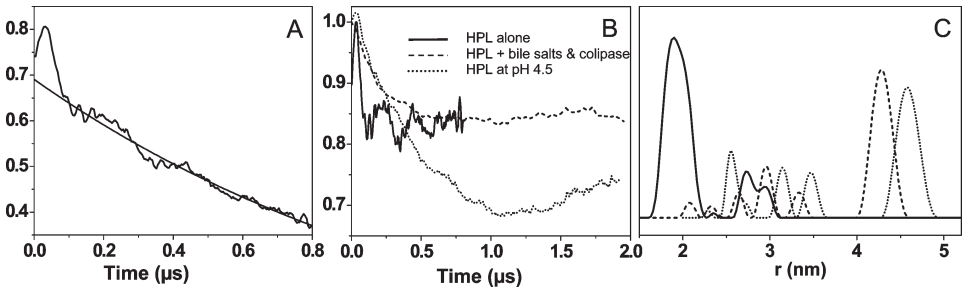


FIGURE 1: DEER experiments. (A) Experimental time domain data for bilabeled HPL alone in solution (black line) and homogeneous background function (gray line) used for correction of the unmodulated part of the spectrum. (B) Corrected experimental time domain data obtained in three conditions: HPL alone (solid line), in the presence of bile salts and colipase (dashed line), and at pH = 4.5 (dotted line). (C) Distance distributions obtained by Tikhonov regularization methods.

Figure 1A displays the experimental time domain data acquired with bilabeled HPL alone in solution and the homogeneous background function used to correct the unmodulated part of the dipolar echo decay (Figure 1B). The interspin distance distribution was extracted from the Tikhonov regularization of the corrected time domain spectrum. When HPL was alone in solution, a single distance distribution of 19 ± 2 Å was found (Figure 1C). The interspin distance distribution was shifted to 42 ± 2 Å in the presence of bile salts and colipase and to 46 ± 2 Å when the pH value was decreased from 6.5 to 4.5, these conditions being known for inducing the lid opening. These results indicated that the lid residue 249 was clearly pushed away from residue 181. The distance distributions obtained from Tikhonov regularization also showed minor peaks between 25 to 35 Å (Figure 1C). These peaks are generally considered as artifacts generated by the mathematical processing of data, but the time domain data contained a fast decaying component which was probably due to distances shorter than 42 Å between some spin labels. These minor peaks might therefore be due to minor intermediate conformations of the HPL lid leading to distances in between 19 Å (closed conformation) and 42 Å (open conformation) between the labels grafted at positions 181 and 249. There is also a possibility that these peaks resulted from a small fraction of unfolded HPL present in the sample as suggested by the simulation of the room temperature CW EPR spectra, but specific activity measurements performed prior to DEER experiments indicated that most HPL was correctly folded and the proportion of unfolded HPL is probably too small to be observed with DEER experiments.

(*B*) *CW EPR Spectroscopy*. Figure 2A′ displays the absorption spectrum at room temperature of the bilabeled HPL (positions 249 and 181) alone in solution compared to the spectrum of an equimolar mixture of C181 and C249 monolabeled HPL molecules. Strong magnetic interactions in the bilabeled HPL are revealed by the broadening of the absorption spectrum compared to the reference one. As the spectra were normalized to their integrated intensities, this broadening was characterized by a decrease of the central peak and by the presence of large outer wings. Since the slope of these enlarged outer wings was weak, the broadening was more visible in the absorption spectra (Figure 2A′) than in the first derivative ones (Figure 2A).

To check whether this broadening of the outer wings was due to a magnetic interaction of the spin labels, the EPR spectrum of the bilabeled HPL was recorded in the presence of colipase at a molar excess of 2 and 4 mM NaTDC to induce the opening of the HPL lid (65% open conformation (*19*)). In this case, the extension of the outer wings disappeared: the magnetic field range where the absorption takes place was nearly identical to the one observed in the absorption spectrum of an equimolar mixture of C181 and C249 monolabeled HPL molecules in the closed conformation (Figure 2A′,B′).

These results show that the opening of the lid dramatically weakens the strong intramolecular magnetic interaction between spin labels observed when the HPL is in its closed conformation. Moreover, the large extension of the outer wings observed for the bilabeled HPL in the closed conformation indicates that the interspin distance is probably lower than 10 Å.

It is worth noting that when the lid opens, the variations of the shapes observed for the low- and high-field peaks are largely due to the influence of the variation of the mobility of the spin label grafted at position 249 when this opening takes place and moderately to the interaction between spin labels (Figure 2B′).

In order to detect only the influence of the magnetic interaction between spin labels, we suppressed the influence of mobility of these labels by recording EPR spectra of frozen solutions (150 K) of the bilabeled HPL in the closed and open conformations (Figure 2C′). A broadening of the absorption spectrum was still observed for HPL in the closed conformation compared to the open one. In the presence of NaTDC and colipase, the absorption spectrum of the bilabeled HPL was found to be nearly identical to those recorded with monolabeled HPL at either position 181 or 249 (data not shown). No significant magnetic interaction between the two spin labels was therefore detectable in the presence of NaTDC and colipase under cryogenic conditions.

Simulation of the Room Temperature CW EPR Spectra of the Spin Label Grafted on the HPL Lid. The simulation of the CW EPR spectrum of the spin label grafted at position 249 on the HPL lid was first performed for the enzyme alone in solution, i.e., with the lid in the closed conformation (*19*). The best fit to the experimental EPR spectrum was obtained with three components (Figure 3A and Table 2). Two narrow-shaped components corresponding to spin labels with moderate mobility were found to largely contribute to the overall spectrum in nearly equal proportions (40% and 50%, respectively). It is worth noting that the term "mobility" takes into account both the rate of the rotational motion (described by τ) and the geometrical restrictions defining the anisotropy of the rotational motion (described by Ω). A third component corresponding to a spin label having a minor contribution to the overall spectrum (10%) was also required to obtain the best fit. This might result from the fact that some spin labels might be grafted on a small fraction of unfolded HPL that could not be detected by the classical biochemical analysis (enzyme activity and protein concentration

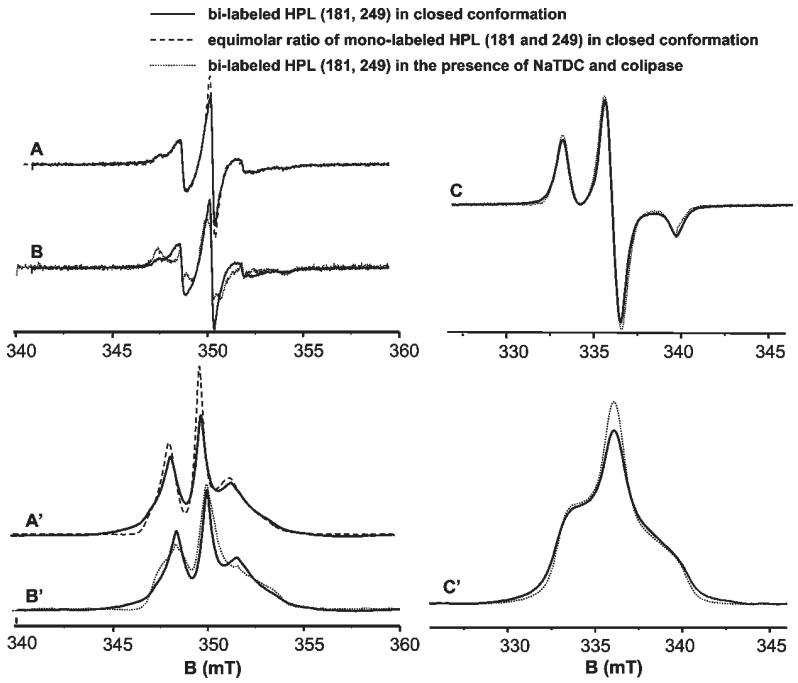


FIGURE 2: CW EPR spectra of mono- and bilabeled HPL. Panels A, B, and C are showing original derivative spectra measured with field modulation and panels A′, B′, and C′ the corresponding absorption spectra obtained from the integration of the conventional derivative spectra. (A, A′) Comparison of the spectra of the bilabeled HPL (positions 181 and 249; solid line) and the sum of the monolabeled HPL in equivalent proportion in the closed conformation (dotted line), recorded at room temperature. (B, B′) Comparison of the spectra of the bilabeled HPL alone in solution (closed conformation; solid line) and in the presence of 4 mM bile salts and colipase at a molar excess of 2 (65% open conformation; dotted line), recorded at room temperature. (C, C′) Comparison of the spectra recorded at 150 K for the bilabeled HPL alone in solution (solid line) and in the presence of bile salts and colipase (dotted line). All spectra have been normalized to their integrated intensities.

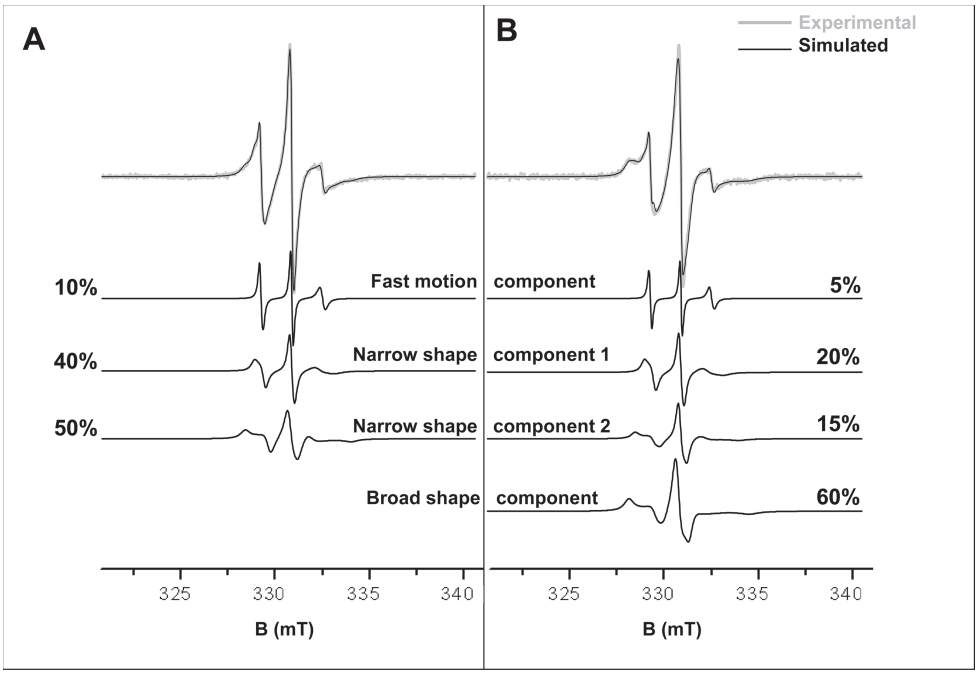


FIGURE 3: Simulation of EPR spectra and decomposition in individual components for the HPL alone in solution (A) and in the presence of 4 mM NaTDC and a molar excess of colipase (B). Simulations of the spectra were performed using the EPRSIM-C software program (*25*). See Table 2 for the parameters describing each spectral component.

assays) used for monitoring the production of HPL samples. The presence of residual free radical was excluded since the mobility would have been much higher.

The simulation of the CW EPR spectrum of the spin label grafted at position 249 on the HPL lid was then performed for the enzyme in the presence of micellar concentration of bile salts (4 mM NaTDC) and colipase in molar excess. Under these conditions, it was previously shown that the respective proportions of HPL molecules with the lid in the closed and open conformations were 35% and 65%, respectively (*19*). Upon

Table 2: Parameters Extracted from the Simulation of the CW EPR Spectra

		simulation parameters			
		τ_c (ns)	Ω	θ_0	φ_0
fast motion component		0.20	0.92	1.57	1.45
closed HPL alone in solution	narrow shape component 1	2.70	0.80	1.35	1.45
	narrow shape component 2	2.80	0.47	0.90	1.30
open HPL in the presence of 4 mM NaTDC and a molar excess of 2 in colipase	narrow shape component 1	2.60	0.74	1.30	1.40
	narrow shape component 2	2.80	0.44	0.90	1.20
	broad shape component	2.85	0.22	0.60	0.90

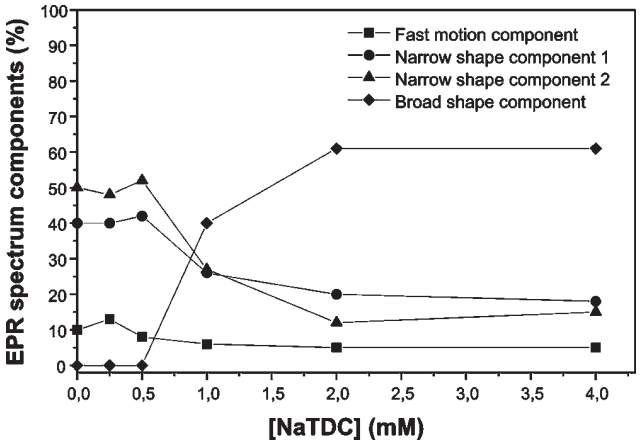


FIGURE 4: Variations with bile salt concentration in the proportion of the individual spectral components obtained by simulation of the experimental spectra.

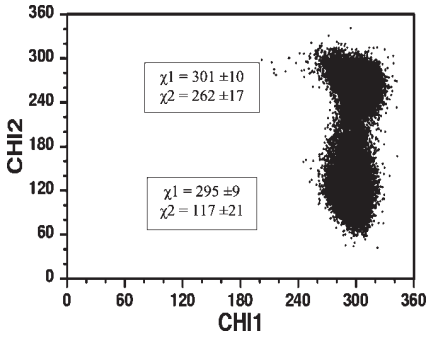


FIGURE 5: Populations of χ_1 and χ_2 dihedral angles obtained by molecular dynamics for the side chain of cysteine modified by MTSL at position 249 of the closed HPL.

simulation of the overall EPR spectrum, four components were required to obtain the best fit (Figure 3B and Table 2). As for the previous case, a minor fast motion component (5%) needed to be introduced. A major broad-shaped component was found to contribute to 60% of the whole spectrum. This component corresponds to the fraction of the HPL in the open conformation, according to our previous study (*19*). Two narrow-shaped components were found to represent 35% of the whole spectrum, and their shapes were nearly identical to those obtained for the EPR spectrum corresponding to HPL with the lid in its closed conformation (Figure 3 and Table 2). These two proportions (60% broad and 35% narrow shapes, respectively) were therefore highly similar to those estimated directly from the experimental spectrum. Hence, both the spectral shapes and the proportion of the two narrow-shaped components indicated that they result from spin labels in the enzyme fraction that has remained with the lid in the closed conformation. These two narrow components might be attributed to the presence of two conformational subensembles of the spin label side chain at position 249 when the lid is closed. On the other hand, the appearance of only one broad-shaped component upon lid opening suggests that only one conformational subensemble is present in the open HPL.

We simulated the EPR spectra obtained when the conformation of the HPL lid was progressively shifted from the closed to the open one using NaTDC in the presence of colipase (see experimental results in ref *19*). The largest proportion of the open conformation that could be obtained was 65%. The variations in each spectral component plotted as a function of NaTDC concentration are shown in Figure 4. The proportions of the two narrow-shaped components were found to decrease simultaneously when the NaTDC concentration and therefore the proportion of HPL in the open conformation were increased. Conversely, the broad-shaped component appeared and increased with NaTDC concentration above 0.5 mM, with a maximum level reached when the proportions of the two narrow-shaped components were the lowest. These results therefore suggest a transition of the spin labels from two distinct environments with moderate mobility to a single environment with low mobility.

Molecular Dynamics of the Cysteine Lateral Chain Labeled with MTSL at Position 249 of the Closed HPL. The root-mean-square deviations (rmsd) for Cα atoms during the time course of the different simulations were between 0.5 and 1 Å, indicating that the simulated system was stable. The X-ray crystallography structure of HPL has shown that the lid thermal factor is high (*7*, *10*), and as expected the root-mean-square fluctuation (rmsf) of the backbone revealed that the spin label is located in a flexible region. MD simulations showed two major conformational subensembles of the labeled C249 side chain characterized by their χ_1 and χ_2 dihedral angles (Figure 5). The average observed value for χ_1 was around 300° for both conformations whereas values of 120° and 260° were found for χ_2. Multimodal distributions were also observed for the other χ_3, χ_4, and χ_5 angles of the spin label in the MD simulations (data not shown). Two representative conformations were extracted from these results to estimate distances between spin labels from 3D models shown in Figure 6.

DISCUSSION

Evidence for Distinct Conformational Subensembles of the Spin Label Grafted to the Closed Lid of HPL. Experiments performed with concentrated samples (>100 μM) of HPL spin labeled at position 249 revealed that the EPR spectrum

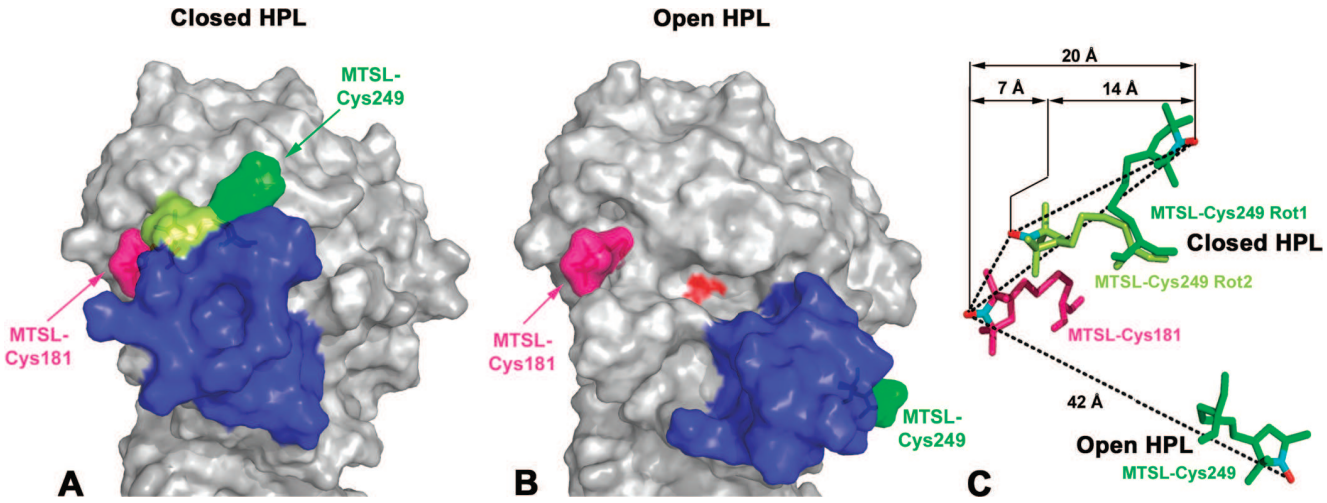


FIGURE 6: Structural modeling of spin labels grafted to cysteine residues in HPL. (A) Surface representation of HPL with the lid (blue surface) in the closed conformation, (B) surface representation of HPL with the lid (blue surface) in the open conformation and the active site accessible to solvent (the surface of the active site serine residue is colored in red, and (C) distances between the oxygen atoms of spin labels measured in silico. Atomic coordinates of HPL in its closed and open conformations were retrieved from the Protein Data Bank (closed HPL, PDB ID 1N8S; open HPL, 1LPB). The spin label at position 181 (MTSL-CYS181) is shown in magenta color in both closed and open HPL. The spin label at position 249 is shown in three different situations: two rotamers representative of the conformational subensembles observed by MD when the lid is closed (MTSL-CYS249-Rot1 colored in green and MTSL-CYS249-Rot2 colored in yellow, panels A and C) and the single conformation modeled in the open structure of HPL (MTSL-CYS249 colored in green in panels B and C). This figure was generated using the PyMOL software program (http://www.pymol.org/).

corresponding to the closed conformation of the HPL lid was in fact composite. The simulation of this EPR spectrum required the introduction of two narrow-shaped components of nearly equal weights. These components correspond to distinct environments and moderate mobilities of the nitroxide spin label (Figure 3A and Table 2) that could result from either two different closed conformations of the lid or the existence of two distinct side chain conformational subensembles in a single closed conformation of the lid.

Concerning the first hypothesis, only one closed conformation of the HPL lid was observed so far by X-ray crystallography (*7*, *33*). An intermediate conformation between the closed (inactive) and open (fully activated) conformations of *Thermomyces* (*Humicola*) *lanuginosa* lipase was however observed (*34*). This intermediate conformation, in which the active site remained not accessible to the solvent, was considered as a step occurring in the first phases of the lipase activation (opening of the lid). Discrete structural changes were observed upon the transition from the closed to this intermediate conformation, such as the isomerization of a disulfide bond synchronized with the flipping of an arginine located in the lid's proximal hinge. Changes of this kind might also occur in HPL and affect the conformation of amino acid side chains in the lid without a full opening of the active site. A spin label grafted to a residue in the lid might then be split into distinct populations showing slightly different but still moderate mobilities, whereas the drastic conformational change that occurs when the lid opens leads to the quasi immobilization of the spin label grafted at position 249 (Figure 3B). As suggested for the lipase of *T. lanuginosa*, an intermediate but still closed conformation of the HPL lid might then evolve toward the open one in a more favorable way than the closed conformation.

The second hypothesis of two distinct side chain conformational subensembles existing in a single closed conformation of HPL was supported by molecular dynamics experiments, showing the existence of two major conformational subensembles for the C249 side chain chemically modified by MTSL (Figures 5 and 6 and Table 1). Using the notation of Lowell (*35*), these two conformational subensembles adopt a m120 state ($\chi_1 = -60°$; $\chi_2 = 120°$) and a m-100 state ($\chi_1 = -60°$; $\chi_2 = -100°$). Since MD were performed at 300 K, one could argue that the distribution within the subensemble may not reflect exactly the population ensemble due to insufficient conformation sampling (*36*, *37*). It has been shown, however, that the use of stochastic dynamics simulations as was performed in this study leads to better conformation sampling (*38*). In addition, the use of CMAP correction with the CHARMM force field improves the treatment of Φ and ψ dihedral angles (*39*). Two conformational subensembles of spin-labeled cysteine lateral chains have been previously observed by X-ray crystallography for T4 lysozyme (*31*, *40*−*42*) and predicted by molecular dynamics for RalGDS-like protein 2 (*43*). The corresponding EPR spectra were found to contain two components which were attributed to these conformational subensembles in both spin-labeled T4 lysozyme (*40*) and RalGDS-like protein 2 (*43*): the conversion time between the two conformational subensembles, governed by the high energy activation of the disulfide bond $S\gamma-S_\delta$ present in the side chain, was sufficiently slow to observe the two components with X band EPR spectroscopy. In T4 lysozyme, these conformational subensembles were always observed in exposed α-helices. In the present study, the spin label conformational subensembles were also identified for an exposed position (residue 249) located in an α-helical region when the HPL lid is in the closed conformation: the short α-helix sitting on the top of the active site entrance (*7*). When the lid opens, residue 249 is still exposed to solvent but becomes located in a turn between two novel α-helices, and we observed only one component for the spin label in this case. The simultaneous disappearance of the two narrow-shaped components when the lid opens (Figure 4) is a strong support for this second hypothesis versus the existence of two distinct conformations of the closed lid.

Distance Estimation between Spin Labels and Amplitude of the Lid Opening in HPL. Distance distributions between the labels grafted on the lid and at position 181 were estimated from

the analysis of their dipolar interaction by the way of the double electron−electron resonance (DEER) method implemented on a pulsed EPR spectrometer. An estimation of these distances was also performed using the CW EPR spectra obtained at room temperature and 150 K.

Since we have two populations of the spin labels at position 249 in the closed conformation of the enzyme and one population for the same spin label in the open conformation (Figure 6), it was expected that we should measure two distances between nitroxides at position 181 (one population) and position 249 (two populations) when HPL is in the closed conformation, and three distances for the sample containing a mixture of closed (35%) and open (65%) conformations of the enzyme in the presence of bile salts and colipase.

With the DEER measurements and HPL in the closed conformation, a single mean distance of 19 ± 2 Å between the spin labels was however measured (Figure 1). The two distances expected between the two populations of the spin label at position 249 and the spin label at position 181 might therefore be either equal or more probably one of the two distances might be lower than 15 Å, the smallest value that can be measured by this technique. Indeed, the broadening observed on the CW EPR spectrum indicated that a fraction of double spin-labeled HPL in the closed conformation contained spin labels involved in a strong magnetic interaction and therefore separated by a short distance (Figure 2). These results are well supported by the structural modeling of spin-labeled HPL using known 3D structures and the results of molecular dynamics, with two values of 20 and 7 Å estimated for the distances between the nitroxide at position 181 and the two representative conformational subensembles at position 249 (Figure 6A,C). The modeling of the side chain of the spin-labeled C181 residue did not require MD simulations since its conformation was found to be drastically constrained by the local environment, and it was adjusted manually.

When the lid opening was induced by bile salts and colipase, the CW EPR spectrum at room temperature did not show this strong interaction anymore (Figure 2B′). Since the concentration of bilabeled HPL was 39 μM and 65% of HPL was in the open conformation, the concentration of the bilabeled and open HPL contributing to the CW EPR spectrum was estimated to be 25 μM. The absence of detectable interaction between spin labels was therefore not due to low concentrations of bilabeled HPL but was merely due to a distance greater than 20 Å between the spin labels, the limit value for detecting magnetic interaction with CW EPR. This is consistent with X-ray crystallography data showing that the maximum value of Cα displacement was 28−30 Å for the lid residues upon opening (Figure 6A,B). On the other hand, since 35% of HPL remained with the lid in the closed conformation, one could expect that some strong interaction between spin labels might still be seen from the CW EPR spectrum. The concentration of bilabeled HPL remaining in the closed conformation was however estimated to be only 14 μM, and only one spin label conformational subensemble at position 249 could give a strong interaction with the spin label at position 181. Since the two spin label conformational subensembles at position 249 were estimated to be in equal proportions (Figures 4B and 5), the concentration of bilabeled and closed HPL that could give a strong interaction between spin labels (distance lower than 10 Å) was found to be close to 7 μM (i.e., 9.4% of spin-labeled HPL). The overall EPR spectrum was therefore resulting from more than 90% of noninteracting or weakly interacting spin labels, and it is not surprising that no strong magnetic interaction was detected from the CW EPR spectrum. Concerning the spectra recorded at 150 K (Figure 2C′), we checked that the spectrum recorded for HPL in the presence of bile salts and colipase was not broadened by dipolar interaction (data not shown). For HPL alone in a frozen solution at 150 K (i.e., in the closed conformation of HPL), the observation of a broadening of the EPR spectrum was thus attributed to a magnetic interaction arising from a fraction of bilabeled molecules with two closed radicals. This broadening was however moderate (Figure 2C′). Since the EPR spectrum resulted from the contribution of 34 μM monolabeled HPL molecules (i.e., 47%), 19.5 μM bilabeled HPL molecules (i.e., 26.5%) with spin labels separated by a mean distance of 19 Å, and 19.5 μM bilabeled HPL molecules (i.e., 26.5%) with spin labels separated by a distance of about 8−10 Å, it was checked whether 26.5% of the spin-labeled HPL molecules with closed and strongly interacting spin labels could have such a moderate effect on the overall EPR spectrum of spin-labeled HPL in a frozen solution, as compared to the EPR spectrum of a frozen solution containing 100% of noninteracting labels. For this purpose, the EPR spectrum of the frozen solution was simulated (see Materials and Methods section) using the above proportions of monolabeled and bilabeled HPL molecules and, in this latter case, interspin label distances equal to 19 and 8−10 Å. This simulation was found to be consistent with the experimental spectrum (result not shown).

The HPL lid opening by bile salts was confirmed by the DEER experiments performed with pulsed EPR. A single mean value of 42 ± 2 Å was estimated for the distance between the spin labels grafted at position 181 and on the lid at position 249 (Figure 1). The distance between these nitroxides was therefore increased by a factor of 2 in the presence of bile salts and colipase. Again, these results are well supported by the structural modeling of spin-labeled HPL using the known 3D structure of HPL in the open conformation, with a value of 42 Å estimated for the distance between the nitroxides at positions 181 and 249 (Figure 6B,C). The various conformations tested for the side chain of the spin label at position 249 all gave distances in the 40 Å range and lower distances appeared to be impossible since this spin label was on the opposite side of the open lid versus the spin label at position 181 (Figure 6B). Here again, it was surprising to observe only one distance corresponding to the open HPL whereas the sample also contained 35% HPL in the closed conformation, and two additional distances were expected in association with the two spin label conformational subensembles at position 249 of the closed HPL lid. We have seen previously that one of these distances (lower than 10 Å) is too small to be estimated by DEER experiments. The distance between the second spin label conformational subensemble at position 249 and that at position 181 was estimated to be 19 Å, and the interaction between these spin labels could therefore be detected by DEER (distances >15 Å). But the residual concentration of bilabeled and closed HPL corresponding to this situation was however estimated to be close to 7 μM, and it is not surprising that this contribution was not detected by DEER for sensitivity reasons.

In conclusion, EPR spectroscopy allowed the measurement of two distances corresponding to the closed and the open conformations of the HPL lid, respectively. The amplitude of the side chain displacement for residue 249 was estimated to be 24 Å, a value similar to that deduced from X-ray 3D structures of HPL (*10*). The DEER method is therefore a valuable tool for measuring the amplitude of the HPL lid opening in a frozen solution. The use of both CW and pulsed EPR spectroscopy, as well as the identification of spin label conformational subensembles, was however requested for interpreting correctly the results.

These experiments supported the existence of two spin label conformational subensembles at position 249 when HPL is in its closed conformation, these spin labels being separated from the spin label at position 181 by distances lower than 10 Å and equal to 19 Å, respectively. These values are compatible with the results of molecular dynamics and X-ray crystallography (Figures 5 and 6).

More importantly, the present findings confirmed the previous attribution of distinct EPR spectra to the closed (narrow-shaped) and open (broad-shaped) conformations of the HPL lid. This attribution was indirectly based on experiments performed with the E600 inhibitor bound to the active site and blocking HPL in the open conformation (*19*). The broad-shaped EPR spectrum of monolabeled HPL at position 249 was always recorded under the same conditions that allowed to measure the maximum amplitude of the lid opening with bilabeled HPL and DEER experiments: either by using bile salts and colipase (42 Å) or by lowering the pH to 4.5 (46 Å), this later condition also promoting the lid opening in HPL (*20*).

The previous work by Belle et al. allowed to quantify the equilibrium between the closed and the open conformation of HPL in solution and to establish that it was a reversible process (*19*). The present work allowed to quantify the lid opening process in solution and in frozen solution by measuring the displacement of one residue located within the lid. Although the distances estimated from EPR experiments are similar to those already deduced from X-ray crystallography, these previous results were obtained in a crystal structure. When various conformations of the lipase lid were observed for the first time, the question was raised whether these conformations could exist in solution or resulted from crystal packing constraints. The data presented here support the existence of HPL conformations in solution identical to those observed in the enzyme crystals. This is an important step forward in the understanding of HPL structure−function relationships under conditions closer to those found in the physiological environment of the enzyme. It is worth noticing that the enzyme concentration (micromolar) in the samples used for EPR experiments is in the same range as the mean HPL concentration measured in samples collected from the small intestine during a meal (250 μg/mL or 5 μM (*44*)).

The next step will consist of experiments performed with lipids to investigate the lid opening process when the HPL binds at a lipid−water interface. The existence of two spin label conformational subensembles with moderate mobility is the fingerprint of the lid in its closed conformation, and this property might be particularly useful for such a study.

ACKNOWLEDGMENT

We thank Dr. Robert Verger for fruitful discussions and continuous interest in this work. We are grateful to Dr. Janez Strancar (Laboratory of Biophysics, EPR Center, "Jozef Stefan" Institute, Ljubljana, Slovenia) for providing the EPRSIM-C software program for EPR spectrum simulation and to Dr. Jessica Blanc for revising the English manuscript.

REFERENCES

1. Carrière, F., Barrowman, J. A., Verger, R., and Laugier, R. (1993) Secretion and contribution to lipolysis of gastric and pancreatic lipases during a test meal in humans. *Gastroenterology 105*, 876–888.
2. Borgström, B., and Brockman, H. L. (1984) Lipases, pp 1−527, Elsevier, Amsterdam.
3. Schmid, R. D., and Verger, R. (1998) Lipases: interfacial enzymes with attractive applications. *Angew. Chem., Int. Ed. 37*, 1608–1633.
4. Wooley, P., and Petersen, S. B. (1994) Lipases: their structure, biochemistry and applications, pp 1−363, Cambridge University Press, Cambridge.
5. Aloulou, A., Rodriguez, J. A., Fernandez, S., Van Oosterhout, D., Puccinelli, D., and Carriere, F. (2006) Exploring the specific features of interfacial enzymology based on lipase studies. *Biochim. Biophys. Acta 1761*, 995–1013.
6. Brady, L., Brzozowski, A. M., Derewenda, Z. S., Dodson, E., Dodson, G., Tolley, S., Turkenburg, J. P., Christiansen, L., Huge-Jensen, B., Norskov, L., Thim, L., and Menge, U. (1990) A serine protease triad forms the catalytic centre of a triacylglycerol lipase. *Nature 343*, 767–770.
7. Winkler, F. K., d'Arcy, A., and Hunziker, W. (1990) Structure of human pancreatic lipase. *Nature 343*, 771–774.
8. Roussel, A., Canaan, S., Egloff, M. P., Riviere, M., Dupuis, L., Verger, R., and Cambillau, C. (1999) Crystal structure of human gastric lipase and model of lysosomal acid lipase, two lipolytic enzymes of medical interest. *J. Biol. Chem. 274*, 16995–17002.
9. Brzozowski, A. M., Derewenda, U., Derewenda, Z. S., Dodson, G. G., Lawson, D. M., Turkenburg, J. P., Bjorkling, F., Huge-Jensen, B., Patkar, S. A., and Thim, L. (1991) A model for interfacial activation in lipases from the structure of a fungal lipase-inhibitor complex. *Nature 351*, 491–494.
10. van Tilbeurgh, H., Egloff, M.-P., Martinez, C., Rugani, N., Verger, R., and Cambillau, C. (1993) Interfacial activation of the lipase-procolipase complex by mixed micelles revealed by X-ray crystallography. *Nature 362*, 814–820.
11. Roussel, A., Miled, N., Berti-Dupuis, L., Riviere, M., Spinelli, S., Berna, P., Gruber, V., Verger, R., and Cambillau, C. (2002) Crystal structure of the open form of dog gastric lipase in complex with a phosphonate inhibitor. *J. Biol. Chem. 277*, 2266–2274.
12. Egloff, M.-P., Marguet, F., Buono, G., Verger, R., Cambillau, C., and van Tilbeurgh, H. (1995) The 2.46 Å resolution structure of the pancreatic lipase-colipase complex inhibited by a C11 alkyl phosphonate. *Biochemistry 34*, 2751–2762.
13. Miled, N., Roussel, A., Bussetta, C., Berti-Dupuis, L., Riviere, M., Buono, G., Verger, R., Cambillau, C., and Canaan, S. (2003) Inhibition of dog and human gastric lipases by enantiomeric phosphonate inhibitors: a structure-activity study. *Biochemistry 42*, 11587–11593.
14. Hermoso, J., Pignol, D., Kerfelec, B., Crenon, I., Chapus, C., and Fontecilla-Camps, J. C. (1996) Lipase activation by nonionic detergents. The crystal structure of the porcine lipase-colipase-tetraethylene glycol monooctyl ether complex. *J. Biol. Chem. 271*, 18007–18016.
15. Hermoso, J., Pignol, D., Penel, S., Roth, M., Chapus, C., and Fontecilla-Camps, J. C. (1997) Neutron crystallographic evidence of lipase-colipase complex activation by a micelle. *EMBO J. 16*, 5531–5536.
16. Prompers, J. J., Groenewegen, A., Hilbers, C. W., and Pepermans, H. A. (1999) Backbone dynamics of *Fusarium solani pisi* cutinase probed by nuclear magnetic resonance: the lack of interfacial activation revisited. *Biochemistry 38*, 5315–5327.
17. Prompers, J. J., van Noorloos, B., Mannesse, M. L., Groenewegen, A., Egmond, M. R., Verheij, H. M., Hilbers, C. W., and Pepermans, H. A. (1999) NMR studies of *Fusarium solani pisi* cutinase in complex with phosphonate inhibitors. *Biochemistry 38*, 5982–5994.
18. Sibille, N., Favier, A., Azuaga, A. I., Ganshaw, G., Bott, R., Bonvin, A. M., Boelens, R., and van Nuland, N. A. (2006) Comparative NMR study on the impact of point mutations on protein stability of *Pseudomonas mendocina* lipase. *Protein Sci. 15*, 1915–1927.
19. Belle, V., Fournel, A., Woudstra, M., Ranaldi, S., Prieri, F., Thome, V., Currault, J., Verger, R., Guigliarelli, B., and Carriere, F. (2007) Probing the opening of the pancreatic lipase lid using site-directed spin labeling and EPR spectroscopy. *Biochemistry 46*, 2205–2214.
20. Ranaldi, S., Belle, V., Woudstra, M., Rodriguez, J., Guigliarelli, B., Sturgis, J., Carriere, F., and Fournel, A. (2009) Lid opening and unfolding in human pancreatic lipase at low pH revealed by site-directed spin labeling EPR and FTIR spectroscopy. *Biochemistry 48*, 630–638.
21. Thirstrup, K., Carriere, F., Hjorth, S., Rasmussen, P. B., Woldike, H., Nielsen, P. F., and Thim, L. (1993) One-step purification and characterization of human pancreatic lipase expressed in insect cells. *FEBS Lett. 327*, 79–84.
22. Jeschke, G. (2002) Distance measurements in the nanometer range by pulsed EPR. *ChemPhysChem 3*, 927–932.
23. Jeschke, G., Chechik, V., Godt, A., Zimmermann, H., Banham, J., Timmel, C. R., Hilger, D., and Jung, H. (2006) DEER analysis 2006—a computational software package for analyzing pulsed ELDOR data. *Appl. Magn. Reson. 30*, 473–498.

24. Tikhonov, A. N., and Arsenin, V. Y. (1977) Solutions of Ill-posed Problems, pp 1−272, John Wiley & Sons, New York.
25. Strancar, J., Koklic, T., Arsov, Z., Filipic, B., Stopar, D., and Hemminga, M. A. (2005) Spin label EPR-based characterization of biosystem complexity. *J. Chem. Inf. Model. 45*, 394–406.
26. Strancar, J. (2007) Advanced ESR spectrocopy in membrane biophysics, in ESR spectroscopy in membrane biophysics (Berliner, L. J., Ed.) pp 49−89, Springer, New York.
27. Morin, B., Bourhis, J. M., Belle, V., Woudstra, M., Carriere, F., Guigliarelli, B., Fournel, A., and Longhi, S. (2006) Assessing induced folding of an intrinsically disordered protein by site-directed spin-labeling electron paramagnetic resonance spectroscopy. *J. Phys. Chem. B 110*, 20596–20608.
28. Brooks, B. R., Brooks, C. L., III, Mackerell, A. D., Jr., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., and Karplus, M. (2009) CHARMM: the biomolecular simulation program. *J. Comput. Chem. 30*, 1545–1614.
29. Sale, K., Song, L., Liu, Y. S., Perozo, E., and Fajer, P. (2005) Explicit treatment of spin labels in modeling of distance constraints from dipolar EPR and DEER. *J. Am. Chem. Soc. 127*, 9334–9335.
30. Sale, K., Sar, C., Sharp, K. A., Hideg, K., and Fajer, P. G. (2002) Structural determination of spin label immobilization and orientation: a Monte Carlo minimization approach. *J. Magn. Reson. 156*, 104–112.
31. Guo, Z., Cascio, D., Hideg, K., and Hubbell, W. L. (2008) Structural determinants of nitroxide motion in spin-labeled proteins: solvent-exposed sites in helix B of T4 lysozyme. *Protein Sci. 17*, 228–239.
32. Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys. 23*, 327–341.
33. van Tilbeurgh, H., Sarda, L., Verger, R., and Cambillau, C. (1992) Structure of the pancreatic lipase-procolipase complex. *Nature 359*, 159–162.
34. Brzozowski, A. M., Savage, H., Verma, C. S., Turkenburg, J. P., Lawson, D. M., Svendsen, A., and Patkar, S. (2000) Structural origins of the interfacial activation in *Thermomyces* (*Humicola*) *lanuginosa* lipase. *Biochemistry 39*, 15071–15082.
35. Lovell, S. C., Word, J. M., Richardson, J. S., and Richardson, D. C. (2000) The penultimate rotamer library. *Proteins 40*, 389–408.
36. Beier, C., and Steinhoff, H. J. (2006) A structure-based simulation approach for electron paramagnetic resonance spectra using molecular and stochastic dynamics simulations. *Biophys. J. 91*, 2647–2664.
37. Sezer, D., Freed, J. H., and Roux, B. (2008) Using Markov models to simulate electron spin resonance spectra from molecular dynamics trajectories. *J. Phys. Chem. B 112*, 11014–11027.
38. Caves, L. S., Evanseck, J. D., and Karplus, M. (1998) Using Markov models to simulate electron spin resonance spectra from molecular dynamics trajectories. *Protein Sci. 7*, 649–666.
39. Buck, M., Bouguet-Bonnet, S., Pastor, R. W., and MacKerell, A. D., Jr. (2006) Importance of the CMAP correction to the CHARMM22 protein force field: dynamics of hen lysozyme. *Biophys. J. 90*, L36–38.
40. Langen, R., Oh, K. J., Cascio, D., and Hubbell, W. L. (2000) Crystal structures of spin labeled T4 lysozyme mutants: implications for the interpretation of EPR spectra in terms of structure. *Biochemistry 39*, 8396–8405.
41. Guo, Z., Cascio, D., Hideg, K., Kalai, T., and Hubbell, W. L. (2007) Structural determinants of nitroxide motion in spin-labeled proteins: tertiary contact and solvent-inaccessible sites in helix G of T4 lysozyme. *Protein Sci. 16*, 1069–1086.
42. Fleissner, M. R., Cascio, D., and Hubbell, W. L. (2009) Structural origin of weakly ordered nitroxide motion in spin-labeled proteins. *Protein Sci. 18*, 893–908.
43. Pistolesi, S., Ferro, E., Santucci, A., Basosi, R., Trabalzini, L., and Pogni, R. (2006) Molecular motion of spin labeled side chains in the C-terminal domain of RGL2 protein: a SDSL-EPR and MD study. *Biophys. Chem. 123*, 49–57.
44. Carriere, F., Renou, C., Lopez, V., De Caro, J., Ferrato, F., Lengsfeld, H., De Caro, A., Laugier, R., and Verger, R. (2000) The specific activities of human digestive lipases measured from the in vivo and in vitro lipolysis of test meals. *Gastroenterology 119*, 949–960.

RESUME

Le nombre considérable d'interactions protéine-protéine (PPIs) existant au sein d'un organisme, ainsi que leur implication cruciale dans la vie cellulaire et dans de nombreuses pathologies, font des PPIs un immense réservoir de cibles potentielles pour la recherche de médicaments. Longtemps délaissées par les compagnies pharmaceutiques ainsi que par les laboratoires académiques en raison de leur difficulté d'approche apparente, les PPIs sont aujourd'hui sur le devant de la scène grâce au développement de méthodologies innovantes et la validation récente de molécules chimiques modulant ces interactions dans des essais précliniques.

L'étude de ces molécules, les modulateurs d'interactions protéine-protéine (PPIM), a des implications tant dans la recherche fondamentale que thérapeutique. Les PPIMs peuvent aider à différencier les multiples fonctions portées par une même protéine, à replacer la protéine dans une cascade de réactions, ainsi qu'à disséquer et reconstituer des réseaux de signalisations protéiques. Elles permettront également de faire émerger de nouvelles familles d'agents thérapeutiques actifs dans diverses pathologies.

Mon travail de thèse a principalement porté sur deux aspects de l'étude de l'inhibition des PPIs. D'une part, l'étude de l'implication des divers paramètres physicochimiques gouvernant une PPI dans sa capacité à être modulée (étude dite de la « druggabilité »), m'a amené à participer à la création d'une base de données structurale des interactions protéine-protéine : $2P2I_{DB}$ (http://2p2idb.cnrs-mrs.fr/).

D'autre part, l'étude de l'espace chimique caractéristique des PPIMs, m'a permis de prendre part à l'élaboration d'un protocole chémoinformatique innovant pour la création de chimiothèques dédiées aux PPIs. Appliqué sur un ensemble de 8,3 millions de petites molécules proposées par les plus grands fournisseurs, ce protocole nous a permis de créer plusieurs librairies de molécules virtuelles dédiées aux PPIs qui seront validées expérimentalement sur des cibles biologiques du laboratoire.