UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

Laboratoire d'Innovation Thérapeutique, UMR 7200

en cotutelle avec

Eskitis Institute for Drug Discovery, Griffith University

THÈSE

présentée par

Noé STURM

soutenue le : **8 Décembre 2015**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/Spécialité : Chimie/Chémoinformatique

Caractérisation de l'empreinte biologique des produits naturels pour des applications de conception rationnelle de médicament assistée par ordinateur

THÈSE dirigée par :

KELLENBERGER Esther	Professeur, Université de Strasbourg
QUINN Ronald	Professeur, Université de Griffith, Brisbane, Australie

RAPPORTEURS :

IORGA Bogdan	Chargé de recherche HDR, Institut de Chimie des Substances Naturelles, Gif-sur-Yvette
GÜNTHER Stefan	Professeur, Université Albert Ludwigs, Fribourg, Allemagne

Acknowledgments

First of all, I would like to thank my two supervisors Professor Kellenberger Esther and Professor Quinn Ronald for their excellent assistance, both academic and personal in nature.

Esther, I remember that day, as I was still an undergraduate student. Next to a coffee machine you asked the open question: “who’s keen for Australia?!” As simple as it sounds, this question changed my life. Since then, you have shaped me, just like biosynthetic enzymes shape natural products and I sense that I will retain this imprint! Now, my candidature has come to an end, it’s time to prove what I learned. I will remember this important period of my life with much nostalgia but I am also looking forward to the future. Really, I would like to express all my gratitude to you for your precious help, your opportunism and also for your patience (I know you would have needed this with me!). I owe you more than my PhD project.

Also, I would like to extend my gratitude to Professor Quinn without whom none of this could have happen. I really enjoyed being part of the Eskitis family. I learned a lot! I particularly enjoyed each of your intervention during group meetings. I’d like to thank you because you gave me the opportunity to study natural products and their biosynthetic origins. It is fascinating. I also would like to express my admiration to you for what you do. I mean, who wouldn’t want to have such an incredible driving power.

I also would like to thank Dr. Rognan Didier who has allowed me to work in his laboratory in the first place. Your gifted nature for science impresses me even today. It was an honor to work with your lab.

I would also like to thank Dr. Campitelli who was an excellent colleague at Eskitis University. Also, as part of Eskitis engineering members, I would like to thanks Stephen Toms for his sympathy and devotion to the group.

Then, I would like to thank my PhD colleagues. I'll start with a special thanks to Jeremy Desaphy and Jamel Meslamani. You guys are awesome! Both of you have guided my first steps. I also would like to thank other colleagues in Illkrich, Guillaume Bret, Franck Da Silva, Ina Slynko and the ones that I am forgetting for their supportive behaviors.

At Eskitis, there are numerous past and present PhD colleagues that were just great with me. I have special thoughts to Liliana Pedro and her devotion to science. Thanks to the French connection, Marie-Laure Vial, Romain "Morvel" Lepage and Fanny Lombard for your efforts in creating social events at Eskitis. Benoît Serive (désolé de te mettre dans le lot des doctorants), Asmaa Boufridi (you are one of us) and Jan Lanz (a swiss is everything), I am glad you came at Eskitis. And last but not least thanks to Shaz for all the chicken drawings he left on my desk.

I would like to express a special thanks to Pauline Fabre for her very supportive and attentive behavior regarding me in all aspects of life. If I am here today, it is also thanks to you.

Last, I would like to thanks the members of my family, for their support during my candidature. It was always hard to leave France mainly because of you.

Résumé

La comparaison de site peut-elle vérifier l'hypothèse: «*Les origines biosynthétiques des produits naturels leurs confèrent des activités biologiques*»? Pour répondre à cette question, nous avons développé un outil modélisant les propriétés accessibles au solvant des sites de liaison. La méthode a montré des aspects intéressants, mais elle souffre d'une sensibilité aux coordonnées atomiques. Cependant, des méthodes existantes nous ont permis de prouver que l'hypothèse est valide pour la famille des flavonoïdes. Afin d'étendre l'étude, nous avons développé un procédé automatique capable de rechercher des structures d'enzymes de biosynthèse de produits naturels disposant de sites actifs capables de lier une molécule de petite taille. Nous avons trouvé les structures de 117 enzymes.

Les structures nous ont permis de caractériser divers modes de liaison substrat-enzyme, nous indiquant que l'empreinte biologique des produits naturels ne correspond pas toujours au modèle « clé-serrure ».

Abstract

Can computational binding site similarity tools verify the hypothesis: *"Biosynthetic moldings give potent biological activities to natural products"*? To answer this question, we designed a tool modeling binding site properties according to solvent exposure. The method showed interesting characteristics but suffers from sensitivity to atomic coordinates.

However, existing methods have delivered evidence that the hypothesis was valid for the flavonoid chemical class. In order to extend the study, we designed an automated pipeline capable of searching natural product biosynthetic enzyme structures embedding ligandable catalytic sites. We collected structures of 117 biosynthetic enzymes. Finally, according to structural investigations of biosynthetic enzymes, we characterized diverse substrate-enzyme binding-modes, suggesting that natural product biological imprints usually do not agree with the "key-lock" model.

Statement of originality

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Brisbane, 7th October 2015

Noé Sturm

Table of contents

Common abbreviations

Standard amino-acid three letter codes

Ala	Alanine	**Leu**	Leucine
Arg	Arginine	**Lys**	Lysine
Asn	Asparagine	**Met**	Methionine
Asp	Aspartic acid	**Phe**	Phenylalanine
Cys	Cysteine	**Pro**	Proline
Glu	Glutamic acid	**Ser**	Serine
Gln	Glutamine	**Thr**	Threonine
Gly	Glycine	**Trp**	Tryptophan
His	Histidine	**Tyr**	Tyrosine
Ile	Isoleucine	**Val**	Valine

Cα	Alpha atom on amino-acid side chain
Cβ	Beta carbon on amino-acid side chain
MOL2	Tripos format of molecular structures
sc-PDB	Screening Protein databank
Å	Angström distance unit (10^{-10} meter)
HET	HETero atom group identifier
H-bond	Non-covalent hydrogen bond
EC	Enzyme Commission number

Résumé (long)

INTRODUCTION

Les produits naturels sont à l'origine des principes actifs de nombreux médicaments.[1] Ces principes actifs prennent effet grâce à la formation d'un complexe protéine-ligand résultant de la reconnaissance moléculaire du ligand par sa cible thérapeutique. Or, dans la nature, les produits naturels sont synthétisés par des enzymes de biosynthèse. Et, lors de la biosynthèse, les précurseurs d'un produit naturel interagissent avec des enzymes, formant ainsi des complexes enzyme-ligand. Partant du principe que deux protéines capables de former un complexe protéine-ligand avec un ligand de structure identique partagent des propriétés structurales locales aux sites de reconnaissance du ligand, nous avons postulé que les bases structurales de la reconnaissance moléculaire d'un principe actif d'origine naturelle par ses enzymes de biosynthèse sont également présentes chez les protéines responsables de son effet thérapeutique (**figure 1**). Cette hypothèse, désignée par « Protein Fold Topology » (PFT), a été initialement formulée par RJ Quinn en 2006 suite à l'observation de points communs dans le mode de reconnaissance de composés de la famille des flavonoïdes par leurs enzymes de biosynthèse et par des kinases.[2–4] L'objectif de ma thèse est d'évaluer si cette hypothèse peut être vérifiée par similarité de site de liaison.

Le travail de thèse comprend une première partie méthodologique, pour la mise en place d'une approche informatisée pour la recherche de PFT à partir des structures de protéines. Ce travail préparatoire comprend un volet de développement du programme

de comparaison de sites SiteAlign, ainsi qu'un volet de test de deux programmes de comparaison de sites (SiteAlign[5] et Shaper[6]) pour caractériser les relations structurales entre les différentes protéines reconnaissant un même ligand.

Dans une deuxième partie, nous avons appliqué la méthode de comparaison de sites sur le seul exemple de PFT décrit (famille des flavonoïdes), démontrant l'efficacité de l'approche pour identifier, à partir de données publiques de structures de protéines, des paires enzymes de biosynthèse/protéines ciblées par le produit naturel.

Enfin, dans une dernière partie, nous avons entrepris un travail exploratoire pour recenser les nouvelles PFT. Pour cela, nous avons tout d'abord collecté des enzymes de biosynthèse dans la Protein Databank[7] et, nous les avons comparé aux protéines « ligandables » de la *sc*-PDB.[8] Les premiers résultats suggèrent de nouvelles PFT mais indiquent aussi les limites de l'approche.

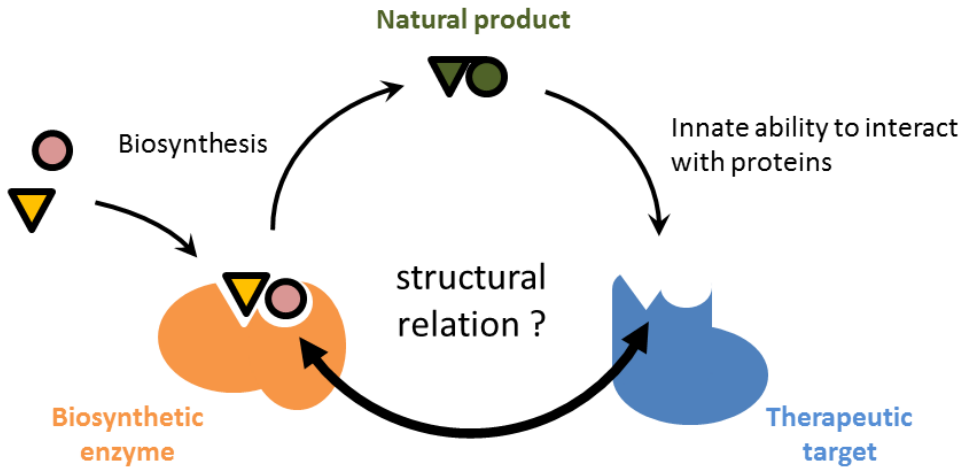


Figure 1. Hypothèse des « protein fold topology ».
Les bases structurales de l'empreinte moléculaire donnée à un produit naturel par une enzyme de biosynthèse se retrouvent-elles chez les protéines cibles de ce produit naturel ?

Chapitre 1.
Développent d'un outil informatique de comparaison de sites.

Les outils de comparaison de sites protéiques sont basés sur des approches géométriques où les représentations simplifiées des protéines sont superposées de manière à optimiser une fonction de score.[9] L'un des programmes de comparaison de site développé au laboratoire, SiteAlign[5], modélise les sites de liaison sur un polyèdre de 80 faces placé au centre de la cavité protéique. Les propriétés de chaque triangle du polyèdre sont obtenues par projection des caractéristiques de chaines latérales des acides aminés constituant le site de liaison (présence de groupements chargés, hydrophobe, aromatiques, de donneurs ou accepteurs de liaison hydrogène, encombrement, orientation vers le site ou enfoui, **figure 2**). Ces caractéristiques sont déterminées pour chaque type d'acides aminés, indépendamment du contexte protéique. Par exemple l'arginine est toujours encodée comme un résidu chargé positivement, même si son groupement guanidinium n'est pas accessible pour interagir avec un ligand. Inversement, une glycine n'a pas de propriété pharmacophorique, même si les groupements NH et CO de sa chaîne principale peuvent être impliqués dans la liaison avec le ligand. Pour réduire la description du site aux groupements chimiques susceptibles d'interagir avec le ligand, nous avons développé une version modifiée du programme SiteAlign, dans laquelle les caractéristiques des acides aminés sont déduites de celles des atomes accessibles au solvant, c'est-à-dire susceptibles d'interagir avec un ligand.

Nous avons utilisé la version originale et la version modifiée de SiteAlign pour rechercher dans la *sc*-PDB,[8] banque des sites ligandables de la Protein Databank,[7] les protéines capables de lier les mêmes ligands qu'un site d'intérêt. Globalement, les performances

des deux versions sont comparables pour cet exercice de criblage virtuel rétrospectif. L'analyse des résultats montre que notre représentation modifiée des sites contient une information plus précise des propriétés pharmacophoriques exposées aux ligands. Cependant, nous avons aussi pu remarquer que notre représentation souffre d'une sensibilité accrue aux coordonnées atomiques, rendant l'exploitation d'un criblage virtuel prospectif difficile.

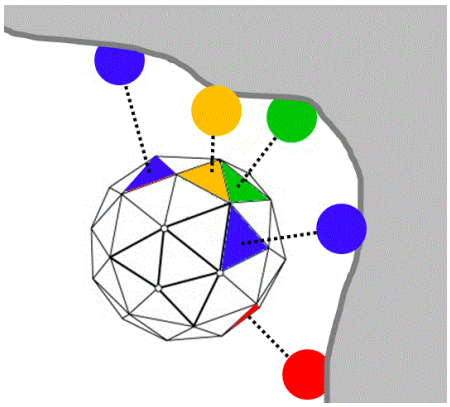


Figure 2. Représentation d'un site de liaison de ligand d'une protéine dans le programme SiteAlign.
La cavité protéique est représentée en gris. Les points de couleurs représentent les acides aminés du site de liaison. Les triangles de couleur représentent les faces du polyèdre sur lesquelles ces résidus sont projetés. Chaque couleur code pour les propriétés géométriques et pharmacophoriques d'un résidu.

Chapitre 2.
Bases structurales de la permissivité des molécules bioactives.

Afin de mieux comprendre les bases moléculaires de la permissivité des composés bioactifs, nous avons analysé la banque de données de structures de protéines (Protein Data Bank),[7] Nous avons identifié 247 molécules drug-like (c'est-à-dire possédant des propriétés physico-chimiques semblables aux principes actifs des médicaments approuvés) en complexe avec au moins deux protéines cibles d'intérêt thérapeutique différentes. Nous avons ainsi composé un jeu de données de 1070 paires de structures tridimensionnelles de complexes différents mais partageant le même ligand (**figure 3**).

La comparaison des structures des différents sites de liaison d'un ligand a révélé que le manque de sélectivité d'un ligand peut être dû au fait que la nature a créé des sites de liaison similaires dans des protéines différentes (y compris si les séquences ne sont pas conservées, et si leurs repliements 3D sont différents). Par exemple, les sites de liaison de l'ATP des kinases ont des structures 3D très similaires même si leurs structures globales sont distantes. Cette caractéristique structurale rend d'ailleurs la conception d'inhibiteur sélectif d'une seule kinase difficile. Notamment, nous avons retrouvé dans notre jeu de données le cas de la 2-morpholin-4-Yl-7-phenyl-4h-chromen-4-one, qui inhibe deux kinases non homologues (<10% d'identité de séquence), Pim-1 et PI3K, en se liant au site de liaison de l'ATP (PDB ID : 1E7V et 1YL3).

La promiscuité d'une molécule bioactive peut également être liée à ses propriétés particulières en tant que ligand. Nous avons démontré que certaines molécules peuvent s'adapter à différents environnements protéiques en modifiant leur conformation. Par exemple, un dérivé phosphorylé de la vitamine B1 adopte des conformations distinctes

dans pas moins d'une trentaine de paires de structures de complexes avec des sites de liaisons différents. D'autres molécules peuvent utiliser des points d'ancrages différents pour interagir avec différentes cibles en adoptant un mode d'interaction éloigné des modèles de complémentarité de forme et d'interaction. C'est le cas de nombreux composés d'origine naturelle, dont la quercétine, molécule rigide polyhydroxylée qui est capable de se lier à des cavités de forme, de taille et de propriétés très différentes en utilisant ou non ses groupements OH dans des liaisons hydrogène intermoléculaires.

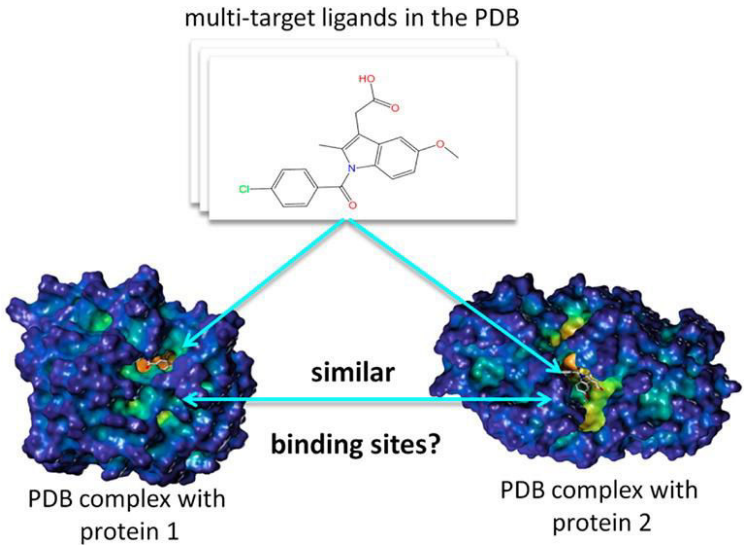


Figure 3. Stratégie d'étude des bases structurales de la permissivité des molécules bioactives.

Chapitre 3

Similarité structurale entre les enzymes de la biosynthèse des flavonoïdes et les protéines ciblées par les flavonoïdes.

Afin d'évaluer si des méthodes de similarité de site permettaient de retrouver la seule PFT connue, nous avons comparé les sites de liaison du ligand de cinq enzymes de biosynthèse de flavonoïdes à un jeu de 8077 sites de liaison de la sc-PDB[10] représentant 3678 protéines d'intérêt thérapeutique. Pour ce faire, nous avons utilisé SiteAlign[5] et Shaper[6], deux programmes informatiques de comparaison tridimensionnelle de site de liaison basés sur des représentations de site différentes. A la différence de SiteAlign, Shaper représente un site de liaison en modélisant sa cavité par un nuage de points annotés de propriétés pharmacophoriques. Tous les criblages réalisés ont permis de retrouver des cibles connues de flavonoïdes par similarité structurale avec des enzymes de biosynthèse. De plus, les calculs répétés pour différentes définitions du site de liaison des enzymes de biosynthèse (présence/absence de molécules d'eau, taille de site variable, ou modifications structurales) ont produit des résultats analogues, démontrant ainsi la robustesse des méthodes. Les cibles retrouvées sont néanmoins caractéristiques de l'enzyme comparée (par exemple les protéines kinase sont préférentiellement retrouvées par comparaison avec la chalcone isomérase), suggérant qu'il existe plusieurs composantes à l'empreinte biologique d'un produit naturel. Enfin, l'analyse détaillée des similarités locales entre les enzymes de biosynthèse des flavonoïdes et les protéines cibles des flavonoïdes révèlent des points d'ancrage communs (c'est-à-dire des groupements capables d'établir le même type d'interactions directionnelles placés dans le même arrangement tridimensionnel) sans qu'il y ait de ressemblance de forme des cavités.

Chapitre 4.

Inventaire des enzymes de biosynthèse de produits naturels dans la Protein Databank.

Forts de nos résultats prouvant que l'empreinte biologique des flavonoïdes peut être retrouvée chez les kinases par similarité de site, nous avons voulu étendre notre étude à toutes les enzymes de biosynthèses dont la structure est connue. La Protein Data Bank (PDB)[7] est la principale ressource publique internationale pour la collecte et la diffusion des structures moléculaires expérimentales de protéines. En 2014, le nombre d'entrées dans la PDB a dépassé la centaine de milliers, fournissant des données structurales pour plus de 35 000 protéines de séquences différentes. Les sites internet donnant accès aux données (« Research Collaboratory for Structural Bioinformatics PDB : RCSB PDB » accessible sur www.rcsb.org, « Protein Data Bank europe : PDBe » accessible sur www.ebi.ac.uk/pdbe/ et « Protein Data Bank japan: PDBj » accessible sur http://pdbj.org/) fournissent diverses annotations et outils d'analyse, cependant aucun d'entre eux ne permet facilement d'identifier les enzymes de biosynthèse de produits naturels.

Par conséquent, nous avons entrepris le développement d'un procédé automatisé pour rechercher dans la PDB les structures d'enzymes de biosynthèse de produits naturels et ce, dans le but de sélectionner celles dont le site actif est « ligandable » (c'est-à-dire prédit comme étant capable de lier avec une haute affinité des molécules « drug-like ») ce qui constitue une donnée importante pour les approches informatiques de conception rationnelle de molécules bioactives d'origine naturelle.

Notre stratégie, résumée sur la **figure 4**, est principalement composée de deux étapes. La première est le filtrage par mots clés des structures de la PDB. La seconde détecte les

sites catalytiques dans les structures issues de la première étape. Ce procédé fait intervenir l'annotation des protéines à partir de données externes (RCSB PDB, UniProt[11] et Catalytic Site Atlas[12]), l'identification des acides aminés catalytiques dans la séquence des structures de protéines via un alignement de séquence réalisé par le programme needle[13] (librairie de programmes EMBOSS[14]), la détection des cavités dans les structures des protéines réalisé par le programme VolSite,[6] la sélection d'une cavité « ligandable » contenant au moins un acide aminé catalytique et l'annotation des enzymes de leurs activités enzymatiques (substrats, produits).

Conjointement, nous avons mené une recherche d'enzymes de biosynthèse de produits naturels à partir de banques de données de voies de métabolismes élucidées expérimentalement (MetaCyc[15] et UniPathway[16] incluse dans la ressource UniProt[11]). Cette recherche parallèle nous a permis d'identifier toutes les enzymes de biosynthèse documentées ainsi que d'en extraire les structures contenant des sites catalytiques identifiés par notre procédé automatique (identique au procédé énoncé précédemment). Enfin, pour vérifier les structures de protéines retrouvées à partir de la PDB, nous avons utilisé les données d'annotation métaboliques accessibles via les ressources d'UniProt[11] et de MetaCyc.[15] La vérification manuelle nous a permis de valider 33 enzymes de biosynthèse en confirmant les voies de biosynthèse dans lesquelles elles sont impliquées ainsi que les réactions qu'elles catalysent. Au total, les données récoltées à partir de la PDB et à partir des banques de données de métabolisme nous ont permis d'identifier automatiquement 117 enzymes de biosynthèse avec un site catalytique « ligandable ». Elles sont impliquées dans les voies de biosynthèse de terpènes, d'isoprènes, de phenylpropanoïdes, de polycétides, d'alcaloïdes, d'antibiotiques, de certains acides gras et d'autres métabolites secondaires. L'analyse des enzymes récupérées nous indique que

les voies de biosynthèse élucidées reflètent les centres d'intérêts de l'industrie pharmaceutique. Par exemple, la classe biosynthétique contenant le plus d'enzymes décrit la biosynthèse d'antibiotiques. Par ailleurs, il est intéressant de constater que certaines voies de biosynthèse disposent de multiples enzymes dont les structures sont connues. Par exemple, la voie de biosynthèse de l'alcaloïde ajmaline est représentée par quatre enzymes différentes impliquées dans des réactions aussi bien en aval qu'en amont de la voie de biosynthèse.

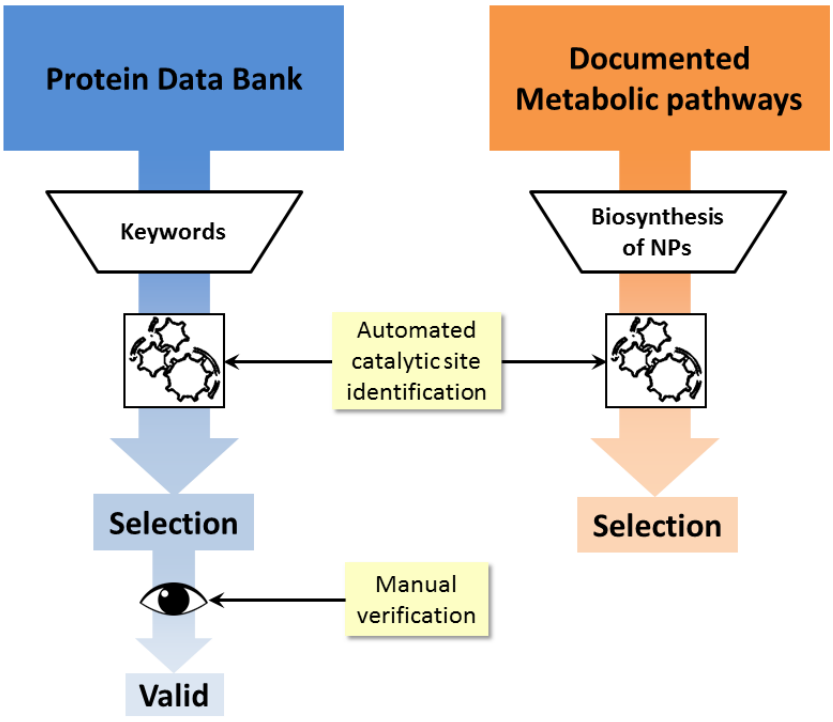


Figure 4. Protocole de création de la banque de donnée d'enzymes de biosynthèse.
D'une part (gauche, bleu), on entreprend un filtrage par mots clés de toutes les entrées de la PDB. S'en suit un procédé automatisé qui permet d'identifier les protéines disposant d'un site catalytique "ligandable". La sélection obtenue a été vérifiée manuellement en utilisant des données extraites à partir des voies de biosynthèse de produits naturels documentées. D'autre part (droite, orange), on entreprend un filtrage des voies de métabolisme pour ne garder que des voies de biosynthèse de produits naturels. S'en suit le même procédé automatique pour identifier les protéines disposant d'un site catalytique "ligandable".

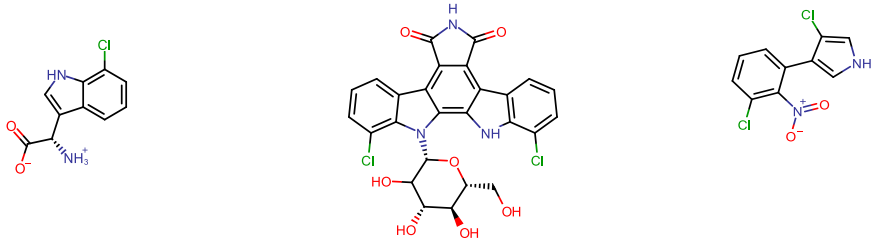
Chapitre 5.
Investigation structurale de l'empreinte biologique des produits naturels

Dans le cadre d'une approche de comparaison de site, le prérequis nécessaire à la recherche d'une protéine cible est que le ligand soit reconnu suivant le principe de complémentarité de forme « clé-serrure ». Dans le contexte de notre étude (recherche de nouvelles PFT), ce prérequis stipule que l'empreinte moléculaire portée par l'enzyme de biosynthèse doit être caractéristique du produit naturel. Cette évaluation a été menée à deux niveaux. Dans un premier temps, nous avons considéré les structures des substrats et produits impliqués dans l'activité enzymatique. Nous avons pu mettre en évidence un ensemble d'enzymes de biosynthèse agissant sur des composés dont la structure chimique n'est pas caractéristique du produit naturel final. Par exemple, les tryptophane halogenases RebH et PrnA sont responsables de la chloration d'un tryptophane, une étape précoce de la biosynthèse des antibiotiques rebeccamycine et pyrrolnitrine[17,18]. Or, comme le montre la **figure 5**, il est difficile de mettre en relation les structures de rebeccamycine et pyrrolnitrine avec le tryptophane initial. Par conséquence, on peut raisonnablement affirmer que les enzymes RebH et PrnA ne portent une empreinte biologique exploitable par similarité de site. Deuxièmement, nous avons considérés les modes de reconnaissance enzyme-substrat. Nous avons pu identifier plusieurs cas de figures. 1/ L'enzyme reconnaît un métabolite proche du produit final avec une complémentarité considérable; 2/ l'enzyme reconnaît un fragment du produit naturel final; 3/ l'enzyme reconnaît un précurseur non-représentatif du produit final (mentioné plus haut); 4/ le mode de reconnaissance n'est pas caractéristique d'un produit naturel; 5/ la structure de l'enzyme n'est pas représentative de l'état fonctionnel de

l'enzyme. Armés de ces connaissances, nous avons pu sélectionner les enzymes de biosynthèse pour lesquelles nous avons le plus de chance de retrouver des PFT. Dans les cas propices et pour poursuivre la validation de l'hypothèse des PFT, nous tout d'abord cherché les protéines connues pour être ciblées par les produits naturels présentes dans notre jeu de données.

Les cibles connues ont été cherchées dans les bases de données DrugBank,[19] ChEMBL[20] et dans la PDB. Suivant les cas où les cibles connues sont présentes dans le jeu de données criblé, nous avons comparé le site catalytique des enzymes portant une empreinte biologique représentative d'un produit naturel à la totalité des structures de sites de liaisons de la sc-PDB[10] avec les programmes SiteAlign[5] et Shaper.[6] Les calculs ont été entrepris sur une centaine de processeurs parallélisés du centre de calcul haute performance de l'IN2P3 de Villeurbanne.

Les criblages virtuels ont permis de retrouver des cibles connues, notamment nous avons identifié une enzyme β-lactamase similaire à une enzyme de biosynthèse de la pénicilline G. Ce résultat suggère qu'une relation structurale entre biosynthèse et β-lactamase serait à l'origine de la résistance bactérienne responsable de l'hydrolyse du cycle β-lactame des pénicillines.[21]



Chloro-L-tryptophane Rebeccamycine Pyrrolnitrine

Figure 5. Un précurseur commun à Rebeccamycine et à Pyrrolnitrine.

Conclusion et perspectives

Nous avons produit un jeu de données de 117 enzymes de biosynthèse, toutes annotées de leur substrats/produits respectifs et pour lesquelles une cavité catalytique a été identifiée. Parmi ces 117 enzymes, nous avons comparé celles qui portent les empreintes biologiques les plus représentatives d'un produit naturel à un jeu de données de plus de huit milles structures de protéines d'intérêt thérapeutique. Les outils développés pendant la thèse permettent le prétraitement automatique des données brutes de criblages (le tri des listes, la définition des seuils de similarité, l'annotation des structures comparées avec les informations relatives à la protéine et l'identification des protéines cibles connues dans les listes de criblage). La totalité des enzymes de biosynthèse, nécessitant une analyse au cas par cas, pourra être supporté par les connaissances produites durant la thèse. Notamment, l'analyse des activités enzymatiques et des modes de reconnaissance de produits naturels nous a permis de caractériser précisément ce qu'est l'empreinte biologique des produits naturels et d'énoncer des critères permettant d'appréhender le potentiel d'une enzyme de biosynthèse pour la recherche de PFT par similarité de site. Par exemple, la considération de ces critères nous a permis d'identifier une relation structurale entre biosynthèse des pénicillines et résistance bactérienne contre les pénicillines. Finalement, cette thèse constitue une base solide pour la recherche d'autres relations structurales biosynthèse/cible, elle contient un inventaire des structures d'enzymes de biosynthèse disponibles ainsi qu'un diagnostic de la pertinence des approches de comparaison de sites pour trouver un lien entre la biosynthèse des composés d'origine naturelle et leurs activités pharmacologiques.

Références

1. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.* **75,** 311–335 (2012).
2. McArdle, B. M., Campitelli, M. R. & Quinn, R. J. A common protein fold topology shared by flavonoid biosynthetic enzymes and therapeutic targets. *J. Nat. Prod.* **69,** 14–17 (2006).
3. McArdle, B. M. & Quinn, R. J. Identification of protein fold topology shared between different folds inhibited by natural products. *Chembiochem Eur. J. Chem. Biol.* **8,** 788–798 (2007).
4. Kellenberger, E., Hofmann, A. & Quinn, R. J. Similar interactions of natural products with biosynthetic enzymes and therapeutic targets could explain why nature produces such a large proportion of existing drugs. *Nat. Prod. Rep.* **28,** 1483–1492 (2011).
5. Schalon, C., Surgand, J.-S., Kellenberger, E. & Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins Struct. Funct. Bioinforma.* **71,** 1755–1778 (2008).
6. Desaphy, J., Azdimousa, K., Kellenberger, E. & Rognan, D. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *J. Chem. Inf. Model.* **52,** 2287–2299 (2012).
7. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
8. Meslamani, J., Rognan, D. & Kellenberger, E. sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinforma. Oxf. Engl.* **27,** 1324–1326 (2011).
9. Kellenberger, E., Schalon, C. & Rognan, D. How to Measure the Similarity Between Protein Ligand-Binding Sites? *Curr. Comput. Aided-Drug Des.* **4,** 209–220 (2008).
10. Desaphy, J., Bret, G., Rognan, D. & Kellenberger, E. sc-PDB: a 3D-database of ligandable binding sites--10 years on. *Nucleic Acids Res.* **43,** D399–404 (2015).
11. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43,** D204–D212 (2015).
12. Furnham, N. *et al.* The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.* **42,** D485–D489 (2014).
13. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443–453 (1970).
14. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277
15. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **42,** D459–D471 (2014).
16. Morgat, A. *et al.* UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.* **40,** D761–D769 (2012).

17. Kirner, S. *et al.* Functions encoded by pyrrolnitrin biosynthetic genes from Pseudomonas fluorescens. *J. Bacteriol.* **180,** 1939–1943 (1998).

18. Yeh, E., Garneau, S. & Walsh, C. T. Robust in vitro activity of RebF and RebH, a two-component reductase/halogenase, generating 7-chlorotryptophan during rebeccamycin biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 3960–3965 (2005).

19. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42,** D1091–1097 (2014).

20. Bento, A. P. *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **42,** D1083–D1090 (2014).

21. Fisher, J. F., Meroueh, S. O. & Mobashery, S. Bacterial Resistance to β-Lactam Antibiotics: Compelling Opportunism, Compelling Opportunity. *Chem. Rev.* **105,** 395–424 (2005).

Summary

Natural products are an inspiring source of drugs.[1] Their active principles take effect thanks to the formation of a protein-ligand complex resulting from the molecular recognition of a ligand by its target protein. Yet, in nature, natural products are synthesized by biosynthetic enzymes. And, during the biosynthetic process, precursors of natural products interact with enzymes, thus forming enzyme-ligand complexes. Considering that two proteins capable of recognizing a ligand with same structure have similar properties embedded in their binding sites, we assumed that the structural basis of the molecular recognition of a natural drug by its biosynthetic enzymes is also embedded in the binding site of the protein targeted by this natural product (**figure 1**). This hypothesis was termed "Protein Fold Topology" (PFT).

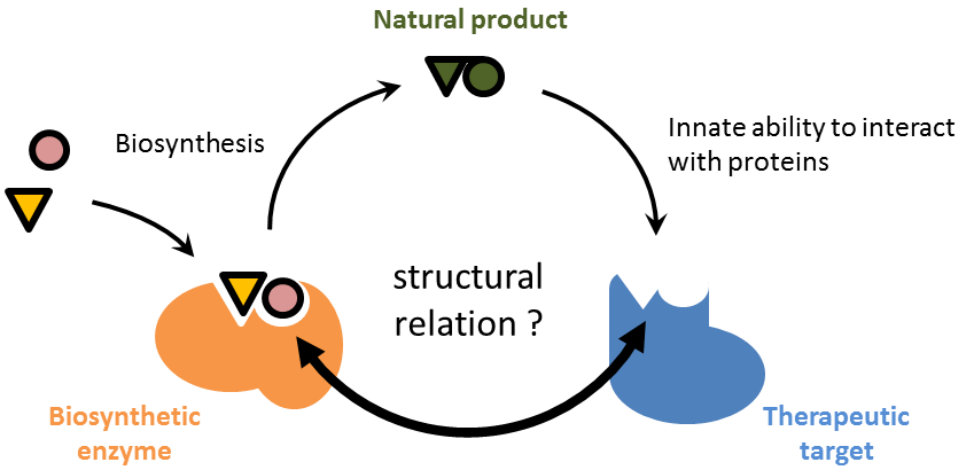


Figure 1. « Protein Fold Toplogy » hypothesis.

In the first part of the thesis, we describe methodologic aspects that were required to address the PFT approach. First, we developed a binding site comparison method derived from SiteAlign.[2] The method models binding sites considering their solvent accessible pharmacophoric properties. Although our new description of binding sites is more representative of potential molecular recognition points, virtual screening experiments showed that it suffers from high sensitivity to atomic coordinates, thus making the exploitation of prospective experiments difficult.

The identification of PFTs implies identifying similar structural features in proteins of different nature, architecture and function. Thus, we have performed the diagnostic of the ability of binding site comparison tools for the identification of unrelated proteins binding a same ligand. We created a dataset containing pairs of different proteins binding to the same ligand. Proteins in pairs of the dataset were compared to each other with three existing binding site comparison tools (SiteAlign,[2] Shaper,[3] FuzCav[4]). Experiments provided evidence that proteins binding ligands with same structures have not always similar binding sites. The results also showed that the ability to bind dissimilar binding sites was most often achieved through ligands' flexible, hydrophobic or low complexity properties. More importantly, the results showed that natural metabolites such as lipids, coenzymes (participating in many biochemical reactions) or the widely distributed flavonoids are all able to bind dissimilar sites.

In the third chapter, we focused on the only described example of PFT and addressed the question: "can binding site similarity describe the relation between

flavonoid biosynthetic enzymes and kinase proteins?". In that, we compared a set of five flavonoid biosynthetic enzymes to about 10 000 binding sites of the sc-PDB.[5] Each screening experiment was able to identify known flavonoid target proteins, thereby demonstrating that computational binding site similarity methods can find relationships between biosynthetic enzymes and target proteins. Results also suggested the existence of multiple component biological imprints, since different biosynthetic enzyme screens yielded in different results. For example, Chalcone Isomerase screens yielded in the list of similar proteins that was most enriched in known flavonoid targets and more particularly kinase proteins, thereby suggesting that Chalcone Isomerase embeds a biological imprint that is most representative of flavonoid molecular core.

Strong with these encouraging results, we undertook the creation of an inventory of biosynthetic enzymes with the underlying aim of searching for new PFTs. Therefore, we designed an automated pipeline capable of searching biosynthetic enzymes in the Protein Databank considering keywords or metabolic data provided by the resources MetaCyc[6] and UniPathway.[7] We could find structures for 117 biosynthetic enzymes of secondary metabolites. The content of our dataset reflects interests of pharmaceutic industry, for example the largest class of biosynthetic enzymes are involved in the biosynthesis of antibiotics.

Lastly, in chapter 5 we have investigated our dataset for application in binding site comparison virtual screenings. In the context of our study, the first requirement for binding site comparison methods to find natural product target proteins is that enzymes must embed structural features that are complementary to the natural

product they shape (following the “key-lock” model). In the first place, the enzyme must interact with a substrate that is significantly similar to a natural product. The analysis of enzymatic activities in our dataset revealed that it is not always the case for biosynthetic enzymes. For example, the chlorination of tryptophan by tryptophan halogenases RebH and PrnA is a premature step of the biosynthesis of the antibiotics pyrrolnitrin and rebeccamycin,[8,9] both unrelated between themselves and with their common precursor. Thus, we can reasonably affirm that RebH and PrnA do not embed a biological imprint that is exploitable by binding site similarity. Secondly, when an enzyme interacts with a compound that is similar to the final natural product, the enzyme must recognize specifically its substrate. Visual inspections of enzymes in complex with their substrate (or analogues) revealed that it is neither the case for all enzymes in our dataset. We observed diverse molecular recognition modes suggesting that binding sites of enzymes in our dataset do not exclusively agree with binding site comparison approaches. However, a subset of our dataset showed potential application in binding site comparison screenings. For instance, a biosynthetic enzyme of penicillin G recognizes the lactam moiety that is responsible for β-lactam penicillins pharmacological activities. Virtual screening experiments yielded in the identification of remote similarities with one β-lactamase. Considering this discovery, it is very tempting to speculate that the origin of bacterial resistance to penicillin (lactam hydrolysis mechanism) could be induced by structural resemblance with penicillin biosynthesis.

We have produced a dataset of 117 biosynthetic enzymes of natural products all annotated with their respective enzymatic activities (substrate/product) and with

catalytic site that could suit drug-like molecules. Amongst these 117 enzymes, we have compared those that, from our point of view, embed biological imprints most representative of natural products to more than 8000 binding sites in complex with a drug-like molecule. The tools designed during the candidature allow automated preprocessing of raw screening results (ordering of screening lists according to similarity score, definition of similarity thresholds specific to each screening, annotation of compared structures with information relative to the protein they belong to, identification of known targets). However, the screening of the full set of enzymes requires a case-by-case analysis that can be supported by the thesis' work. In particular, the analysis of enzymatic activities and of the binding-modes of enzymes with their substrates have enabled us to characterize precisely what is a biological imprint and to define a set of criterion to apprehend the potential of biosynthetic enzymes to find new PFTs by binding site similarity. For example, the consideration of these criterion have allowed us to identify a structural relationship between the biosynthesis of penicillins and bacterial resistance to penicillins. Finally, this thesis constitute a strong ground for the search of other biosynthesis/targets relations, the thesis contains an inventory of available biosynthetic enzyme structures as well as a diagnostic of the pertinence of binding site comparison approaches to find a link between natural products and their pharmacologic activities.

References

1. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.* **75,** 311–335 (2012).
2. Schalon, C., Surgand, J.-S., Kellenberger, E. & Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins Struct. Funct. Bioinforma.* **71,** 1755–1778 (2008).
3. Desaphy, J., Azdimousa, K., Kellenberger, E. & Rognan, D. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *J. Chem. Inf. Model.* **52,** 2287–2299 (2012).
4. Weill, N. & Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein−Ligand Binding Sites. *J. Chem. Inf. Model.* **50,** 123–135 (2010).
5. Desaphy, J., Bret, G., Rognan, D. & Kellenberger, E. sc-PDB: a 3D-database of ligandable binding sites--10 years on. *Nucleic Acids Res.* **43,** D399–404 (2015).
6. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **42,** D459–D471 (2014).
7. Morgat, A. *et al.* UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.* **40,** D761–D769 (2012).
8. Kirner, S. *et al.* Functions encoded by pyrrolnitrin biosynthetic genes from Pseudomonas fluorescens. *J. Bacteriol.* **180,** 1939–1943 (1998).
9. Yeh, E., Garneau, S. & Walsh, C. T. Robust in vitro activity of RebF and RebH, a two-component reductase/halogenase, generating 7-chlorotryptophan during rebeccamycin biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 3960–3965 (2005).

Introduction

My PhD project took place in a partnership involving Eskitis Institute for drug-discovery, in Brisbane, Australia and the Structural Chemogenomics Laboratory, in Illkirch, France. Eskitis Institute focuses mainly on natural products drug-discovery, therefore the original idea onto which my project relies came from Australia. The Chemogenomics Laboratory is specialized in structural bioinformatics and has a strong expertise in binding site comparison. It was therefore called to provide his knowledge for the project. The collaboration was born upon discussion of the computational tool SiteAlign between my two supervisors.

In nature, chemical compounds are synthesized through a series of reactions carried out by biosynthetic enzymes. These chemical compounds have often been related to biological functions involved in organism lives. Most commonly, they are classified into two categories: primary metabolites and secondary metabolites (the latter also referred to as natural products). Primary metabolites are associated with known functions, ensuring the economy of an organism. However, the majority of secondary metabolite functions remain unclear.[1] Nevertheless, since the early 80s, chemical ecology provides increasing evidence that the production of natural products are a result of an evolutionary processes aiming at improving organisms' survival functions.[2] For example, pathogenic compounds can affect growth of attacked organisms or alternatively, toxins provide defensive mechanisms against predators.[3] This fact tells us that nature designs chemical compounds that are able to interact within the biological world. Natural products' immense diversity and their potent biological activities have already attracted

many drug-discovery programs. In the last few decades, several reports have described the importance of natural products in the creation of marketed drugs.[4] In fact, nearly half of all approved drugs are natural products, natural product derivatives or have been designed on the basis of natural product models. Since natural products have long been used to benefit human health, it is worth understanding their origins and why they exhibit attractive features to treat human disease.

In earlier studies,[5–7] it has been suggested that the biosynthetic origins of natural products could be responsible for their potent biological activities. The idea behind this suggestion is that, while being synthesized, natural products interact with biosynthetic enzymes and thus, "memorize" a biological imprint within their architecture. Following this idea, McArdle et al. investigated molecular recognition of flavonoid by biosynthetic enzymes and compared recognition patterns with flavonoid target proteins.[6] Thanks to the observation of several crystallographic structures, the team concluded that, despite not sharing similar folds, the enzymes shared similar topological features with flavonoid target proteins in the proximity of the bound ligands. Indeed, kinases exhibit the kinase-like fold whereas the studied flavonoid biosynthetic enzyme exhibit the thiolase-like fold, but both structures show remarkable traits in the way their active sites are formed. The study pointed out similar local arrangements of secondary structure elements, providing an interaction pattern adapting to flavonoids molecular core. This similarity was termed Protein Fold Topology (PFT). Having set a possible relation between biosynthetic enzymes and therapeutic targets, McArdle and al carried out a second study,[5] in which they investigated if shared PFT between therapeutic targets of different folds was the underlying factor of the molecular recognition of a same natural product inhibitor. They used the zincin-like fold as a starting point to investigate compounds

recognized by multiple protein folds. According to the structural classification of protein database (SCOP),[8] there was 64 inhibitors of proteins exhibiting the zincin-like fold. From these, 28 were known to be inhibitors of other folds. Bestatin was the only natural product recognized by different folds and with available complex crystallographic structures. A second PFT was identified between two targets of bestatin inhibitor, leukotriene A4 hydrolase aminopeptidase, the zincin-like fold protein and an aminopeptidase, exhibiting the phosphorylase/hydrolase-like fold. Similarly to the flavonoid-kinase example, visual inspection of the crystallographic complexes (PDB ID: 1TXR and PDB ID: 1HS6 respectively) allowed the identification of similar arrangements of secondary structure elements. Again, the arrangements in the two folds provide equivalent molecular recognition points. Throughout those two studies, McArdle et al. suggested that fold topology relationship linking biosynthetic enzymes to therapeutic targets could be used as a tool to discover novel targets of natural products.

In 2011, Kellenberger and al. precisely reviewed the possibility of using PFT as a drug discovery approach.[7] In that, they investigated if kinase-flavonoid PFT could be identified by pharmacophore models focusing on protein-ligand H-bond interactions. A common patterns in H-bond interactions was searched in 21 complexes involving flavonoids bound to biosynthetic enzymes and kinases. In order to avoid trivial matches between intermolecular H-bonds, focus was made on complexes involving at least three hydrogen bonds located on two different rings of the flavonoid ligands only. This would ensure the presence of a biological imprint representative of flavonoid molecules within the binding sites. The systematic comparison of the considered biosynthetic enzymes with kinases showed that flavonoid H-bond patterns are more diverse than conserved. However, despite not being able to identify an overall conserved H-bonding pattern, a

small number of similar patterns could be identified between pairs of proteins, revealing a potential application in computer-aided drug discovery. Hence, the statement was set that in drug-discovery, the PFT approach could bridge biosynthetic and the therapeutic spaces using biosynthetic enzymes binding sites as a biological imprint in identifying target proteins by binding site similarity. As concluded in the previously mentioned study,[7] the approach needs more extensive evaluation both in terms of methods for the representation of biological imprints and in terms biosynthetic enzyme variety. The subject of my thesis is to evaluate the ability of computational binding site comparison methods to capture biological imprints of a natural product within target protein binding sites and to extend the study of PFTs to a wider range of biosynthetic enzymes.

References

1. Williams, D. H., Stone, M. J., Hauck, P. R. & Rahman, S. K. Why are secondary metabolites (natural products) biosynthesized? *J. Nat. Prod.* **52,** 1189–1208 (1989).
2. Haslam, E. Secondary metabolism – fact and fiction. *Nat. Prod. Rep.* **3,** 217–249 (1986).
3. Gershenzon, J. & Dudareva, N. The function of terpene natural products in the natural world. *Nat. Chem. Biol.* **3,** 408–414 (2007).
4. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.* **75,** 311–335 (2012).
5. McArdle, B. M. & Quinn, R. J. Identification of protein fold topology shared between different folds inhibited by natural products. *Chembiochem Eur. J. Chem. Biol.* **8,** 788–798 (2007).
6. McArdle, B. M., Campitelli, M. R. & Quinn, R. J. A common protein fold topology shared by flavonoid biosynthetic enzymes and therapeutic targets. *J. Nat. Prod.* **69,** 14–17 (2006).
7. Kellenberger, E., Hofmann, A. & Quinn, R. J. Similar interactions of natural products with biosynthetic enzymes and therapeutic targets could explain why nature produces such a large proportion of existing drugs. *Nat. Prod. Rep.* **28,** 1483–1492 (2011).
8. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540 (1995).

Chapter 1.

Development of a Computational Tool for Binding Site Comparison

This chapter was originally edited as a report included in my 1st year's candidature confirmation, a required milestone at Griffith University. However, in order to keep my thesis manuscript as concise as possible, I shortened and reviewed the report for incorporation into my thesis.

SiteAlign[1] is a binding site comparison tool that has been developed in Illkirch, France. The program compares binding site amino-acids spatial arrangements in order to detect common three-dimensional patterns. The attractive aspect of SiteAlign for the identification of new PFTs is in particular the fact that it models binding site residues by projecting then from their Cβ atoms onto a discretized sphere at the center of binding sites. In doing so, the program superimposes residues with similar physico-chemical properties and thus, matches molecular recognition points together, much similarly to what was done manually and visually by McArdle et al. to identify the flavonoid-kinase PFT.[2] However, we have rapidly identified a discrepancy between SiteAlign's model and the PFT similarities observed by Bernadette McArdle. The PFT description reported similar molecular recognition points provided by atoms at the origin of H-bonds interacting with the ligand. For example, a H-bond was provided by a carboxyl oxygen in the backbone of the flavonoid biosynthetic enzyme.[2] SiteAlign does not in its binding site representation. Moreover, residues are represented with their Cβ. In the case of large residue side chains such as Lys or Arg, interacting atoms might be located remotely to Cβ and therefore the binding site representation might be inaccurate. Therefore, we have decided to dedicate a part of my PhD to modify SiteAlign by adjusting its binding site representation to biological imprints of natural products in order to find new PFT relations.

In the first part of this chapter, I reviewed how SiteAlign works and focused specifically on binding site representation features. In a second part, I describe the modifications I made to represent binding sites with atoms at the origin of biological imprints rather than with a set of predefined descriptors modeled from Cβ atoms. Last, I discuss about virtual screening experiments that we performed to benchmark our modified version of SiteAlign against the original version.

1. SITEALIGN

1.1. Overview of SiteAlign

SiteAlign is a program that has first been described by C. Schalon and al.[1] The program models residues spatial arrangement onto a discretized sphere placed within a ligand binding site. According to the protein environment around the polyhedron, each face of the discretized sphere is assigned the set of descriptors encoding topological and physico-chemical properties of the residue that is facing it. Binding sites are compared to each other by searching the polyhedron superimposition that maximizes overlap between the descriptors.

1.2. SiteAlign input

SiteAlign requires two input files: the list of residues in the binding site and a protein structure file (MOL2 or PDB format). The list of residues is used for appropriate computing memory allocation while the protein structure is used to extract spatial coordinates of Cα, Cβ and identifiers of binding site residues only.

1.3. The sphere representation of protein binding sites

The basic idea behind SiteAlign is to model binding sites onto a spherical polyhedron. The polyhedron has 1Å radius, is initially placed at the geometrical center of binding site Cα atoms and is composed of 80 triangles uniformly scattered around the surface (**Figure 1**). Spatial arrangement of residues is modeled onto the polyhedron by projecting the Cβ atoms (from Cα if Gly residue) towards the center of the polyhedron. A face of the polyhedron is associated to the closest residue that projects onto it and is thereby assigned eight descriptors encoding topological and physico-chemical properties.

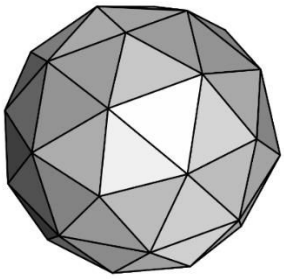


Figure 1. SiteAlign polyhedron.
Pentakis icosidodecahedron

The full polyhedron can be regarded as a fuzzy map, as each triangle encodes the space that is facing it regardless of the exact position of the projected residue.

I investigated the number of residues present in binding sites of the sc-PDB (v.2010).[3] Binding sites structures are defined as the set of residues located within a 6.5Å cutoff around heavy atoms of the bound ligand. As shown in **figure 2**, 50% of the binding sites contain 36 to 50 residues while only 10 contained more than 88 residues. For an average value of 43 residues, nearly half of the triangles are needed to model the binding site. The analysis of an extreme case (PDB ID: 3NLC, 96 residues) revealed that

up to 37 residues could be masked by another residue located closer to the center of the polyhedron. Duplicated projections are generally due to site definition. The number of residues is roughly proportional to ligand size and thus, it happens that the residue selection can expand beyond the binding site surface.

In order have a better idea of the fuzziness that is encoded into each triangle, I visually inspected the spatial volume covered by a triangle. As shown in **figure 3**, at 10Å from the center of the polyhedron, (roughly the average distance between Cα of residues and binding site centers) one triangle covers an area of approximately 5Å^2, which fits relatively well to the size of one residue.

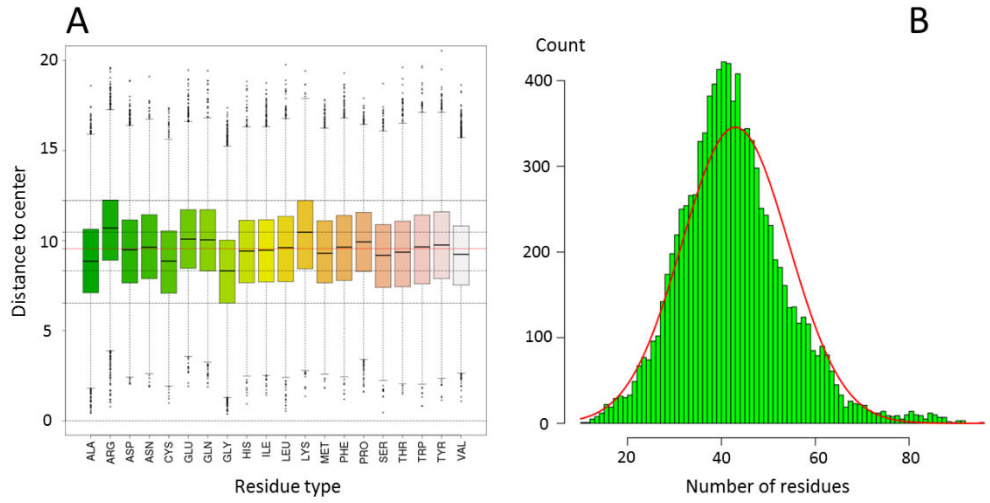


Figure 2. sc-PDB binding site population and distance to center.
A: distances from Cβ to the center of the binding site of the 9877 binding sites in the sc-PDB. Each boxplot represents a residue type. The red horizontal line represents the mean distance. B: distribution of the number of residues present in 6.5Å based binding site definition. The green bars represent the count of residues for the 9877 binding sites of the sc-PDB whereas the red line represents the density plot of these residue counts.

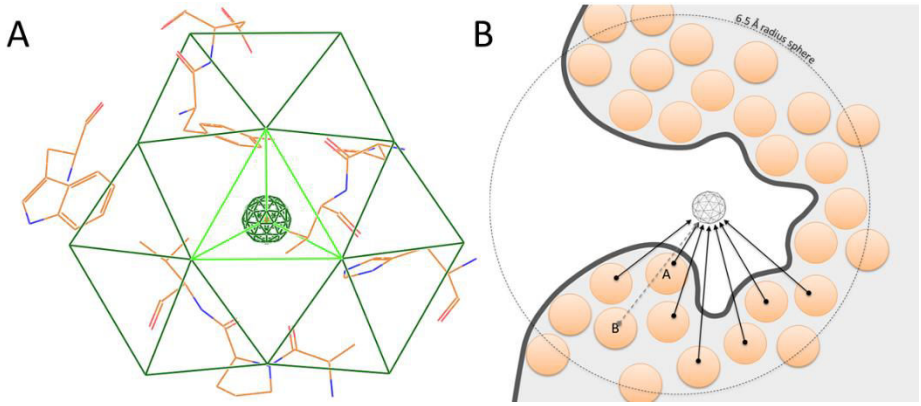


Figure 3. SiteAlign polyhedron within a binding site.
A: Spatial coverage of a section of the discretized sphere. The view is oriented as if one would look from outside the binding site, the sphere being behind the orange colored residues wires. The polyhedron is represented with dark green wires. Other green lines represent the projection of a section of the polyhedron at 10Å, roughly overlaying the residues. B: Residue projection scheme. The dark grey line represents the surface of a protein. Light grey area indicates the protein medium and orange circles represent residues. The polyhedron is located at the center of the binding site. Black arrows represent residue projections onto the polyhedron. Residue (B) is masked by residue (A).

1.3.1. Sphere implementation

The polyhedron is described by a set of 42 vertices, each described by Cartesian coordinates directly encoded into SiteAlign source code. Each triangle of the polyhedron is described as a vector of three vertices. Triangles are stored in an array of 80 elements describing the whole polyhedron. Accessing a triangle of the polyhedron and thus, manipulating residue descriptors, is done by calling the corresponding elements of the array.

1.3.2. Physico-chemical descriptors

In SiteAlign, residues are described by five types of physico-chemical descriptors according to the properties of residue side chains: count of H-bond acceptors, count of H-bond donors, presence of aromatic group, presence of an aliphatic chain and charge at pH7. Values for each physico-chemical descriptor a given in **table 1**.

	Aliphatic	Donor	Acceptor	Aromatic	Charge
Ala	1				
Arg		3			+1
Asn		1	1		
Asp			2		-1
Cys	1				
Glu			2		-1
Gln		1	1		
Gly					
His		1	1	1	
Ile	1				
Leu	1				
Lys		1			+1
Met	1				
Phe				1	
Pro	1				
Ser		1	1		
Thr	1	1	1		
Trp		1		1	
Tyr		1	1	1	
Val	1				
amplitudes	*[0-1]*	*[0-3]*	*[0-2]*	*[0-1]*	*[-1 - +1]*

Table 1. SiteAlign physico-chemical descriptors.
The last line represents the ranges of values that each descriptor can take.

1.3.3. Topological descriptor

In addition to the five physico-chemical descriptors, residues are described by three topological descriptors. At the difference to the fixed set of physico-chemical descriptors, two of the three topological descriptors can vary depending on the position of the residue in the binding site. The topological descriptors are: distance from Cβ to the center of the polyhedron (from Cα if the residue is a Gly), orientation of the residue and size of the residue. Distances are discretized into several bins with 0.5Å intervals. Orientation descriptor encodes whether residue side chains point towards or outwards the binding site, taking respectively the value 1 or 2. Orientation is determined by comparing the distances to the center of the polyhedron from Cα and Cβ (Gly residues have an arbitrary value). The size descriptor is predefined for each residue type as one of three following sizes: small, medium and large (taking respectively the values 1, 2, or 3). Correspondences between residue and sizes are shown in **table 2**.

Size	Residues	Descriptor value
Small (0-3 heavy atoms)	Ala, Cys, Gly, Pro, Ser, Val	1
Intermediate (4-6 heavy atoms)	Asp, Asn, His, Ile, Lys, Leu, Met, Asp, Gln	2
Large (7-10 heavy atoms)	Thr, Phe, Arg, Tyr	3

Table 2. Size descriptors per residue type.

1.3.4. Association of descriptors with triangles

Each triangle of the polyhedron is associated to an array of eight values initially set to 0 (empty). This array will be referred as a fingerprint. A fingerprint is assigned the eight descriptors of the residue that projects onto it. **Table 3** describes fingerprints format.

0	1	2	3	4	5	6	7
Aliphatic	Donor	Acceptor	Aromatic	Distance	Size	Orientation	Charge
1	0	0	0	12	2	1	0

Table 3. Format of fingerprint arrays.
The top line represents the index of each descriptor in the fingerprint. The second line describes the descriptor type. The third line represents the descriptors of a Leu residue located at 6Å from the center of the polyhedron. The color code distinguishes the nature of descriptors. Physico-chemical descriptors are represented with light blue, topological descriptors with green.

1.3.5. Map of fingerprints

The polyhedron combines all fingerprints into a map representing the binding site. Initially, the map is composed of 80 empty fingerprints. After projection of binding site residues, the map contains a subset of populated triangles (assigned to residue descriptors). In practice, the map is encoded as an array of 80 fingerprints (**figure 4**) and thus, the resulting binding site map contains 640 integers. It is noteworthy to mention that the order of the fingerprints in

the map does not follow any special rule, thus it is difficult to know which fingerprints are adjacent on the polyhedron.

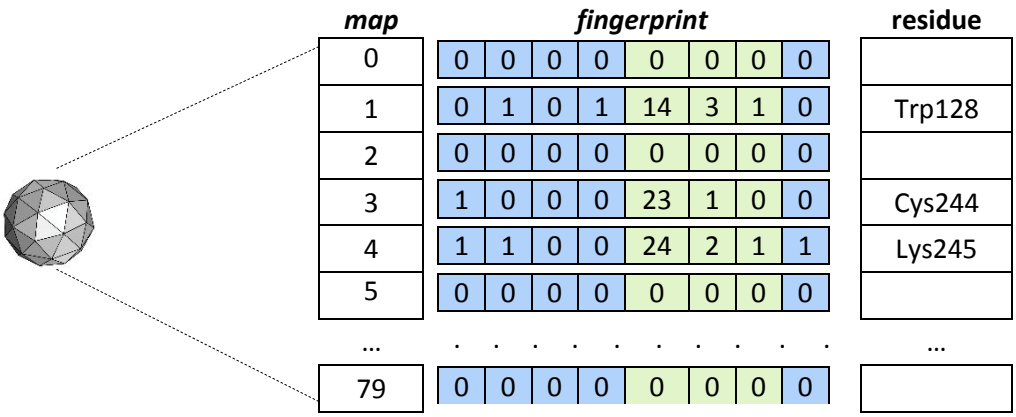


Figure 4. Binding site map carrying residue descriptors.
The polyhedron is encoded into an array of fingerprints mapping the binding site. Each fingerprint contains or not residue descriptors according to the spatial arrangement of residues.

1.4. Scoring similarity between maps

For binding site comparison, the program SiteAlign eventually has to compare maps. The maps are compared to each other by systematic pairwise comparison of the 80 fingerprints. The comparison of two fingerprints involves the computation of a score specific to each descriptor. In turn, the mean of descriptor scores is calculated to obtain a fingerprint score. This process is performed for each fingerprint pair, thus resulting in 80 fingerprint scores. These scores are ultimately combined together in order to obtain scores characterizing the similarity between the two maps. Scoring functions of each descriptor type are given in the following section.

1.4.1. Specific descriptor scores

Nomenclature:

- $v_{t,m}^{(d)}$ is the value of the descriptor with index *d* in the fingerprint associated to the triangle *t* of the map *m*

- $v_{t,m} = \left(v_{t,m}^{(1)}, \dots, v_{t,m}^{(8)}\right)$ is the set of descriptors (i.e. the fingerprint) in the triangle *t* of the map *m*. If the triangle does not contain any information $v_{t,m} = 0$.

- $s_t^{(d)}$ is the score between $v_{t,1}^{(d)}$ and $v_{t,2}^{(d)}$.

- s_t is the score of triangle *t* between two maps.

- $M^{(d)}$ is the maximal amplitude of the descriptor *d*

Physico-chemical descriptors scores - $\boldsymbol{s_t^{(i)}}$, $i \in (1,2,3,4,8)$

Where (1) = aliphatic

(2) = H-bond donor

(3) = H-bind acceptor

(4) = aromatic

(8) = charge

$$s_t^{(i)} = 1 - \frac{\left|v_{t,1}^{(i)} - v_{t,2}^{(i)}\right|}{M^{(i)}}$$

The distance to center descriptor score - $s_t^{(5)}$

$$d = \left| v_{t,1}^{(5)} - v_{t,2}^{(5)} \right|$$

$$s_t^{(5)} = \begin{cases} 1 - \frac{d}{30} & if\ d \leq 30 \\ 0 & else \end{cases}$$

The orientation descriptor score - $s_t^{(6)}$

$$s_t^{(6)} = \begin{cases} 1 & if\ \ v_{t,1}^{(6)} = v_{t,2}^{(6)} \\ 0 & else \end{cases}$$

The orientation descriptor score is either one when two orientations are the same, either zero in the opposite case.

The size descriptor score - $s_t^{(7)}$

$$s_t^{(7)} = 1 - \frac{\left| v_{t,1}^{(7)} - v_{t,2}^{(7)} \right|}{2}$$

As discussed earlier in this report, the size descriptor varies between 1 and 3. Its maximal amplitude is 3 – 1 = 2.

1.4.2. Fingerprint score calculation

Fingerprint's specific descriptor scores are summed together to obtain the mean score of the fingerprint.

$$s_t = \frac{1}{8}\sum_{d=1}^{8} s_t^{(d)}$$

1.4.3. Global score calculation

Fingerprint scores are summed together in order to obtain global scores S_1 and S_2. In S_1, the sum of scores is divided by the number of fingerprint pairs with at least one populated fingerprint (N_1). In S_2, the sum of scores is divided by the number of fingerprint pairs without empty fingerprints at all (N_2). Finally, D_1 and D_2 are defined as the complementary values of the global scores S_1 and S_2 and thus, are distance scores rather than similarity scores. In such a way, a value of 1 represents different binding sites whereas the value 0 represents identic binding sites.

$$S_1 = \frac{1}{N_1}\sum s_t \qquad S_2 = \frac{1}{N_2}\sum s_t$$

$$D_1 = 1 - S_1 \qquad D_2 = 1 - S_2$$

D_1 distance accounts for a global distance score, considering all residues, whether or not they match a residue of the compared maps, thus affecting the result negatively if the binding site sizes are different in terms of residue count. In contrast, D_2 accounts for local alignments since it considers only contribution of superimposed residues. According to the definition of SiteAlign thresholds, two binding sites are considered similar if $D_1 < 0.6$ AND $D_2 < 0.2$.

1.5. Structural alignment

The alignment is based on a systematic search algorithm as follows:

(1) A polyhedron is placed at the center of the reference binding site.

(2) Binding site residues are projected onto the polyhedron and the reference map is generated.

(3) Another discretized polyhedron searches the compared binding site. Six parameters (3 translation axes and 3 rotation axes) are explored systematically so that the polyhedron explores a satisfactory number of positions. By default, the search explores a cubic grid of 4Å width by increments of 0.25Å. Each position is tested with 16 rotations along each of the three axis of space using π/8 increments. After each transformation, residues in the compared binding site are projected onto the polyhedron in order to compute D_1 and D_2 scores.

(4) The three best polyhedron positions (defined by 6 parameters each) are stored into memory for further refinement.

(5) A refinement is performed around the three best positions. Increments of exploration parameters are decreased (by default the cubic grid is reduced to 0.5Å) for refined search. Again, after each iteration, D_1 and D_2 scores are computed.

(6) The best polyhedron position of the refined search is selected.

(7) The transformation that led the compared polyhedron to the best solution is applied to the structure of the compared protein.

2. SOLVENT ACCESSIBLE BINDING SITE DEFINITION

Since binding site definition is not a trivial task, it was separated from binding site comparison. Original binding site structures present in the sc-PDB are defined as the residues within a 6.5Å cutoff around any heavy atom of the co-crystalized ligand.[3] The radius value has been determined statistically in order to suit most of drug target protein binding sites. However, this binding site description is not representative of biological imprints (or molecular recognition points in enzymes). For instance, large ligands lead to binding sites that might contain buried residues, which are unlikely to interact with a ligand (as seen in section 1.3). Moreover, this definition requires the presence of a ligand in the binding site, which is not necessarily the case in biosynthetic enzyme structures. Therefore, we had to ask the question "how to define a binding site representing biological imprint of natural products, especially if the protein structure is free of ligand?". We developed a side module coded in Perl to identify: (1) residues located within 6.5Å around any heavy atom of the ligand; (2) residue with at least one atom exposed to the solvent, (3) solvent accessible residues lining on the binding site cavity surface; (4) charged residues located on the binding site mouth edge.

2.1. Solvent accessibility filter

Solvent accessible surface areas have been introduced by Lee & Richards.[4] There are computed by rolling a virtual probe on the Van der Waals surface of the molecular structure (**figure 5**). Typically the probe has a radius that simulates a water molecule

(1.4 Å by default) and therefore the calculated contact surface can be considered as the solvent accessible surface.

In a first attempt, we used M.L Connolly's molecular surface package[5] with the command msroll. The program was not stable enough for very large macromolecules (over 10 000 atoms). For this reason we decided to choose a more recent program, namely naccessV2.1.1.[6] The program naccess takes as input a molecular structure file in PDB format and returns a structure file containing an additional field: the computed atomic solvent accessible surface area ($Å^2$). Original PDB files corresponding to the sc-PDB entries were taken from the protein databank (PDB) repository.[7] However, before solvent accessibility computation, protein structure files were filtered as they may contain various molecules displacing the solvent. I programmed the following parsers in the Perl module script.

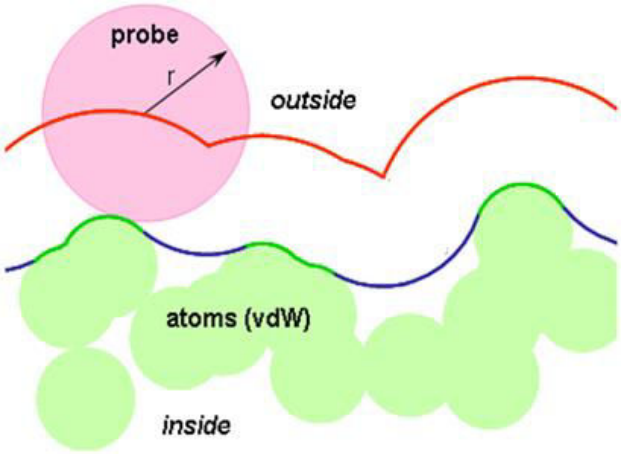


Figure 5. Solvent accessible surface definition.
Light green circles represent Von der Waals surfaces of atoms that occupy the binding site. The outside environment is located at the top of the green circles. The pink circle indicates the probe which is rolled on atomic surface. As the probe rolls on the atomic surfaces, its center traces the solvent accessible surface (red line). The contact surface of the probe with atomic surfaces is indicated by the green lines whereas blue lines represent reentrant surfaces.

Unwanted molecule filter

In raw PDB files, protein chains are not the only molecular objects. Solvent molecules and non-covalently bound hetero-atom groups such as co-factors, prosthetic groups or ligands. These molecules were systematically removed. In order identify unwanted hetero atom groups, the Perl module requires the list of unwanted HET codes used in sc-PDB (v.2010). Each unwanted hetero-atom group was systematically inspected for eventual covalent bonds with the protein. If a hetero-atom was bound, then we kept it in the structure. Ligands were identified according to their HET codes provided by an sc-PDB annotation file.

Alternate position filter

Proteins are dynamic objects and can adopt many different conformations. Some crystallographic structures describe multiple alternate positions of residue side chains. We selected the alternate position that is the most populated only by systematically checking occupancy factors. If a protein structure was solved by NMR experimental method, only the first model was selected.

Protein chain extractor

Protein chain(s) of all residues in binding site were identified from atom lines of the PDB file. Any protein chain not involved in the formation of the binding site was removed. Binding site residue were considered on the basis of sc-PDB binding sites.

Metal atom extractor

At the difference to binding sites in the sc-PDB, divalent metal ions were kept within binding sites (we plan to consider them in binding site comparison). We have considered biologically relevant metal ions only (Ca, Fe, Zn, Mg, Mn, Co, Gd).

Naccess input file writer

After all the previous filtering steps, a final parser writes all selected molecules in an updated PDB file used for input in naccess.

Atomic solvent accessibility inserter

In MOL2 files, solvent accessibility values can be stored as the 8th element of an atom line (usually used for atomic charge description). Thus, once computed, each atomic solvent accessible surface area value was extracted from naccess output file and used for re-insertion into the protein MOL2 file. Since residue numbering in sc-PDB MOL2 files did not follow the residue numbering scheme of the PDB, I had to identify residues in MOL2 files considering spatial coordinates of Cα atoms to match with their corresponding ones in naccess output file. Ultimately, residue atoms were tagged with either the accessible surface area value or with a negative value if inaccessible, thus "switching" them off during binding site comparison. An illustration of the resulting binding site is represented in the **figure 6**.

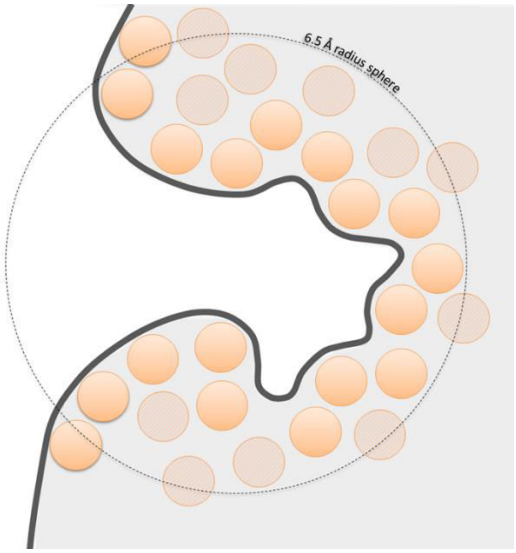


Figure 6. Solvent accessibility filtering of binding site residues.
The original binding site is represented by orange discs, all contained within the 6.5Å distance cutoff from the ligand (dotted line). Solvent accessible residues of the binding site are represented by orange circles whereas inaccessible residues are represented by shaded circles.

2.2. Binding site delimitation

At this stage, we "switched off" residues without any solvent accessible atoms. However, even with this definition some residues are still "switched on" even if they are irrelevant for the characterization of molecular recognition points because too remote from binding site cavity. It is mainly the case of residues gaining solvent exposure from the surface of the protein, outside of the binding site cavity. In order to delimit the binding site we calculated binding cavities using VolSite.[8]

Briefly, VolSite uses a lattice containing regular cells (by default they are 1.5Å wide). A cell is defined "in the protein" if any protein atom is less than 2.5 Å away from the cell's center. All remaining cells are investigated for buriedness by inspecting 120 different directions around them. If more than 40 directions intersect an "in the protein" cell, then the cell under investigation is considered within the binding site cavity. At the end, a grid of points derived from the cell centers describes the cavity. The interesting thing about VolSite cavities is that they are a good mean for us to

define the binding site boundaries. Together with customized VolSite parameters (cell sizes = 0.8, burriedness threshold=40, 6Å truncation from ligand heavy atoms), we used an empirical distance cutoff of 4Å to identify solvent accessible atoms beyond the boundaries of the cavity (**figure 7)**. Thereby, we "switched off" the atoms of residues when they are too remote by tagging them with a negative value instead of their solvent accessible surface area.

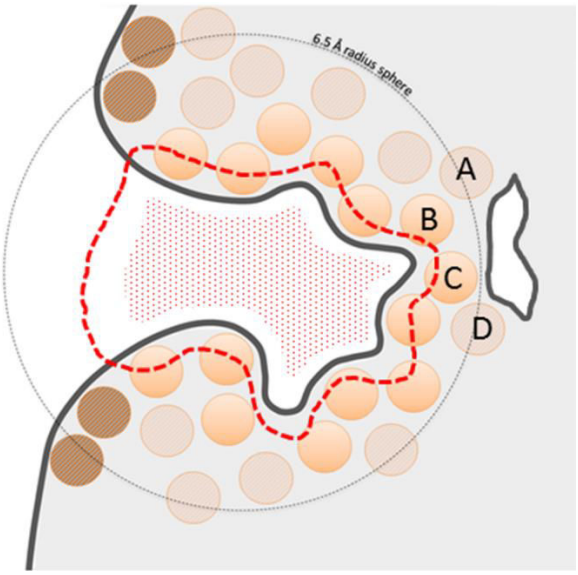


Figure 7. Binding site delimitation based on VolSite cavity grid points.
Red dots represent VolSite cavity. The red dashed line represents the 4Å distance cutoff delimiting the solvent accessible binding site. A small adjacent pocket is represented to illustrate why this delimitation was set. The dark shaded circles represent solvent accessible residues located on the external surface of the binding site. Residue A and D are exposed to solvent because they are in the proximity of an adjacent small pocket.

2.3. Binding site mouth detection

Considering that the strength of the electrostatic interaction depends on the polarity of the environment of interacting atoms, residues located on the periphery of a ligand binding site are less important than residues deeply buried into the cavity. We used a polyhedron (identic to SiteAlign's) placed at the center of binding sites to detect the mouth of the binding site. Residue projections onto the sphere were coded in the Perl

module similarly to SiteAlign. The idea is that, after residue projections, an empty region (without any residue projection) on the polyhedron faces the binding site mouth (**figure 8**). However, single empty triangles can face buried part of the binding site. In order to fill empty triangles facing the buried binding site and thus to facilitate the identification of the binding site mouth, any residue from the 6.5Å binding site was projected onto the sphere (even those that are inaccessible to the solvent). We defined the largest empty zone (containing several triangles) as a marker to locate the mouth of the site. Indeed, residues of the binding site mouth generally project on triangles sharing two vertices with a triangle of the largest empty zone (**figure 8**). Any charged residue projected onto a triangle directly surrounding the largest empty region was "switched off" with a special value in the MOL2 protein file.

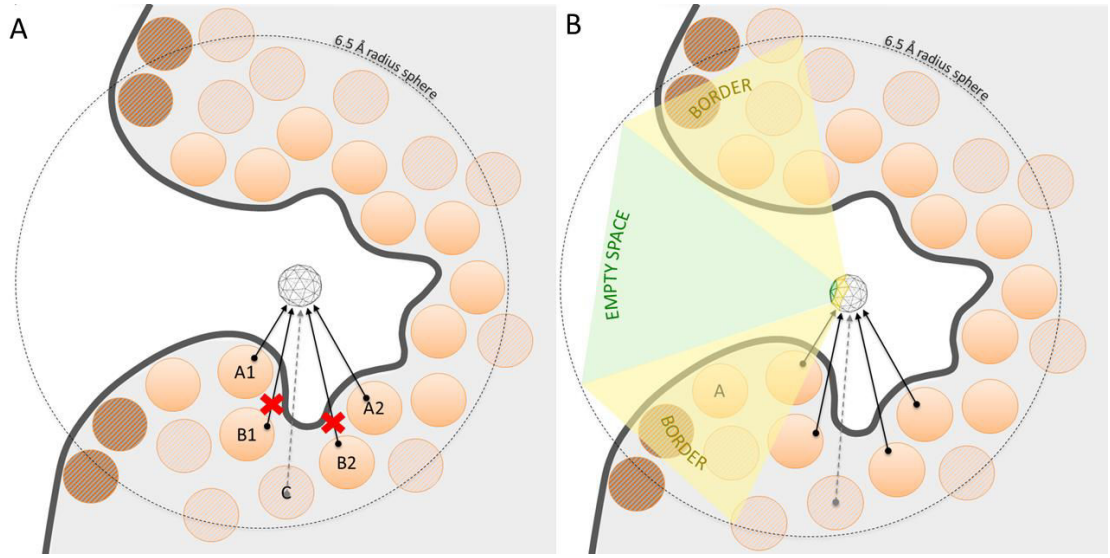


Figure 8. Identification of binding site mouth.
A: This figure illustrates why inaccessible residues were included onto the polyhedron. Residues B1 and B2 are masked by residue A1 and A2 and therefore they are not projected onto the polyhedron, leaving an empty triangle. The residue C was projected onto the polyhedron to fill the empty triangle. B: The green surface faces the largest empty group of triangles onto the polyhedron. The yellow surface faces triangles directly surrounding the largest empty region.

2.4. Analysis of solvent accessible binding sites

As shown in **figure 9**, the number of residues in sc-PDB sites decreased when considering solvent exposure and distance to binding cavity cutoff. The solvent accessibility filter has discarded 4 residues on average (43 residues in sc-PDB sites, 39 in solvent accessible sites), indicating that sites defined by the 6.5Å distance cutoff from the ligand are mainly composed of solvent accessible residues. However, the cavity delimitation has reduced the number of residues more dramatically indicating that the sc-PDB binding sites contain an average of 10 residues located at more than 4Å from any cavity points in the considered VolSite cavity. The average number of residues in our final representation of binding sites is 33. Lastly, according to the **figure 9**, binding site definition based on VolSite cavities is more constant, since the standard deviation in the size of final binding sites is lower than for sc-PDB sites.

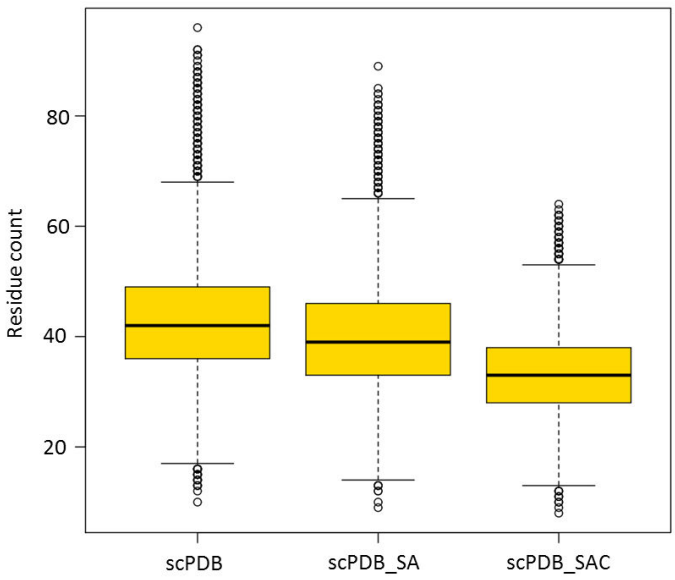


Figure 9. Residues counts in sites of the sc-PDB according to different definition.
ScPDB: original binding site of the sc-PDB. scPDB_SA: solvent accessible binding sites. scPDB_SAC: solvent accessible binding sites, after cavity delimitation. Distribution was generated for the 9877 entries of the sc-PDB. Yellow boxes represent 50% of the binding sites. Vertical dashed lines represent first and third quartiles. Circles represent outliers.

We addressed the question: is solvent exposure a good indicator to identify atoms providing potential molecular recognition points? Therefore, we detected protein interacting atoms using an early version of IChem.[9] Basically, a set of chemico-geometrical rules scans protein-ligand complexes and detects protein-ligand atom pairs susceptible to interact. The **figure 10** shows that about 50% of interacting atoms have solvent accessible surface area between 5 and 15 $Å^2$ (except metal). When compared to solvent accessible surface area of all accessible atoms, one can clearly see that interacting atoms are generally more exposed. Thereby, we can say that our modified definition of binding sites contains higher proportion of interacting atoms and that it is likely to embed a better representation of biological imprints. However, a few interacting atoms are inaccessible to the solvent. This is due to the fact that crystal structures are not evenly accurate and that solvent accessibility is highly dependent on atomic coordinates.

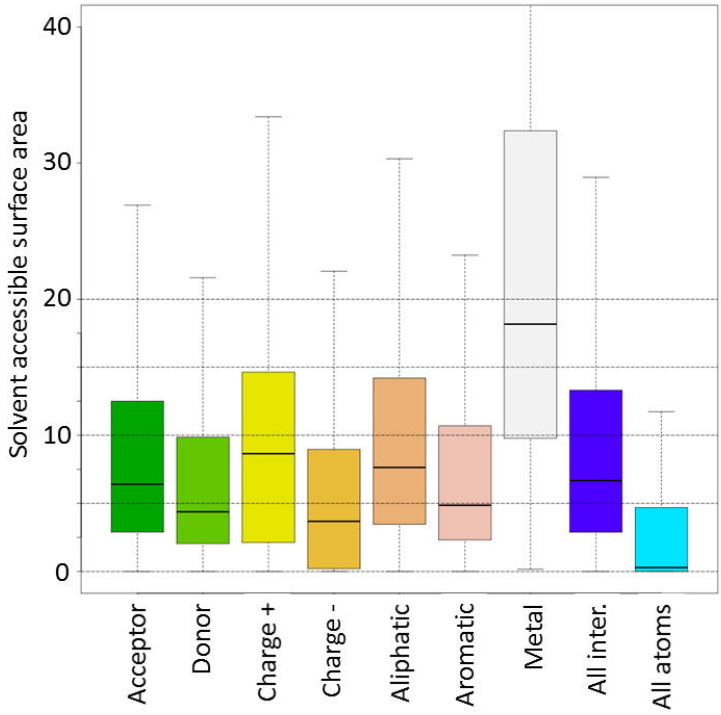


Figure 10. Atomic solvent accessible areas per interaction type.
All entries in sc-PDB were considered. Boxes indicate the range of values for 50% of the detected interactions. Horizontal lines in boxes represent the median values whereas vertical dashed lines represent the third and first quartiles. Acceptor: H-bond acceptor atoms. Donor: H-bond donor atoms. Charge+: positively charged atoms in ionic bond. Charge-: negatively charged atom in ionic bond. Aliphatic: carbon atom in hydrophobic contact. Aromatic: aromatic atom in π-stacking. Metal: divalent metal ion in ionic interaction. All inter.: all interacting atoms regardless of the interaction they do. All atoms: all atoms in binding sites. Solvent surface area are expressed in $Å^2$.

3. MODIFICATION OF SITEALIGN

We have defined a representation of binding sites embedding potential molecular recognition points contributing to biological imprints in biosynthetic enzymes. Nevertheless, the original version of SiteAlign is not able to interpret the atomic tags that we have inserted in the protein MOL2 file. Therefore, we tuned SiteAlign's source code to consider the previously encoded solvent exposure information. In essence, we modified the format of the fingerprints and the methods that add the descriptors to fingerprints on the polyhedron. For a given residue type, the set of physico-chemical descriptor was given the ability to vary depending on atomic solvent

exposure. In addition, the point at the origin of residue projection was shifted towards solvent exposed atoms in order to represent molecular recognition points.

3.1. Physico-chemical descriptors

The physico-chemical descriptors used in the original version of SiteAlign represent the commonly used pharmacophoric features necessary to describe the binding-mode of a ligand to a protein. Hence, we kept them all. However, we added a new descriptor for atoms bearing the features H-bond acceptor and donor at the same time. In SiteAlign, each residue type is represented by an invariant set of descriptors encoding the pharmacophoric features of the residue side chains. We gave to descriptors the possibility to encode the polar features of protein backbones and added a conceptual sense to the information descriptors carry. Our descriptors exclusively represent solvent accessible pharmacophoric features. As shown in **figure 1** of **annex 1,** each atom bearing a pharmacophoric feature was assigned (a) particular descriptor(s). Descriptors of polar interactions (H-bond acceptor, H-bond donor and H-bond donor/acceptor) represent the count of solvent accessible atoms providing the interaction. Aromatic, charge and aliphatic descriptors are defined with an integer that is “switched on” (value different to 0) when at least one atom bearing the feature of interest in the residue is accessible to solvent. Values of residue descriptors are listed in **table 4**. It is noteworthy to mention that we also consider metal ions in binding sites.

	Aliphatic	Donor	Acceptor	AD	Aromatic	Charge
Ala	{0 , 1}	[0 - 1]	[0 - 1]			
Arg	{0 , 2}	[0 - 4]	[0 - 1]			{0 , +1}
Asn	{0 , 1}	[0 - 2]	[0 - 2]			
Asp		[0 - 1]	[0 - 3]			{-1 , 0}
Cys	{0 , 1}	[0 - 1]	[0 - 1]			
Glu	{0 , 1}	[0 - 1]	[0 - 3]			{-1, 0}
Gln	{0 , 1}	[0 - 2]	[0 - 2]			
Gly		[0 - 1]	[0 - 1]			
His	{0 , 1}	[0 - 3]	[0 - 3]		{0 , 1}	
Hip	{0 , 1}	[0 - 3]	[0 - 1]		{0 , 1}	{0 , +1}
Ile	{0 , 3}	[0 - 1]	[0 - 1]			
Leu	{0 , 3}	[0 - 1]	[0 - 1]			
Lys	{0 , 2}	[0 - 2]	[0 - 1]			{0 , +1}
Met	{0 , 2}	[0 - 1]	[0 - 1]			
Phe	{0 , 2}	[0 - 1]	[0 - 1]		{0 , 2}	
Pro	{0 , 3}	[0 - 1]	[0 - 1]			
Ser		[0 - 2]	[0 - 2]	[0 - 1]		
Thr	{0 , 1}	[0 - 2]	[0 - 2]	[0 - 1]		
Trp	{0 , 2}	[0 - 2]	[0 - 1]		{0 , 2}	
Tyr	{0 , 2}	[0 - 2]	[0 - 2]	[0 - 1]	{0 , 2}	
Val	{0 , 2}	[0 - 1]	[0 - 1]			
Metal						{0 , +2}
amplitudes	*[0 - 3]*	*[0 - 4]*	*[0 - 3]*	*[0 - 1]*	*[0 - 2]*	*[-1 - 2]*

Table 4. Possible values of residue descriptors.
AD: atom providing H-bond acceptor/donor interaction. Descriptors between curly brackets only take the specified values. Descriptors between square brackets can take any integer value within the specified interval.

3.2. Topological descriptor

As we aim at representing binding sites by potential molecular recognition points, size and orientation topological descriptors are not relevant any more. Hence we discarded them. However, the distance to the center of the polyhedron is still an important descriptor necessary for the representation of the spatial arrangement of residues in binding sites. In SiteAlign, residues are projected onto the polyhedron from their Cβ but this point can be distant to the interacting atom of a residue, especially when residues have large side chains. Therefore, we shifted the origin of the projection to the geometrical center of solvent accessible atoms in the residue.

3.3. Residue fingerprint

Our set of descriptors contains six physico-chemical descriptors and one topological descriptor. Accordingly, we adapted SiteAlign fingerprints. In SiteAlign, fingerprints are filled with the set of predefined descriptors depending on the residue nature only. Our, physico-chemical descriptors are added into fingerprints if atoms bearing the features are accessible to the solvent and "switched on" only. The **table 5** illustrates the format of "solvent accessible" fingerprints.

0	1	2	3	4	5	6
Aliphatic	Donor	Acceptor	Donor/Acceptor	Aromatic	Distance	Charge

Table 5. Solvent accessible fingerprint of a residue.
Blue boxes represent physico-chemical descriptors. The green box represents a topological descriptor. The first line represents the index of each descriptor in the fingerprint array.

3.4. Scoring function

The scoring functions used in SiteAlign are well suited for the comparison of the fingerprints. Since our fingerprints resemble the original fingerprints, we based our scoring functions on SiteAlign's. The major part of the scoring function was not modified, except coefficients used for normalization as the amplitudes of descriptors are different. Similarly to other descriptors, we incorporated a scoring function for the newly added acceptor/donor descriptor.

In SiteAlign, each descriptor has an equivalent contribution in the final score. In our case, the introduction of a third polar descriptor has prompted us to introduce weight coefficients in order to modulate each descriptor contribution. Therefore, we are able

to counter-balance the contribution of polar features with apolar features. The coefficients are incorporated into the scoring function that computes the fingerprint score using the following formula:

$$S_t = \frac{\sum_{i=1}^{7} w_i . S_t^i}{\sum_{i=1}^{7} w_i}$$

Where,

S_t is the score between two matched triangles.
w_i is the weight coefficient of *i*-th descriptor in the fingerprint.
S_t^i is the score between the *i*-th compared descriptors.

3.5. Characterization of modifications impacting binding site comparison

3.5.1. Lengthways modification of the topological descriptor

Modification of the topological descriptor intend to shift the points that represent residues towards the solvent accessible surface. In order to characterize how the modification affects the topological descriptor, we have compared position of Cβs to position of "accessible centers" (geometrical center of atoms tagged with solvent accessible surface area value). Distances were computed using the initial position of the polyhedron in solvent accessible binding site (center of residue Cα). As shown in **figure 11,** "accessible centers" are closer to the center of binding sites for nearly 60% of the considered residues, which supports our expectation. However, solvent accessible binding sites (containing 33 residues on average) have a considerable number of residues with "accessible centers" more distant than the Cβ of the corresponding residue. In this category, the most different distances (yellow points under the diagonal) suggest the presence of hydrophilic side chains pointing outwards

the cavity and thus, exposed to solvent from the protein surface outside of the binding cavity (yet still in the 4Å cutoff from VolSite cavity points). This striking fact highlights the limit of "accessible centers" to represent molecular recognition points. In fact, when a residue side chain points outwards the cavity, the backbone atoms are pointing towards the cavity. Thereby, polar backbone atoms are most susceptible to provide molecular recognition points and should be the representative ones to project the residue. Unfortunately, our modification projects these residues onto the polyhedron from a point that is worst that the Cβ. However, as illustrated by the purple points on **figure 11**, the vast majority of the residues place "accessible centers" and Cβs at a very similar distance to the center of the cavity. Thereby, the vast majority of the topological distances have an insignificant but existing impact on the binding site alignment.

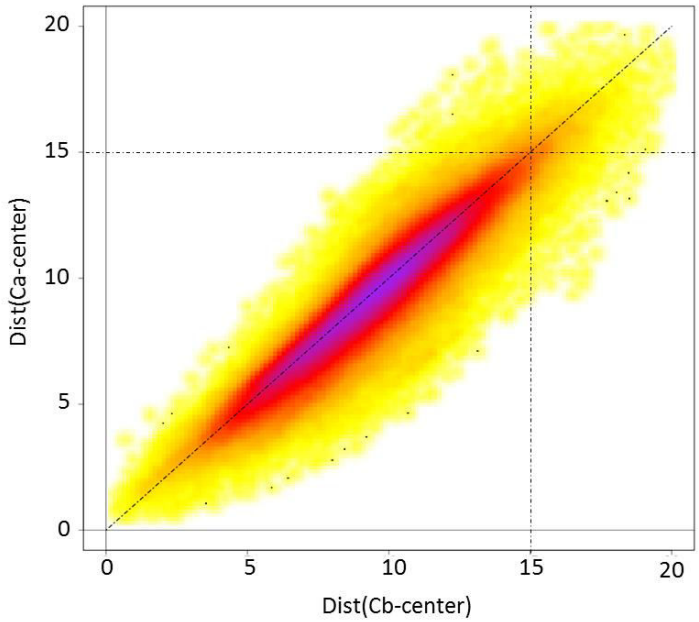


Figure 11. Variation of distance between Cβs and accessible center of residues for all 9877 binding sites in the sc-PDB. Dist(Ca-center): distance between accessible center of residues to center of the discretized sphere. Dist(Cb-center): distance between Cβ and center of discretized sphere. Each point of the plot represents one residue. The color codes for point density as follows: yellow < orange < red < purple.

3.5.2. Transversal modification of the topological descriptor

Displacement of the origin of residue projections affects the topological descriptor lengthways, but it also affects the triangles onto which residues are projected. In order to characterize the impact of the displacement, we measured how frequently each residue type was assigned a different triangle when projected from Cβ and from "accessible center". We considered solvent accessible residues in all binding sites of the sc-PDB. Not surprisingly, the residues that are most often projected onto different triangles are the largest residues. At least 50% of Trp, Tyr, Phe and Arg are assigned a different triangle (**figure 12**). Other residues project onto different triangles in about 30% to 40% of their respective frequencies. Thereby, we can assume that the displacement of the projection will have a significant impact on the binding site comparison. It is interesting to see that smaller residues, such as Ala, Gly of Pro are assigned different triangles in 30% of their relative population. Since residues with small side chains do not have many possibilities to displace the origin of the projection, we can assume that the newly assigned triangle of small residues are more representative of protein backbone atoms.

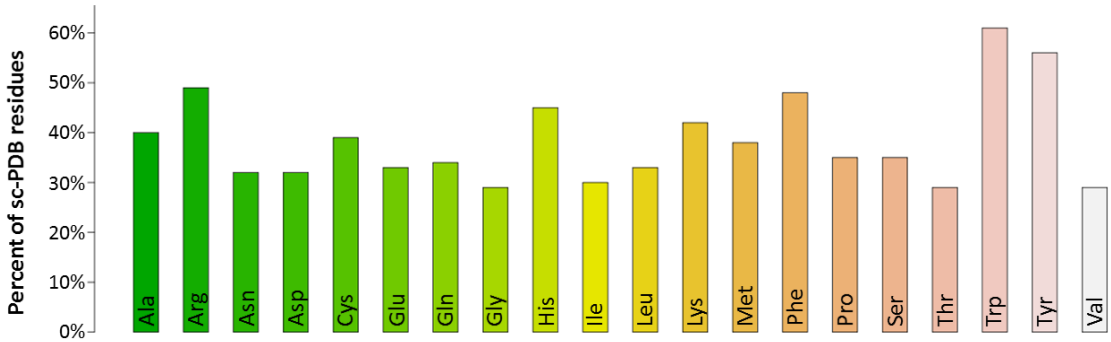


Figure 12. Percentage of different triangle assignment per residue type.
Residue percentage is relative to each residue type.

3.5.3. Fingerprints contents

Alongside with the topological descriptor modifications, new definition of physico-chemical descriptors is also prone to affect comparison outcomes because of differences in content. In order to characterize how physico-chemical descriptors vary upon modification, we measured the presence of our descriptors in “solvent accessible” fingerprints for all accessible residues in the sc-PDB entries. It turns out that all fingerprints in the original version of SiteAlign are often inaccurate to describe potential molecular recognition points. As shown in **figure 13**, for each residue type, there is at least one solvent accessible atom carrying a pharmacophoric feature that is not encoded in original SiteAlign fingerprints. For example, about 75% of the solvent accessible Gly residues expose donor or acceptor features to the solvent although Gly fingerprint is empty in SiteAlign. This observation is similar for 15 out of the 20 residue types, which definitively demonstrates that our fingerprints will have a significant impact on binding site comparison. Moreover, we assigned an aliphatic descriptor to not less than about 70% of the solvent accessible residues whose side chain contain high proportion of carbon atoms (Arg, Glu, Gln, His, Lys, Phe, Trp and Tyr) whereas these same residues have null hydrophobic descriptors is SiteAlign. In addition, about 75% of Ser, Thr and Tyr expose an atom providing donor/acceptor pharmacophoric feature, which suggests that the oxygen of hydroxyl groups is not always accessible to the solvent even though the original fingerprints always contain donor and acceptor descriptors. Lastly, even if the descriptors in the original fingerprints are at least present in 60% of our fingerprints, we can assume that our fingerprints are more representative of potential molecular recognition points, especially if polar atoms of

the protein backbone are exposed to the solvent. However, at this stage we did not know if the overall impact of our modifications was positive of negative on binding site comparisons.

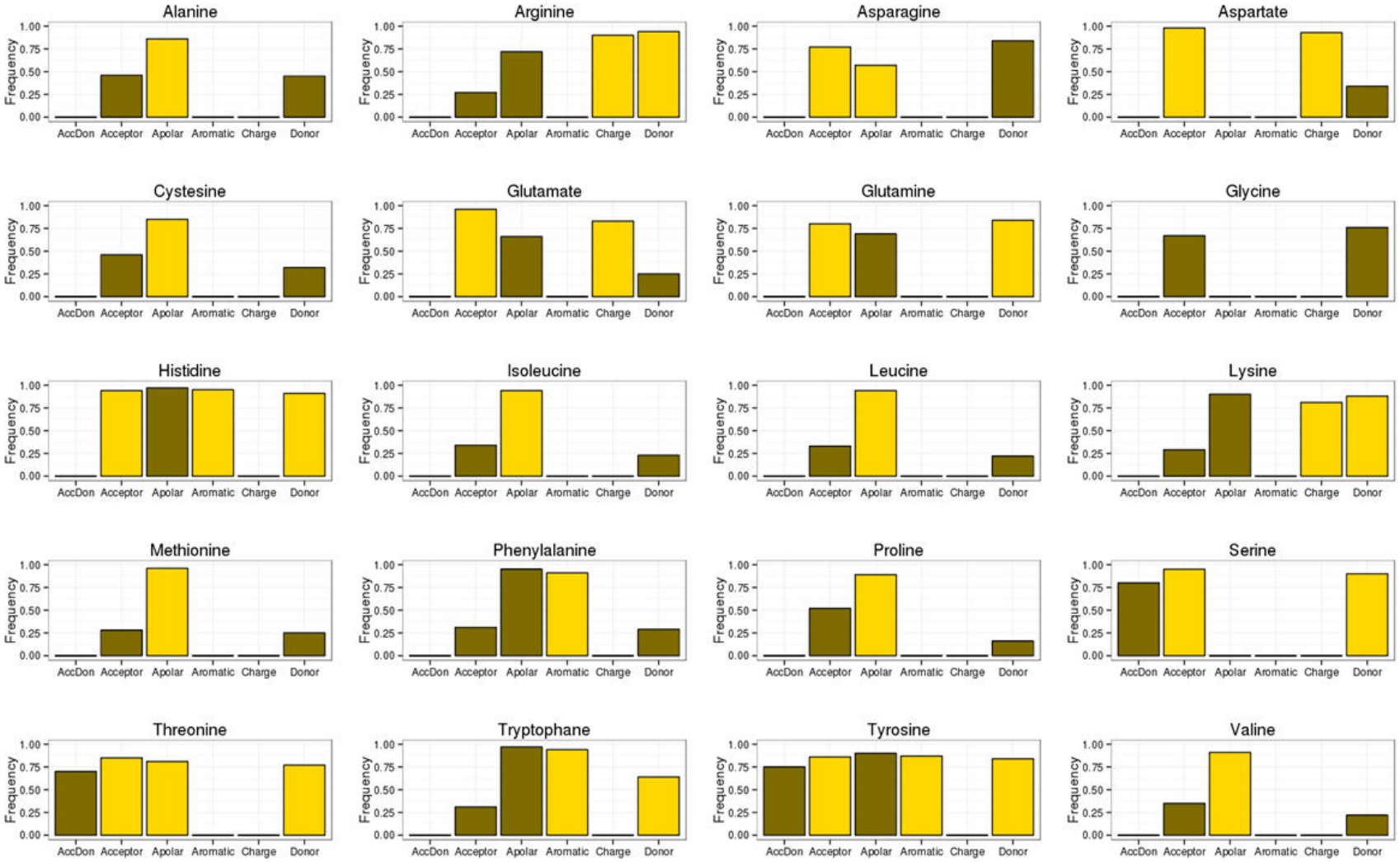


Figure 13. Physico-chemical properties of solvent accessible atoms in the modified representation of the binding site. Bars represent frequencies of each descriptor calculated by considering 5584 sc-PDB sites randomly chosen. Descriptor bars are shown in the following order. AccDon: atom providing acceptor/donor H-bond interaction. Acceptor: atom providing acceptor H-bond interaction. Apolar: carbon atom providing hydrophobic contacts. Charge: charged atom providing ionic interaction. Donor: atom providing H-bond donor interaction. Yellow bars represent descriptors of SiteAlign whereas dark yellow bars represent solvent accessible descriptors.

4. BENCHMARKING VERSIONS OF SITEALIGN

We modified inputs of SiteAlign in order to define binding sites more representative of potential molecular recognition points. Alongside with these modifications, we adapted SiteAlign to our newly defined binding sites. We gave evidence that our definition of binding sites and the adapted set of descriptors have an impact on binding site comparisons. However, this evidence did not tell us if our modifications have a benefic or a negative impact. In this section, we will focus on the comparison of the original version of SiteAlign, from now called SiteAlign-4, and our modified version, from now called SiteAlign-5. We tested two versions of SiteAling-5 (5.1 and 5.2). In SiteAlign-5.1, solvent accessible residues lining the cavity surface are considered only whereas in SiteAlign-5.2, binding site comparison is computed including the contribution of the topological descriptor (distance of the residue to the center of the polyhedron) of buried residues present in original sites of the sc-PDB. Following tests were performed for each version of SiteAlign.

4.1. Similarity threshold definition

Before virtual screening experiments, we defined SiteAlign-5 similarity thresholds, required to discriminate similar from dissimilar sites. To that end, we used a training set defined in an earlier study.[10] The training set is composed of 1336 pairs of binding sites. Out of them, 649 pairs are assumed to be dissimilar whereas the 687 others are assumed similar. Similar binding sites (with different co-crystalized ligands) have been chosen amongst proteins sharing same UniProt name[11] and were predicted similar using SiteAlign. Dissimilar sites where randomly chosen by ensuring different first level of EC

numbers[12] and SiteAlign dissimilar prediction. Solvent accessible binding sites of the 1336 pairs were prepared according to the previously described method. Binding site similarity was computed for each pair of the training set using SiteAlign-5. In order to determine the consensus score (D1 and D2) that best discriminates similar from dissimilar sites, we tested 10 000 classification models by varying systematically D1 and D2 with increments of 0.01. Each classification model was evaluated using the following F-measure:

$$F_{measure} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Where

TP = True Positive: the count of similar sites correctly classified.

FP = False Positive: the count of dissimilar sites incorrectly classified.

FN = False Negative: the count of similar sites incorrectly classified.

Precision is a coefficient that characterizes the predictive value of classification models (value between 0 and 1). *Recall* represents the ratio between the count of similar sites correctly classified and the total number of similar sites. Precision and recall are combined together into the F-measure in order to characterize the tradeoff between precision and recall in the classification model. Basically, the higher the F-measure is, the better the tradeoff is between precision and recall. The best tradeoff was found for the consensus threshold D1<0.59 AND D2<0.17, the predictive value of the model being 0.97, with 88% of the similar sites correctly classified. It is not fair to compare F-measure

outcomes between SiteAlgn-4 and SiteAlign-5 because the training set was made using SiteAlign. However, qualitative assessment shows that SiteAlign-5 and SiteAlign-4 scoring methods are different. As shown in **figure 14**, there is larger overlap between scores of similar and dissimilar sites when using SiteAlign-5. As seen on the figure, a large proportion of the "dissimilar" sites pass the defined threshold values, suggesting that there are some similarity within "dissimilar" sites that was not captured by SiteAlign-4.

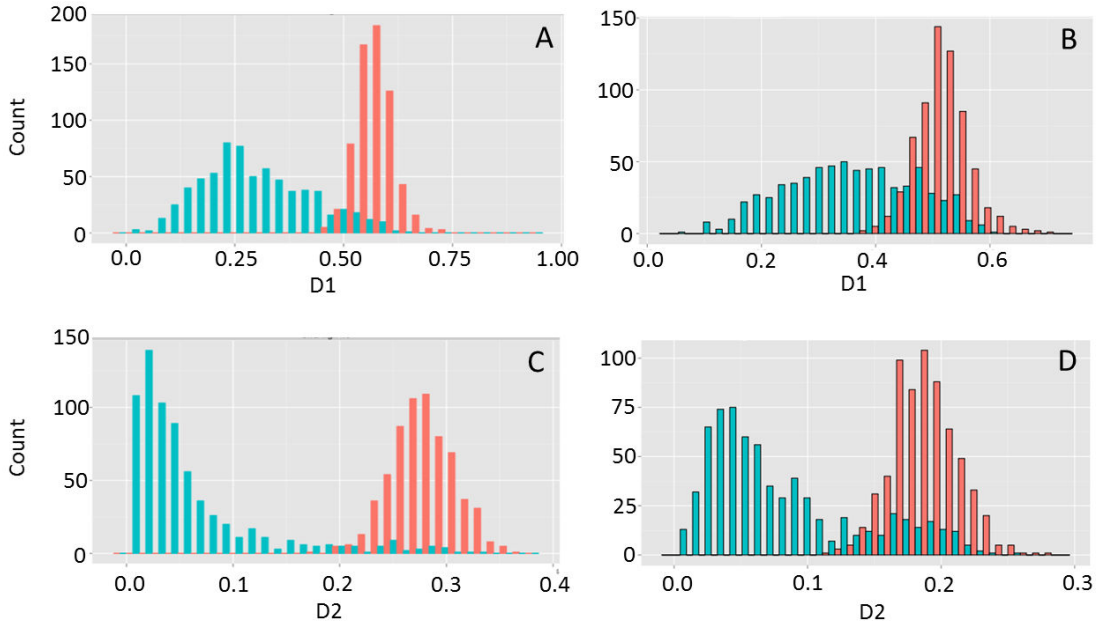


Figure 14. Distributions of distance scores D1 and D2 for SiteAlign-4 and SiteAlign-5.1
A: distribution of D1 using SiteAlign-4. B: distribution of D1 using SiteAlign-5.1, C: distribution of D2 using SiteAlign-4. D: distribution of D2 using SiteAlign-5.1. Blue bars correspond to similar binding site pairs whereas red bars represent dissimilar binding site pairs.

4.2. Testing SiteAlign-5 against SiteAlign-4

In order to decipher if our modification of SiteAlign have a benefic or a negative impact on virtual screening outcomes, we performed two experiments. Focus was given to statistical evaluation, therewith characterizing positive or negative effects. In that, we compared prototypical binding sites of a serine protease protein and of a kinase protein to all entries in the sc-PDB. We then analyzed SiteAlign-5's classification regarding serine

and kinase protein families and other protein families known to recognize compounds interacting with serines or kinases respectively. For binding site comparison using SiteAlign-5, all solvent accessible binding sites were prepared from all entries in the sc-PDB according to the previously described method. Experiments were repeated independently with SiteAlign-4 and SiteAlign-5 variants.

4.2.1. Binding site similarity across different fold families

The first experiment is a diagnostic of SiteAlign-5's prediction across the serine protease family inspired by earlier benchmarking studies.[1,13,14] Serine proteases are interesting for binding site comparison tests because their inhibitors exhibit a broad specificity for different fold types[15] and substrate cleavages[16] and thus, they are all true positives when compared to a prototypical binding site. We classified entries in the sc-PDB (v.2010) according to four categories of folds and substrate cleavage. The first category (270 entries) represents trypsin-like folds and trypsin specific substrate cleavage. The second category (15 entries) represents trypsin-like folds but with substrate cleavage different to that of trypsin. The third category (14 entries) represents subtilisin-like folds. The fourth category (5 entries) represents α/β hydrolase folds. The last category is composed of the 5284 remaining entries. We used a prototypical binding site in bovine trypsin (PDB ID: 1AQ7) as query for comparison with sc-PDB entries. Sensitivity and specificity of screenings were evaluated by computing a Receiver-Operating Characteristic (ROC) plot[17] specific to each fold category. An area under the ROC curve (ROCAUC) higher than 0.5 indicates sensitivity and specificity of the scoring method. The higher the value is over 0.5, the better the performances are. As opposed, a ROCAUC equal to 0.5 indicates no sensitivity/specificity of the scoring method (random

selection). We computed each ROC curve considering proteins passing D1 threshold only and trimmed protein lists to obtain equivalent number of proteins in the screening outcomes of the two methods.

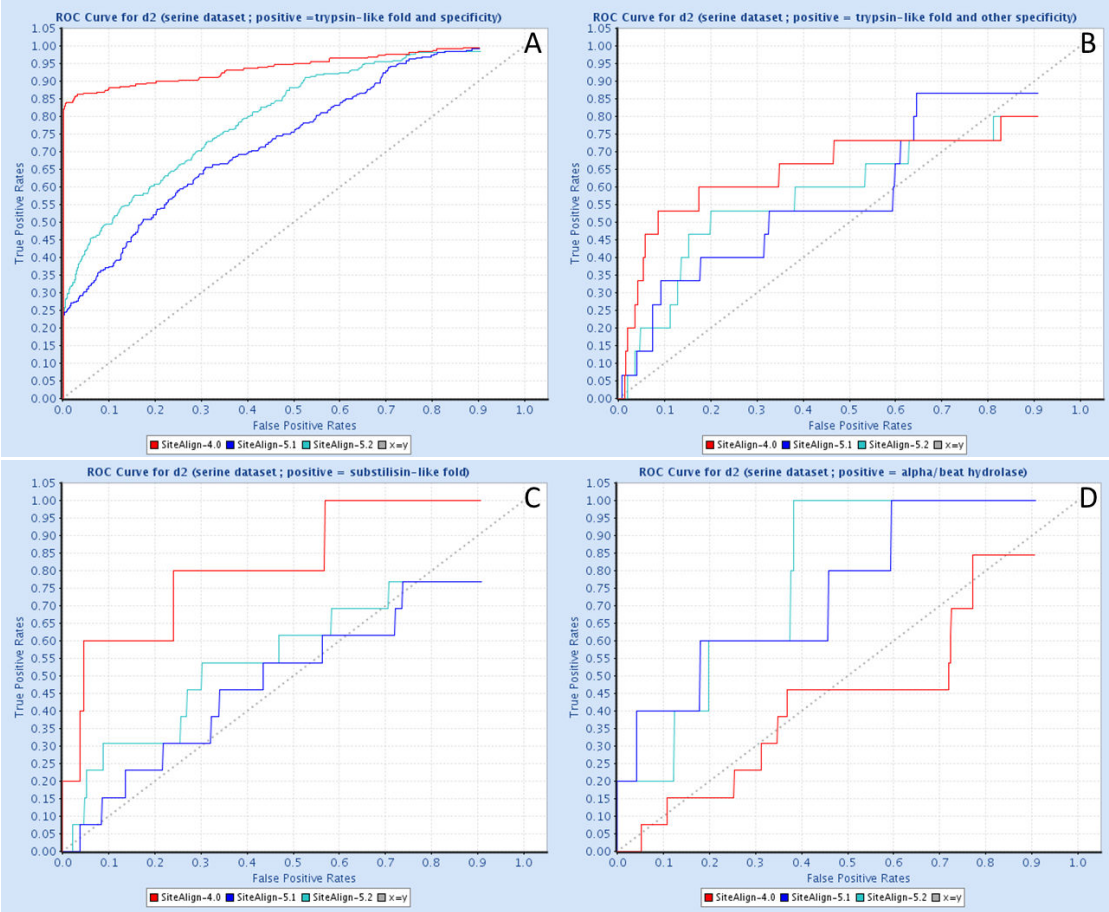


Figure 15. ROC curves of different SiteAlign versions for serine protease screening tests.
A: true positive proteins exhibit trypsin-like folds with trypsin substrate specificity. B: true positive proteins exhibit trypsin-like folds with substrate cleavage different to that of trypsin. C: true positive proteins exhibit subtilisin-like folds. D: true positive proteins exhibit α/β-hydrolase folds. Red curves represent SiteAlign4. Blue curves represent SiteAlign5.1. Light-blue curves represent SiteAlign5.2. The diagonal dotted line represent the random classification.

As shown in **figure 15A**, ROCAUCs indicate that SiteAlign-5 performance relative to trypsin-like fold and trypsin substrate cleavage is just acceptable (ROCAUC near 0.7) compared to SiteAlign-4 (ROCAUC over 0.8). However, as indicated by the steep early slope, SiteAlign-5 predicted proteins of the first category (trypsin-like fold, trypsin substrate cleavage) with the highest similarity scores even though the background noise

is affecting results quicker than using SiteAlign-4 (**figure 2** of **annex 1**). A reasonable explanation for SiteAlign-5 failure resides in its sensitivity to small conformational changes. As shown in **figure 15B**, SiteAlign-5's performance relative to proteins exhibiting trypsin-like folds and substrate cleavage different to that of trypsin are lower than SiteAlign-4. However, given that fact that the query belongs to the first category (trypsin-like fold, trypsin substrate cleavage), variations of physico-chemical properties are expected when compared to the query, which was indeed captured by SiteAlign-5. In **figure 15C**, SiteAlign-5's predictions are comparable to a random selection (ROCAUC near 0.5), indicating that SiteAlign-5 was not able to capture similar molecular recognition points in trypsin and subtilisin-like folds. Lastly, the **figure 15D** indicates that SiteAlign-5 screenings outcome was the most enriched in proteins exhibiting α/β hydrolase folds, thereby suggesting that inhibitor recognition might be induced by similar molecular recognition points in α/β hydrolase folds and trypsin substrate cleavage.

4.2.2. Similarity of permissive ligand binding sites

The second experiment aimed at evaluating SiteAlign-5's prediction for very permissive ligand binding sites such as adenine tri-phosphate (ATP) recognition sites.[18] In that, we have classified proteins of the sc-PDB into four categories. The first category (510 entries) represents protein kinases (EC numbers 2.7.10.- , 2.7.11.- , 2.7.18.- , 2.7.13.- or 2.7.99.-). The second category (177 entries) represents ATP-binding sites of other miscellaneous kinase proteins. The third category (263 entries) represents non-kinase proteins co-crystalized with ATP/ADP ligands. The last category is composed of the

remaining entries of the sc-PDB. We used a prototypical ATP-binding site of protein kinase pim-1 (PDB ID: 1YHS) as query for screening the sc-PDB.

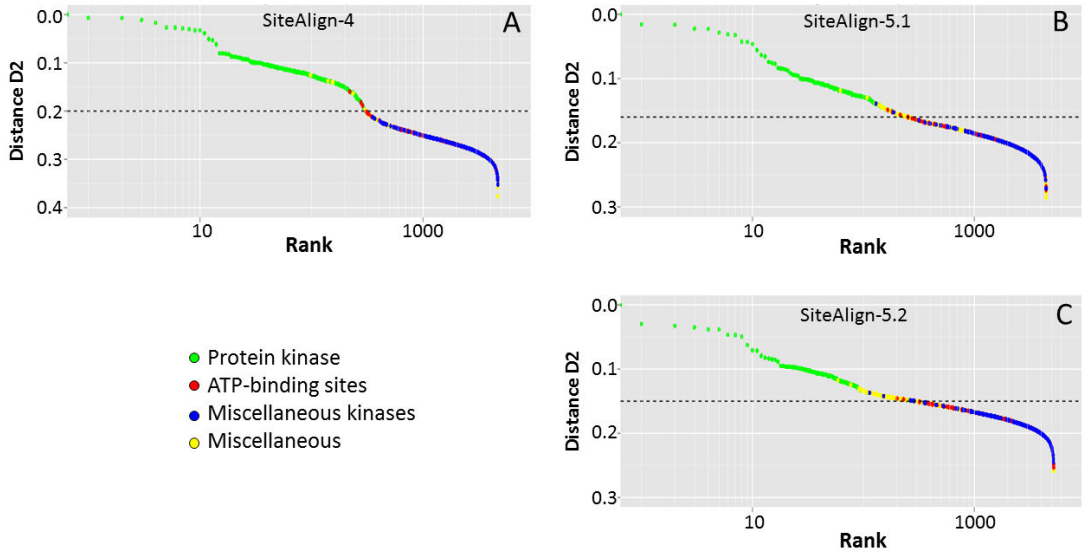


Figure 16. Rank plots of different SiteAlign versions for permissive ligand binding site screening test.
A: rank plot of SiteAlign-4. B: rank plot of SiteAlign-5.1. C: rank plot of SiteAlign-5.2. Ranks describe the position of the proteins in the list of screened proteins sorted by decreasing similarity scores. Dotted lines represent the D2 threshold value. Rank axis is represented by logarithmic scale. Color codes of points in the plots are given in legend. Because blue points are overlapping other points, they hide yellow, red and green points under D2 threshold values.

As shown in **figure 16**, the tested scoring methods all give the highest scores to protein kinase binding sites whereas miscellaneous kinase ATP-binding sites and ATP-binding sites of other proteins are generally ranked behind. This was expected given the variety of kinases, their flexibility and the promiscuity of ATP/ADP ligands. Nevertheless, SiteAlign-4 was the method that enriched the most protein kinases in the list of proteins predicted as similar. SiteAlign-5 only predicted about three to two times less protein kinases as similar. As mentioned above, the failure of SiteAlign-5 can be explained by small conformational changes of residue side chains, resulting in different sets of descriptors encoding solvent accessible pharmacophoric features. At the difference of SiteAlign-4, our modified version has predicted about 100 entries from the miscellaneous category as similar to the query, which clearly suggests the potential of our approach to detect remote similarities between unrelated proteins.

CONCLUSION

We have modified an existing 3D binding site comparison tool with the aim of capturing common molecular recognition patterns between proteins of unrelated folds. We have been able to define a binding site representation that incorporates information on potential molecular recognition points using atomic solvent accessibility and adapted SiteAlign to the new representation. The presented method appeared to be more detailed and less dependent to protein folds than SiteAlign-4, which suggests a potential to capture remote similarities. However, virtual screening tests have demonstrated that our method suffer from sensitivity to small conformational changes, thus making the exploitation of prospective virtual screening results difficult because relevant hits tend to get lost in background noise.

REFERENCES

1. Schalon, C., Surgand, J.-S., Kellenberger, E. & Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins Struct. Funct. Bioinforma.* **71,** 1755–1778 (2008).
2. McArdle, B. M., Campitelli, M. R. & Quinn, R. J. A common protein fold topology shared by flavonoid biosynthetic enzymes and therapeutic targets. *J. Nat. Prod.* **69,** 14–17 (2006).
3. Meslamani, J., Rognan, D. & Kellenberger, E. sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. *Bioinforma. Oxf. Engl.* **27,** 1324–1326 (2011).
4. Lee, B. & Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55,** 379–400 (1971).
5. Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* 709–713 (1983).
6. Hubbard, S. & Thornton, J. M. *NACCESS*. (Biochemistry and Molecular Biology, University College London, 1993).
7. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
8. Desaphy, J., Azdimousa, K., Kellenberger, E. & Rognan, D. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *J. Chem. Inf. Model.* **52,** 2287–2299 (2012).
9. Da Silva, F., Desaphy, J., Bret, G. & Rognan, D. IChemPIC: A Random Forest Classifier of Biological and Crystallographic Protein-Protein Interfaces. *J. Chem. Inf. Model.* **55,** 2005–2014 (2015).
10. Weill, N. & Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein−Ligand Binding Sites. *J. Chem. Inf. Model.* **50,** 123–135 (2010).
11. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43,** D204–D212 (2015).
12. Webb, E. C. *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. (Academic Press, 1992).
13. Shulman-Peleg, A., Nussinov, R. & Wolfson, H. J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **339,** 607–633 (2004).
14. Schmitt, S., Kuhn, D. & Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **323,** 387–406 (2002).
15. Bartoli, L., Calabrese, R., Fariselli, P., Mita, D. G. & Casadio, R. A computational approach for detecting peptidases and their specific inhibitors at the genome level. *BMC Bioinformatics* **8 Suppl 1,** S3 (2007).

16. Igarashi, Y. *et al.* CutDB: a proteolytic event database. *Nucleic Acids Res.* **35,** D546–549 (2007).

17. Triballeau, N., Acher, F., Brabet, I., Pin, J.-P. & Bertrand, H.-O. Virtual screening workflow development guided by the 'receiver operating characteristic' curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **48,** 2534–2547 (2005).

18. Kahraman, A., Morris, R. J., Laskowski, R. A. & Thornton, J. M. Shape variation in protein binding pockets and their ligands. *J. Mol. Biol.* **368,** 283–301 (2007).

ANNEX 1

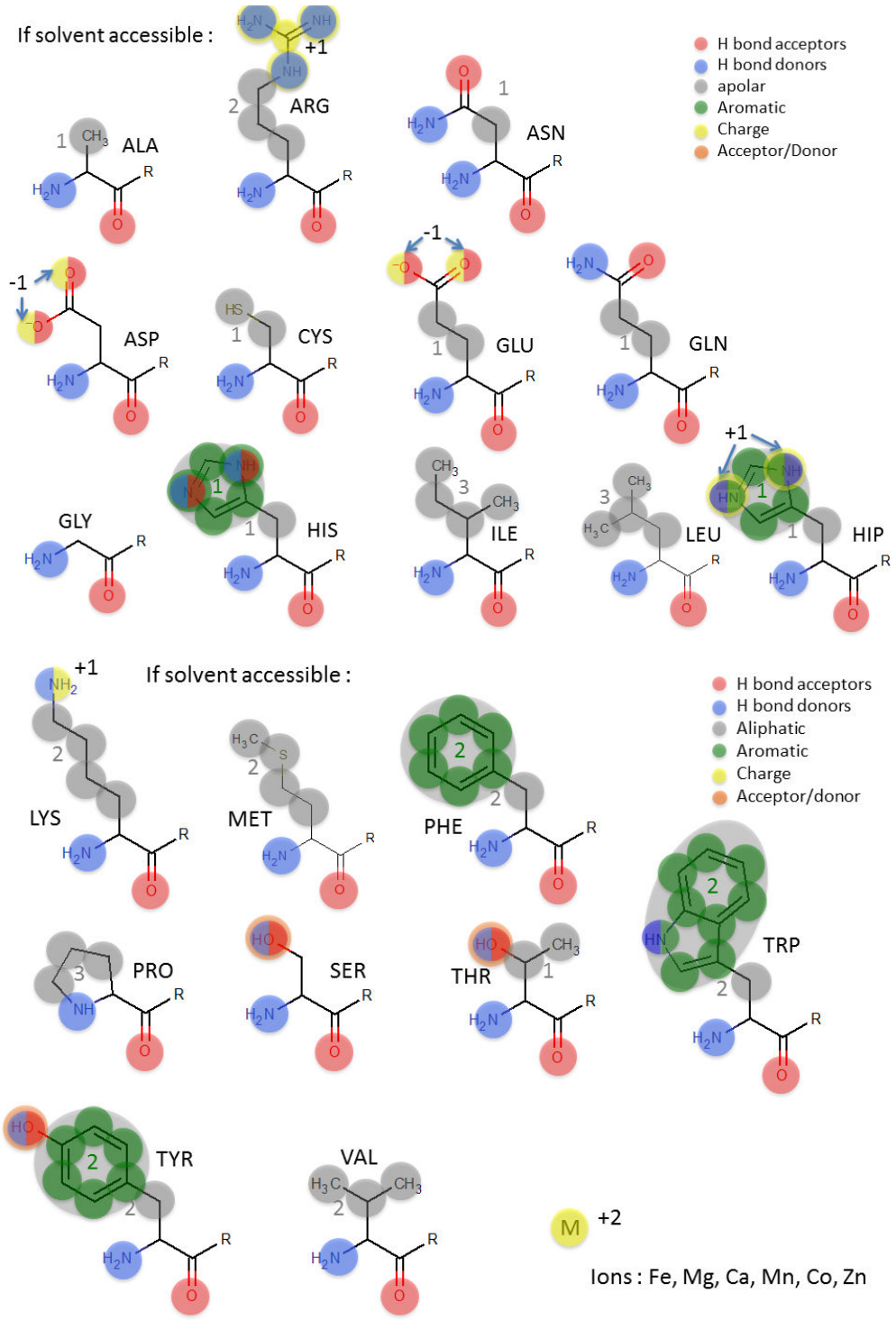


Figure 1. Pharmacophoric features of atoms in amino-acids as considered in SiteAlign-5.
Pharmacophoric features assigned to atoms of standard amino-acids are shown with colored discs. The color code is given by the legend.

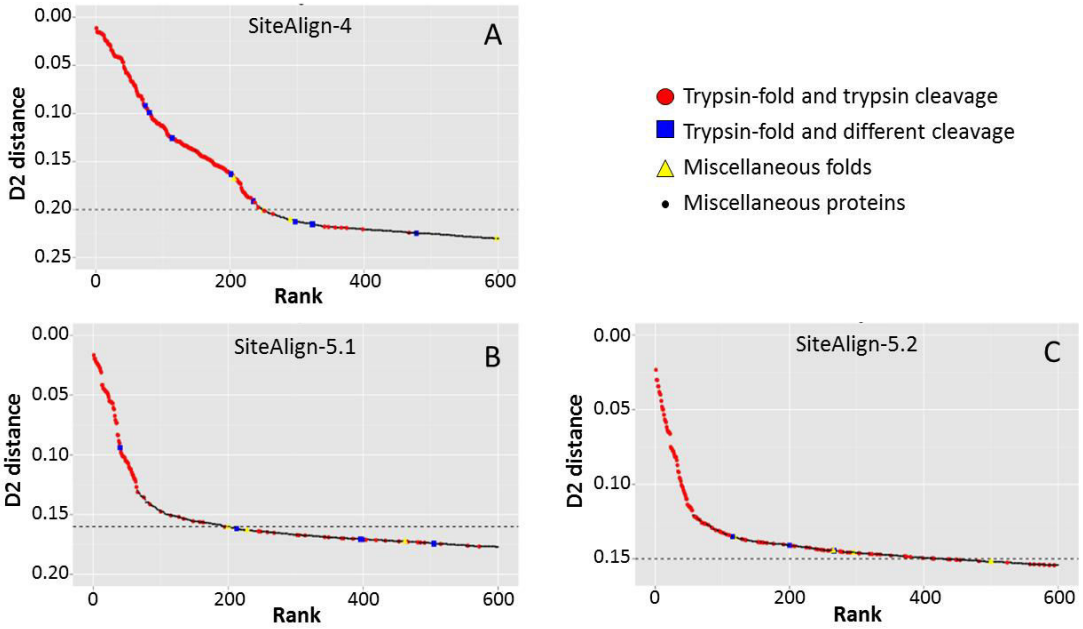


Figure 2. Rank plot of virtual for the serine dataset and different version of SiteAlign.
The plots are focusing on the 600 first ranked proteins. Color codes of the points is given by the legend. A: screening experiment using SiteAlign-4. B: screening experiment using SiteAlign-5.1. C: screening experiment using SiteAlign-5.2

Chapter 2.

Structural Insights into the Molecular Basis of the Ligand Promiscuity

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Structural Insights into the Molecular Basis of the Ligand Promiscuity

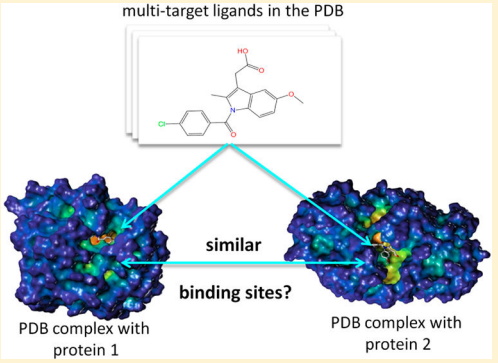
Noé Sturm,[†,‡] Jérémy Desaphy,[†] Ronald J. Quinn,[‡] Didier Rognan,[†] and Esther Kellenberger[†,*]

[†]UMR 7200 CNRS/Université de Strasbourg, MEDALIS Drug Discovery Center, 74 route du Rhin, 67401 Illkirch, France

[‡]Eskitis Institute, Griffith University, Brisbane, Qld 4111, Australia

Supporting Information

ABSTRACT: Selectivity is a key factor in drug development. In this paper, we questioned the Protein Data Bank to better understand the reasons for the promiscuity of bioactive compounds. We assembled a data set of >1000 pairs of three-dimensional structures of complexes between a "drug-like" ligand (as its physicochemical properties overlap that of approved drugs) and two distinct "druggable" protein targets (as their binding sites are likely to accommodate "drug-like" ligands). Studying the similarity between the ligand-binding sites in the different targets revealed that the lack of selectivity of a ligand can be due (i) to the fact that Nature has created the same binding pocket in different proteins, which do not necessarily have otherwise sequence or fold similarity, or (ii) to specific characteristics of the ligand itself. In particular, we demonstrated that many ligands can adapt to different protein environments by changing their conformation, by using different chemical moieties to anchor to different targets, or by adopting unusual extreme binding modes (e.g., only apolar contact between the ligand and the protein, even though polar groups are present on the ligand or at the protein surface). Lastly, we provided new elements in support to the recent studies which suggest that the promiscuity of a ligand might be inferred from its molecular complexity.



INTRODUCTION

Achieving target selectivity is often desirable in drug discovery in order to minimize side effects and possible adverse reactions due to binding to unintended targets. In recent years, much effort has been put into development of computational methods to predict all possible targets of all possible compounds,[1,2] based on the following assumptions: similar compounds share the same targets,[3] drugs with similar side-effect phenotypes share the same targets,[4] and similar protein−ligand binding sites recognize the same compounds.[5] The empirical approaches, which have benefited notably from the availability of ever growing databases collecting structure and activity data of bioactive compounds,[6,7] have proved to be successful in the identification of new targets for drugs and have also contributed to improving the understanding of the main mechanism of action of drugs as well as mechanisms of their adverse reactions.[8−13] For example, the anti-HIV drug Rescriptor, an inhibitor of the viral reverse transcriptase, was predicted and experimentally confirmed to bind to the histamine H4 receptor, thereby suggesting molecular basis for the painful rashes associated with this drug.[10] In binding and functional experiments, we recently demonstrated that some but not all protein kinase inhibitors affect the neurotransmitter release in the synapse through the binding to synapsin I, whose ATP-binding site was beforehand identified as similar to the staurosporine-binding site in Pim-1 kinase.[9]

In pharmaceutical research, the off-target activities of a compound can be characterized from in vitro testing of the compound against a panel of proteins. For example, large-scale profiling experiments are performed at the CEREP, which provides data for >2000 drugs and bioactive compounds tested in >200 assays in the BioPrint database.[14] Comprehensive analyses of BioPrint have suggested link between the chemical properties of a compound and its effects at multiple targets (i.e., its promiscuity). In particular a strong correlation was observed between lipophilicity and promiscuity.[15] The positive ionization,[16] a high number of aromatic rings,[15] and the predominance of ring systems in the compound[17] were also shown to have negative effect on compound selectivity. An independent study on data generated by GlaxoSmithKline (800 compounds tested in >490 assays) confirmed the importance of lipophilicity and aromaticity in the promiscuity of compounds.[18]

In the study presented in this paper, we investigated the reasons for which a compound can target different proteins from the structural point of view. In particular, we sought to know if the promiscuity of a compound was the consequence of the presence of similar binding sites in different proteins, or if it is due to specific characteristics of the compound itself. To this purpose, we exploited the information in the Protein Data Bank (PDB)[19] to identify ligands involved in complexes with different proteins. We then compared the different sites for the promiscuous ligands and we showed that different proteins exhibit the same binding pocket, and that some compounds can adapt to different protein cavities. We finally investigated which

Received: April 20, 2012
Published: August 25, 2012

molecular properties might prompt a compound to bind to dissimilar binding sites.

MATERIALS AND METHODS

Identification in the sc-PDB of Promiscuous Ligands and Their Targets. The sc-PDB[20] repository is a database built from the Protein Data Bank.[19] It exclusively contains complexes between a low molecular weight compound and its bound protein. Practically, the selection of complexes depends on physicochemical criteria for the ligand (e.g., 140 ≤ molecular weight ≤ 810, >1 carbon atoms, >1 oxygen or nitrogen atoms, <20 rotatable bonds), functional criteria for the protein (e.g., no cytochromes or immunoglobulins), and topological criteria for the binding mode (e.g., number of residues in site >7, buried surface area of the ligand >50%). Each sc-PDB entry consists of a ligand, a protein, and the corresponding binding site, which is defined as all residues with at least one atom within a 6.5 Å radius sphere centered on the ligand center of mass. The sc-PDB coordinate files include hydrogen atoms, thereby fully defining the ionization and the tautomeric state of the ligand.[21] The different proteins in the sc-PDB could be distinguished unequivocally by their name, which derived from the Uniprot[22] recommended name. The different ligands in the sc-PDB could be distinguished unequivocally by their canonical SMILES representation. The sc-PDB is a nonredundant database: for a given pair of protein and ligand, only the PDB entry with the best resolution is considered.

The data set was created from the 8166 entries of sc-PDB, release 2010. In total, 518 ligands were found in at least two complexes with different proteins. About half of them were discarded due to their high similarity with nucleic acids, peptides, monosaccharides, oligosaccharides, or fatty acids (The filtering rules are given in the Supporting Information, Table S1). The data set contains 247 promiscuous ligands.

2D-Description of Ligands. The following chemical descriptors were computed for ligands using PipelinePilot8 (Accelrys Inc., San Diego, CA, USA): molecular weight, number of hydrogen bond (H-bond) donors or acceptors, number of rotatable bonds, molecular polar surface area, ALogP (Ghose/Crippen group-contribution estimate for logP), circular FCFP_4 fingerprints, FCFP_4 density (FCFP_4 size/number of non hydrogen atoms), H-bonding propensity (number of H-bond donors and acceptors/total number of atoms), and three-dimensionality (number of sp3 carbon atoms/total number of carbon atoms).

The 247 compounds of the data set were clustered using the Jarvis-Patrick algorithm in MOE2011 (Chemical Computing Group Inc., Montreal, Canada). The MACCS keys were compared using the Tanimoto coefficient. The similarity threshold was set to 0.65 for the creation of the lists of similar compounds and for the comparison of lists ("cluster overlap" parameter).

Conformational Variability of Protein-Bound Ligands. Protein-bound ligand structures were first compared by computing the root-mean-square deviation (rmsd) of the positions of the ligand heavy atoms after the best-fit superposition of the two sets of coordinates. The rmsd was computed using the 'Match' routine of Sybyl-X1.3 (Tripos, Inc., St. Louis, MO, US), which takes into account topological symmetry within molecules. Although rmsd values are commonly used and easy to interpret, they may be biased toward low values for small molecules or toward high values if one or more of the paired atoms are at a great distance from each other.[23] In the present study, the rmsd values may be misleading for ligands which do not interact totally with their target proteins (for example a high rmsd value may be observed if the ligand moiety which interacts with the protein has a well conserved structure in the two compared complexes, whereas the ligand moiety which points outward has different structures). To overcome this limitation, we evaluated the shape similarity of the ligand part that contacts the bound protein as follows: all ligand atoms involved in nonbonded interactions with the protein were identified as previously described;[24] their coordinates were written in MOL2 format using a simplified atom typing based on the nature of protein−ligand interactions (C.3 for any atom engaged in a hydrophobic contact, N.Am for a H-bond donor, O.2 for a H-bond acceptor, N.4 for a positively charged atom, O.Co2 for a negatively charged atom) Two sets of atoms originating from the complexes of a ligand with two different proteins were 3D-aligned by optimizing the volume overlap from Gaussian functions representing the atoms.[25] The alignment routine was written using the OEChem and OEShape toolkits (OpenEye, Inc., Santa-Fe, CA, U.S.A.). The overlap of atoms was scored with a Tanimoto coefficient (shTc):

$$\mathrm{shTc}_{A,B} = \frac{\sum_i O_{A,B}}{\sum_i I_A + \sum_i I_B - \sum_i O_{A,B}}$$

Where, for each of the five above-mentioned atom types i, $O_{A,B}$ is the overlap volume between conformers A and B, and I is the self-overlap volume of each entity A and B. The shTc score is normalized and quantifies the conservation in the two complexes of the protein-interacting moiety of the ligand. For example, if all protein-interacting atoms of a ligand in complex A represent 60% of all the protein-interacting atoms of the ligand in complex B (or vice versa), the shTc value is equal to 0.6. Alternatively, if the total numbers of protein-interacting atoms of the ligand are identical in complexes A and B and if 75% of the protein-interacting atoms of the ligand are identical in the two complexes, then the shTc value is equal to 0.6 too. The Tversky coefficient (shTv) was computed in order to distinguish the different scenarios:

$$\mathrm{shTv}_{A,B} = \frac{\sum_i O_{A,B}}{\alpha \sum_i I_A + \beta \sum_i I_B - \sum_i O_{A,B}}$$

where, for each of the five above-mentioned atom types i, $O_{A,B}$ is the overlap volume between conformers A and B, I is the self-overlap volume of each entity A and B, and α and β are weights so that $\alpha \neq \beta$ and $\alpha + \beta = 1$. By contrast to a Tanimoto index ($\alpha = \beta = 1$), the Tversky index gives more importance to either the reference or the fit object by assigning different weights to the self-overlap volumes I_A and I_B. The retained Tversky coefficient was the maximal value obtained for either of the two parameter sets $\alpha = 0.05/\beta = 0.95$ or $\alpha = 0.95/\beta = 0.05$.

2D Comparison of the Targets of Promiscuous Ligands. The protein sequences in fasta format were downloaded from the RCSB PDB.[26] The comparisons of the protein sequences were performed using the default parameters of the Needle routine in the EMBOSS package.[27] Only the protein chains which form the ligand binding site were considered. If several comparisons were made for a given pair of proteins, only the highest sequence identity value was retained. A sequence identity above 30% is a good indicator of protein homology.[28] In the present analysis, we considered that

an evolutionary link exists between two proteins aligned over more than 100 residues with a sequence identity above 25%.

3D Comparison of the Targets of Promiscuous Ligands. The comparisons of the protein structures were performed using the default parameters of the CE program.[29] This program identifies the longest combination of pairs of fragments which are structurally equivalent in the two protein chains (a fragment represents the Cα atoms of 8 consecutive residues) and calculates the statistical significance of the structural alignment by evaluating the probability of finding such an alignment from a random comparison of structures (Z-score). The input files were the structure files which were downloaded from the RCSB PDB. Only the protein chains which form the ligand binding site were considered. If several comparisons were made for a given pair of proteins, only the result with the highest Z-score was retained. A Z-score value higher than 4 denotes the conservation of the overall fold of the two proteins under investigation.

3D Comparison of Binding Sites for Promiscuous Ligands. The comparisons of the binding sites were performed using three in house programs, Volsite/Shaper,[30] SiteAlign4.0,[31] and Fuzcav.[32] The sc-PDB binding site coordinates in MOL2 format were used as input files. The comparisons of sites using Shaper were repeated for hydrated binding sites. Hydrated sites were prepared using Sybyl-X1.3 and include all crystallographic water molecules whose oxygen atom is closer than 3.5 Å from any ligand polar atom and closer than 3.5 Å from at least three binding site residues. The position of water hydrogen atoms was optimized to maximize the number of H-bonds made with the protein.

In SiteAlign,[31] eight topological and physicochemical attributes are projected from the Cβ-atom of cavity-lining residues to an 80 triangle-discretized polyhedron placed at the center of the binding site, thus defining a cavity fingerprint of 640 integers. 3D alignment is performed by moving the sphere within the target binding site while keeping the query sphere fixed. After each move, the distance of the newly described cavity descriptor is compared to that of the query, the best alignment being that minimizing the distance between both cavity fingerprints. The similarity is evaluated by a "global" score which is computed by considering the pairs of aligned triangles with non null properties in the mobile sphere or the fixed sphere (D1) and a "local" score which is computed by considering only triangles with non null properties in the mobile and the fixed spheres (D2). D1 and D2 scores lower than 0.6 and 0.2, respectively, indicate that the geometry and the chemical nature of residues are similar in the two sites which are compared.[31]

From a known protein−ligand complex, Volsite[30] converts the site into a regular lattice of pseudoatoms filling the cavity. The pseudoatoms farther than 6 Å from any ligand heavy atom were discarded. To each pseudoatom is assigned a pharmacophoric type, depending on the nature of the closest protein atom (H-bond acceptor, H-bond donor, H-bond acceptor and donor, negative ionizable, positive ionizable, hydrophobic, aromatic, or none if there is no protein atoms within a 4 Å distance). Shaper then aligns two sets of cavity points using Gaussian functions (see above) and then scores the alignment according to the quality of the overlap.[30] In practice, we demonstrated that a similarity score (S) higher than 0.35 indicates that the cavity shape and pharmacophoric properties are similar in the two sites which are compared.

FuzCav[32] annotates the Cα atoms of cavity-lining amino acids with the pharmacophoric properties of its parent residue (H-bond donor, H-bond acceptor, positive ionizable, negative ionizable, aromatic, aliphatic), then enumerates all triplets of Cα (three properties, three distances ≤14.3 Å) to populate a vector of 4833 integers which encode all possible combinations of triplets. The comparison of two sites consists in the direct computing of the distance between two numerical fingerprints (it does not generate a 3D alignment of sites). The benchmarking of the program revealed that a similarity score higher than 0.16 reflects the conservation of spatial arrangement and physicochemical properties of amino acids in the two sites which are compared.

RESULTS AND DISCUSSION

Setting up a Data Set of Promiscuous Ligands and Their Bound Proteins. To better understand the molecular basis for ligand promiscuity, we searched for ligands whose crystal structure is available for complexes with two or more different proteins. We restricted our analysis to proteins which are potentially able to bind small compounds with high affinity (from here on called *druggable*)[33] and to ligands whose molecular weight ranges from 140 to 800. We did not considered monosaccharides, because they are usually weak binders and their binding sites are poorly druggable. We neither studied nucleotides nor peptides, because they are highly flexible and known to recognize conformer-specific binding pockets.[34]

Among the 4229 unique ligands in the sc-PDB, we identified 247 promiscuous ligands. The chemical diversity of the set was evaluated by a nonhierarchical clustering based on MACCS keys compared using the Tanimoto coefficient. The similarity threshold of 0.65 yielded 145 clusters which correctly grouped compounds according to biochemical scaffolds. For example, it was observed that all thiamine derivatives define a single cluster (Figure 1A). In fact, about one-third of the ligands in the data set correspond to natural lipids, amino acids, and protein cofactors, or their close analogs.

The distribution of key physicochemical properties in the data set is given in Figure 1B. The molecular size was evaluated using the molecular weight. The average molecular weight in the data set is 367 and about 90% of all the 247 promiscuous ligands have molecular weight ranging from 200 to 500. The molecular flexibility was evaluated using the number of rotatable bonds. The ligands in the data set have up to fourteen rotatable bonds. Only fifteen ligands are fully rigid whereas ten compounds have more than 10 rotatable bonds. Last, the molecular polarity was evaluated using the number of H-bond donors and acceptors, the polar surface area (PSA), and the LogP (not shown). The cumulated number of H-bond donors and acceptors ranges from 2 to 20, and approximately one-third of ligands is distributed in each of the [2; 5], [6; 10], and [11; 20] intervals. The PSA ranges from 20 to 321 Å^2, and approximately 60% of ligands is in the [50; 150] interval. A quarter of the ligands have a PSA exceeding 150 Å^2. The LogP ranges from −10.7 to +7.3 and respectively 30% and 60% of ligands are distributed in the [−5; 0] and [0; +5] intervals. About 94% of the 247 ligands comply with the Lipinski's rules of five.[35] Altogether, the area of molecular property space occupied by the molecules in the data set overlap that occupied by orally absorbed drugs (from here on this characteristics will be called *drug-like*, for a comprehensive review on drug-likeness see ref 36). There seems however to be a bias in the data set

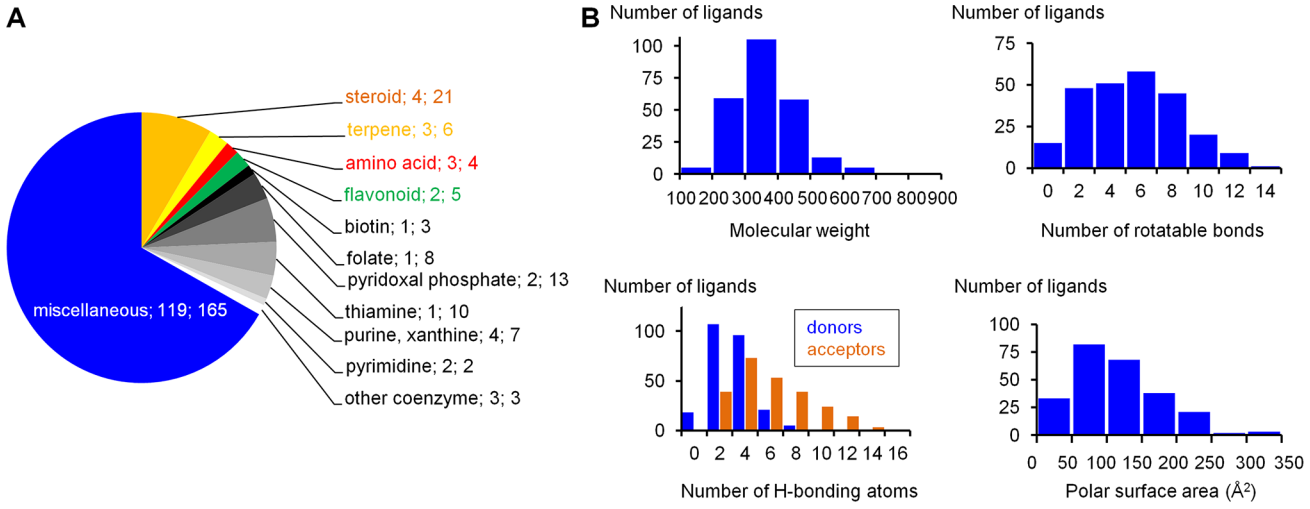


Figure 1. Description of the ligands in the data set. (A) Biochemical classification of the 247 promiscuous ligands. The name of each class is followed by the number of chemical clusters within the class, and then by the total number of members in the class. (B) Distribution of physicochemical descriptors.

toward polar compounds. To further evaluate the *drug-like* property of the 247 promiscuous ligands, we compared them to 959 drugs selected in Drugbank (FDA-approved small molecule drugs with molecular weight lower than 900, no nutraceuticals, biologics, or experimental drugs).[37] We could hence identify five drugs in our data set, namely diethylstilbestrol, progesterone, novobiocin, trimetrexate, and trimethoprim. In addition, 46 of the promiscuous ligands were found similar to 33 known drugs using circular FCFP_4 fingerprints, Tanimoto coefficient, and a similarity threshold of 0.5.

The 247 promiscuous ligands correspond to 689 PDB complexes, but only to 393 different proteins which nevertheless cover a wide range of biological functions (Figure 2A). About 65% of the ligands bind two different proteins (Figure 2B). Other ligands have up to 7 different targets, with the exception of the nonselective kinase inhibitor staurosporine which was found in complex with 23 different members of this enzyme family.

The total number of protein pairs in the data set is equal to 1070. The pairs were categorized according to their sequence identity and their global three-dimensional structure similarity: (i) 264 pairs are made of two proteins which have high sequence identity (>25% with >100 aligned residues) and a common fold (CE Z-score >4); we named them the *homologous pairs*; (ii) 478 pairs are made of two proteins which have low sequence identity but a common fold; we named them the *convergent pairs*; and (iii) 328 pairs are made of two proteins which have no sequence or fold similarities; we named them the *distant pairs*.

Binding Sites Which Accommodate the Same Ligand Are Not Necessarily Similar. We analyzed our data set in order to understand the molecular basis of the promiscuity of drug-like ligands, assuming that a ligand can associate with two different targets for one of the two following reasons: (i) the ligand-binding sites in the two proteins are similar or (ii) the ligand is able to adapt to two different binding sites. To test the first of the two hypothesis, we evaluated the similarity between the sites in each protein pairs using three different approaches: the 3D alignment of icosahedrons encoding the position and the pharmacophoric properties of the binding site-lining amino acids (SiteAlign),[31] the 3D-alignment of grid points which

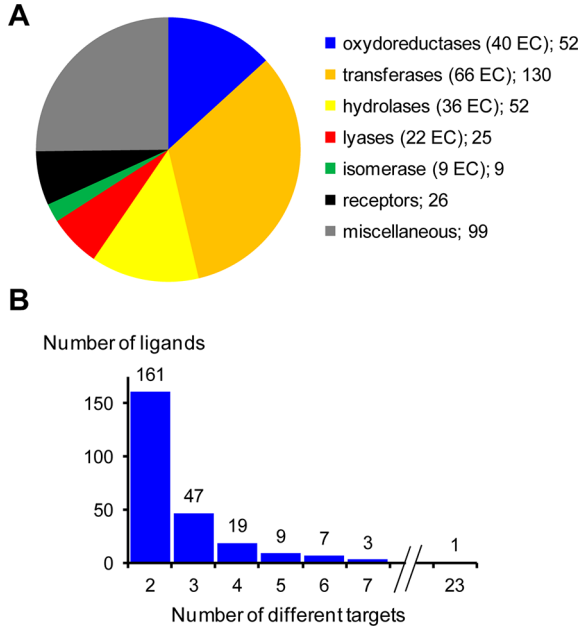


Figure 2. Description of the proteins in the data set. (A) Functional classification of the 393 proteins that are targeted by promiscuous ligands. The name of each class is followed by the number of members in the class. For the classes that group enzymes, the number of different subclasses as described by the Enzyme Commission (EC) is indicated too. (B) Level of promiscuity across the data set indicated by the number of different targets per ligand.

represent the cavity shape and the pharmacophoric properties at site surface (Shaper),[30] and the comparison of 3D-pharmacophoric fingerprints (Fuzcav).[32] Two sites were considered similar if they met the similarity criteria of at least one of the three approaches.

About three-quarters of the 1070 binding sites pairs were predicted to be similar. Similar sites were identified in all three categories of protein pairs: homologous, convergent, and distant pairs (Figure 3). Site similarity was detected in almost all homologous pairs of proteins (97%) and in about 80% of the convergent pairs. As shown in Figure 3, our data clearly revealed that the topology of the sites tend to be preserved in

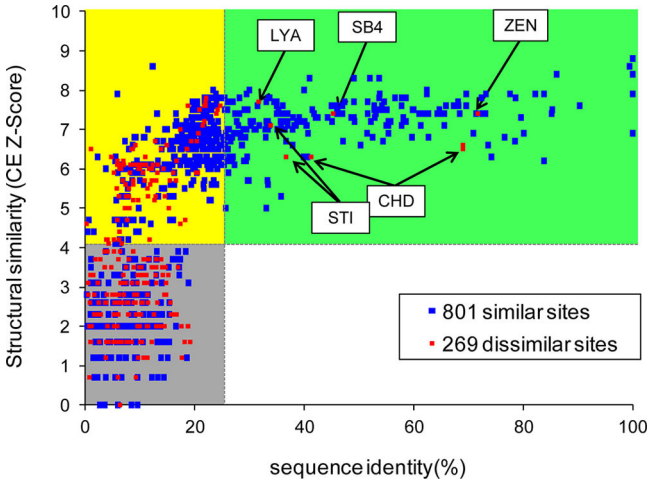


Figure 3. Sequence, fold, and site similarities in pairs of the target proteins. Three categories of protein pairs are highlighted with the different background colors: homologous proteins (green), convergent proteins (yellow), and distant proteins (gray). Boxes give the HET code of the five ligands associated to pairs of dissimilar sites of homologous proteins.

two proteins which share more than 25% of sequence identity and have a conserved fold. This observation is in line with bioinformatics studies, which demonstrated that the key functional amino acids are generally well conserved across the proteins of a functional family.[38] Interestingly, binding site similarity was also observed in about half of the distant pairs, meaning that proteins with no genetic evolutionary relationship can have a common local three-dimensional structure. This finding underlines the potency of site comparison methods to predict the ligand binding capability of a protein.

A quarter of the 1070 binding site pairs were found dissimilar with all three programs. Most of them correspond to protein pairs which have distinct sequence and fold characteristics. We cannot exclude expressly that the absence of similarity between two sites is due to methodological aspects. We nevertheless verified that crystallographic water molecules located in the cavity only have a marginal influence on site comparisons. In our data set, we found that one or several crystallographic water molecules mediate interactions between ligand and protein in 335 out of 689 PDB complexes, representing 628 pairs of sites. We repeated all Shaper calculations using as input the sites including these bridging water molecules. We observed that although the similarity score was modified in most of the comparisons (98%) which involve a "hydrated" site, the overall proportion of dissimilar and similar sites pairs in the data set was not changed upon consideration of water molecules (Table 1).

Table 1. Water at Binding Interfaces Hardly Affects Site Comparison Using Shaper

		number of similar sites pairs		
pairs of proteins	total number of site pairs	with and without water	only without water	only with water
homologous	264	242	0	0
convergent	478	331	12	29
distant	328	143	8	12
all	1070	716	20	41

In summary, the ligand information available in the PDB revealed that the promiscuity of a ligand can be explained by the presence of the same binding pocket in different proteins. However, a significant number of dissimilar sites were observed among the investigated pairs of complexes between a ligand and two different proteins, thereby supporting the assumption that the promiscuity of a ligand may solely originate from its physicochemical properties. We identified 76 ligands which have the capacity to bind to dissimilar protein sites.

Multiple Binding Modes Explain Why a Ligand Can Bind to Dissimilar Sites. In order to understand why a ligand can bind to two dissimilar sites, we compared the corresponding protein-bound ligand conformations. In particular, we scored the overlay between all heavy atoms of the ligand in the two sites (rmsd) and between the subset of heavy atoms in direct interaction with the protein (shTc). We hence could define three categories of pairs, depending on the structural and binding characteristics of the ligand: (i) in the class called *flexible ligand*, the ligands adopt different conformations in the two sites of a pair (high rmsd, low shTc), (ii) in the class called *bianchor ligand*, the ligands exhibit similar conformations in the two sites but use different sets of atoms to interact with the protein (low rmsd, low shTc), and (iii) in the class called *difficult to rationalize*, the ligands use the same moieties to interact with the two sites (high shTc). Figure 4 shows that, depending on the threshold used for rmsd and shTc, the exact number of pairs assigned to each category

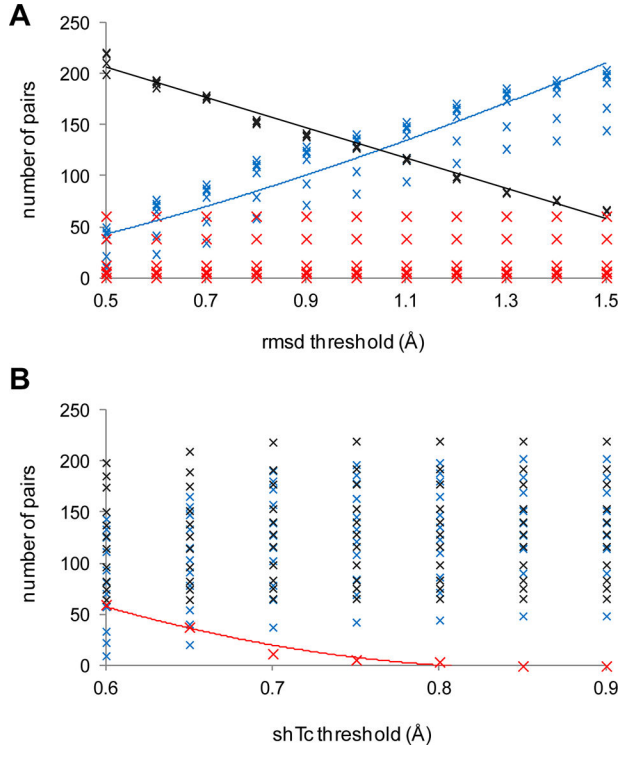


Figure 4. Structural characteristics of ligands in the pairs of dissimilar sites. The 269 pairs of dissimilar sites were classified as flexible ligand (black crosses), bianchor ligand (blue crosses), and difficult to rationalize (red crosses), for all possible combinations of thresholds for rmsd ranging from 0.5 to 1.5 Å (0.1 Å increment) and shTc ranging from 0.6 to 0.9 (0.1 increment). The linear, power, and order 2 polynomial trendlines were plotted for the flexible ligand (R^2 = 0.98), bianchor ligand (R^2 = 0.75), and difficult to rationalize (R^2 = 0.97) series, respectively.

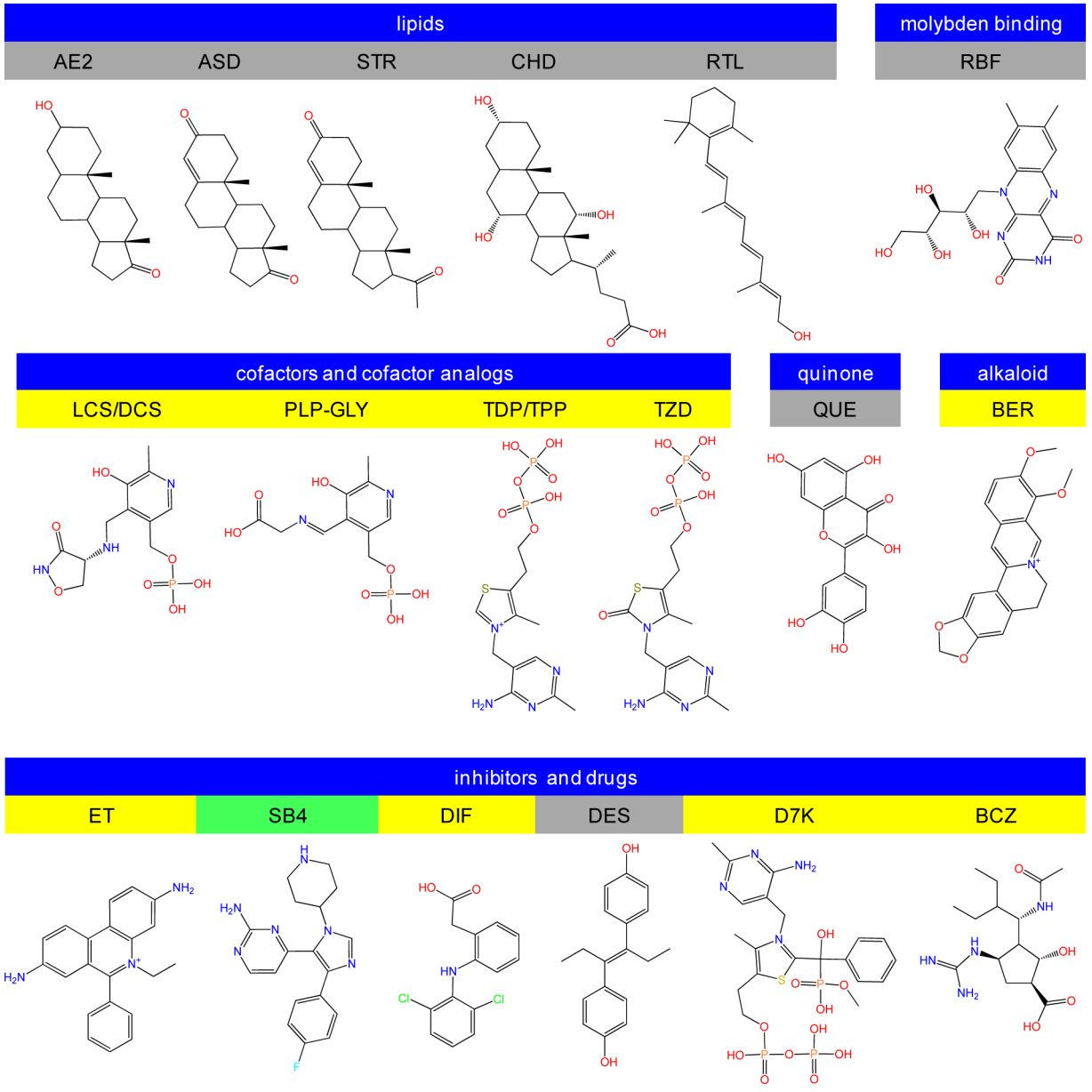


Figure 5. Chemical structure of superpromiscuous ligands. Ligands are labeled using their HET code, whose shading indicates the category of their parent pairs (green if homologous, yellow if convergent, and gray if distant), and are ordered according to their biochemical nature or their biological function.

varies, but the trends remain constant. In particular, most of the pairs of dissimilar sites correspond to ligands which adapt to the protein environment by changing their three-dimensional structure and/or their binding mode. Figure 4 indicates that the number of cases difficult to rationalize represents about 22% of the dissimilar pairs if shTc is equal to 0.6, that is, if at least 60% of the ligand atoms in interaction with one site are found among the ligand atoms in interaction with the other site of the pair. This number becomes zero if shTc is equal or higher than 0.85, meaning that at least 10% of the ligand atoms in contact with the protein are different in the two complexes.

These observations, which are based on 76 different ligands and 269 pairs of dissimilar sites, suggested that the ability of a ligand to bind to dissimilar sites principally results from its capability to modify its conformation. In addition, in about half of the pairs of dissimilar sites, the interacting atoms of the ligand in one complex constitute a subset of the interacting atoms of the ligand in the other complex (as indicated by shTv $\geq$ 1.5shTc), thus indicating that the ligand has different degrees of burial into the two proteins. The lack of similarity between sites is accordingly due to the limited size of the common ligand recognition area. In the remaining half of the pairs of dissimilar sites, the multiple possibilities of the ligand to form nonbonded interactions with a protein explain why it can bind to two topologically different sites.

Interestingly, 18 of 76 ligands were shown to use almost the same chemical moieties to bind to different sites (shTc $\geq$ 0.6, pairs which are difficult to rationalize in Figure 4), suggesting that different binding modes may be established from the same set of atoms of the ligand without significant conformational adaptation. From here on, we will call them the *superpromiscuous* ligands. Noteworthy, these ligands, with the exception of one of them, were in complex with nonhomologous proteins. They correspond to a limited number

dx.doi.org/10.1021/ci300196g | J. Chem. Inf. Model. 2012, 52, 2410−2421

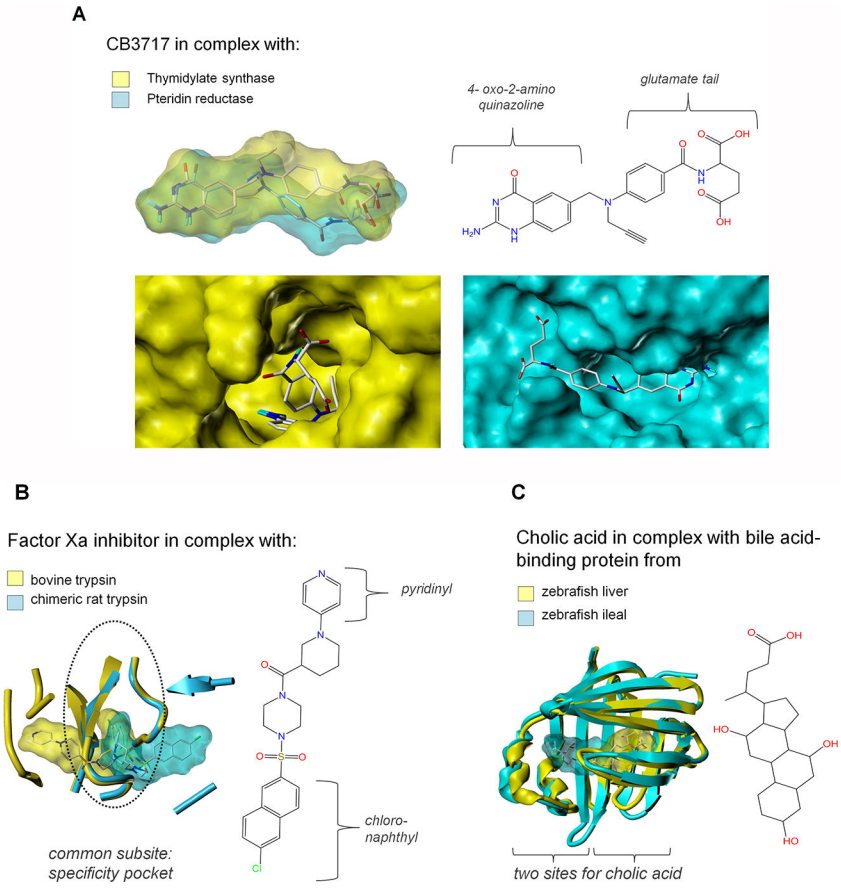


Figure 6. Examples of ligand bound to dissimilar sites. (A) HET code: CB3. PDB codes: 1an5 and 2bfa. (B) HET code: ZEN. PDB codes: 1ql8 and 1j17. (C) HET code: CHD. PDB codes: 2qo4 and 3elz. In A, B, and C, the ligand shapes are delimited by transparent solvent-excluded surfaces. In A, the protein shapes are delimited by solid solvent-excluded surfaces. In B and C, the three-dimensional structures of complexes are aligned for the best-fit of the protein backbones, as represented by ribbons (Sybyl X1.3, Tripos, Inc., St. Louis, MO, U.S.A).

of chemotypes: lipids (retinol and steroids), two cofactors, and a coenzyme which binds molybden, two natural products (a quinone and an alkaloid), and six enzyme inhibitors or drugs (Figure 5).

Examples of Multiple Binding Modes of a Ligand. An example of a ligand able to bind to dissimilar sites of a distant pair is given in Figure 6A. The ligand CB3717 binds to thymidylate synthase and pteridine reductase. The three-dimensional alignment of the two active ligand structures overlays the 4-oxo-2-amino quinazoline moiety, thus evidencing large variations in the rest of the molecule. This observation is in line with the experimental binding modes. Thymidylate synthase buries the entire CB3717 into its cofactor binding site, although the precise location of the 4-oxo-2-amino quinazoline moiety depends on the presence of a substrate.[39] In the complex between pteridine reductase and CB3717, the substrate-binding site mainly establishes nonbonded interactions with the 4-oxo-2-amino quinazoline while the glutamate tail of the ligand stretches out of the protein surface.[40]

Interestingly, we also observed ligands able to bind to dissimilar sites of homologous pairs (Figure 3). For example, the specific human factor Xa inhibitor (HET code: ZEN) is an inhibitor of bovine trypsin and of a rat trypsin mutant which was engineered to mimic factor Xa. The two enzymes have the same fold (rmsd of Cα atoms = 0.64 Å), and their amino acid sequence is highly conserved in the active site, thus defining virtually identical binding cavities. However changes in the nature of a few residues control the positioning and the affinity of ligand, so that the enzyme specificity pocket is occupied by the pyridine ring of the ligand in bovine trypsin[41] whereas it is occupied by a chloronaphthyl group in the chimeric rat trypsin (Figure 6B).[42] As a consequence, the two binding sites have only 13 residues in common, which represent only half of each site.

Alternate binding modes were also observed for the protein kinase inhibitor imatinib (HET code STI) in different tyrosine-protein kinases. Here substantial conformational changes at the secondary structure level induced either the tight binding of the inhibitor in an extended-conformation or a weaker binding to a more compact conformation (PDB codes: 1xbb, 2oiq).[43,44] In tyrosine kinases, these structural changes are involved in enzyme activation/inactivation. Changes in sequence and structure also explain the poor similarity between SB4 inhibitor-binding sites in Mitogen-activated protein (MAP) kinase 14 and MAP kinase 1 (PDB codes: 1bl7, 3erk), and between the antifolate LYA-binding site in human and protozoan thymidylate synthases (PDB codes: 1juj, 3k2h).The last example of a ligand bound to dissimilar sites in homologous pairs is cholic acid (HET code: CHD) which occupies different parts of a well conserved binding pocket in two homologue fatty acid-binding proteins (Figure 6C). Actually, in this family of enzymes, the number of cholate molecules per binding site is either one or two, and the

stoichiometry is finely tuned by the presence or not of a single disulfide bridge.[45]

Superpromiscuous Ligands Have Extreme Binding Modes. The first chemical class of superpromiscuous ligands (Figure 5) is made of lipids. In the studied complexes, we observed that the nonbonded interactions between lipids and their target protein involve principally hydrophobic contacts (from 9 to 18 amino acids establishing apolar contacts, none or a single amino acid establishing polar contacts). As an example, 4-androstene-3-17-dione was cocrystallized with two dehydrogenases of the steroid metabolism. The two binding sites have equivalent size and are almost exclusively lined by apolar residues, yet sequence variations cause differences in site enclosure (Figure 7A). Similar observations were made for the other steroid examples, and for retinol.

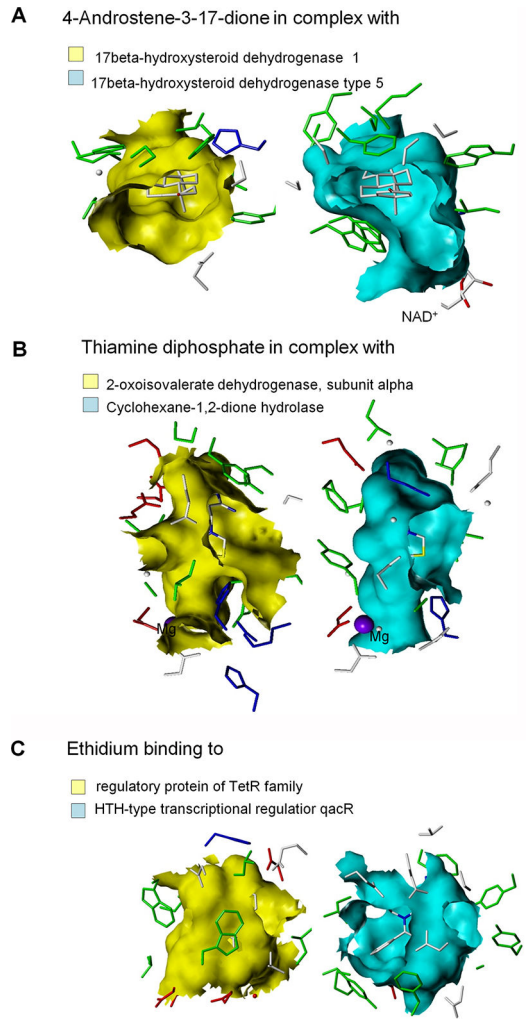


Figure 7. Examples of dissimilar sites which accommodate the same ligand while using similar binding modes. (A) HET code: ASD. PDB codes: 1qyx and 1xf0. The adenosine moiety of NAD+ cofactor is not depicted. (B) HET code: TPP. PDB codes: 1umb and 2pgn. (C) HET code: ET. PDB codes: 2zoz and 3br3. The orientation of two complexes corresponds to a fixed position of the ligand. Protein cavity shapes are delimited by solid solvent excluded surfaces. The side chains of ligand-interacting amino acids are represented by capped sticks and colored according to their property (red for acidic, blue for basic, white for neutral polar, and green for apolar). The bound ligands are represented by CPK-colored capped sticks.

Other essential natural metabolites were described in Figure 5, in particular the cofactors thiamine diphosphate (TDP, TPP, and TDZ) and pyridoxal (LCD/DCS and GLY-PLP). In the studied complexes, we observed that the molecular recognition involves numerous H-bonds and ionic interactions (more than 7 amino acids establishing polar contacts, these residues representing from 36% to 61% of the total number of residues in interaction with the ligand). The example of thiamine diphosphate in complex with a dehydrogenase and a hydrolase is given in Figure 7B. The binding pockets of the two enzymes have similar size and overall shape yet they have drastically different electrostatic properties. The binding mode is preserved because the intermolecular H-bonds involve protein backbone atoms, which anchor the aminopyrimidine moiety and a phosphate group of the coenzyme. The two enzymes also have in common a magnesium ion coordinated by the alpha and beta-phosphate groups of the cofactor.

Among the superpromiscuous ligands are also two natural products (Figure 5), the flavonoid quercetin and the alkaloid berberine, which both have a marked aromatic character. In the studied complexes, we observed that the two compounds establish none or one single H-bond to their target, even though quercetin contains 7 H-bond donors and acceptors.

These examples suggest that the interaction between natural molecules, which have numerous biological functions, and their multiple targets corresponds to an extreme binding mode, very hydrophobic or, on the contrary, very hydrophilic. Similar observations could not be made for the others superpromiscuous ligands (inhibitors and drugs). For example, ethidium which is a fluorescent DNA intercalating agent but also an antitrypanosomiasis drug, was found in two complexes with different bacterial transcriptional regulators. In the two complexes, we observed that ethidium establishes both apolar contacts and electrostatic interactions (three or eight aromatic stacking and H-bonds) with the protein, but that the two networks of nonbonded interactions are different. By considering the proteins, we could notice that the binding pockets share similar geometric features (the size and the overall shape are the same, although the opening are different) but exhibit very different electrostatic properties (Figure 7C).

At this point it is worth mentioning that crystallographic water molecules mediate up to eight intermolecular H-bonds between the superpromiscuous ligands and their target proteins. Furthermore, we noticed that the consideration of water molecules in protein sites yields a significant increase of similarity between sites for ethidium, pyridoxal, and quercetin. In detail, seven pairs of sites having as ligand ethidium, quercetin, or pyridoxal were predicted dissimilar if water is not included in proteins, whereas only three of them were predicted dissimilar if water is included in the proteins.

Do the Promiscuous Ligands Have Specific Characteristics? In this study, we considered 247 ligands capable of binding to different target proteins. Among them, 76 were demonstrated to be able to recognize different protein environments, and 18 of them were called superpromiscuous because they use almost the same anchor atoms in a preserved conformation to bind to different proteins. We have already mentioned that the superpromiscuous ligands are of limited chemical diversity. We here investigated whether the ligands in our data set possess specific chemical features that distinct them from other drug-like ligands. In particular, we compared the 247 promiscuous ligands with ligands in two other data sets: one composed of 959 approved drugs and the second one of

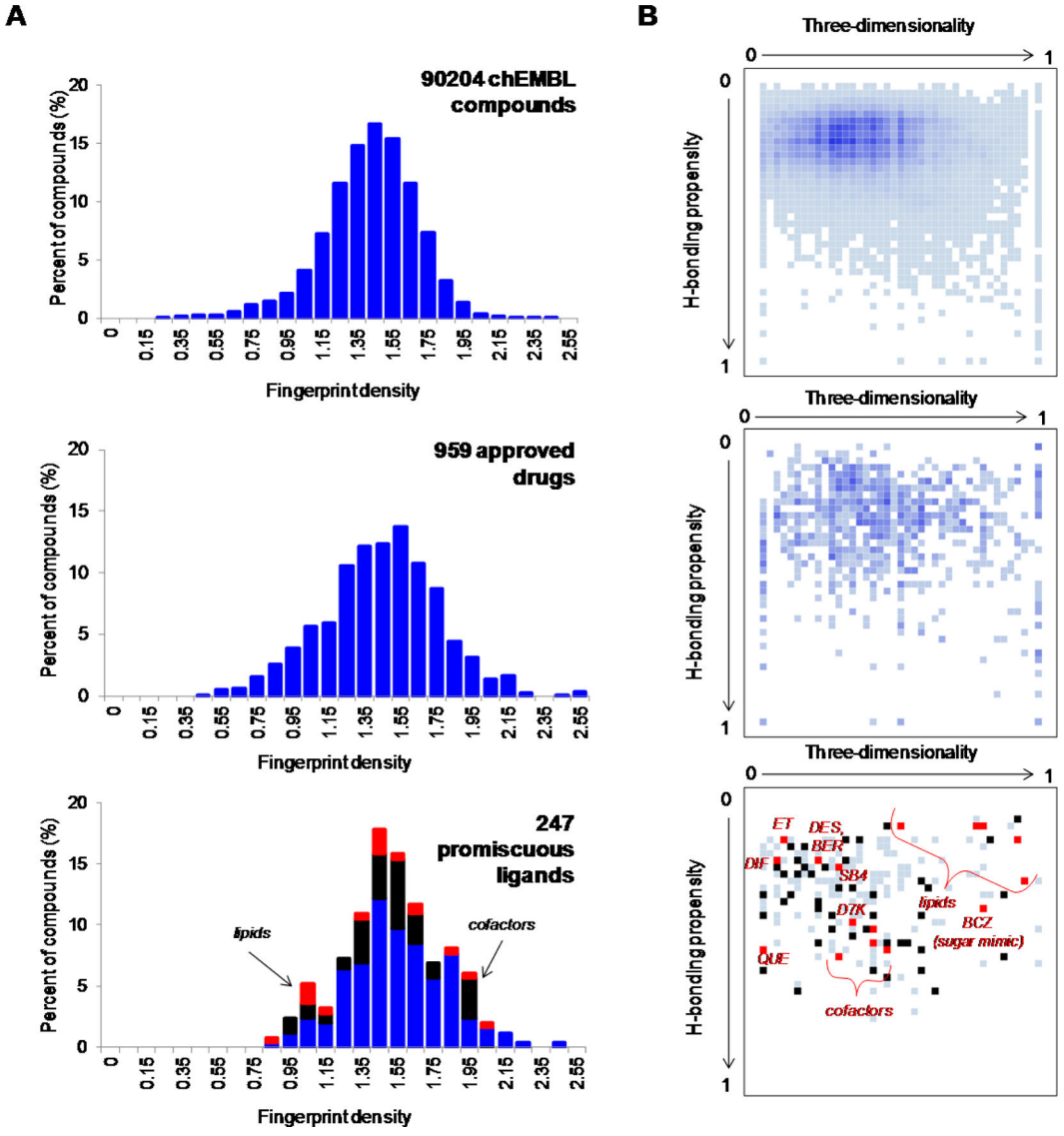


Figure 8. Comparison of the promiscuous ligands in the data set (bottom panels) with drugs (medium panels) and with bioactive compounds for which a single target is known (top panels). (A) Molecular complexity. (B) H-bonding propensity and molecular three-dimensionality. In the top and middle panels of B, the intensity of the color reflects the number of ligands in each bin. In the bottom panel of A and B, the 76 promiscuous ligands which adapt to different protein environments are highlighted in black, except the 18 superpromiscuous ligands which are colored in red. The plots were generated using the php library gg2.0.34.

90 204 bioactive compounds. The drugs were retrieved from Drugbank by querying FDA-approved "small molecule" drugs (whose molecular weight is lower than 900), but not nutraceuticals, biologics, or experimental drugs. The bioactive compounds were retrieved from ChEMBL[46] by querying compounds for which a single target has been reported and whose affinity for its target is higher than 6 (as expressed by the logarithm of a dissociation or inhibition constant).

We analyzed three molecular descriptors, molecular complexity expressed as the circular fingerprint density,[47] H-bonding propensity, and three-dimensionality (see Material and Methods). The molecular complexity in the data set of bioactive compounds follows a normal distribution (Figure 8A). In the data set of approved drugs, the molecular complexity is in the same value range as the data set of bioactive compounds, but values are more scattered around the mean value than in a normal distribution. This trend is even more pronounced in the data set of promiscuous ligands: it thus appears that this data set is rich in molecules of low complexity (including the lipids) and in molecules of high complexity (including the cofactors), which both have been classified as the superpromiscuous ligands. To further delineate molecular complexity, we partitioned the data sets according to H-bond propensity and three-dimensionality (Figure 8B). Again, the chemical space defined by the extreme values of the two properties is common to all three data sets, but the distributions of points varies significantly. The data set of promiscuous ligands especially occupies regions which are not highly populated in the two other data sets. More precisely, we could spot superpromiscuous ligands in regions of high three-dimensionality and low H-bonding propensity (including lipids), in regions of high three-dimensionality and high H-bonding propensity (including the sugar mimic BCZ), in regions of very low three-dimensionality and H-bonding propensity (including the inhibitors ET and DIF), and in

regions medium three-dimensionality and high H-bonding propensity (including cofactors).

Altogether, our findings suggested that drug-like compounds are able to adapt to different protein environments if they are flexible or if they possess specific chemical features, such as a high proportion of aromatic rings with few or no aliphatic hydrophobic groups, or on the opposite, a high proportion of aliphatic hydrophobic groups with few or no polar atoms.

Do the Promiscuous Ligands Have Similar Affinity for Their Different Targets? Promiscuity can be defined as the ability of a compound to exert its effects through multiple biological targets. In high-throughput screening assays, a binding affinity of 10 μM (K_i, IC_{50}) is commonly used to detect hits and therefore to consider a protein as a target. Because X-ray diffraction can observe much lower affinity complexes (mM), we investigated the binding affinity values associated with PDB files in our data set. Upon parsing bindingMOAD[48] and bindingDB[49] databases, binding affinity data could be retrieved for 265 pairs (Figure 9). All but 40 pairs

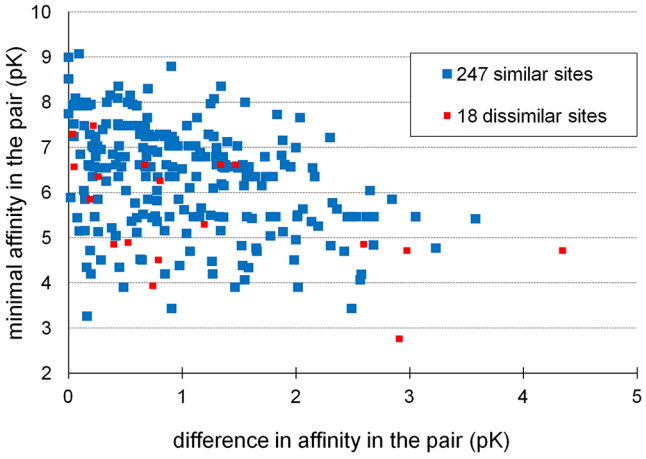


Figure 9. Ligand affinity in 265 pairs of targets for promiscuous ligands. For each ligand/protein complex, affinity represents the average value among pIC_{50}, pEC_{50}, pK_i, and pK_D data retrieved from BindingMOAD and binding databases. Differences in affinity and maximal affinities are expressed in pk units.

met the affinity threshold of 10 μM for the two targets. Although these data are not sufficient to establish robust statistics, it appeared that neither a low affinity for both sites nor a high affinity difference is somehow correlated with the degree of similarity between the corresponding binding sites.

CONCLUSIONS

In the present study, we addressed the issue of ligand promiscuity from a structural point of view. Such an approach was already carried out for a small number of primary metabolites (glucose, nucleotides, heme, estradiol) capable of binding to many different proteins.[34,50] We focused herein on *drug-like* ligands and *druggable* proteins. We proposed the critical analysis of a wide and diverse data set of ligands which are present in PDB complexes with two or more different proteins.

By comparing the different proteins targeted by a ligand at the level of their sequence, structure or binding-site conservation, we demonstrated that ligand promiscuity is either due to the presence of similar binding cavities in different proteins which do not necessarily share other evolutionary relationships (conservation of amino acid or of overall fold) or to specific characteristics of the ligand itself. The conformational flexibility of the ligand frequently explained why discrepancies in size, shape, and physicochemical properties were observed between the binding sites of different targets. Accordingly, we could also observe substantial variations in the number and nature of the ligand atoms in direct interaction with protein. Lastly, we identified a small number of ligand chemotypes that are able to remarkably adapt to different protein environments. In particular, we provided evidence that compounds of low complexity and compounds of very high complexity are prone to bind to dissimilar sites even though their conformation and their chemical moiety bound to proteins are conserved. Noteworthy, natural metabolites were the most promiscuous compounds in the studied data set, thereby suggesting that Nature has developed diverse protein architectures to bind metabolically important ligands ("hubs") like lipids, coenzymes (e.g., thiamine diphosphate participates to many enzymatic reactions like dehydrogenation, decarboxylation or transketolasation), ancient metabolites (e.g., pyridoxal[51]), and widely distributed natural products (e.g., quercetin is present in large quantity in many plants). Last, our findings are consistent with the suggested link between the lipophilicity of a compound and its prosmiscuity[15] and with the importance of three-dimensionality in advancing drug candidates to late clinical stages.[52]

ASSOCIATED CONTENT

Supporting Information

The rules for the biochemical classification of the sc-PDB ligands (Table S1). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Tel.: +333 68 85 42 21. Fax: +333 68 85 43 10. E-mail: ekellen@unistra.fr.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The CC-IN2P3 (Villeurbanne) and GENCI (Project x2011075024) are acknowledged for providing computational resources to this study. We thank the Eskitis Institute at Griffith University for financial support.

ABBREVIATIONS

MAP, Mitogen-activated protein; PDB, Protein Data Bank; rmsd, root-mean-square deviation; shTc, shape Tanimoto coefficient; shTv, shape Tversky coefficient

REFERENCES

(1) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **2007**, *152*, 38−52.

(2) van der Horst, E.; Peironcely, J.; van Westen, G.; van den Hoven, O.; Galloway, W.; Spring, D.; Wegner, J.; van Vlijmen, H.; Ijzerman, A.; Overington, J.; Bender, A. Chemogenomics approaches for receptor deorphanization and extensions of the chemogenomics concept to phenotypic space. *Curr. Top. Med. Chem.* **2011**, *11*, 1964−1977.

(3) Vidal, D.; Garcia-Serna, R.; Mestres, J. Ligand-based approaches to in silico pharmacology. *Methods Mol. Biol.* **2011**, *672*, 489−502.

(4) Campillos, M.; Kuhn, M.; Gavin, A.-C.; Jensen, L. J.; Bork, P. Drug Target Identification Using Side-Effect Similarity. *Science* **2008**, *321*, 263−266.
(5) Kellenberger, E.; Hofmann, A.; Quinn, R. Similar Interactions of Natural Products with Biosynthetic Enzymes and Therapeutic Targets could explain why Nature produces such a Large Proportion of Existing Drugs. *Nat. Prod. Rep.* **2011**, *28*, 1483−1492.
(6) Nicola, G.; Liu, T.; Gilson, M. K. Public Domain Databases for Medicinal Chemistry. *J. Med. Chem.* **2012**, DOI: 10.1021/jm300501t.
(7) Wassermann, A.; Bajorath, J. BindingDB and ChEMBL: online compound databases for drug discovery. *Expert. Opin. Drug. Discov.* **2011**, *6*, 683−687.
(8) Bender, A.; Scheiber, J.; Glick, M.; Davies, J. W.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. L. Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure. *Chem. Med. Chem.* **2007**, *2*, 861−873.
(9) Defranchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan, D. Binding of protein kinase inhibitors to synapsin I inferred from pair-wise binding site similarity measurements. *PLoS One* **2010**, *5*, e12214.
(10) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L. H.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175−181.
(11) Kinnings, S. L.; Liu, N.; Buchmeier, N.; Tonge, P. J.; Xie, L.; Bourne, P. E. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* **2009**, *5*, e1000423.
(12) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Cote, S.; Shoichet, B. K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **2012**, *486*, 361−367.
(13) Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24*, 805−815.
(14) Krejsa, C.; Horvath, D.; Rogalski, S.; Penzotti, J.; Mao, B.; Barbosa, F.; Migeon, J. Predicting ADME properties and side effects: the BioPrint approach. *Curr. Opin. Drug Discov. Devel.* **2003**, *6*, 470−480.
(15) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* **2007**, *6*, 881−890.
(16) Peters, J.-U.; Hert, J.; Bissantz, C.; Hillebrecht, A.; Gerebtzoff, G. g.; Bendels, S.; Tillier, F.; Migeon, J.; Fischer, H.; Guba, W.; Kansy, M. Can we discover pharmacological promiscuity early in the drug discovery process? *Drug Discovery Today* **2012**, *17*, 325−335.
(17) Yang, Y.; Chen, H.; Nilsson, I.; Muresan, S.; Engkvist, O. Investigation of the relationship between topology and selectivity for druglike molecules. *J. Med. Chem.* **2010**, *53*, 7709−7714.
(18) Young, R. J.; Green, D. V. S.; Luscombe, C. N.; Hill, A. P. Getting physical in drug discovery II: the impact of chromatographic hydrophobicity measurements and aromaticity. *Drug Discovery Today* **2011**, *16*, 822−830.
(19) Berman, H. M.; Kleywegt, G. J.; Nakamura, H.; Markley, J. L. The Protein Data Bank at 40: Reflecting on the Past to Prepare for the Future. *Structure* **2012**, *20*, 391−396.
(20) Meslamani, J.; Rognan, D.; Kellenberger, E. sc-PDB: a database for identifying variations and multiplicity of "druggable" binding sites in proteins. *Bioinformatics* **2011**, *27*, 1324−1326.
(21) Kellenberger, E.; Foata, N.; Rognan, D. Ranking targets in structure-based virtual screening of 3-D protein libraries: Methods and Problems. *J. Chem. Inf. Model.* **2008**, *48*, 1014−1025.
(22) The UniProt, C., Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2012**, *40*, D71−D75.
(23) Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Comparing protein−ligand docking programs is difficult. *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 325−332.
(24) Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2006**, *47*, 195−207.
(25) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653−1666.
(26) Bourne, P. E.; Beran, B.; Bi, C.; Bluhm, W. F.; Dimitropoulos, D.; Feng, Z.; Goodsell, D. S.; Prlić, A.; B. Quinn, G.; W. Rose, P.; Westbrook, J.; Yukich, B.; Young, J.; Zardecki, C.; Berman, H. M. The evolution of the RCSB Protein Data Bank website. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 782−789.
(27) Mullan, L. J.; Bleasby, A. J. Short EMBOSS User Guide. European Molecular Biology Open Software Suite. *Brief Bioinform.* **2002**, *3*, 92−94.
(28) Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **1999**, *12*, 85−94.
(29) Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998**, *11*, 739−747.
(30) Desaphy, J.; Azdimoussa, K.; Kellenberger, E.; Rognan, D. Comparison and prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *J. Chem. Inf. Model.* **2012**, DOI: 10.1021/ci300184x.
(31) Schalon, C.; Surgand, J.-S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins: Struct., Funct., Bioinf.* **2008**, *71*, 1755−1778.
(32) Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50*, 123−135.
(33) Gashaw, I.; Ellinghaus, P.; Sommer, A.; Asadullah, K. What makes a good drug target? *Drug Discovery Today* **2012**, *17*, S24−S30.
(34) Stockwell, G. R.; Thornton, J. M. Conformational diversity of ligands bound to proteins. *J. Mol. Biol.* **2006**, *356*, 928−944.
(35) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliver. Rev.* **2001**, *46*, 3−26.
(36) Ursu, O.; Rayan, A.; Goldblum, A.; Oprea, T. I. Understanding drug-likeness. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 760−781.
(37) Wishart, D. S. DrugBank and its relevance to pharmacogenomics. *Pharmacogenomics* **2008**, *9*, 1155−1162.
(38) Galperin, M. Y.; Koonin, E. V. Divergence and Convergence in Enzyme Evolution. *J. Biol. Chem.* **2012**, *287*, 21−28.
(39) Stout, T. J.; Sage, C. R.; Stroud, R. M. The additivity of substrate fragments in enzyme-ligand binding. *Structure* **1998**, *6*, 839−848.
(40) Schüttelkopf, A. W.; Hardy, L. W.; Beverley, S. M.; Hunter, W. N. Structures of Leishmania major Pteridine Reductase Complexes Reveal the Active Site Features Important for Ligand Binding and to Guide Inhibitor Design. *J. Mol. Biol.* **2005**, *352*, 105−116.
(41) Stubbs, M. T.; Reyda, S.; Dullweber, F.; Möller, M.; Klebe, G.; Dorsch, D.; Mederski, W. W. K. R.; Wurziger, H. pH-Dependent Binding Modes Observed in Trypsin Crystals: Lessons for Structure-Based Drug Design. *Chem. Bio. Chem.* **2002**, *3*, 246−249.
(42) Reyda, S.; Sohn, C.; Klebe, G.; Rall, K.; Ullmann, D.; Jakubke, H.-D.; Stubbs, M. T. Reconstructing the Binding Site of Factor Xa in Trypsin Reveals Ligand-induced Structural Plasticity. *J. Mol. Biol.* **2003**, *325*, 963−977.
(43) Jacobs, M. D.; Caron, P. R.; Hare, B. J. Classifying protein kinase structures guides use of ligand-selectivity profiles to predict inactive conformations: Structure of lck/imatinib complex. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 1451−1460.
(44) Atwell, S.; Adams, J. M.; Badger, J.; Buchanan, M. D.; Feil, I. K.; Froning, K. J.; Gao, X.; Hendle, J. r.; Keegan, K.; Leon, B. C.; Müller-Dieckmann, H. J.; Nienaber, V. L.; Noland, B. W.; Post, K.;

Rajashankar, K. R.; Ramos, A.; Russell, M.; Burley, S. K.; Buchanan, S. G. A Novel Mode of Gleevec Binding Is Revealed by the Structure of Spleen Tyrosine Kinase. *J. Biol. Chem.* **2004**, *279*, 55827−55832.
(45) Capaldi, S.; Guariento, M.; Saccomani, G.; Fessas, D.; Perduca, M.; Monaco, H. L. A Single Amino Acid Mutation in Zebrafish (Danio rerio) Liver Bile Acid-binding Protein Can Change the Stoichiometry of Ligand Binding. *J. Biol. Chem.* **2007**, *282*, 31008−31018.
(46) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(47) Selzer, P.; Roth, H.-J. r.; Ertl, P.; Schuffenhauer, A. Complex molecules: do they add value? *Curr. Opin. Chem. Biol.* **2005**, *9*, 310−316.
(48) Benson, M. L.; Smith, R. D.; Khazanov, N. A.; Dimcheff, B.; Beaver, J.; Dresslar, P.; Nerothin, J.; Carlson, H. A. Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res.* **2008**, *36*, D674−D678.
(49) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.
(50) Kahraman, A.; Morris, R. J.; Laskowski, R. A.; Favia, A. D.; Thornton, J. M. On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1120−1136.
(51) Kim, K. M.; Qin, T.; Jiang, Y.-Y.; Chen, L.-L.; Xiong, M.; Caetano-Anollés, D.; Zhang, H.-Y.; Caetano-Anollés, G. Protein Domain Structure Uncovers the Origin of Aerobic Metabolism and the Rise of Planetary Oxygen. *Structure* **2012**, *20*, 67−76.
(52) Lovering, F.; Bikker, J.; Humblet, C. Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *J. Med. Chem.* **2009**, *52*, 6752−6756.

Chapter 3.

Similarity between Flavonoid Biosynthetic Enzymes and Flavonoid Protein Targets Captured by Three-Dimensional Computing Approach

Similarity between Flavonoid Biosynthetic Enzymes and Flavonoid Protein Targets Captured by Three-Dimensional Computing Approach

DOI 10.1055/s-0035-1545697
Planta Med 2015; 81: 467–473

Publisher and Copyright:
© 2015 by
Georg Thieme Verlag KG
Rüdigerstraße 14
70469 Stuttgart
ISSN 0032-0943

Reprint with the permission by the publisher only

Thieme

Similarity between Flavonoid Biosynthetic Enzymes and Flavonoid Protein Targets Captured by Three-Dimensional Computing Approach

Authors Noé Sturm[1,2], Ronald J. Quinn[1], Esther Kellenberger[2]

Affiliations
[1] Eskitis Institute for Drug Discovery, Griffith University, Brisbane, Australia
[2] Laboratory of Therapeutic Innovation, Medalis Drug Discovery Center, Université de Strasbourg, Illkirch, France

Key words
- flavonoid
- biosynthetic enzyme
- natural product
- binding site similarity

received July 6, 2014
revised January 19, 2015
accepted January 23, 2015

Bibliography
DOI http://dx.doi.org/10.1055/s-0035-1545697
Published online February 26, 2015
Planta Med 2015; 81: 467–473

ISSN 0032-0943

Correspondence
Dr. Esther Kellenberger
Laboratory of Therapeutic Innovation
Medalis Drug Discovery Center
UMR 7200 CNRS-University of Strasbourg
74 Route du Rhin
67400 Illkirch
France
Phone: + 33 3 68 85 42 21
Fax: + 33 3 68 85 43 10
ekellen@unistra.fr

Correspondence
Prof. Dr. Ronald James Quinn
Eskitis Institute for Drug Discovery
Griffith University
Innovation Park, 46 Don Young Road
Nathan, Brisbane, QLD 4111
Australia
Phone: + 61 7 37 35 60 00
Fax: + 61 7 37 35 60 01
r.quinn@griffith.edu.au

Abstract

Natural products are made by nature through interaction with biosynthetic enzymes. They also exert their effect as drugs by interaction with proteins. To address the question "Do biosynthetic enzymes and therapeutic targets share common mechanisms for the molecular recognition of natural products?", we compared the active site of five flavonoid biosynthetic enzymes to 8077 ligandable binding sites in the Protein Data Bank using two three-dimensional-based methods (SiteAlign and Shaper). Virtual screenings efficiently retrieved known flavonoid targets, in particular protein kinases. A consistent performance obtained for variable site descriptions (presence/absence of water, variable boundaries, or small structural changes) indicated that the methods are robust and thus well suited for the identification of potential target proteins of natural products. Finally, our results suggested that flavonoid binding is not primarily driven by shape, but rather by the recognition of common anchoring points.

Abbreviations

Bed-ROC:	Boltzmann-enhanced distribution ROCAU
CHI:	chalcone isomerase
CHS:	chalcone synthase
3D:	three-dimensional
DFR:	dihydroflavonol-4-reductase
FBE:	flavonoid biosynthetic enzyme
LAR:	leucoanthocyanidin reductase 1
PDB:	Protein Data Bank
2,3QD:	quercetin-2,3-dioxygenase
RAC:	ras-related C3 botulinum toxin substrate
ROC:	receiver operating characteristics
ROCAU:	receiver operating characteristics area under the curve

Supporting information available online at http://www.thieme-connect.de/products

Introduction

Natural products are chemical compounds synthetized by living organisms. Secondary metabolites are those which are dispensable for survival but give particular species their characteristic features. Secondary metabolites have a broad range of functions, for example, toxins and repellants are used as weapons against prey or predators and attractants are used to attract symbiotic organisms [1]. If they have an extrinsic action on other living organisms, natural products usually disturb an important pathway or trigger a specific biological activity. At the molecular scale, they exert their effect as a drug by interacting with biological macromolecules, especially proteins.

Natural products occupy a diverse chemical space and are involved in a large variety of functions, and therefore represent a rich source of therapeutically useful compounds. Around half of all approved drugs are natural products or their derivatives [2]. Discovery of therapeutic natural products is nevertheless challenging. Extraction, purification, and structure characterization are complex tasks. The determination of potential biological activities is also demanding, requiring many biological assays in a trial and error approach.

Computational approaches have recently been proposed to facilitate the identification of targets for a compound of interest. Ligand-based methods, which are based on the assumption that similar compounds bind to the same target, have

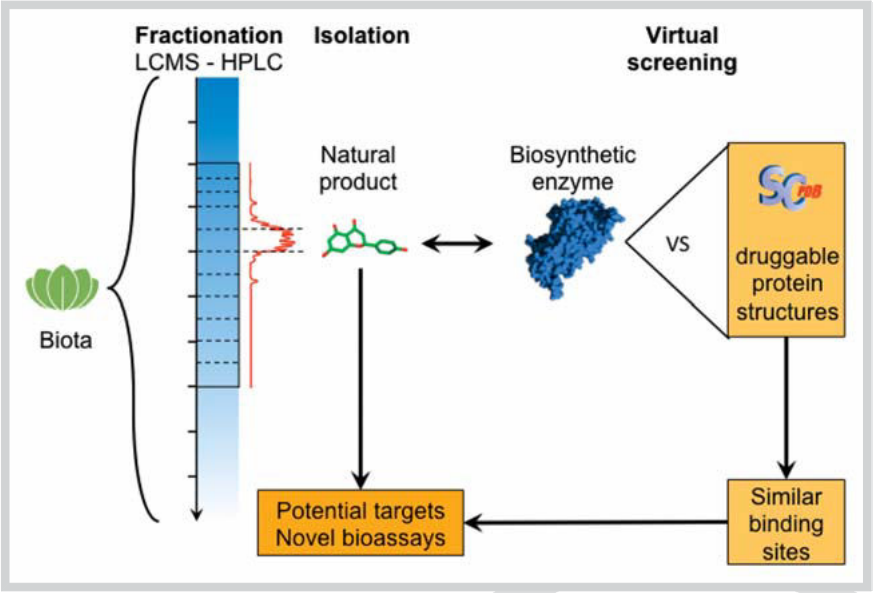


Fig. 1 Ligand-free three-dimensional computing approach to target identification for natural products. (Color figure available online only.)

been successful in drug repositioning and ligand profiling [3]. However, models are predictive only if the biological activity of the explored chemical space is already characterized, thus preventing their application to a novel chemical structure. Structured-based methods in principle circumvent this problem because they interpret the 3D structure of proteins, and do not rely on a training dataset. Docking of a given compound into a series of protein binding sites could efficiently prioritize compounds for experimental testing. A direct comparison of binding sites has also allowed the identification of common ligands of different proteins, assuming that similar binding sites accommodate the same ligand. This second approach is of special interest because it does not depend on a ligand conformational search and gives a robust prediction even if proteins undergo small structural changes [4].

Natural products are made by nature through interaction with biosynthetic enzymes and therefore embed a biological imprint [5,6]. In the present study, we addressed the question "can computing methods find similarity between the active site of biosynthetic enzymes and the binding site of drug targets?". To establish the proof of concept, we focused on flavonoids because different compounds of this class of natural products have been co-crystallized with several biosynthetic enzymes as well as with several protein targets, in particular kinases. The active sites of five different FBEs were used as a query to search the PDB [7] using two different site comparison methods, namely SiteAlign and Shaper (**Fig. 1**).

Results and Discussion

In this study, five different proteins were chosen to represent the family of FBEs: CHS, CHI, 2,3QD, DFR, and LAR from the flowering plant *Medicago sativa* (CHS and CHI), the fungus *Aspergillus japonicus* (2,3QD) and the grape vine *Vitis vinifera* (DFR and LAR). These proteins act on nine different substrates in five different pathways of flavonoid metabolism (**Fig. 1S**, Supporting Information) [8], and, therefore, are expected to constitute a representative panel of the possible modes of flavonoid recognition. In support of this hypothesis, the size and composition in amino acids largely differ in the five enzymes (**Fig. 2**). In addition, active sites in the different enzymes are dissimilar, with a single exception (CHS vs. DFR compared using Shaper, **Table 1S**, Supporting Information). The query dataset contains a total of ten different 3D structures, because CHI, 2,3QD, and DFR enzymes were co-crystallized with up to three different flavonoids (**Table 1**). Of note, all copies of a given protein site were found to be similar despite slight changes in the site definition and description (**Table 1S**, Supporting Information).

The ten FBE active sites were compared to 8077 protein sites which were selected from the PDB according to their predicted ability to accommodate a small molecular weight ligand with high affinity [9]. The searched set of binding sites, from here on called the screening dataset, represents 2379 proteins (as defined by UniProt identifiers [10]) and 967 enzymatic activities (as described by unique Enzyme Commission numbers [11]). Each protein in the screening dataset was annotated as (1) a FBE if it belonged to the set of query proteins, or (2) a flavonoid target if it was crystallized in complex with a flavonoid (**Table 2S**, Supporting Information) or if a micromolar or better affinity for a flavonoid was reported in the ChEMBL database [12] (IC_{50} or $K_i \leq 10\,\mu M$, **Table 3S**, Supporting Information), or (3) a decoy. Among the 71 flavonoid targets identified, kinases were frequently encountered because the screening dataset is highly enriched in kinases (22% of entries) and in protein kinases (77% of the kinases). Also, flavonoids have been suggested to function as anticancer agents due to the inhibition of protein kinases [13–17]. Several types of steroid receptors, phosphodiesterases, and carbonic anhydrases are also targeted by flavonoids.

Site comparisons were performed using two different methods, namely Shaper and SiteAlign [9,18]. A total of 20 virtual screening experiments were analyzed. Overall performances were assessed by plotting ROC curves [19,20]. The x-axis of ROC curve represents the false positive rate, i.e., selectivity. The y-axis of ROC curve represents the true positive rate, i.e., sensitivity. Here we considered that the number of true positives is the count of FBE and flavonoid targets in the selection and the number of false positives the count of decoys in the selection. Random picking in the screening dataset theoretically produces a diagonal line with an area under the curve (ROCAU) equal to 0.5. Whatever the query site and the comparison method, we observed that ranking

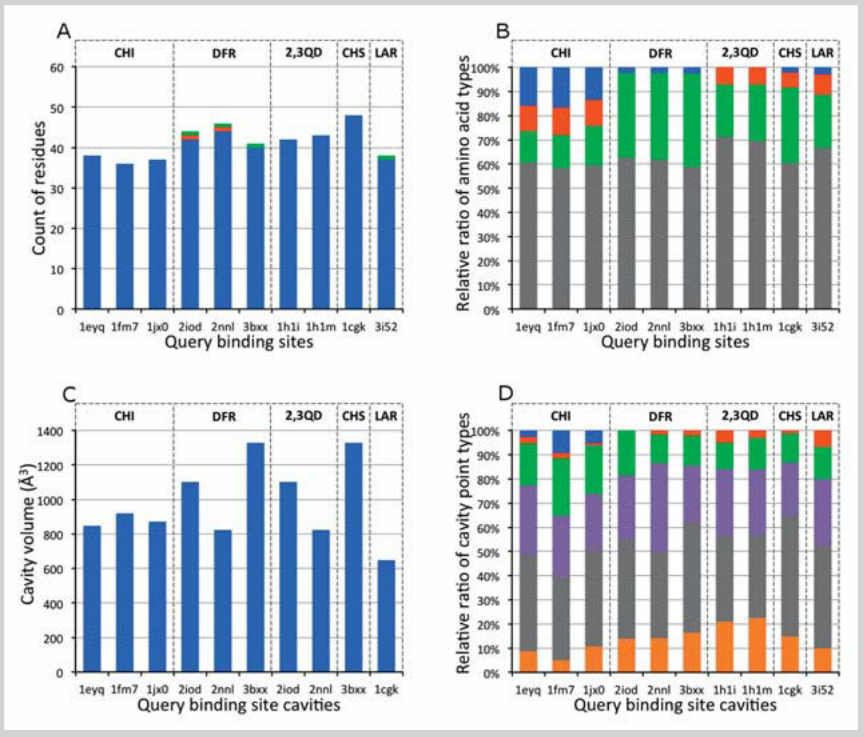


Fig. 2 Description of flavonoid biosynthetic enzyme active sites. **A** Number of amino acids, water molecules, and cofactors in site. Amino acids are colored in blue, water molecules in red, cofactors in green. **B** Composition in amino acids of site. Apolar residues are colored in grey, negatively charged residues in red, positively charged residues in blue, and other polar residues in green. **C** Volume of cavity (Å^3) computed using VolSite. **D** Pharmacophoric description of cavity. Aromatic property is colored in orange, hydrophobic property in grey, hydrogen-bond acceptor in purple, hydrogen-bond donor in green, positive charge in blue, and negative charge in red. (Color figure available online only.)

Table 1 Flavonoid biosynthetic enzymes. Enzyme Commission number indicates the type of reaction catalyzed by the enzyme. UniProt ID is a unique sequence identifier. PDB code is the 3D structure identifier.

Protein *Species*	Enzyme commission	UniProt ID	Ligand name	PDB code
Chalcone isomerase (CHI) *Medicago sativa*	5.5.1.6	CFI1_MEDSA	Naringenin 5-deoxyflavonol 5-deoxyflavonol	1eyq 1fm7 1jx0
Dihydroflavonol-4-reductase (DFR) *Vitis vinifera*	1.1.1.219	P93 799_VITVI	Myricetin Dihydroquercetin Quercetin	2iod 2 nnl 3bxx
Quercetin 2,3-dioxygenase (2,3QD) *Aspergillus japonicus*	1.13.11.24	QDOI_ASPJA	Quercetin Kaempferol	1h1i 1h1 m
Chalcone Synthase (CHS) *Medicago sativa*	2.3.1.74	CHS2_MEDSA	Naringenin	1cgk
Leucoanthocyanidin reductase 1 (LAR) *Vitis vinifera*	1.17.1.3	Q4W2K4_VITVI	(+)-Catechin	3i52

by similarity is significantly better than random picking (**Fig. 3**). The range of ROCAU values was between 0.60 and 0.78 (**Table 4S**, Supporting Information), meaning that predictions were fair to good, respectively.

Comparing methods, we observed that, overall, SiteAlign performed better than Shaper, with ROCAUs in the 0.68–0.78 and 0.60–0.72 ranges, respectively. Since shape superimposition is determinant in predictions made using Shaper while more emphasis is given on pharmacophoric features in SiteAlign, we could postulate that flavonoid binding to flavonoid targets is not primarily driven by shape complementarity, but rather by the recognition of common anchoring points.

For CHI, three 3D structures of the active site were tested as query, yielding almost identical ROC curves and ROCAUs (**Fig. 3**; **Table 4S**, Supporting Information). Consistent results were also obtained for the two screenings using DFR queries, and for the three screenings using 2,3QD queries, further demonstrating that small changes in the size and composition of a query site did not affect the quality of predictions made using SiteAlign and Shaper. Consequently, we concluded that site comparison methods are robust and that there is no quantitative benefit in repeating virtual screening using several similar structures of FBE active site.

To further challenge the methods, we investigated the impact of water molecules on screening results obtained using Shaper (**Table 4S** and **Fig. 2S**, Supporting Information). Noteworthy is that only tightly bound water molecules were included in the sites (more precisely water molecules establishing two or more hydrogen bonds with the protein). FBE sites contained between 0 and 1 water molecules, representing less than 1.3 % of the atoms exposed at the protein site surface. Consequently, water only marginally affected the global description of the query site, with variations in shape and of physicochemical properties being limited to a few spots. These local changes were not sufficient to af-

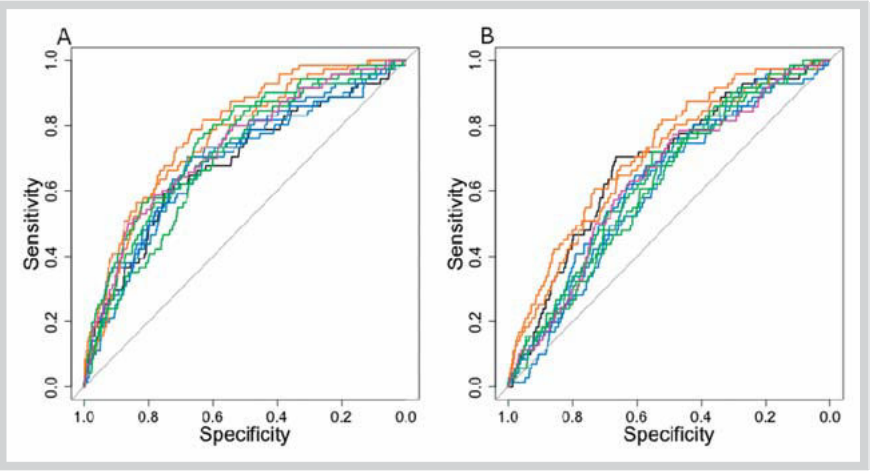


Fig. 3 Receiver operating characteristics curves. **A** SiteAlign. **B** Shaper. Curves are colored according to FBE proteins: CHI in blue, DFR in green, 2,3QD in orange, CHS in black, and LAR in pink. (Color figure available online only.)

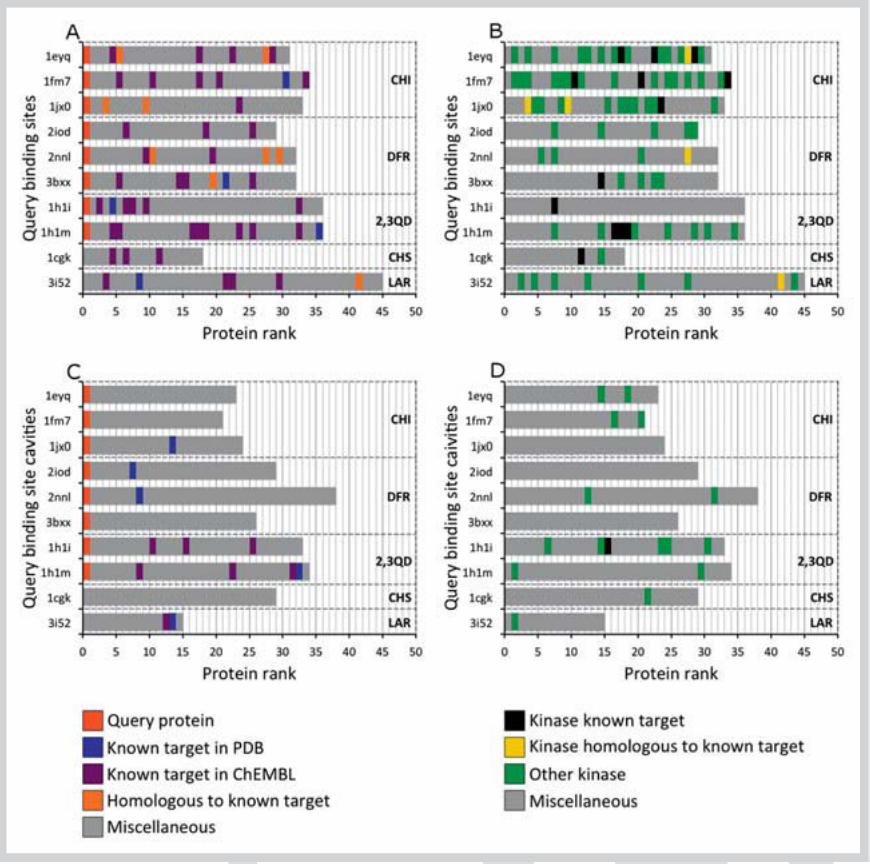


Fig. 4 Composition of hit list. **A** FBE and flavonoid targets in SiteAlign lists. **B** Kinase proteins in SiteAlign hit lists. **C** FBE and flavonoid targets in Shaper lists. **D** Kinase protein in Shaper lists. In **A** and **C**, copies of FBE query are colored in red. Flavonoid targets are colored in blue or purple according to experimental evidence sources (PDB or ChEMBL, respectively). Protein homologs to flavonoid targets are colored in orange. In **B** and **D**, flavonoid targets are colored in black. Kinases homologous to flavonoid targets are colored in yellow. Other kinases are colored in green. (Color figure available online only.)

fect virtual screening results. ROCAU obtained with and without water in the query sites were highly similar.

Given that we aimed at selecting a small number of proteins for experimental testing, methods for virtual screening not only have to be sensitive and selective, i.e., with ROCAUs close to 1, but also have to achieve the early recognition of true targets. Bed-ROC, which increases the weight of true positives in the early fraction of the selection (here the 40 top-ranked entries), indicated that SiteAlign addressed the early recognition of flavonoid targets up to 11 times better than Shaper (**Table 4S**, Supporting Information), as also suggested by the initial slopes of ROC curves (**Fig. 3**). The analysis of ROCAU and Bed-ROC revealed that the ability to discriminate FBE and flavonoid targets from decoys also depends on the query site. Virtual screening experiments using 2,3QD as a query indeed identified the highest number of true positives among top scorers, and exhibited the highest selectivity and sensitivity as well.

In a prospective screening exercise, only top-ranked proteins are submitted for experimental validation. We therefore analyzed hit lists obtained in the retrospective screening exercises. Hit lists were built assuming that similarity is significant if it differs by more than 2.5 standard deviations from the mean value of the distribution of scores. All distributions of scores were unimodal and could be approximated to the normal distribution with a slight skew on the tails (**Fig. 3S-6S**, Supporting Information). All 20 hit lists had relatively small and consistent sizes (between 18

and 45 using SiteAlign, and between 15 and 38 using Shaper, see ◉ **Fig. 4**). A few nonselective flavonoid targets were found in several hit lists. Steroid receptors were present in all SiteAlign lists. These proteins have promiscuous binding sites [21]. For example, human peroxisome proliferator-activated receptor γ [22] was found in seven different hit lists (SiteAlign combined with CHI or 2,3QD, Shaper combined with CHI, DFR, or LAR). Carbonic anhydrase 2 [23] was also frequently encountered in hit lists.

Detailed analysis of each hit list showed that the composition was characteristic of each FBE screening. We especially observed FBE-specific flavonoid targets, thereby suggesting that there is not a single flavonoid imprint across the FBE family. Some flavonoid targets were found in only one FBE query. For example, human RAC-α serine/threonine protein kinase [24], human mitogen-activated protein kinase 1 [25], and human phosphatidylinositol 4,5-biphosphate 3-kinase catalytic subunit γ isoform [17] were only present in CHI hit lists. Many kinases, and more specifically serine/threonine protein kinases, were actually present in CHI hit lists, but not in other hit lists (◉ **Fig. 4 B, D**). The flavonoid biological imprint embedded in CHI thus constituted a good bait to identify kinases which potentially bind flavonoids. CHI is involved in the formation of the isoflavan scaffold by catalyzing ring closure on chalcone substrates, and thus may retain an imprint of the complete isoflavan scaffold (**Fig. 1S**, Supporting Information). In addition, the active site composition in CHI differs from that in other FBEs. Especially CHI, like the kinases retrieved from the screening dataset, contains more charged residues than other FBEs (◉ **Fig. 2**).

Considering that all the proteins homologous to flavonoid targets in the SiteAlign hit lists are putative true positives, the performance of retrospective screenings was probably underestimated. For example, proto-oncogene tyrosine-protein kinase Src from both humans and chickens [24] were present in the CHI hit list (1eyq), while only the human enzyme was marked as a flavonoid target. Androgen receptors from both humans and chimpanzees were identified in the CHI hit list (1eyq), while only the human enzyme was marked as a flavonoid target.

Finally, we asked the question "can similarity score be interpreted into common structural features?". To that end, we displayed the 3D alignment for a selection of similar pairs and observed that secondary structure elements are well superimposed although the protein global 3D structures are different. As shown on ◉ **Fig. 5**, the active site of CHI is formed by α1 and α2 helices and a β1 three-stranded sheet and β2 strand. The similar binding site in RAC-α serine/threonine protein kinase is made of α3 and α4 helices that well superimpose to α1 and α2 in CHI. In addition, the β3 three-stranded sheet and α5 helix in the kinase well match β1 and β2 in CHI. Interestingly, secondary structure elements with a conserved position in space do not necessarily match secondary structure elements of the same type, as illustrated by the superimposition of the β2 strand from CHI to the α5 helix in the kinase.

In this retrospective study, we were able to use FBE as bait to retrieve flavonoid targets from a large set of ligandable proteins. Protein similarity based on shape (Shaper) returned hit lists with up to 14.7% of flavonoid targets. We demonstrated that shape-based similarity is not the method of choice, especially with promiscuous natural products in particular flavonoids. In this study, protein similarity based on molecular anchoring points (SiteAlign) returned hit lists containing up to 27% of flavonoid targets. SiteAlign successfully identified alternate domains of a helix and a β-sheet as possible equivalent anchoring points. The diversity of flavonoid targets and other proteins retrieved using different FBE queries suggested that the biological imprint gained during biosynthesis of natural products is unique to each biosynthetic enzyme (here, FBE) rather than there being a single unique flavonoid biological imprint across the FBE family. All FBE queries retrieved known flavonoid targets as well as a set of non-related flavonoid targets. This methodology promises to deliver non-related flavonoid targets as an enriched bioassay screening set.

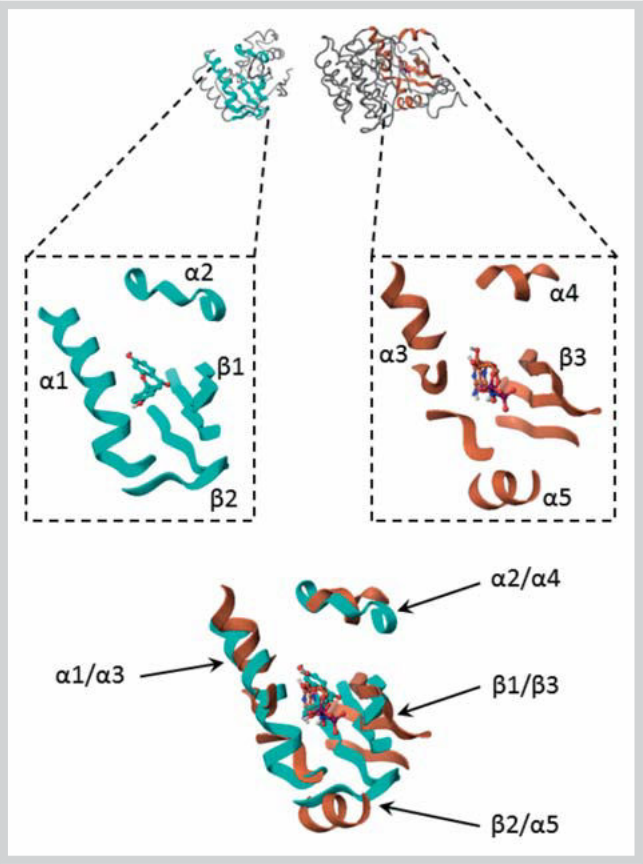


Fig. 5 Three-dimensional alignment of sites in chalcone isomerase and Ras-related C3 botulinum toxin substrate-α serine/threonine protein kinase. The active site of CHI (pdb code: 1fm7) is represented by cyan ribbons and the ATP-binding site of RAC-α serine/threonine protein kinase (pdb code: 4ekk) by orange ribbons. Ligands are rendered with a ball and stick. Sites were aligned using SiteAlign. (Color figure available online only.)

Material and Methods

Three-dimensional structures of protein binding sites

FBEs and the screening dataset were extracted from the 2012 release of the sc-PDB database [26]. The sc-PDB provides an all-atom description of complexes between a small molecular weight ligand and a ligandable protein, which includes all protein chains, metal ion(s), cofactor(s), and water molecule(s) (establishing at least two hydrogen bonds with the protein chains) in the vicinity of the ligand. For each protein, the binding site was defined as all protein residues delimiting the cavity detected using Volsite [9] and with at least one heavy atom distant from less than 6.5 Å from any ligand heavy atom. Last, we verified that the FBE active site was consistent with the amino acid sequence of the native protein as described in the UniProt database [10].

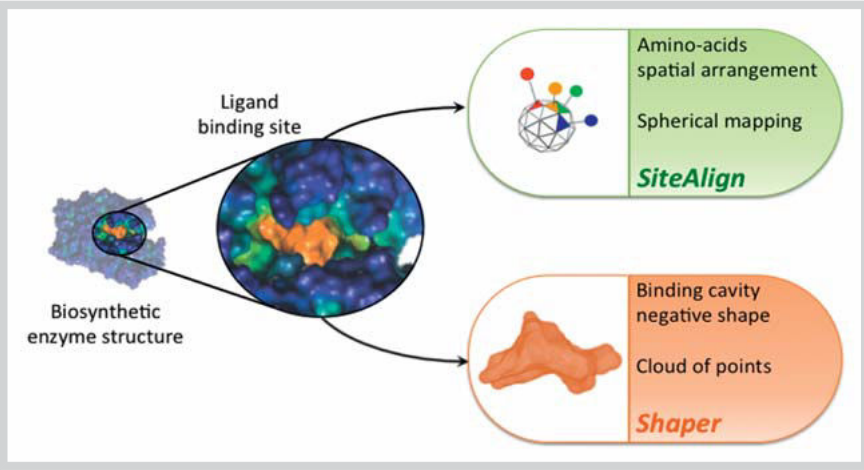


Fig. 6 Principle of protein binding sites comparison in SiteAlign and Shaper. (Color figure available online only.)

Binding site comparison

Site similarity was evaluated using two programs based on different methods, SiteAlign [18] and Shaper [9] (**Fig. 6**). Briefly, SiteAlign represents a binding site with an 80-triangle polyhedron centered on the protein cavity. Physicochemical properties of binding site amino acids are projected onto triangles of the polyhedron (cofactors, metal ions, and water molecules are ignored). Null property is assigned to triangles not hit by the projection of an amino acid. Binding sites are aligned by optimizing the superimposition of two polyhedrons for the best match of physicochemical properties. SiteAlign quantifies site similarity using two distances, whether considering all matched triangles (*D1* score) or only matched triangles with non-null properties in the two polyhedrons (*D2* score).

In the present study, the *D1* score was used as a filter; two sites were dissimilar if *D1* was lower than 0.6. The *D2* score was used to rank solutions.

Shaper represents the negative image of a binding site, including amino acids, cofactor(s), and water molecule(s); 1.5 Å-spaced grid points filling the cavity are annotated with pharmacophoric properties of the nearest protein atoms. Binding sites are aligned by maximizing the geometric overlap of grids. Shaper quantifies site similarity by computing the proportion in the query site of the grid points with position and properties common to that in the compared site (*RefTversky* score).

Virtual screening

FBE active sites were compared to all the 8077 entries of the sc-PDB using Shaper and SiteAlign. Each screening experiment yielded a ranked list of 8076 binding sites, sorted by decreasing similarity to the query. For a given query, a hit list was obtained by selecting all proteins with at least one copy having a similarity score better than the mean of the distribution plus 2.5 standard deviations.

ROCAUs were computed using the package *pROC* [27] in R. Bed-ROC values were computed using the package *enrichvs* in R. The alpha coefficient for Bed-ROC was set to 200.

Supporting information

Tables showing the similarity between active sites of FBEs, sc-PDB proteins in a complex with a flavonoid, proteins with a micromolar or better affinity for flavonoids, as well as ROCAU and Bed-ROC values are available as Supporting Information. Also, figures displaying the biosynthetic reactions catalyzed by FBEs, ROC curves for site comparison using Shaper, distribution of SiteAlign distances, as well as SiteAlign score and Shaper similarity score distributions can be found in this section.

Acknowledgements

The Calculation Center of the IN2P3 (CNRS, Villeurbanne, France) is acknowledged for allocation of computing time.

Conflict of Interest

The authors declare no conflict of interest.

References

1 *Demain AL, Fang A.* The natural functions of secondary metabolites. Adv Biochem Eng Biotechnol 2000; 69: 1–39
2 *Newman DJ, Cragg GM.* Natural products as sources of new drugs over the 30 years from 1981 to 2010. J Nat Prod 2012; 75: 311–335
3 *Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, Jenkins JL, Lavan P, Weber E, Doak AK, Cote S, Shoichet BK, Urban L.* Large-scale prediction and testing of drug activity on side-effect targets. Nature 2012; 486: 361–367
4 *Kellenberger E, Schalon C, Rognan D.* How to measure the similiarty between protein ligand-binding sites. Curr Comput Aided Drug Des 2008; 4: 209–220
5 *McArdle BM, Campitelli MR, Quinn RJ.* A common protein fold topology shared by flavonoid biosynthetic enzymes and therapeutic targets. J Nat Prod 2006; 69: 14–17
6 *Kellenberger E, Hofmann A, Quinn RJ.* Similar interactions of natural products with biosynthetic enzymes and therapeutic targets could explain why nature produces such a large proportion of existing drugs. Nat Prod Rep 2011; 28: 1483–1492
7 *Gutmanas A, Alhroub Y, Battle GM, Berrisford JM, Bochet E, Conroy MJ, Dana JM, Fernandez Montecelo MA, van Ginkel G, Gore SP, Haslam P, Hatherley R, Hendrickx PM, Hirshberg M, Lagerstedt I, Mir S, Mukhopadhyay A, Oldfield TJ, Patwardhan A, Rinaldi L, Sahni G, Sanz-Garcia E, Sen S, Slowley RA, Velankar S, Wainwright ME, Kleywegt GJ.* PDBe: Protein Data Bank in Europe. Nucleic Acids Res 2014; 42: D285–D291
8 *Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Pujar A, Shearer AG, Travers M, Weerasinghe D, Zhang P, Karp PD.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 2012; 40: D742–D753
9 *Desaphy J, Azdimousa K, Kellenberger E, Rognan D.* Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. J Chem Inf Model 2012; 52: 2287–2299

10 *Consortium TU.* Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res 2014; 42: D191–D198
11 *Bairoch A, Apweiler R.* The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000; 28: 45–48
12 *Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Kruger FA, Light Y, Mak L, McGlinchey S, Nowotka M, Papadatos G, Santos R, Overington JP.* The ChEMBL bioactivity database: an update. Nucleic Acids Res 2014; 42: D1083–D1090
13 *Sak K.* Site-specific anticancer effects of dietary flavonoid quercetin. Nutr Cancer 2014; 66: 177–193
14 *Peer WA, Murphy AS.* The science of flavonoids. In: Grotewold E, editor. Flavonoids as signal molecules: targets of flavonoid action. New York: Springer; 2006: 239–268
15 *Lu X, Jung J, Cho HJ, Lim DY, Lee HS, Chun HS, Kwon DY, Park JH.* Fisetin inhibits the activities of cyclin-dependent kinases leading to cell cycle arrest in HT-29 human colon cancer cells. J Nutr 2005; 135: 2884–2890
16 *Havsteen BH.* The biochemistry and medical significance of the flavonoids. Pharmacol Ther 2002; 96: 67–202
17 *Walker EH, Pacold ME, Perisic O, Stephens L, Hawkins PT, Wymann MP, Williams RL.* Structural determinants of phosphoinositide 3-kinase inhibition by wortmannin, LY294002, quercetin, myricetin, and staurosporine. Mol Cell 2000; 6: 909–919
18 *Schalon C, Surgand JS, Kellenberger E, Rognan D.* A simple and fuzzy method to align and compare druggable ligand-binding sites. Proteins 2008; 71: 1755–1778
19 *Swets JA, Dawes RM, Monahan J.* Better decisions through science. Sci Am 2000; 283: 82–87
20 *Hawkins PC, Warren GL, Skillman AG, Nicholls A.* How to do an evaluation: pitfalls and traps. J Comput Aided Mol Des 2008; 22: 179–190
21 *Sturm N, Desaphy J, Quinn RJ, Rognan D, Kellenberger E.* Structural insights into the molecular basis of the ligand promiscuity. J Chem Inf Model 2012; 52: 2410–2421
22 *Puhl AC, Bernardes A, Silveira RL, Yuan J, Campos JL, Saidemberg DM, Palma MS, Cvoro A, Ayers SD, Webb P, Reinach PS, Skaf MS, Polikarpov I.* Mode of peroxisome proliferator-activated receptor gamma activation by luteolin. Mol Pharmacol 2012; 81: 788–799
23 *Ekinci D, Karagoz L, Ekinci D, Senturk M, Supuran CT.* Carbonic anhydrase inhibitors: *in vitro* inhibition of alpha isoforms (hCA I, hCA II, bCA III, hCA IV) by flavonoids. J Enzyme Inhib Med Chem 2013; 28: 283–288
24 *El Amrani M, Lai D, Debbab A, Aly AH, Siems K, Seidel C, Schnekenburger M, Gaigneaux A, Diederich M, Feger D, Lin W, Proksch P.* Protein kinase and HDAC inhibitors from the endophytic fungus *Epicoccum nigrum*. J Nat Prod 2014; 77: 49–56
25 *Tasdemir D, Mallon R, Greenstein M, Feldberg LR, Kim SC, Collins K, Wojciechowicz D, Mangalindan GC, Concepcion GP, Harper MK, Ireland CM.* Aldisine alkaloids from the Philippine sponge *Stylissa massa* are potent inhibitors of mitogen-activated protein kinase kinase-1 (MEK-1). J Med Chem 2002; 45: 529–532
26 *Desaphy J, Bret G, Rognan D, Kellenberger E.* sc-PDB: a 3D-database of ligandable binding sites – 10 years on. Nucleic Acids Res 2015; 43: D399–D404
27 *Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011; 12: 77–84

Chapter 4.

Inventory of Natural Product Biosynthetic Enzymes from the Protein Databank

INTRODUCTION

In the 1980s and 1990s, only a limited number of resources for protein information was available. Pioneering databases such as the protein identification resource[1], Swiss-Prot[2] and DNA data bank of Japan[3] provided high quality protein annotation with experimental observations. Since then, significant advance in genome sequencing has dramatically increased genomic data.[4] As a result, myriads of databases have emerged. EcoCyc database[5] (originally focusing on *Escherichia coli*), Saccharomyces genome database[6] or the human genome database[7] all provided data for specific genomes. The increase of genomic data has been accompanied by significant progresses of bioinformatic tools for gene prediction which opened the era of proteomics. In response to the plethora of existing databases, collaborative projects have emerged to centralize the data. Today, the main proteomic collaboration is the UniProt consortium, which is composed of the European Bioinformatics Institute, the Swiss Institute of Bioinformatics and the Protein Information Resource (formerly known as the protein identification resource). It aims at a comprehensive integration of reliable protein sequences with high quality information on protein features such as protein function, domain structure, redundancy across proteomes and many more. The UniProt database[8] currently contains more than 80 million protein sequences. Since it started, in 1986, more than half a million of protein sequences have been manually annotated/verified (Swiss-Prot) whereas other protein sequences, generated at a higher pace, are computationally annotated (TrEMBL). In the meantime, other initiatives have been focusing on enzymes functional data. Characterization of

enzymes plays an essential role in the study of cellular machineries. In particular, understanding how cells are regulated by enzymes has a tremendous potential in the field of disease understanding and treatment. For example, pathologies are often related to miss-regulated signaling pathways. Identification of their origins, validation of target proteins and development of drug molecules require the knowledge of each metabolic steps involved in the disease. The Braunschweig Enzyme Database[9] (BRENDA), the Kyoto Encyclopedia of Genes and Genomes[10] (KEGG) and MetaCyc[11] appear amongst the pioneering databases providing metabolic information. These resources integrate protein sequences with enzymatic activities and thus, assign biochemical reactions to enzymes. BRENDA contains abundant data extracted from literature for more than 77 000 enzymes. The database is also renowned for its classification of enzymes according to the Enzyme Commission number.[12] For better understanding of enzymatic steps imbrication in metabolic processes, KEGG and MetaCyc provide manually drawn metabolic networks. Since more recently, other resources aim at helping the scientific community to understand metabolism. For example, human metabolic networks are available in the Reactome knowledge base,[13] a community-based annotation project. Besides the UniProt consortium is now also integrating metabolic information using the small molecule ontology described in ChEBI[14], manually curated biochemical reactions available in Rhea[15] and hierarchical representation of metabolic pathways provided by UniPathway[16].

Small molecules involved in cells metabolism are the functional ends of genes. These molecules are synthesized by nature in a process that is often termed biosynthesis. The biosynthesis of a small molecule is composed of multiple enzyme-catalyzed reactions

where substrate molecules are converted into more complex molecules. Yet biosynthesis supporting essential functions of infectious organisms have been used as drug targets. For example, penicillins inhibit the synthesis of cross-links essential to bacterial cell walls.[17] But biosynthetic enzymes also constitute interesting proteins in the field of natural products drug discovery. In particular, they provide pharmaceutical industry potential ways to produce complex compounds in large quantities or to synthesize new natural products. For example, fungal polyketide synthase has been modified to reprogram it production and thus explore new product variants.[18] To these extents, biosynthetic enzyme structures play an important role. More recently, the initiative Natural Product Biosynthesis (NatPro) was established in order to reveal biosynthetic enzyme structures related to human health and disease. So far, the results show some 64 structures (http://www.natprobio.org/). Determination of biosynthetic enzyme structures has also been a field of interest over that last decades within the scientific community. As for all publicly available protein structures, biosynthetic enzyme structures were all submitted into the Protein Data Bank[19] archive (PDB) which has collected over hundred thousands of structures since creation.

Since there are no comprehensive databases dedicated to natural product biosynthetic enzyme structures we had to mine the PDB. In practice, we have tested two approaches. First, we searched the PDB following a top-down approach using keywords as a filter and secondly, we performed a knowledge-based approach using metabolic data provided in MetaCyc and in UniProt (**figure 1**).

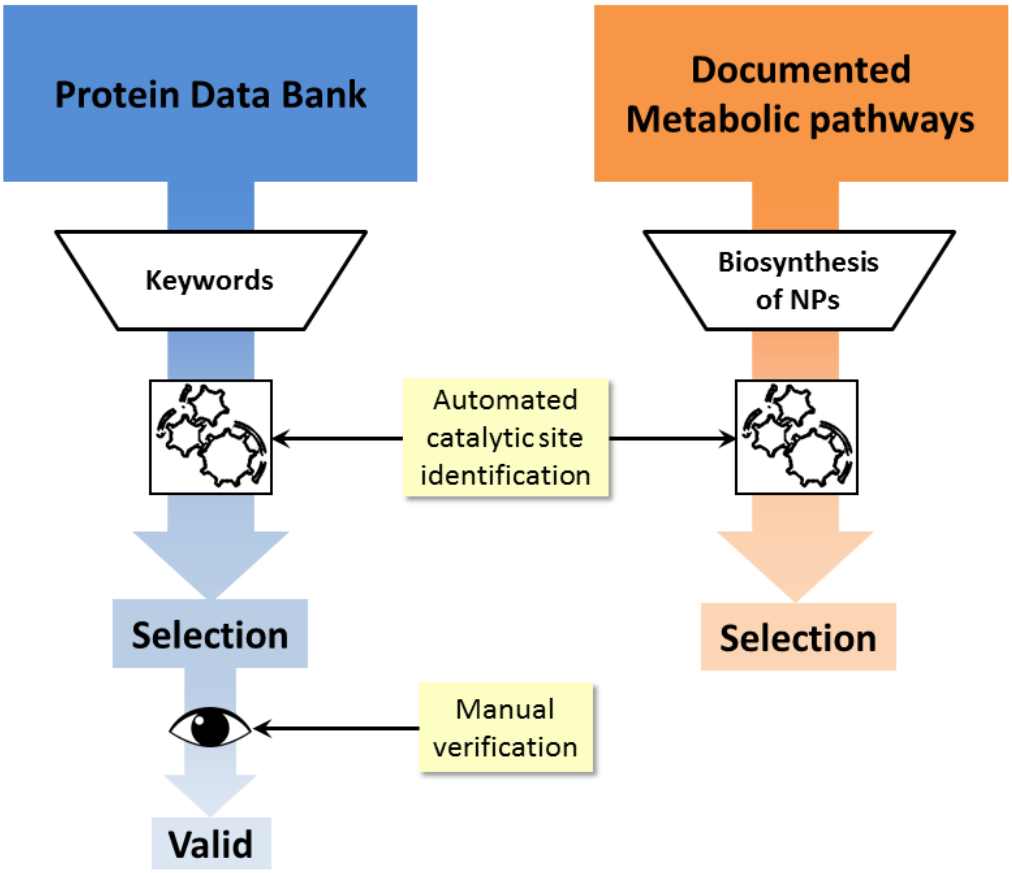


Figure 1. Overall process of top-down and knowledge-based strategy.
On the left side and colored with blue is represented the overall workflow of the top-down strategy. It starts by filtering structures of the protein data bank using keywords. An automated catalytic site identification step allows to obtain a selection of proteins with ligandable catalytic sites. This selection is ultimately verified by manual inspections using documentation of natural products biosynthetic pathways from knowledge-based databases. On the right side and colored with orange is represented the overall workflow of the knowledge-based strategy. Natural product biosynthetic pathways are selected from UniProt and MetaCyc knowledge-based databases of metabolic pathways. Relative enzymes are passed into the automated catalytic site identification process, which returns a selection of biosynthetic enzymes with ligandable catalytic sites.

1. METHODS AND MATERIALS

1.1. Knowledge-based strategy for collecting the biosynthetic enzymes

1.1.1. Overall flowchart

In the knowledge-based strategy, we collected biosynthetic enzymes of natural products based on metabolic data elaborated by experts. We searched two high-quality resources freely available on internet, namely UniProt[8] and MetaCyc.[11] We directly extracted and linked protein names in metabolic database with protein structures in the Protein Data Bank.[19] Structure files were then downloaded and submitted to an automated process for the identification of catalytic sites. The ligandability of all cavities containing catalytic residues was then assessed.

UniProt and MetaCyc were investigated independently and results were pooled together, while removing duplicates. Noteworthy all collected structures are assigned information on the enzyme as given in the two source databases.

1.1.2. Searching UniProt

The UniProt consortium's database provides manually curated documentation on metabolic and biosynthetic pathways. The complete documentation of a pathway describes all known enzymes and their catalyzed reactions.

We selected a subset of pathways that we assumed to be representative of the natural products biosynthesis. More precisely, we selected pathways related to antibiotics, terpenes, steroids, phenylpropanoids, alkaloids, polyketides and pigments (a detailed list is

provided in **annex 4, table 1**). We ignored the so-called miscellaneous pathways, which includes all pathways related to primary metabolites (such as amino-acids, carbohydrates, cofactors, or nucleotides), proteins, cell-wall constituents as well as all degradation pathways (a detailed list is provided in **annex 4, table 2**). From thereon, all proteins in miscellaneous pathways are called miscellaneous proteins. Of note, the sum of two lists do not reflect current knowledge on metabolic pathways, because pathways involving enzymes with known structures are considered only.

In practice, the list of all documented biosynthetic and metabolic pathways was downloaded as text file from http://www.uniprot.org/docs/pathway.txt (release 2015_08, 22nd July 2015). This list indexes pathways to their associated enzymes. We filtered the list using our selection of pathways and collected for each remaining enzyme its Uniprot accession codes, which in turn were used to find related PDB accession codes. The UniProt-PDB correspondence was made using the entry mapping summary (release of July 2015) provided by the structure integration with function, taxonomy and sequence[20] (SIFTS initiative).

Catalytic activities of selected enzymes were searched at the -!- CATALYTIC ACTIVITY lines of the comment section in UniProt protein description files. Catalytic activity lines contain a description of the enzymatic reaction including substrates, cofactors and product molecule names following the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) as published in Enzyme Nomenclature.[12] Chemical structures of compounds involved in the reactions were collected from the database ChEBI[14] using the nomenclature of the IUBMB.

1.1.3. Searching MetaCyc

MetaCyc is a knowledge-based database of experimentally elucidated metabolic pathways. It provides information on reactions, enzymes, genes, and species amongst others. We focused our analysis on all pathways taking place in the secondary metabolites biosynthesis, assuming that secondary metabolites are natural products (i.e natural products are a.k.a secondary metabolites). Thus, we considered the "secondary metabolites biosynthesis" section of the database. Of note, some pathways in this section are also present in other sections (unrelated to the biosynthesis of natural products) of the database. From thereon, all discarded pathways and corresponding enzymes are called miscellaneous pathways and miscellaneous proteins.

In practice, the metabolic and biosynthetic pathways provided by MetaCyc database are arranged in a hierarchical tree. In order to collect biosynthetic enzymes of natural products, we inspected all pathways under the "biosynthesis of secondary metabolites" branch in July 2015. We wrote a sequence of scripts that performed the steps illustrated in **figure 2**. The complete list of secondary metabolites pathways was retrieved using BioCyc REST-based web service and a recursive algorithm scanned all children pathways of the "SECONDARY-METABOLITE-BIOSYNTHESIS" node to obtain all pathway identifiers in the considered sub-branches of the tree.

The selected pathway identifiers were then used to collect gene identifiers. Using gene identifiers, we accessed the gene descriptions pages of the website (http://metacyc.org/) and looked for UniProt protein accession codes. Related enzyme structures were then obtained using the entry mapping summary provided by the SIFTS initiative.

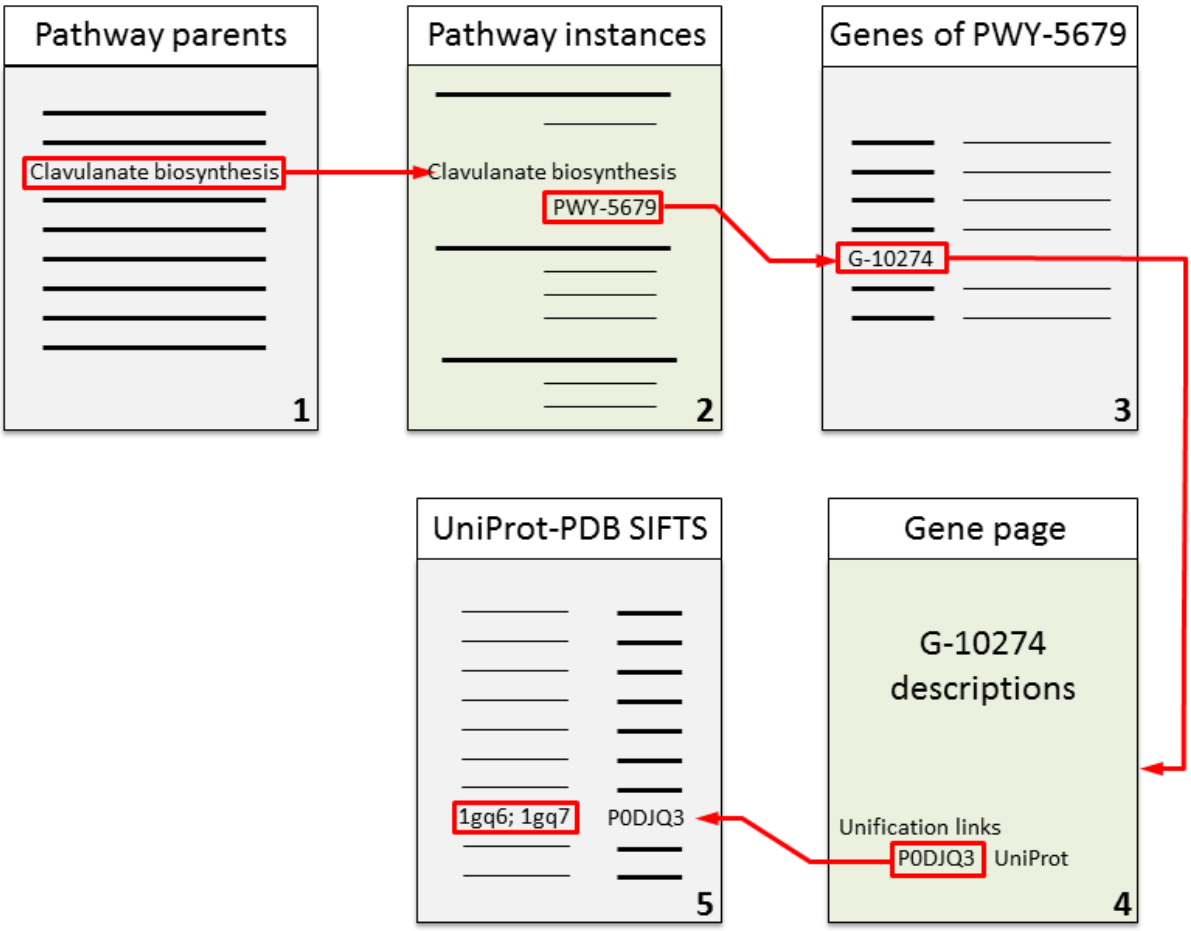


Figure 2. Biosynthetic enzymes collection process from MetaCyc database.
The figure illustrates the search of clavulanate biosynthetic enzymes structures. Green boxes represent web based pages whereas grey boxes represent text files. 1/ Complete list of secondary metabolites biosynthesis pathway parents in the hierarchical classification was returned. 2/ All pathway instances are scanned recursively from the list of pathway parents. 3/ Genes taking place in pathway instances are searched. 4/ Gene description pages are parsed to obtain related protein accession code in unification links. 5/ SIFTS entry mapping summary is used to obtain related enzyme PDB structures. Generic query URLs are provided in **annex 4, table 3**.

MetaCyc gives the detailed chemical structure of substrates and products for each enzymatic reaction described in a pathway. We collected chemical information as follows: Enzymatic activities were searched using reaction identifiers associated to the genes; Reaction identifiers in turn allowed us to parse reaction description pages provided by BioCyc web service; The reaction description pages provide the identifiers of the compound page which contain smiles structure of the compounds (**figure 3**).

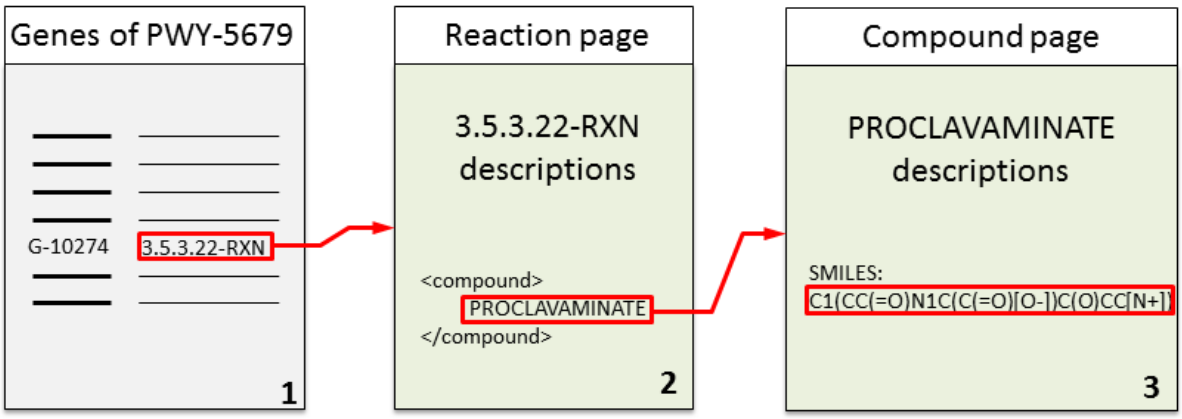


Figure 3. Collection of the chemical structure of substrates and products in MetaCyc.
The figure illustrates the search for proclavaminate smiles structure. Green boxes represent web based pages whereas grey boxes represent text files. 1/ Reaction identifiers of related genes are searched. 2/ Reaction identifiers are used to load reaction description pages, which are parsed to obtain compound identifiers. 3/ Compound identifiers are used to load compound description page that provides smiles structure. Generic query URLs are provided **in annex 4, table 4.**

1.2. Top-down strategy for collecting the biosynthetic enzymes

1.2.1. Overall flowchart

In this approach, we directly explored the PDB archive (www.rcsb.org) in October 2014. A first filter selected entries maching keywords related to the biosynthesis of natural products. In a second step, we discarded all structures that have not been solved by X-ray crystallography or that obviously not describe a biosynthetic enzyme of natural product. After protein annotation, structure files were submitted to an automated process for the identification of catalytic site. The ligandability of all cavities containing catalytic residues was then assessed. Lastly, each entry was validated, or discarded, based on manual checks with enzymatic reaction data found in UniProt, MetaCyc and in the literature.

1.2.2. Step 1: text-mining

For each PDB entry, we created a textual data file containing: keywords, structure title, article title and literature reference(s) found in PDB file header lines tagged with KEYWDS, TITLE and JRNL respectively. In addition, we added the full content of the publication abstract recorded from PubMed (http://www.ncbi.nlm.nih.gov/pubmed) using the extracted journal reference(s). Each textual data file was then searched for four text motifs: "*biosynth"*, "*natural product"*, *"secondary metabol"* and *"plant defense".* PDB entries that did not match any of the four text motifs were discarded from the pool of entries.

1.2.3. Step 2: primary metabolism filtering

In order to discard protein structures obviously involved in miscellaneous metabolic and biosynthetic pathways, we used the annotation provided by "-!- PATHWAY" lines of UniProt protein description files. Any protein structure associated to a miscellaneous pathway was removed from the pool of entries. Of note, at this stage, we also conserved all proteins without pathway annotation or not in the miscellaneous pathways.

1.2.4. Step 3: protein annotation

Protein chains in each PDB file were annotated with Uniprot identifiers, recommended protein names, gene names, species and EC numbers found in UniProt protein description files (see section 1.3.1 on page 117).

1.2.5. Last step: manual validation of entries

We used EC numbers to search MetaCyc website. In practice, we extracted from the database all enzymatic reactions containing the EC number of the collected proteins and searched corresponding chemical compounds (substrates and products), genes and species. Four criteria were considered to establish a link between a metabolic reaction and a protein structure: 1/ a partial or exact match of gene names; 2/ a strong evolutionary relationship between species (assumed when species fall within the same phylogenic branch, see **figure 1** in **annex 4**); 3/ manual validation of the enzymatic reaction as part of secondary metabolites biosynthesis; 4/ the presence of an enzymatic activity description in UniProt. Last, we manually validated relevant enzymatic activities in the context of the study by analyzing individually chemical compound structures or enzyme names. In particular, we favored reactions involving compounds containing the molecular scaffold of the end product in the biosynthetic pathway.

1.3. Automated process for catalytic site identification

This process includes annotation of each protein chain, mapping of catalytic residues in the structure file, detection of all protein ligandable cavities and selection of the catalytic cavity. The same process is applied to the two collection approaches.

1.3.1. Protein annotation

Description report of each structure was accessed programmatically via the RESTful web service of the RCSB (July 2015). Obsolete entries and entries without Uniprot accession codes (e.g., structure of nucleic acids) were systematically discarded. For each protein, data in RCSB report were compared to the description provided by UniProt consortium. Provided an exact match of both the protein name and EC numbers, the protein in structure file was annotated with recommended name, protein identifiers and accession codes, gene name, species and pathway names found in Uniprot protein file.

If a structure file contains two or more proteins, each protein was assigned its own annotations as described above. Nevertheless, green fluorescent proteins and other fusion found in chimera were systematically ignored. We also ignored house-keeping proteins such as ribosome constituents. Proteins were further considered only if they belong to the set of known biosynthetic enzymes.

Next, we looked for catalytic residues in each protein. In order to identify them, we retrieved the number of all residues in ACT_SITE lines of UniProt files. If no information was found in the UniProt file, we searched the catalytic site atlas[21] (CSA). CSA is available as a flat file with one catalytic residue per line (the file is accessible at: https://www.ebi.ac.uk/thornton-srv/databases/CSA/Downloads.php). A catalytic motif is

made of several residues observed in the enzyme structure, is represented by a PDB accession code and the chain identifier and sequence number in the PDB file. CSA can provide multiple motifs for a protein. In such a case, we assigned the different motifs to the protein. Protein structures with documented catalytic motifs were discarded from the pool of entries.

1.3.2. UniProt-to-PDB mapper

A catalytic residue number retrieved from UniProt[1] represents the position of the amino acid in the full length precursor protein sequence, which is generally derived from genomic data. Unfortunately, this number does not necessarily match that in PDB structure files of the same protein. We have designed an in-house UniProt-to-PDB mapper to renumber UniProt residues according to PDB numbering scheme. The operation is performed by aligning the amino-acid sequence of the protein structure to the UniProt amino-acid sequence. The amino-acid sequence from the structure was built following the SEQRES section of PDB files. Sequence alignment then proceeds chain by chain using the global sequence alignment algorithm[22] (Needleman-Wunsch) implemented in the EMBOSS package.[23] If the structure contained multiple copies of a protein, the mapper yielded in one mapping per protein chain. **Figure 4** shows an example of sequence mapping.

The mapping step is a prerequisite for the identification of catalytic residues in protein structures.

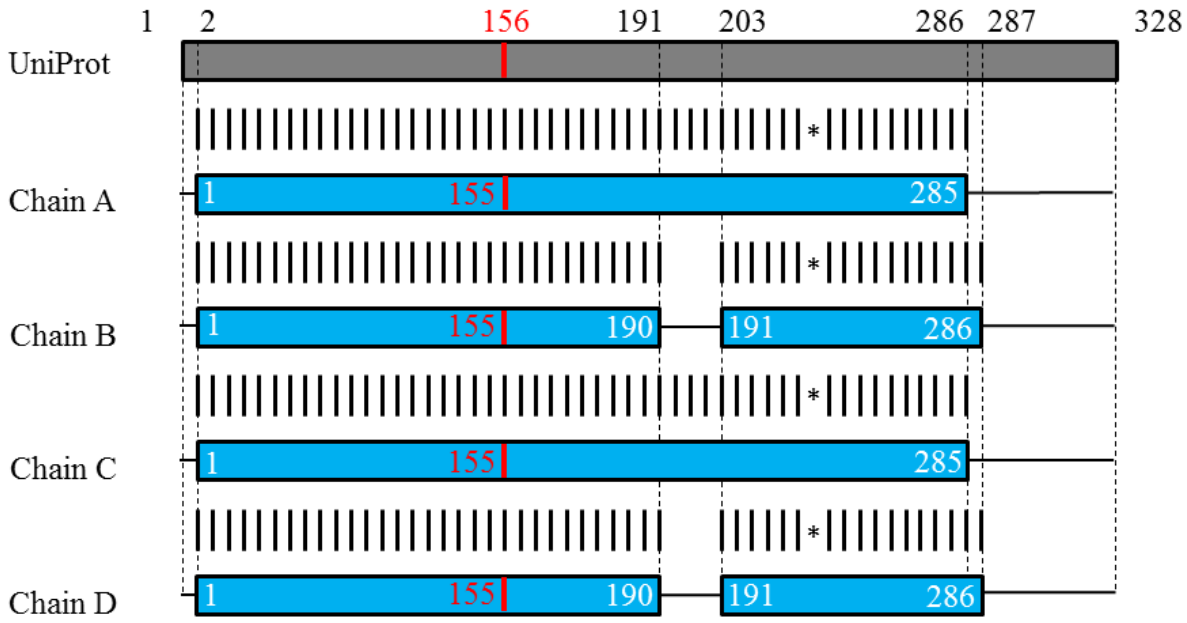


Figure 4. Example of UniProt-to-PDB sequence renumbering.
The figure illustrates the alignment of a structure of 17-β-hydroxysteroid hydrogenase (PDB ID: 1FDV) to its UniProt sequence. Grey bar represents amino-acid sequence from UniProt[1]. Blue bar represents the amino-acid sequence derived from the crystallographic structure chains. A horizontal line along a blue bar represents a gap in the crystallographic amino-acid sequence. Vertical hashes represent correctly aligned amino-acids. When a mutation was introduced in the crystallographic structure, it is spotted by a star. Black numbers above the grey bar represent the residue position in UniProt sequence. White numbers in blue bars represent the residue sequence numbers in the crystallographic structure. Red numbers represent active site residue numbers from UniProt annotation with UniProt identifiers (top) and with PDB residue sequence numbers.

1.3.3. Cavity generation and identification of the catalytic site

PDB structure files were converted into MOL2 format files using UCSF Chimera[24] without alteration on the coordinate section. Cavities were detected in the protein structures using the program VolSite[25] with default parameters and without ligand specification. VolSite considers standard amino acids and cofactors as part of the protein, whereas it is blind to solvent molecules, HET ligands and prosthetic groups. Only cavities with a ligandability score higher than "-1" were further considered.

We then searched catalytic residues in each of the cavities of a given protein structure as follows:

- The list of cavities of generated by VolSite for a protein was sorted by decreasing size.

- Each cavity was transformed into a list of residues lying within 4Å from any cavity point. Then, residue lists were systematically searched for the presence of active site residues.

- The largest cavity containing all catalytic residues was selected. If no cavities contained all active site residues, the largest cavity containing the highest number of active site residues was selected. Protein cavities with no catalytic residues were discarded. The process was iterated independently over all catalytic motifs.

- If there was multiple copies of the protein in the structure, one cavity was selected per protein chain. In such a case, we selected the cavity that contains the largest catalytic motif. If all cavities contained catalytic motifs of same size, we selected the cavity formed by residues with the lowest average temperature factor.

- If two different proteins were present in the structure file, two cavities were chosen.

1.3.4. Ligandability assessment

Proteins with identified catalytic sites were assessed for ligandability using the program VolSite. In theory, a positive ligandability prediction value corresponds to a cavity with physico-chemical properties likely to accommodate a drug-like ligand as opposed to negative prediction values. All protein assembly structures were assessed for ligandability, ensuring that the selected catalytic cavity was ligandable. If the catalytic cavity was not predicted ligandable, then we discarded it structure from the pool of entries.

2. RESULTS

2.1. Statistics for the knowledge-based approach to collect biosynthetic enzymes

2.1.1. Searching UniProt

The number of proteins and corresponding PDB structures passing each step of the collection process is summarized in **table 1**.

In July 2015, the index of metabolic and biosynthetic pathways documented by the UniProt[8] consortium contained 719 pathway descriptions for a total of 122 146 proteins. Out these proteins, only 3 360 have solved structures up to date, accounting for a total of 13 374 structures. Filtering proteins that are not involved in natural products biosynthesis resulted in a list of 4 855 enzymes present in 72 pathways. Only 214 of these enzymes have known three-dimensional structures, accounting for a total of 1 034 PDB entries. All but 8 of these entries passed the protein annotation process. Annotation failures were the consequence of missing or obsolete protein accession code, inconsistency between RCSB and UniProt, or unwanted protein. In addition, we removed 4 proteins from hetero-dimeric enzymes, because they did not belong to the set of known biosynthetic enzymes.

In order to characterize the active site in enzyme structures, we looked for catalytic residues in proteins. At least one documented catalytic residue was found for 103 enzymes, corresponding to 449 structures (a maximum of 15 residues were tagged as active for squalene—hopene cyclase's structure with PDB ID: 2SQC). About two thirds of catalytic residue residue annotations were found in UniProt protein files (66 enzymes, 240

structures). Catalytic residues deriving from UniProt were detected in 201 of the 240 structures. In 199 structures, the full set of catalytic residues was mapped to the structure whereas for 2 other structures, only a subset of the annotated catalytic residues was mapped. We could not find any catalytic motif in 39 structures (mainly because the structure does not describe the catalytic domain). The last third of catalytic residue annotations were found in the Catalytic Site Atlas (27 enzymes, 209 structures). In CSA, catalytic residue documentation is composed by two types of annotations, namely literature and homology annotations. Literature annotations provide catalytic residues from literature articles of protein structures. Homology annotations have been inferred from literature entries by identification of homologous catalytic motifs in the protein sequences. Catalytic residues documented in the Catalytic Site Atlas[21] derived directly from PDB structure residue identifiers and thus, they all mapped to their enzyme structure successfully.

Assuming that active site is located in a cavity, the structures were analyzed using the program VolSite. No cavities were detected on 73 structures. Cavities were detected in the 376 other structures, but only 334 of them have a cavity containing one or more catalytic residue. Because we aim at drug design application, we lastly filtered non-ligandable cavities (i.e., with a low likelihood to accommodate a drug-like ligand), thereby yielding a final dataset of 69 enzymes for 280 PDB structures.

	Protein counts	Structure counts
Proteins from all pathways	122146	13374
Miscellaneous pathways filtering	4855	1034
Biosynthetic enzymes with structures	214	1034
Protein annotation	210	1026
Active site residue annotation	103	449
Active site	66	240
Homology CSA	25	197
Literature CSA	12	12
Protein cavity generation	89	376
Catalytic site identification	77	334
Ligandability assessment	69	280

Table 1. Statistics searching UniProt

2.1.2. Searching MetaCyc

The number of proteins and corresponding PDB structures passing each step of the collection process is summarized in **table 2**.

The secondary metabolite biosynthesis section of MetaCyc[11] was parsed in July 2015. The database contained a total of 2 363 metabolic pathways organized in a hierarchical classification. The branch “secondary metabolites biosynthesis” of the hierarchical tree contains 696 pathways grouped into 16 classes. The 696 pathways are linked to 4 156 genes, each described on a page that contains cross references with Uniprot, allowing us to retrieve the accession codes for 1729 proteins. 3D-structures were available for almost

half of these proteins (representing 145 pathways, **annex4**, **table 3**). Protein annotation failed in two cases, and 18 additional entries were discarded (hetero-dimeric structures).

We identified catalytic motifs in 83 enzymes (corresponding to 318 structures). Again, the majority of the matched motifs originate from Uniprot (52 proteins, 178 structures). We successfully mapped all catalytic residues of the motifs in 159 structures whereas only partial match of the motif was found for 28 structures. Additional matched motifs originated from catalytic site atlas (31 proteins, 140 structures). In total, 299 structures were assigned a catalytic motif.

In the next step, cavity detection only succeeded for 276 structures (72 proteins), among them 238 structures (62 proteins) contained at least one catalytic residue within a detected cavity. The remaining set of biosynthetic enzymes from MetaCyc was ultimately filtered according to ligandability values, thus removing 69 structures. The final dataset contains 53 enzymes for 169 PDB structures.

	Protein counts	Structure counts
NP Biosynthetic enzymes	1729	907
NP Biosynthetic enzymes with structures	206	907
Protein annotation	204	905
Active site residue annotation	83	318
Active site	52	178
Homology CSA	21	130
Literature CSA	10	10
Protein cavity generation	72	276
Catalytic site identification	62	238
Ligandability assessment	53	169

Table 2. Statistics searching MetaCyc.

2.1.3. Comparison of the two searches

The number of PDB structures passing each step of the collection process, starting from the two source databases, UniProt and MetaCyc, is summarized in **table 3**.

Altogether, knowledge-based flowcharts yielded in a total of 1 436 PDB files. Nearly half of initial sets of structures is common to MetaCyc and UniProt workflows. An identical process was applied to "Uniprot initial set" and to "MetaCyc initial set". At each step of the two workflows, approximately the same proportion of structures was discarded. At the end of the workflows, 72.9% and 81.4% of structures were discarded from "Uniprot initial set" and from "MetaCyc initial set" respectively. In the two searches, the most drastic cut occurred at the catalytic residue annotation step: 56% of the structures

were hence discarded because no information about catalytic site was found in UniProt or in the Catalytic Site Atlas.

	UniProt	Common	MetaCyc
NP biosynthesis selection	1034	505	907
Protein annotation	1026	505	905
Active site residue annotation	449	188	318
Protein cavity generation	376	*NC*	276
Catalytic site identification	334	146	238
Ligandability	280	107	169

Table 3. Comparison of UniProt and MetaCyc searches.
Numbers are counts of PDB structures. NC: not calculated

2.1.4. Description of "UniProt final set"

The 280 structures of "UniProt final set" represent 14 different classes of pathways. Only two of the 16 investigated classes were finally not associated to structures. Two enzymes in the biosynthesis of carotenoids (dehydrosqualene synthase and phytoene desaturase) did not have catalytic residue annotation and, the only enzyme in the biosynthesis of mycotoxin (noranthrone synthase) failed the step of catalytic residue mapping. **Figure 5** illustrates the 14 classes of pathways that we could populate with ligandable structures. **Table 4** indicates the number of structures present in each class. The biosynthesis of antibiotics was the most populated pathway class with 109 structures. The pathways for the biosynthesis of isoprenoids, secondary metabolites and steroid ranked well behind with

41, 37 and 40 structures, respectively. The **figure 5** also shows that seven pathway classes contain structures also present in "MetaCyc final set". Most populated pathways shared the highest number of structures common to "MetaCyc final set" (**table 4**). Biosynthesis of antibiotics, isoprenoids and secondary metabolites shared 42, 26 and 25 structures, respectively, with "MetaCyc final set". Interestingly, 10 out of 11 structures associated to the biosynthesis of alkaloids are contained within "MetaCyc final set". One can notice that our natural products biosynthetic pathways selection was not exactly representative of the secondary metabolite biosynthesis section in MetaCyc. For example, none of the 20 structures associated to the biosynthesis of steroids and none of the 16 structures associated to the biosynthesis of lipids are contained within "MetaCyc final set".

Some enzymes were found in several pathways (compare counts in **table 1** and **4**). For example, pentalene synthase from *Streptomyces exfoliates*, responsible for the cyclization of farnesyl diphosphate into pentalene, was classified in the biosynthesis of antibiotics and in the biosynthesis of sesquiterpenes (PDB ID: 1HM4 and 1HM7). Phenylalanine aminomutase from *Taxus canadensis* is involved in the in the metabolism of phenylpropanoids and in the biosynthesis of alkaloids (PDB ID: 3NZ4). It is responsible for the conversion of phenylalanine into trans-cinamate, a key precursor common to many phenylpropanoids, and responsible for the conversion of phenylalanine into β-phenylalanine as a biosynthetic step in the preparation of Taxol's 13C side chain.

As a note, 62 structures of "MetaCyc final set" did not fit UniProt pathways classification and were therefore not mapped into bubbles of the **figure 5**.

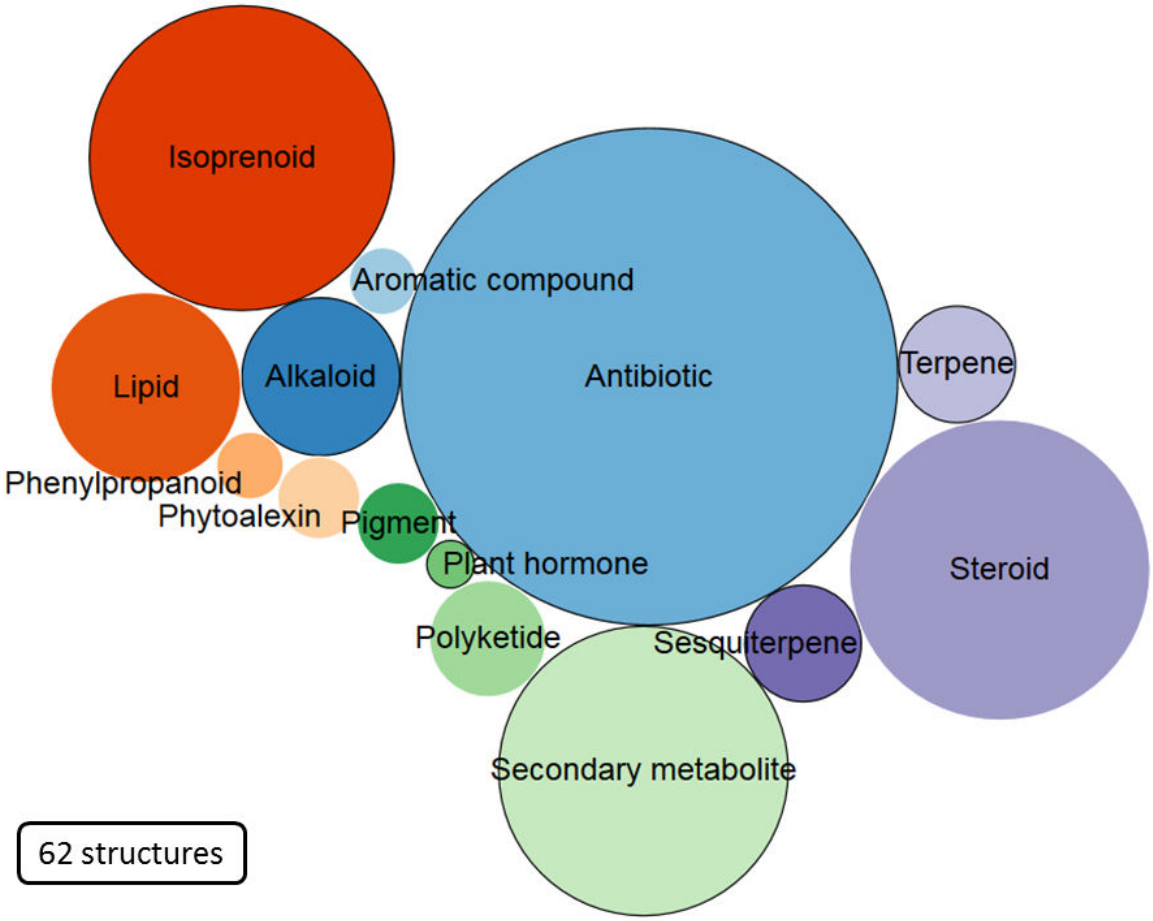


Figure 5. Classes of UniProt pathways populated with enzymes of known 3D-structure.
Bubbles represent Uniprot classes of secondary metabolite biosynthetic pathways. Bubble size is proportional to the number of structures in the "UniProt final set". Black outline indicates which classes are also populated with structure of the "MetaCyc final set". The box in the bottom left corner gives the count of structures found in the "MetaCyc final set" that do not belong to any of the represented pathways.

Classes of UniProt Pathways	number of Structures in the "UniProt final set"	Proteins in the "UniProt final set"	Total protein in UniProt	Structures in the "MetaCyc final set"
Alkaloid	11	6	10	10
Antibiotic	109	22	77	42
Aromatic compound	2	1	4	0
Carotenoid	0	0	2	0
Isoprenoid	41	9	59	26
Lipid	16	7	12	0
Mycotoxin	0	0	1	0
Phenylpropanoid	2	2	6	0
Phytoalexin	3	2	3	0
Pigment	3	1	6	0
Plant hormone	1	1	1	1
Polyketide	6	1	2	0
Secondary metabolite	37	8	15	25
Sesquiterpene	6	3	5	4
Steroid	40	5	8	0
Terpene	6	3	5	1

Table 4. Classes of UniProt pathways populated with enzyme of known 3D-structure.

2.1.5. Description of the "MetaCyc final set"

The 169 structures contained within "MetaCyc final set" were associated to 12 of the 16 biosynthetic pathway classes of MetaCyc secondary metabolites biosynthesis documentation. We did not identify any structure in these four missing classes mainly because we failed in the characterization of active sites. For example, no catalytic residues

were found for the only enzyme of the biosynthesis of ergothioneines. The same scenario happened to enzymes in the biosynthesis of insecticides, and in the biosynthesis of sulfur-containing-secondary compounds. Besides, biosynthesis of xanthones was empty because none of its enzymes has a known structure.

Figure 5 summarized the repartition of the "MetaCyc final dataset" into the 12 classes of pathways. Biosynthesis of terpenoids and antibiotics are the two most populated classes of pathways, with 64 and 48 structures, respectively. Two third of the classes share structure(s) with "UniProt final set". Structures of "MetaCyc final set" are thus more present in UniProt pathways than structures from "UniProt final set" are present in MetaCyc pathways. The number of structures in the "UniProt final set" largely exceeds that in the "MetaCyc final set", thereby 173 structures did not fit into any class of MetaCyc pathway.

Similarly to enzymes in "UniProt final set", some of the enzymes from "MetaCyc final set" are present in several pathways (**tables 2, 3 and 4**). For example, 5-epi-aristolochene synthase from *Nicotiana tabacum* (PDB ID: 1HXC), which catalyzes the cyclization of farnesyl diphosphate into (+)-5-epiaristolochene during the biosynthesis of capsidiol, is associated to the biosynthesis of phytoalexins and to the biosynthesis of terpenoids.

Another example is polyneuridine-aldehyde esterase from *Rauvolfia serpentine,* which catalyzes the formation of a precursor in ajmaline, a nitrogen-containing secondary compound also involved in the biosynthesis of terpenoids.

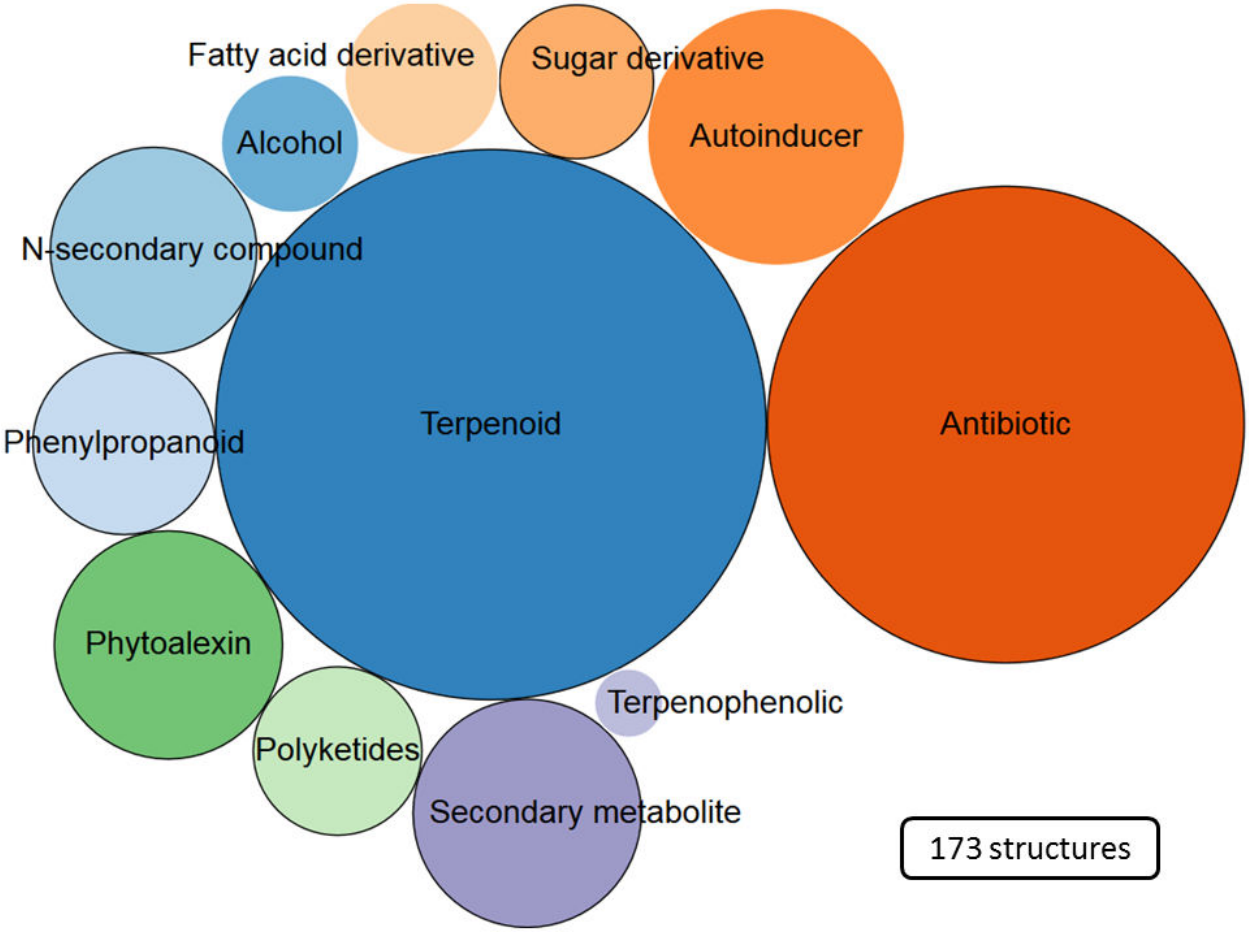


Figure 5. Classes of MetaCyc pathways populated with enzymes of known 3D-structure.
Bubbles represent MetaCyc classes of secondary metabolite biosynthetic pathways. Bubble size is proportional to the number of structures in the "MetaCyc final set". Black outline indicates which classes are also populated with structure of the "Uniprot final set". The box in the bottom left corner gives the count of structures found in the "UniProt final set" that do not belong to any of the represented pathways.

Classes of MetaCyc pathways	Number of structures in the "MetaCyc final set"	Proteins in the "MetaCyc final set"	Total Proteins in MetaCyc	Structures in the "UniProt final set"
Alcohol	4	2	76	0
Antibiotic	48	17	496	42
Autoinducer	14	3	12	0
Ergothioneine	0	0	7	0
Fatty-acid derivative	5	3	18	0
Insecticide	0	0	9	0
Nitrogen-containing secondary compound	9	4	161	9
Phenylpropanoid	7	1	199	7
Phytoalexin	11	1	37	11
Polyketides	6	3	67	5
Sulfur-containing secondary compound	0	0	13	0
Secondary metabolites	11	5	141	5
Sugar derivatives	5	4	77	1
Terpenoid	64	13	494	43
Terpenophenolic	1	1	5	0
Xanthone	0	0	1	0

Table 5. Classes of MetaCyc pathways populated with enzyme of known 3D-structure.

2.2. Statistics for the Top-down approach to collect biosynthetic enzymes

We mined 103 993 entries in RCSB PDB archive (October 2014), searching for four text motifs (i.e. "*biosynth*", "*natural product*", *"secondary metabol"*, *"plant defense"*). Text mining detected a total of 7 608 structures, accounting for 2 949 proteins. **Figure 6** shows

occurrences of the different matched text motifs. Most frequent text motif was "*biosynth*" (6 629 structures matched). In addition, the text motif "*biosynth*" was found in 48.6% of the occurrences matching "*secondary metabol*" and 60.5% of the occurrences matching "*natural product*" while it was found in only 12.6% of the occurrences matching "*plant defense*".

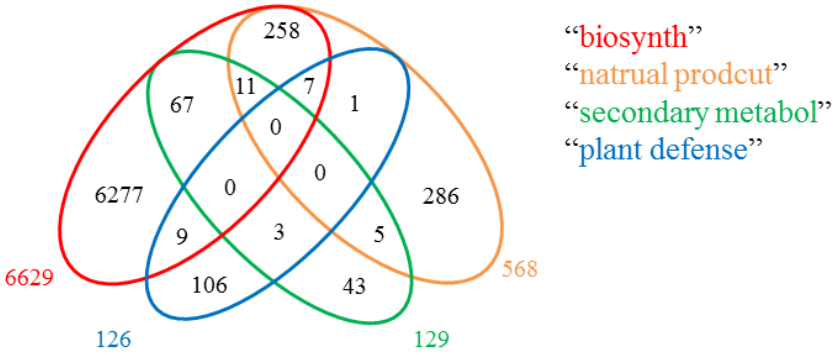


Figure 6. Text motif matches from textual data in PDB.

Proteins that matched a text motif were further annotated. About 5% of entries (399 structures) were discarded because of missing or obsolete protein accession code, inconsistent annotation when comparing PDB and UniProt, or because the protein was obviously not a biosynthetic enzyme of natural products (**table 6**).

After protein annotation, 3024 different proteins (in 7209 structures) we submitted to the catalytic motif identification. We could find a catalytic motif for only a third of the submitted proteins (749 proteins -2382 structures- with UniProt active site residue(s) and 397 additional proteins -1002 structures- with motif(s) of Catalytic Site Atlas).

Searching catalytic motifs in structures slightly reduced the dataset. Considering catalytic residue annotations originating from UniProt, 2101 of the 2382 searched structures (687 of the 749 proteins) contained the full catalytic motif in at least one protein chain. A partial match was observed for 148 additional structures. The 133 remaining structures did not contain any residue in the catalytic motif. As expected, we did not encounter any difficulties to detect active residues into the 1002 structures with catalytic residue annotation from Catalytic Site Atlas.

At this stage, aiming at speeding up coming calculations and manual analysis, we decided to filter non relevant entries of the dataset using the list of proteins associated to the 401 miscellaneous pathways (**annex 4, table 2**), thereby discarding 48.5% of the structures and leaving us with a set of 1642 structures.

Out of these structures, at least one cavity was found in 1319 structures. In turn, catalytic residues were found in 84.7% of the structures.

We aim at using the selected structures for binding site similarity experiments in drug design applications and thus, we have limited the dataset to structures with ligandable active sites and associated to enzymatic reactions involving mature metabolites (similar to the final products in the pathway). This filter discarded a quarter of the proteins.

Last, we manually validated entries by assessing their enzymatic activity with the help of UniProt, MetaCyc and the literature. Only 33 of the 323 checked proteins were indeed natural products biosynthetic enzymes interacting with mature metabolites.

Steps	Protein counts	Structure counts	Relative deletion (count)		Cumulative deletion
PDB structure downloads	~35000	103993	0 %	(0)	/
Keyword search	*NC*	7608	92.7 %	(96385)	100%
Protein annotation	3024	7209	5.2 %	(399)	5.2%
Active site residue annotation	1146	3384	53.1 %	(3825)	55.5%
Active site	749	2382			
Homology CSA	374	934			
Literature CSA	68	68			
Miscellaneous metabolism filtering	605	1642	51.5 %	(1742)	78.4%
Protein cavity generation	497	1319	18.6 %	(305)	82.4%
Catalytic site identification	422	1118	15.2 %	(201)	85.3%
Ligandability assessment	323	760	32.0 %	(358)	90.0%
Manual validation	33	138	81.8 %	(622)	98.2%

Table 6. Statistics for the top-down approach. NC: not calculated

2.3. Comparison of top-down and knowledge-based strategies

2.3.1. Comparison of the statistics

The top-down and knowledge-based strategies only differ in their first steps. The Top-down strategy considered all PDB entries, and collected structures based on keywords. Consequently, it yielded a large pool of structures containing many false positive. The knowledge-based strategy benefited from expert annotation provided in high quality resources of metabolic pathways (UniProt and MetaCyc) and thus, only a limited number of well annotated true positives were further parsed. In order to compare the number of structures passing each step in top-down versus knowledge-based approaches, we pooled "initial sets" of MetaCyc and UniProt together and re-calculated numbers of structures passing each step (**table 7**). Not surprisingly, the number of structures selected by text-

mining five-fold exceeded that collected in the knowledge-based approach. The step that applied the largest cut in the structure pool is the identification of catalytic residues. It discarded 53.1% and 59.4% of annotated protein structures in top-down and knowledge-based flowcharts, respectively. After completion, the top-down strategy collected a total 33 enzymes whereas the knowledge-based strategy collected a pool of 105 distinct enzymes (67 in "MetaCyc final set" and "77 in UniProt final set"). The combination of "top-down final set" and "knowledge-based final set" yielded in a "global final set" of 117 enzymes. Interestingly, 12 enzymes are neither in "UniProt final set" nor in "MetaCyc final set".

Top-Down	Number of structures	Shared entries	Number of structures	Knowledge-based
Keyword search	7608	871	1436	**NP biosynthesis selection**
Protein annotation	7209	867	1426	**Protein annotation**
Active site residue annotation	3384	387	579	**Active site residue annotation**
Miscellaneous metabolism filtering	1642	378	579	**/**
Protein cavity generation	1319	*NC*	*NC*	**Protein cavity generation**
Catalytic site identification	1118	290	427	**Catalytic site identification**
Ligandability assessment	760	244	342	**Ligandability**
Manual validation	138	115	342	**/**

Table 7. Top-down strategy compared to knowledge-based strategy.
NC: not calculated. Primary metabolism filtering step is made in the first step of the knowledge-based process whereas it takes place at the fourth position in top-down process. All the other steps are equivalent. The last line is specific to the top-down strategy.

2.3.2. Classification of “top-down structures” in UniProt pathways

The 138 structures contained within the “top-down final set” are associated to 10 of the 16 biosynthetic pathway classes considered in UniProt database. The two classes that did not include any structures from the knowledge-based strategy (biosynthesis of carotenoids and mycotoxins) remained empty upon classification of the “top-down structures”. **Figure 7** illustrates which of the other 14 classes of pathways in UniProt contain the “top-down structures”. No structures were classified in pathways for the biosynthesis of isoprenoids, lipids, phenylpropanoids and polyketides.

The top-down strategy mainly identified proteins involved in the biosynthesis of antibiotics (54 structures), the biosynthesis of secondary metabolites (22 structures) and the biosynthesis of steroids (22 structures) (**table 8**). Other classes of pathways only contained a few structures of the “top-down final set”. Strikingly, potential true positives were discarded by the manual validation step. For example about a half of structures in the class of antibiotic synthesis and all structures in the class of isoprenoid biosynthesis were discarded manually. Two explanations account for manual deletions. Enzymatic activities in UniProt description were absent, thus invalidating the entries, or the involved molecules did not contain the scaffold of the final product in the pathway. By contrast, manual validation retained most of structures in the biosynthesis of aromatic compounds, phytoalexin, pigment, plant hormone, secondary metabolite, sesquiterpene, terpene and steroid.

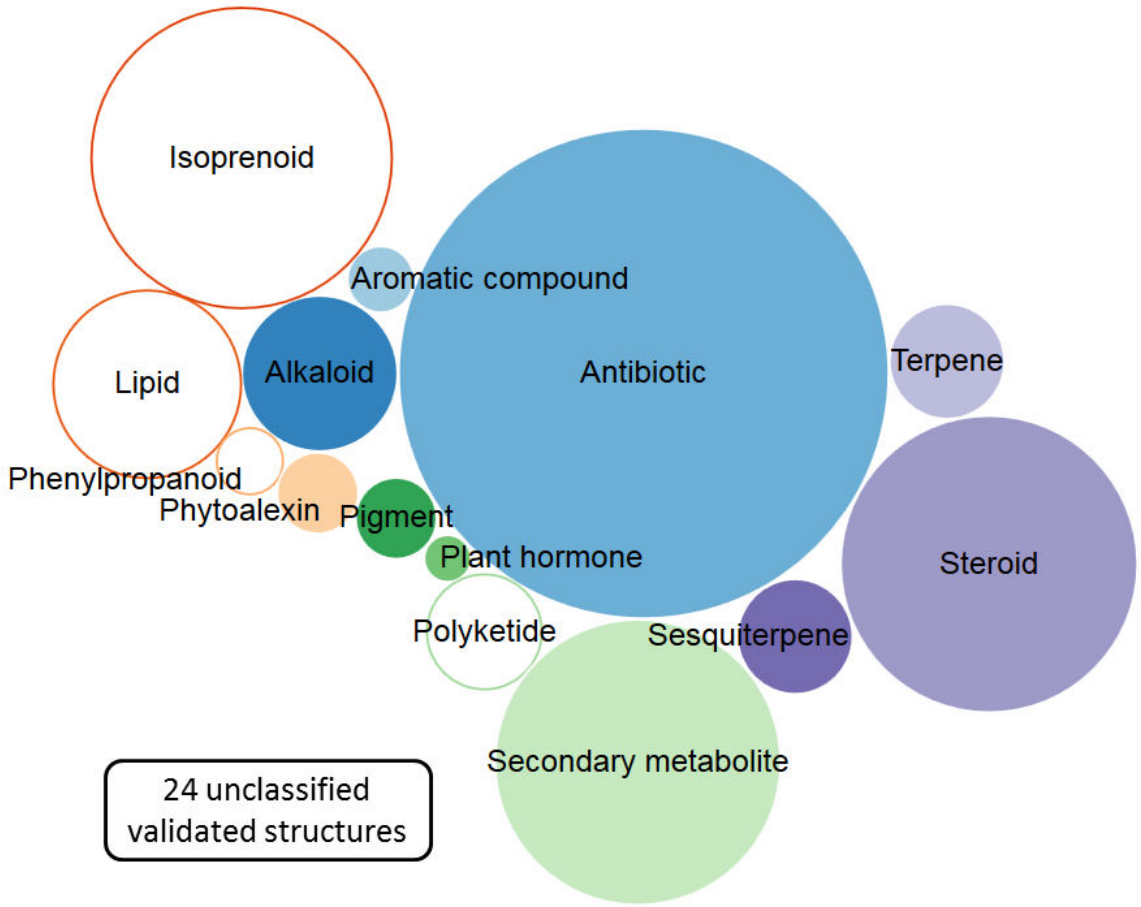


Figure 7. Classes of UniProt pathways populated with enzymes from the top-down approach.
Bubbles represent Uniprot classes of secondary metabolite biosynthetic pathways. Bubble size is proportional to the number of structures in the "UniProt final set". The bubbles with white bodies represent a pathway into which no ligandable enzyme structure was found in the PDB. The box in the bottom left corner gives the count of structures found in the "Top-down final set" that do not belong to any of the represented pathways.

Classes of UniProt Pathways	Number of Structures in the "Top-down final set"	Proteins in the "Top-down final set"	Total Proteins in the class	Structures in the "UniProt final set"
Alkaloid	4 (10)	2	10	11
Antibiotic	54 (98)	8	77	109
Aromatic compound	2 (2)	1	4	2
Carotenoid	0 (0)	0	2	0
Isoprenoid	0 (32)	0	59	41
Lipid	0 (3)	0	12	16
Mycotoxin	0 (0)	0	1	0
Phenylpropanoid	0 (1)	0	6	2
Phytoalexin	1 (1)	1	3	3
Pigment	3 (3)	1	6	3
Plant hormone	1 (1)	1	1	1
Polyketide	0 (5)	0	2	6
Secondary metabolite	22 (25)	4	15	37
Sesquiterpene	3 (3)	2	5	6
Steroid	20 (25)	2	8	40
Terpene	5 (5)	2	5	6

Table 8. Classes of UniProt pathways populated with enzymes from the top-down approach.
Numbers in brackets give the counts of structures before manual selection (last step of the process).

2.3.3. Top-down resulting structures in MetaCyc classification

The 138 structures contained within the "top-down final set" were associated to 5 of the 16 biosynthetic pathway classes considered in MetaCyc database (in the secondary metabolites biosynthesis section). The four classes that did not include any structures from the knowledge-based strategy (biosynthesis of ergothioneines, insecticides, sulfur-

containing secondary compounds and xanthones) remained empty upon classification of the “top-down structures”. **Figure 8** illustrates which of the other 16 classes of pathways in MetaCyc contain the “top-down structures”. No structures were classified in pathways of the biosynthesis of alcohols, fatty acid derivatives, sugar derivatives, autoinducers, polyketides, secondary metabolites and terpenophenolics.

The top-down strategy mainly identified proteins involved in the biosynthesis of antibiotics (16 structures), terpenoids (11 structures), phenylpropanoids (7 structures) and phytoalexins (6 structures) (**table 9**). Manual selection in the last step of the top-down flowchart discarded many biosynthetic enzymes due to missing UniProt enzymatic activity description and to different substrate/product structures when compared to final products of the pathways. For example 20 out of the 36 structures of enzymes involved in antibiotic biosynthesis were discarded manually. We also excluded 36 out of the 47 structures of enzymes involved in the biosynthesis of terpenes.

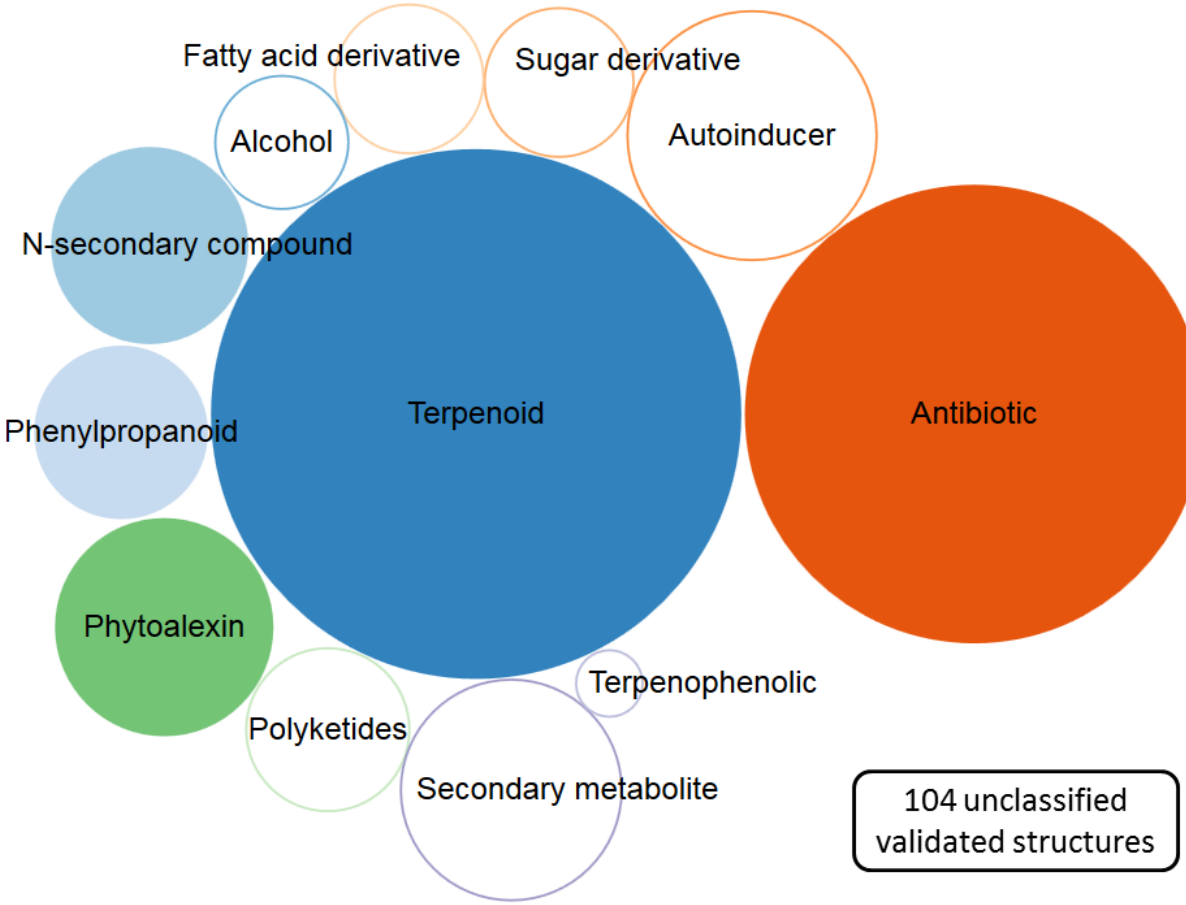


Figure 8. Classes of MetaCyc pathways populated with enzymes from the top-down approach.
Bubbles represent MetaCyc classes of secondary metabolite biosynthetic pathways. Bubble size is proportional to the number of structures in the "MetaCyc final set". The bubbles with white bodies represent a pathway into which no ligandable enzyme structure was found in the PDB. The box in the bottom left corner gives the count of structures found in the "Top-down final set" that do not belong to any of the represented pathways.

Classes of MetaCyc Pathways	Number of Structures in the "Top-down final set"	Proteins in the "Top-down final set"	Total Proteins in the class	Structures in the "MetaCyc final set"
Protein data bank biosynthetic enzymes				
Pathways	**Structures**	**Enzymes**	**Total enzymes**	**Structures in MetaCyc**
Alcohol	0 (0)	0	76	4
Antibiotic	16 (36)	7	496	48
Autoinducer	0 (9)	0	12	14
Ergothioneine	0 (0)	0	7	0
Fatty-acid derivative	0 (0)	0	18	5
Insecticide	0 (0)	0	9	0
Nitrogen-containing secondary compound	3 (8)	1	161	9
Phenylpropanoid	7 (7)	1	199	7
Phytoalexin	6 (6)	1	37	11
Polyketides	0 (0)	0	67	6
Sulfur-containing secondary compound	0 (0)	0	13	0
Secondary metabolites	0 (5)	0	141	11
Sugar derivatives	0 (1)	0	77	5
Terpenoid	11 (47)	4	494	64
Terpenophenolic	0 (0)	0	5	1
Xanthone	0 (0)	0	1	0

Table 9. Classes of MetaCyc pathways populated with enzymes from the top-down approach.
Numbers in brackets give the counts of structures before manual selection (last step of the process).

3. DISCUSSION

3.1. MetaCyc final set" and "UniProt final set" are overlapping but distinct because source data are classified differently

Biological processes in living organism are difficult to describe as a unique partitioned network of pathways. There is no ontology for metabolic pathways, whose definition itself is ambiguous. For example, definition of the end points in pathways depends on decisions made by the curators who design metabolic databases. UniProt's pathways are directly derived from UniPathway database,[16] which is a collaborative project between the Swiss Institute of Bioinformatics (SIB), the French National Institute for Research in Computer Science and Control (INRIA Rhone-Alpes) and the Laboratory of Alpine Ecology of Grenoble, France. MetaCyc's pathways are the result an initiative lead by Stanford Research Institute international (SRI) which eventually collaborated with the department of Plant biology of Carnegie Institution, Stanford, USA and the Thompson Institute for Plant Research, Ithaca, USA. Being curated by different teams, UniPathway and MetaCyc databases are ruled by different concepts, and consequently contain pathways of different nature. On one hand, MetaCyc database has a smaller compartmentation intended to avoid overlap between the pathways, whereas on the other hand, UniPathway compartmentation is much larger, allowing pathways to overlap each other.

In addition, "MetaCyc initial set" and "UniProt initial set" are distinct because we made a selection of pathways in UniProt that is not representative of pathways of the secondary metabolites biosynthesis in MetaCyc (**annex4, table 1**). For example, steroids such as

zymosterol, cholesterol, estrogen and lanosterol belong to the terpenes in UniProt but did not fall in the secondary metabolites biosynthesis section of MetaCyc. However, MetaCyc secondary metabolites section does not exclude steroids entirely since the hopanoid biosynthetic pathway is represented in UniProt and MetaCyc final sets. This examples illustrates the overlap of steroid-like compounds and natural products which explains why we selected cholesterol, zymosterol and estrogen biosynthetic pathways in UniProt pathways.

Some other pathways were clearly common between the two resources, for example the biosynthesis of antibiotics and the biosynthesis of alkaloids. Alkaloids pathways include the biosynthesis of taxol, ajmaline, (s)-scoulerine, 3α(S)-strictosidine biosynthetis, all represented in MetaCyc and UniProt final sets. Nevertheless, the pathways content differs in the two resources. For example in UniProt, taxol biosynthesis include all biosynthetic steps that are involved in the preparation of every component of taxol. However, in MetaCyc, the preparation of taxol's 13C-side chain ensured by phenylalanine aminomutases[27] (UniProt ID: Q6GZ04 and Q68G84, translocation of an amide on phenylalanine) is affiliated to a pathway that is branched to taxol biosynthesis but distinct in itself. This example highlights the different compartmentation of pathways in UniPathway and MetaCyc and illustrates the overall smaller pathway sizes of MetaCyc pathways.

The biosynthesis of antibiotics is one of the largest class of pathways in both resources. We can mention the biosynthesis of clavulanic acid, cephalosporin C, erythromycin, kanosamine, daunorubicin and penicillin amongst others. Most of antibiotics pathways are

represented in UniProt and MetaCyc final sets. Nevertheless half of the structures in "UniProt final set" are not in "MetaCyc final set". Here again MetaCyc pathways compartmentation excludes peripheral biosynthetic steps, such as the synthesis of L-arginine via L-ornithine (a precursor of clavulanic acid) which involves glutamate N-acetyltransferase 2[28] (UniProt ID: P0DJQ5). In other cases, biosynthetic enzymes contained within "UniProt final set" are not affiliated to a biosynthetic pathway in MetaCyc, even though their enzymatic reaction is described as a standalone reaction. It is the case of nebramycin 5' synthase (UniProt ID: Q70IY1) and aclacinomycin methylesterase RdmC (UniProt ID: Q54528) respectively involved in biosynthesis of kanamycin and aclacinomycin according to UniProt. Alternatively, representative species of biosynthetic enzymes also differ in the two resources. For example, an enzyme from penicillin biosynthesis, isopenicillin N synthase from *Emericella nidulans* (UniProt ID: P05326), is referenced in UniProt whereas MetaCyc provides the enzymes from *Acremonium chrysogenum* (UniProt ID: P05189) and *Amycolatpsis lactamdurans* (UniProt ID: P27744) both performing isopenicillin N synthases. For the record, the two later enzymes, do not have known structures. Lastly, some antibiotic biosynthetic pathways are not described in MetaCyc (vancomycin biosynthesis).

3.2. Top-Down strategy compared to knowledge-based strategy

The two strategies are composed of two main steps (**figure 1**). A first step intended to filter natural product biosynthetic enzymes from an initial database and a second step for the identification of ligandable catalytic sites in enzymes structures. Although both strategies

are conceptually similar, they have different philosophies. The top-down strategy was designed to emancipate from knowledge-based databases. Unfortunately, it was impossible to automatically validate that a protein in the dataset was indeed a natural product biosynthetic enzyme. In the very last step of top-down strategy, we thus verified proteins one by one using knowledge-based resources. Although extremely time-consuming, this manual step allowed the filtering of enzymes catalyzing reactions involving metabolites which are either much smaller or much less complex than the end product of the pathway and thus, they are irrelevant for natural products repositioning by binding site similarity.

The knowledge-based strategy yielded in 3 times more natural product biosynthetic enzymes than the top-down strategy. This higher number is partly due to the fact that we did not manually check “knowledge-based final set”, thereby keeping enzymes acting on metabolites which are either much smaller or much less complex than the end product of the pathway (e.g., enzymes in the isoprenoid biosynthesis).

Importantly the top-down strategy identified 12 enzymes not present in the “knowledge-based final set”, demonstrating a potential to find enzymes undocumented in metabolic resources. These enzymes were unambiguously assigned to natural product biosynthesis in the literature. Their sequence and biological functions were described in UniProt, but no links were given to biosynthetic pathways yet.

3.3. Chemical diversity of natural products in the dataset

In this section we discuss the chemical diversity of metabolites interacting with the 117 biosynthetic enzymes in the dataset. We focused on substrates and final products of reactions. We ignored cofactors, ions, byproducts or adducts. A detailed description of enzymatic reactions, biosynthetic pathway intermediates and final products is given annex 4. Seven general categories were considered: hydrocarbons (15 members, **figure 10**), phenylpropanoids and polyketides (16 members, **figure 11**), nitrogen-containing compounds (7 members, **figure 12**), aminoglycosides and relatives (6 members, **figure 13**), β-lactams (4 members, **figure 14**), macrolides (4 members, **figure 15**) and precursors and small molecules (12 members, **figure 16**). In figures 10 to 12, template structures represent all the compounds sharing a common scaffold.

3.3.1. Hydrocarbons

Hydrocarbons are mainly composed of sesquiterpenes. Templates in **figure 10** represent (+)-camphor (**1**), pentalenene (**2**), aristolochen (**3**), albaflavenone (**4**), E-α-bisabolene (**5**) and the molecular core of Taxol (**6**). These compounds can be related to interesting pharmacologic and ecologic activities. Pentalenene is a precursor of pentalenolactone, which has antibiotic activity and was reported as an inhibitor of the glucose metabolism by inactivation of the enzyme glyceraldehyde-3-phospahte dehydrogenase,[29–31] However, the activity of pentalenolactone requires the epoxide and the lactone groups,[32,33] which are not included in the pentalenene template. Aristolochen template represents a precursor structurally related to the phytoalexin capsidiol. By definition, phytoalexins are produced

by higher plants in response to pathogenic infections.[34] For example, capsidiol is synthesized by *Nicotiana tabacum* and *Capsicum annuum* plants when challenged by fungus such as *Phytophthora capsic.*[35] A recent study showed that capsidiol affects the growth of two pathogenic fungus differently, suggesting that the defense mechanism is specific to the host in the way it interacts with the attacking organism.[36] This result suggest that the defense mechanism of capsidiol could occur via interactions with a specific target. The hydrocarbons class also includes steroids such as lanosterol (**8**), 17β-estradiol (**9**) , hop-22(29)-one (**10**) and a compound involved in androstenedione degradation (**12**). However, the nature of these compounds and their presence in the human body renders them questionable for natural product repositioning. Lovastatin (**11**) is a polyketide which inhibits cholesterol synthesis regulation[37,38] and is therefore used as hypolipidemic drug. It hints at a novel PFT example because targeted proteins are known and the catalytic cavities of its biosynthetic enzymes are ligandable. Templates **13**, **14** and **15** represent jasmonic acid, a precursor of CAI-1 autoinducer and a mycolic acid, respectively. They all contain long and flexible linear carbon chains.

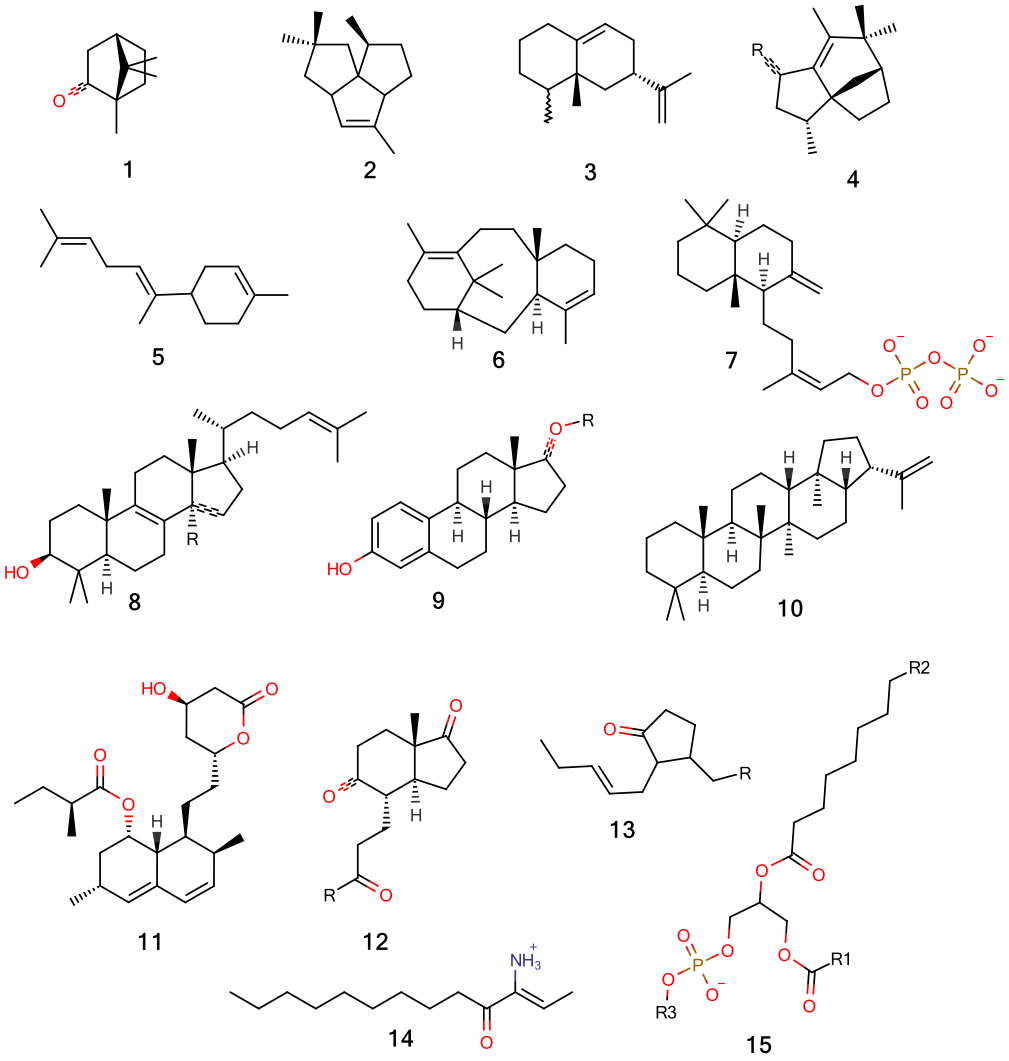


Figure 10. Hydrocarbon templates in the final dataset.

3.3.2. Phenylpropanoids and polyketides

Phenylpropanoids and polyketides (**figure 11**) represent a variety of early precursors of natural products as well as compounds structurally related to compounds with pharmacological and ecological activities. Feruloylacteyl-CoA (**18**), a trans-caffeat representative template (**19**), a methyl-naphthoic acid (**26**) and a common precursor of polyketides (**30**) are examples of early precursor of natural products.

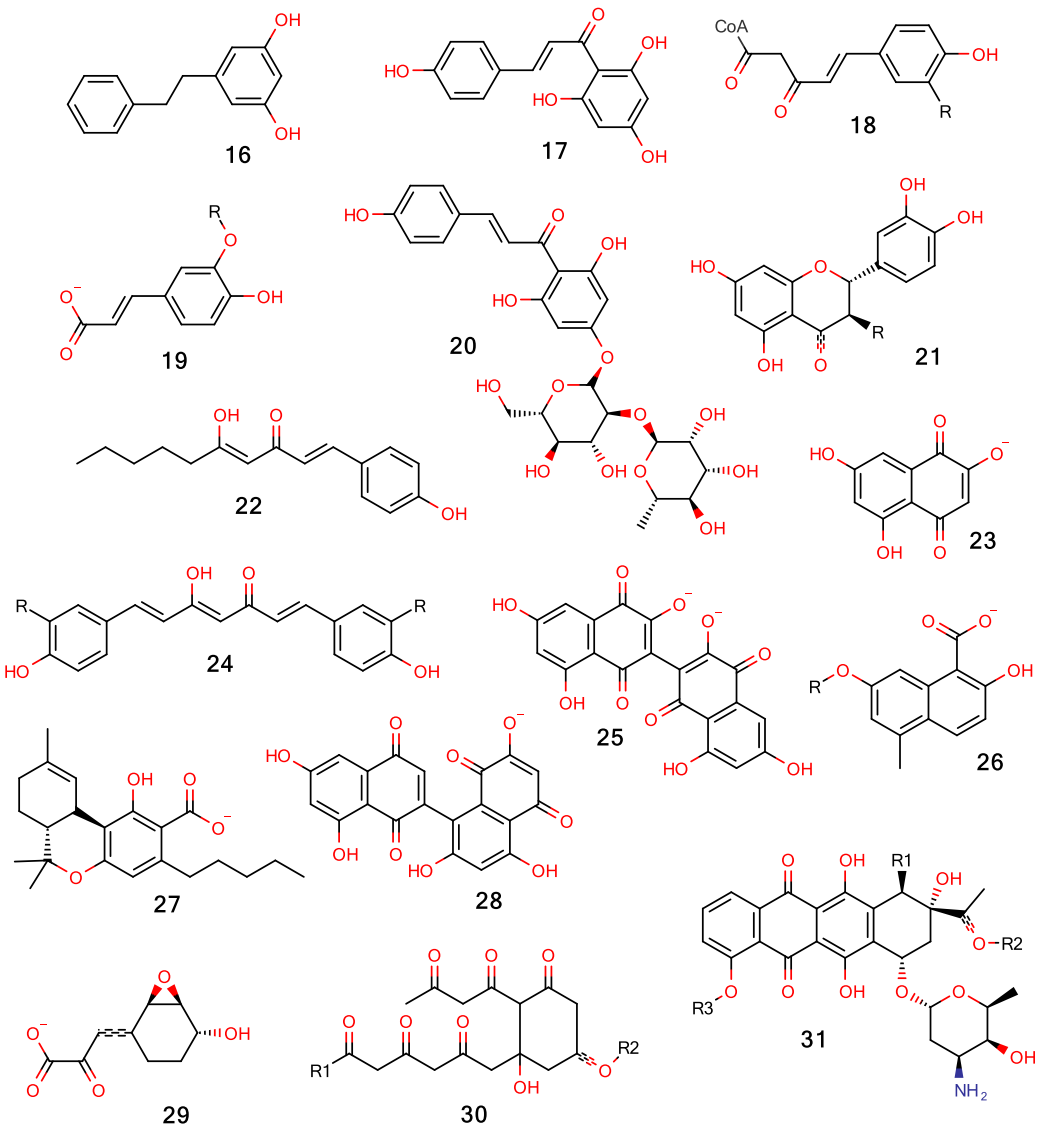


Figure 11. Phenylpropanoids and polyketides templates in the final dataset.

Chalcones (**17**, **20**), flavonoids (**21**), curcuminoids (**24**), Δ^9-tetrahydrocannabinol (**27**), bacilysin intermediate (**29**) or anthracylines (**31**) all have known pharmacological and ecological activities. For example, anthracylines are used as anticancer agents, because they intercalating between DNA and RNA strands,[39,40] thus preventing cell growth and inhibits topoisomerases II.[41] The precursor of bacilysin (L-anticapsin, (**29**)) is an antibiotic.

It causes cell wall peptidoglycan disruption due to irreversible inhibition of glutamine—fructose-6-phosphate transaminase.[42] L-anticapsin contains an epoxide with a carboxyl group, present on the shown intermediate, which is suggestive of a reaction with a thiol group of the inhibited enzyme.[43] Another example of class of compounds associated to numerous health benefits is the class of curcuminoids, which has antioxidant, anti-tumor, anti-inflammatory properties.[44]

3.3.3. Nitrogen-containing compounds

Nitrogen-containing compounds (**figure 12**) are composed of alkaloids with known pharmacological activities except phenazine-1-carboxylate precursor (**33**). The template representing tropine (**32**) is present in atropine and in scopolamine, two approved drugs. The two compounds exhibit an activity on mammalian nervous system and more precisely, they target muscarinic acetylcholine receptors.[45,46]

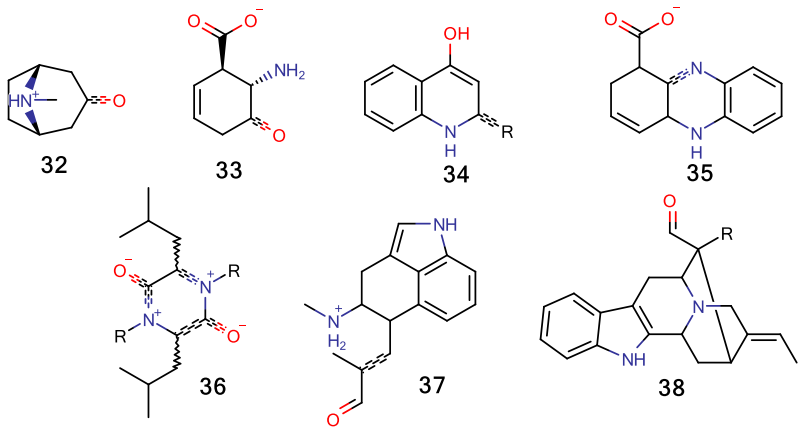


Figure 12. Nitrogen-containing compounds in the final dataset.

Templates **33**, **34** and **35** represent phenazine-1-carboxylate and precursors. Phenazines have antibiotic activities taking place in the ecology of their producing organism[47] and have recently been reported with potential antimicrobial and anticancer activities.[48] Precursors of fumigaclavine C and ergotamine are represented by the template (**37**). Fumigaclavine C was suggested to exhibit vasorelaxant activity by blockage of calcium channels.[49] Ergotamine is an approved drug for headache treatment. It targets at least 15 proteins, including adrenergic/dopamine receptors or 5-hydroxytryptamine receptors.[50] Lastly, ajmaline precursor (**38**) is an approved drug. Its antiarrythmic effect is the consequence of the binding to sodium channel protein type 5 subunit α.[50]

3.3.4. Aminoglucoside and relatives

Aminoglycosides class contains antibiotic and precursors such as paromamine (**42**), tobramycin (**44**), streptomycin (**42**), kanosamine (**39**, **42**) and myo-inositol (**40**). Except tobramycin, the precursors are not highly similar to their final natural product, since they miss additional rings. Tobramycin is a precursor of kanamycin A, an approved drug blocking protein synthesis in bacteria because of its interaction with ribosomal RNA.[51]

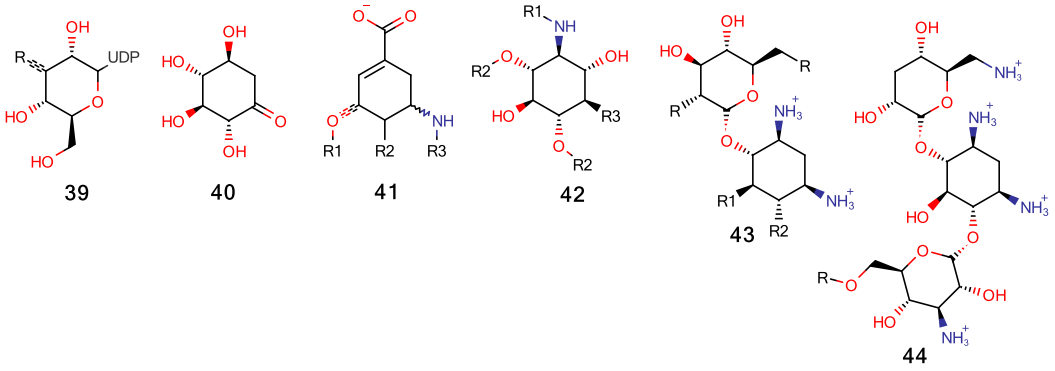


Figure 13. Aminoglycosides and relative compounds in the final dataset.

3.3.5. β-lactam antibiotics

The β-lactams penicillin G (**47**) and cephalosporin (**48**) inhibit peptidoglycan formation in bacterial cell walls.[17] Both molecules are approved drugs widely used as antibiotics during the past decades. Many bacteria have showed systems to circumvent lactam-containing drugs. In particular they produce β-lactamases, which hydrolyze penicillin-like antibiotics, disabling them to target penicillin binding proteins.[17] Clavulanic acid (**46**) is an irreversible inhibitor of β-lactamase.[52] The template (**45**) represents a precursor of clavulanic acid.

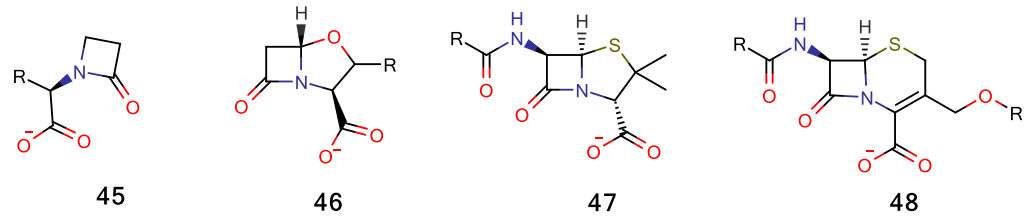


Figure 14. Lactam-containing compounds in the final dataset.

3.3.6. Macrolide compounds

Macrolides, i.e. macrocycles containing lactone group, constitute another class of antibiotics. Several of them, e.g., erythromycin (**50**) are approved drugs. They inhibit protein synthesis in Bacteria and can target different proteins (e.g., 50S ribosomal proteins, ribosomal RNA, cytochrome P450 and lanosterol 14-α demethylase).[50]

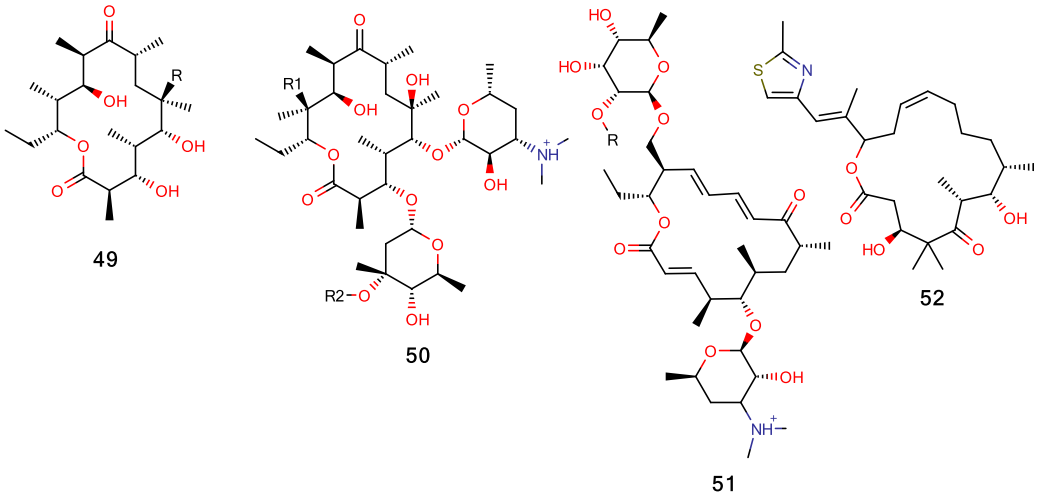


Figure 15. Macrolide compounds in the final dataset.

3.3.7. Precursors and low molecular weight compounds

Precursors and low molecular weight compounds (**figure 16**) are mostly involved in early stages of the biosynthesis of natural products. For example, coumaryl-CoA (**56**) is a precursor of many phenylpropanoids. Mevalonate (**57**) and 4-CDP-2-C-methylerytritol (**58**) are two important early precursors of phosphorylated isoprenes (isopentenyl diphosphate and dimethylallyl diphosphate (**60)**). L-tryptophane's template (**59**) is an early precursor of

the antibiotics rebeccamycin and pyrrolnitrin. Compound **64** represents O-acteyl-L-serine, a precursor of D-cycloserine.

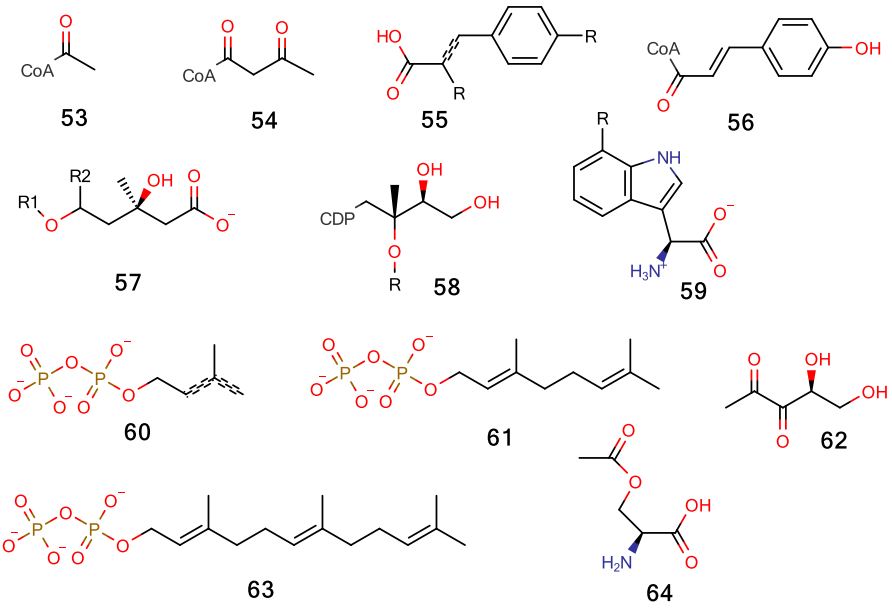


Figure 16. Precursor and low molecular weight compounds in the final dataset.

4. PERSPECTIVES

4.1. Top-down

4.1.1. Text mining

Keywords such as "*biosynth*" are not the most suited because they also match primary metabolite biosynthesis or protein biosynthesis. For example, a structure of 3-isopropylmalate dehydrogenase (PDB ID: 1A05) matched the term "*biosynth*" although it is involved in the biosynthesis of the amino-acid leucine. In addition, the term "natural product" can match publication abstracts of any protein structure that is co-crystalized with a natural product. It is the case of a structure of GTP-binding nuclear protein Ran (PDB ID: 4HB2) which is inhibited by the polyketide leptomycin B. For future investments, it is recommended to, either search with smarter keywords such as natural product names in combination with biosynthetic related keywords, or to extrapolate knowledge-based data instead.

4.1.2. Active site detection

In order to identify active pocket in biosynthetic enzymes, we assumed that all, or at least most of the catalytic residues belong to the substrate binding site. The hypothesis was not always correct and we found three problematic scenarios.

Firstly, an ambiguity occurs when an active site residue is located between two cavities. **Figure 17** illustrates nicely the case; Prostaglandin G/H synthase 1 (PDB ID: 1HT8) catalyzes the cyclization of arachidonate into prostaglandin. Tyr385 is the catalytic residue responsible of cyclooxygenase activity.[53] The Tyr residue is in the binding pocket of natural

substrate, arachidonic acid, but is it also adjacent to another catalytic site. In this second site, there is a second catalytic residue, His207, responsible for peroxidase activity. Thereby, our algorithm selects the active site corresponding to the peroxidase activity because it contains more catalytic residues than the active site corresponding to cyclooxygenase activity.

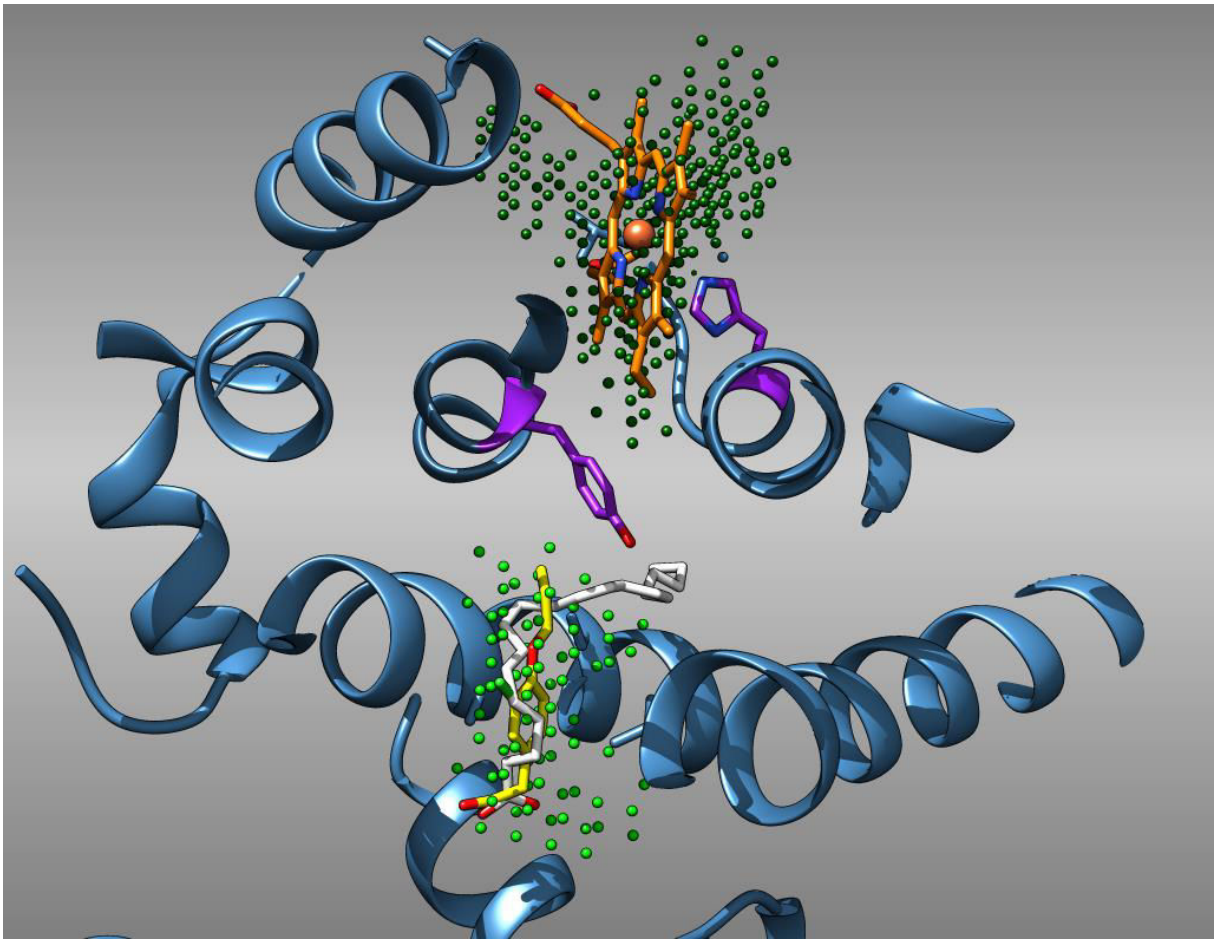


Figure 17. Problem of catalytic cavity detection illustrated with PDB file 1HT8.
View of the structure of prostaglandin G/H synthase (steel blue ribbons) in complex with the inhibitor methyl flurbiprofen (yellow sticks). The natural substrate (arachidonic acid, white sticks), was extracted from another co-crystal structure of prostaglandin G/H synthase (PDB ID: 1DIY) after superimposition of the two enzyme chains (0.44Å on 438 atom pairs in Chimera). Active site residue side chains are represented with purple sticks (Tyr385 and His207). The clouds of green points represent two distinct cavities generated by VolSite. A heme (orange sticks), is present in the upper cavity.

Secondly, cavities are generated without ligand specification in VolSite.[25] The ligand specification provides a reference for a distance cutoff to truncate the cavities. Without ligand, this limit is replaced by geometric parsing of the space around each cavity point. For this reason, some cavities extend outside of the convex-hull of the protein by taking the shape of a “mushroom”. We observed that cavities in our dataset are generally larger to enzymes cavities in sc-PDB (defined around bound ligand). For example, a structure of erythromycin C-12 hydroxylase (PDB ID: 2V59), part of erythromycin’s biosynthesis was assigned a large cavity that exceeds the limits of erythromycin molecular recognition (**figure 18**). A narrow well is present at the bottom of the cavity under the location of the heme group, and the upper part of the cavity extends in a tunnel above the substrate recognition site. An excessively large cavity does not constitute an optimal bait to search ligandable sites of PDB targets. In order to tackle this issue, we suggest to trim the cavities down to keep what recognizes the natural product substrate only.

Lastly, some cavities were not detected because they are too buried. The **figure 19** illustrates the scenario. This problem is inherent to the automated approach of binding site detection.

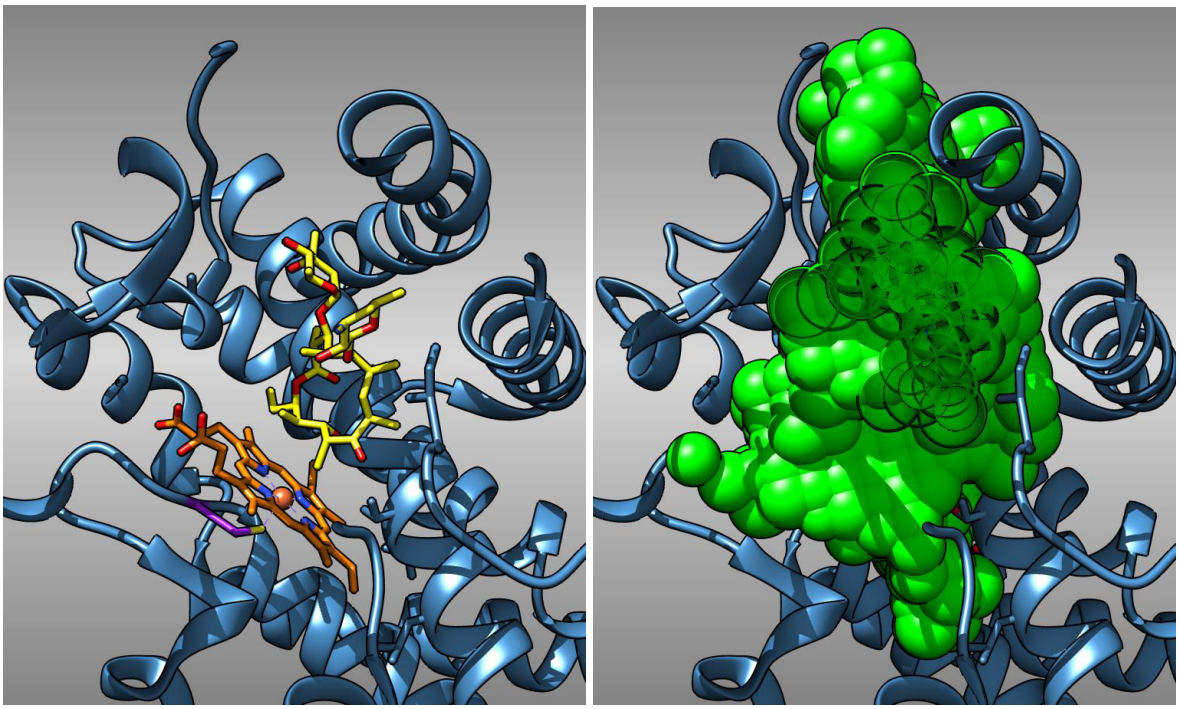


Figure 18. Clipped view of a cavity representation in erythromycin C-12 hydroxylase (PDB ID: 2JJO).
Left: the enzyme backbone is represented with steel blue ribbons. The co-crystalized ligand is the natural substrate, erythromycin precursor, represented with yellow sticks. The orange sticks represent a heme group while it is coordinating an iron ion. The amino-acid side chain of the catalytic residue is represented in purple (Cys35). Right: cavity points generated by VolSite are represented with large green spheres.

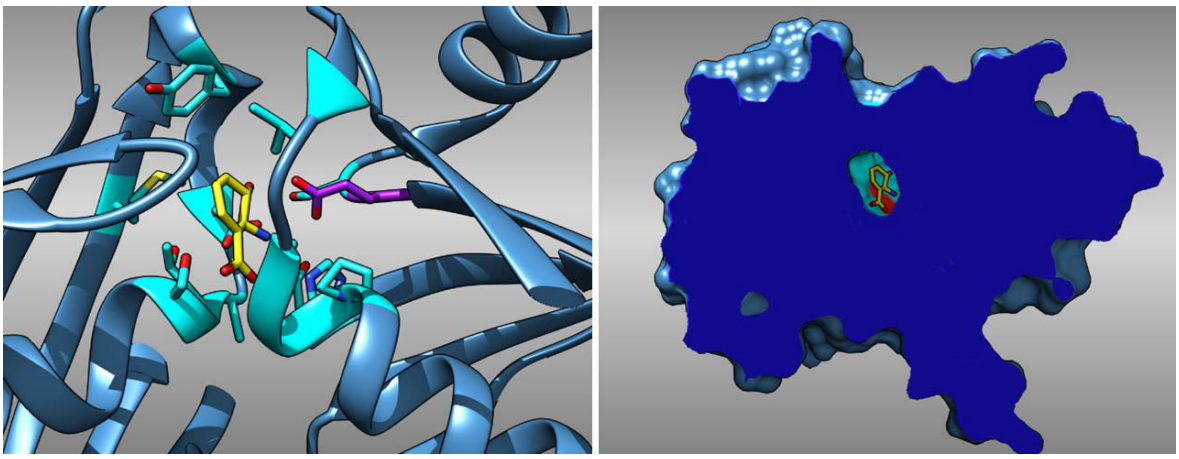


Figure 19. Catalytic site of a structure of trans-2,3-dihydro-3-hydroxyanthranilate isomerase (PDB ID: 1U1X) an enzyme part of phenazine biosynthesis. Left: The protein backbone is presented with steel blue ribbons. The co-crystalized ligand is a phenazine precursor (yellow sticks). Side chains of protein residues proximal to the ligand are represented with cyan sticks. The active site residue is represented with purple sticks. On the right side is shown a clipped view of the protein surface. One can see the buried cavity occupied by the phenazine precursor.

Significant improvement of the method for active cavity detection may be expected in the future if catalytic residue annotations were considered with additional information for the catalytic site identification (e. g. the position of a natural substrate or analogue compound). During the manual validation process, many enzyme structures were visually inspected and their catalytic sites were identified using information obtained in literature only. If enzymatic activity is known, then the natural substrates and products supposed to bind within the cavity are known. If an enzyme structure is available with a substrate, a product or an analog, then the location of the catalytic cavity is identifiable. Similarly, if a structure is co-crystalized with a cofactor only, the position of contributing atom(s) of the cofactor generally points towards the catalytic cavity. Besides, literature often gives insights into catalytic mechanisms. These points are illustrated in **figure 20**. Dihydropinosylvin synthase (UniProt ID: Q02323) concatenates three malonyl-CoA groups and a cinnamoyl-CoA group to form a dihydropinosylvin while releasing four CoA.[54] In the structure (PDB ID: 1XET), the active site contains a CoA and a ligand mimicking a natural product precursor. The positions of the sulfur atom in CoA and the ligand, mimicking the substrate, well define the catalytic cavity. However, automation of the identification of catalytic cavities considering these elements is a challenging task.

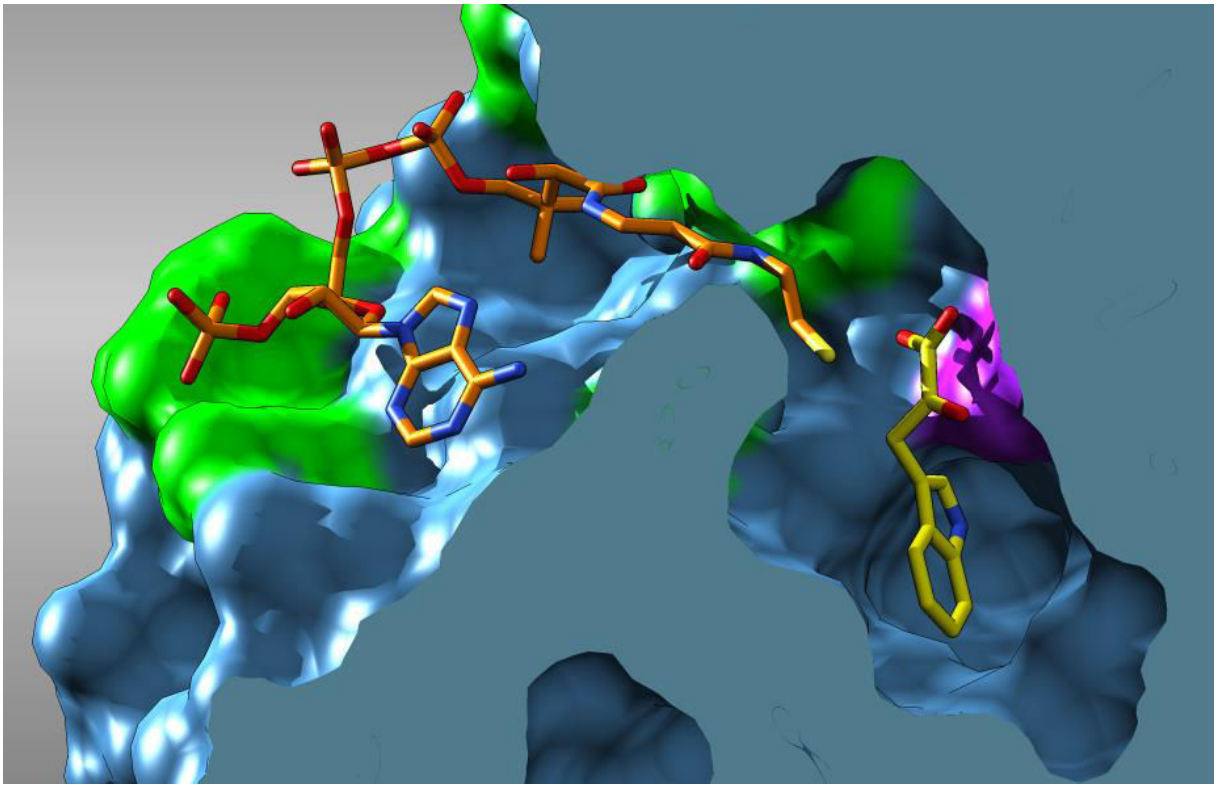


Figure 20. Clipped view of a catalytic site of dihydropinosylvin synthase.
The figure depicts a clipped view of the catalytic cavity in a structure of dihydropinosylvin synthase (PDB ID: 1XET). The enzyme surface is represented in steel blue. Green surfaces represent UniProt documented substrate binding residues. Here CoA represented with orange sticks, mimics the released cofactor. Purple surface represents the active site residue. The yellow molecule, 3-(1H-indol-3-yl)-2-oxopropanoic acid, mimics the positon of dihydropinosylvin, a phytoalexin precursor of hydropinosylvin. The identification of the catalytic site is deducible from green colored residues, from the cofactor sulfur atom, or from the ligand mimicking the natural product.

4.1.3. Catalytic templates

In an attempt to identify catalytic sites of enzymes without active site residue annotations, we have tested a method based on catalytic template graph matching. As reported by Torrance et al,[55] CSA 3D-motifs built from reference enzymes are very close to corresponding CSA "homologous" 3D-motifs extracted from other enzymes (<1Å root mean square deviation), even if sequence similarity between the two enzymes is low. Assuming that catalytic sites may be detected using 3D-motifs, we used Cα and Cβ atoms of catalytic residues to create a dataset of graphs from CSA motifs (**figure 21**). Protein cavities were

then scanned for the presence of these graphs. Due to the complexity of preliminary results and because of time constraints, this method was not pursued. In brief, retrospective tests successfully identified known catalytic motifs, but most proteins cavities matched many small graphs, making prospective searches difficult.

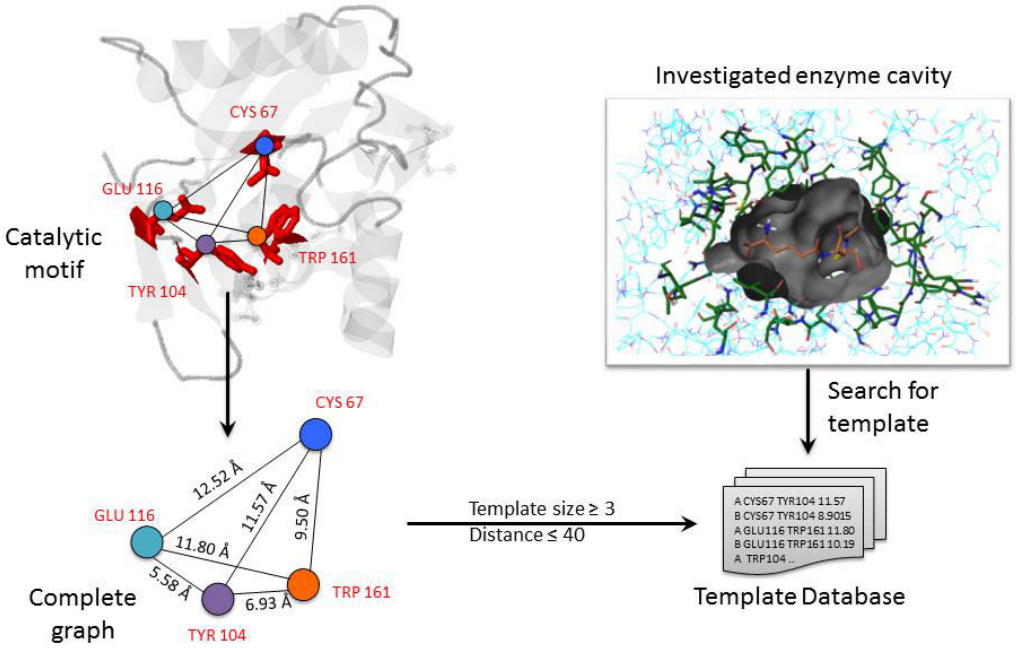


Figure 21. Method of 3D catalytic template search in enzyme cavities.
The figure depicts the search for catalytic templates in one enzyme cavity. Investigated residues are represented by green sticks whereas the rest of the enzyme residues are represented with cyan wires. Beforehand, a graph database was set up. It contains distances between all α-carbons and between all β-carbons characterizing a complete graph formed by 3D catalytic motif documented in the catalytic site atlas, shown here with red sticks residues in an enzyme structure (PDB ID: 3HYQ) represented by grey transparent ribbons. The set of distances of the catalytic template graph is then indexed into a database. In order to investigate enzyme cavities, the database was queried with pairs of amino-acids and a distance tolerance. If the query returned all the pairs of amino-acids from a catalytic motif, we searched the maximum common graph in the cavity. A match was assumed only if the complete graph formed by the catalytic motif was found in the cavity.

4.1.4. UniProt-to-PDB mapping

Not all searched active site residues are presentative of the catalytic sites we are looking for. For example, noranthrone synthase from *Aspergillus parasiticus* (UniProt ID: Q12053,) is a multi-domain enzyme that catalyzes the iterative formation of norsolorinate anthrone,

a mature precursor of aflatoxins, through a series of reactions and an ultimate cyclization.[56–58] The enzyme has four available structures (PDB IDs: 3HRR, 3HRQ, 3ILS, 2KR5) out of which three represent parts of the product template (PT) domain and one represents Thioesterase/Claisen cyclase domain (**figure 22**). UniProt annotation provides three active residues describing distinct activities. β-ketoacyl synthase activity induced by Cys543, acy/malonyl transferase activity induced by Ser993 and thioesterase activity induced by Ser1937.[59] Although in the process, the structures with PDB ID: 3HRR, 3HRQ, and 2KR5 were assigned three catalytic residues, none of them could possibly map the structures because the protein sequences in the structures do not contain the residues of interest. Thioesterase activity is the only activity that is embedded within a structure (PDB ID: 3ILS), but at the time we parsed the data, UniProt annotation did not include the active site residue. However, literature investigations led to evidence that the structures of the PT domain with PDB ID: 3HRR and 3HRQ contains a catalytic motif responsible for the last step cyclization. Crawford et al. suggested its structural basis nicely with site mutagenesis and docking studies.[60] Therefore, we can consider that the molecular recognition of norsolorinate anthrone is embedded in the PT domain. This example also shows that active residues documented in UniProt and CSA are not exhaustive.

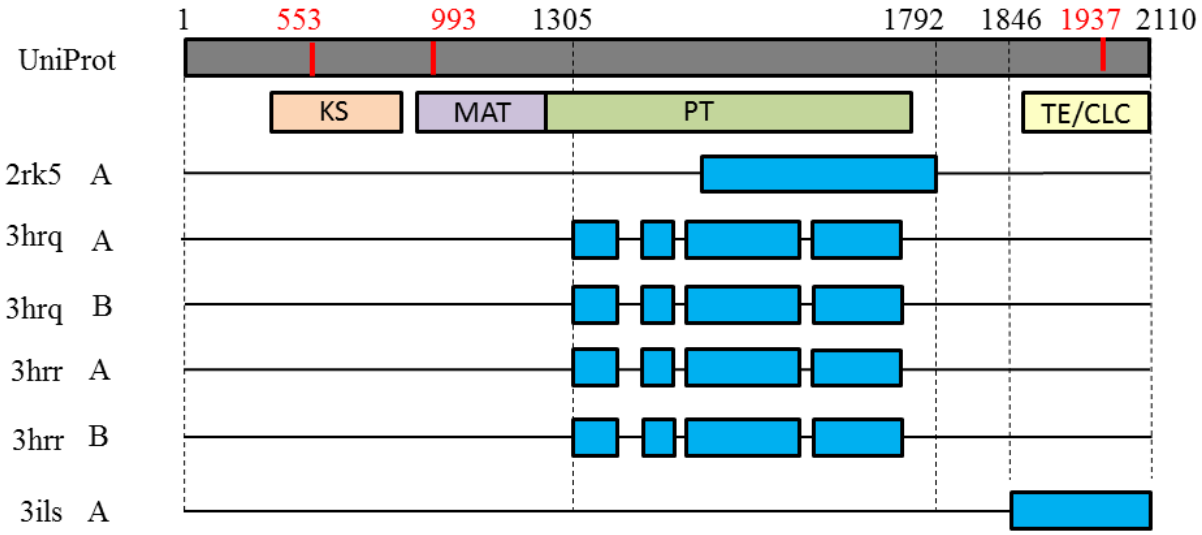


Figure 22. Qualitative sequence alignments of Noranthrone synthase to available structures.
The grey bar represents UniProt sequence. Colored bars underneath UniProt sequence represent different domains of the enzyme. Orange represents the β-ketoacyl synthase domain (KS), purple the *m*alonyl-CoA:ACP transacylase domain (MAT), green the product template domain (PT) and yellow the Thioesterase/Claisen cyclase domain (TE/CLC). Blue bars represent the chain sequences of the different available structures of Noranthrone synthase. Note that the scale was deliberately extended in the area where the structures are known. Active residues from UniProt documentation are denoted with red lines and red numbers.

4.2. Perspectives for ligandable natural products biosynthetic enzyme structures collection.

In this study, we have searched a method to collect ligandable biosynthetic enzymes structures using two different approaches. The top-down workflow has showed that a simple keyword search is not sufficient for efficient search of biosynthetic enzymes but that it has the potential to find structures not described in documented biosynthetic pathways. Here we suggest a workflow combining the different approaches that we tested. The workflow requires development of an extrapolation method able to mine the PDB.

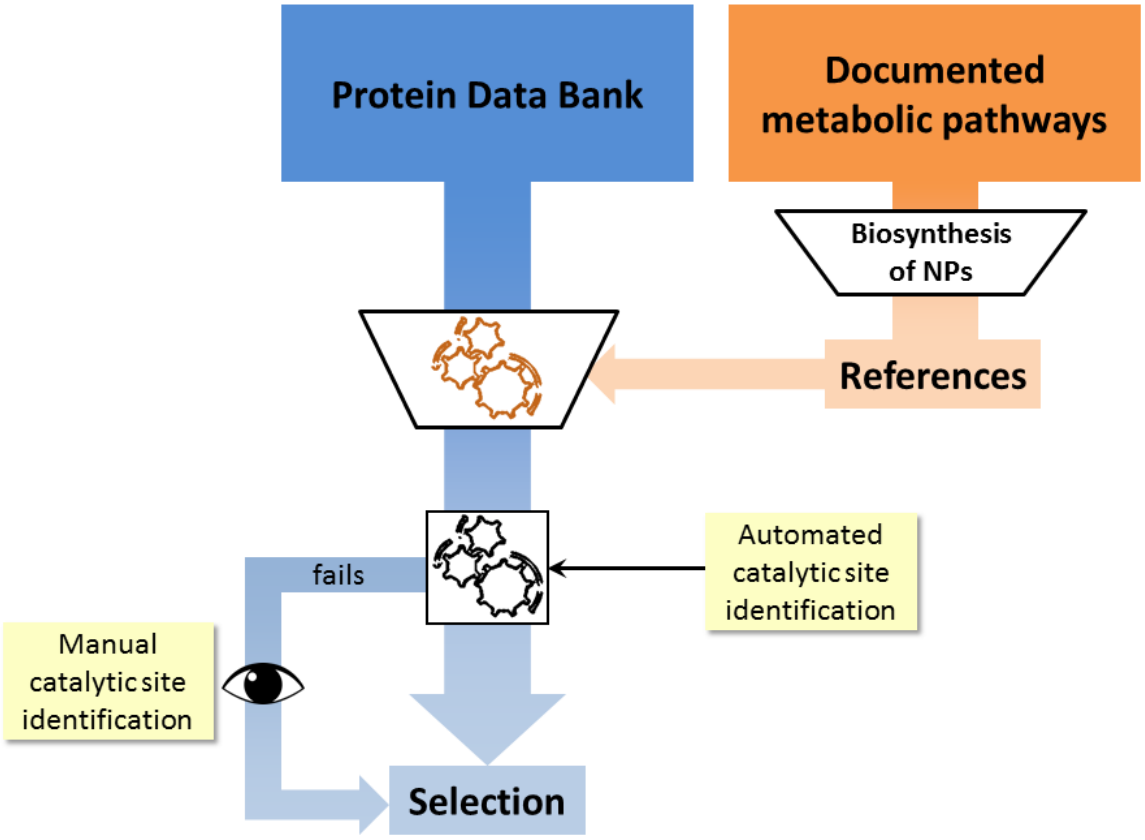


Figure 23. Perspective for the collection of natural products biosynthetic enzymes.
This picture illustrates a single method for the collection of natural products biosynthetic enzymes with ligandable catalytic sites embedding the natural product molecular recognition. All structures of the protein data bank are filtered using a set of predefined knowledge-based rules extrapolated from documented metabolic pathways. Enzymes with enzymatic activities involving mature intermediates are selected first and identified in the PDB. Eventual undocumented orthologous enzymes are selected if they perform an enzymatic activity taking place in the secondary metabolism and if they are expressed in a species with significant evolutionary relationship to documented biosynthetic enzymes of natural products. Follows the automated catalytic site identification process is guided by documented data related to enzymatic activities and positions of relevant ligands. Fails are visually inspected in order to recover eventual missed structures.

Known biosynthetic enzymes provide a valuable asset for extrapolation methods. It is possible to use them for catalytic site identification, using known biosynthetic enzymes are reference catalytic templates. It is not the method that is missing, structural methods such as catalytic site identification are being used and accessible through web servers such as the one provided by Biochemical and biophysical systems group (http://catsid.llnl.gov/).[61]

CONCLUSION

We tested and compared the results of two workflows for collecting ligandable natural product biosynthetic enzymes. In the top-down workflow, the keyword search has yielded high numbers of false positives whereas, as expected, the knowledge-based workflow directly yielded true biosynthetic enzymes. Nevertheless, the top-down workflow has shown its potential for the extrapolation of known biosynthetic enzymes since it was able to collect 12 enzymes undocumented in biosynthetic pathways. The knowledge-based workflows showed that different resources of metabolic information differ in content suggesting the use of other resources for more comprehensive data collection. Moreover, differences in content highlight the lack of a universal pathway ontology.

In this study, we designed an automated catalytic site detection algorithm able to treat approximately half of the retrieved structures and highlighted difficulties linked to biosynthetic enzymes catalytic site identification. Moreover, the automated procedure has consistently reduced manual curation. In practice, only enzymes for whose no active site residue is documented in databases require manual curation. Besides, ligandability filter showed that not all biosynthetic enzymes structures are suited to structure based drug design.

We have identified 117 enzymes involved in the biosynthesis of antibiotics, terpenes, isoprenes, phenylpropanoids, polyketides, alkaloids and other secondary metabolites and more are to be collected through manual curations. As suggested by the literature references, elucidated biosynthesis of natural products have often been extensively studied for their potent pharmacological activities (antibiotics). Moreover, we have

identified a spectrum of enzymes synthetizing precursors at the origin of diverse natural products. Although these enzymes are not suitable for natural product repositioning purposes, ligandability predictions of their catalytic sites suggest that they represent interesting targets for inhibition of biosynthetic pathways in pathogenic species.

Molecular graphics of protein crystallographic structures and analyzes were performed with the UCSF Chimera package. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311).

REFERENCES

1. George, D. G., Barker, W. C. & Hunt, L. T. The protein identification resource (PIR). *Nucleic Acids Res.* **14,** 11–15 (1986).
2. Bairoch, A. & Boeckmann, B. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* **19 Suppl,** 2247–2249 (1991).
3. Tateno, Y., Miyazaki, S., Ota, M., Sugawara, H. & Gojobori, T. DNA data bank of Japan (DDBJ) in collaboration with mass sequencing teams. *Nucleic Acids Res.* **28,** 24–26 (2000).
4. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30,** 418–426 (2014).
5. Keseler, I. M. *et al.* EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res.* **33,** D334–337 (2005).
6. Dwight, S. S. *et al.* Saccharomyces genome database: underlying principles and organisation. *Brief. Bioinform.* **5,** 9–22 (2004).
7. Letovsky, S. I., Cottingham, R. W., Porter, C. J. & Li, P. W. D. GDB: The Human Genome Database. *Nucleic Acids Res.* **26,** 94–99 (1998).
8. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43,** D204–D212 (2015).
9. Chang, A. *et al.* BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res.* gku1068 (2014). doi:10.1093/nar/gku1068
10. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42,** D199–205 (2014).
11. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **42,** D459–D471 (2014).

12. Webb, E. C. *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. (Academic Press, 1992).

13. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42,** D472–D477 (2014).

14. Hastings, J. *et al.* The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* **41,** D456–D463 (2013).

15. Morgat, A. *et al.* Updates in Rhea—a manually curated resource of biochemical reactions. *Nucleic Acids Res.* **43,** D459–D464 (2015).

16. Morgat, A. *et al.* UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.* **40,** D761–D769 (2012).

17. Fisher, J. F., Meroueh, S. O. & Mobashery, S. Bacterial Resistance to β-Lactam Antibiotics: Compelling Opportunism, Compelling Opportunity. *Chem. Rev.* **105,** 395–424 (2005).

18. Xu, Y. *et al.* Rational reprogramming of fungal polyketide first-ring cyclization. *Proc. Natl. Acad. Sci.* **110,** 5398–5403 (2013).

19. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

20. Velankar, S. *et al.* SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.* **41,** D483–D489 (2013).

21. Furnham, N. *et al.* The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.* **42,** D485–D489 (2014).

22. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443–453 (1970).

23. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277

24. Pettersen, E. F. *et al.* UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).

25. Desaphy, J., Azdimousa, K., Kellenberger, E. & Rognan, D. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *J. Chem. Inf. Model.* **52,** 2287–2299 (2012).

26. Green, M. L. & Karp, P. D. The outcomes of pathway database computations depend on pathway ontology. *Nucleic Acids Res.* **34,** 3687–3697 (2006).

27. Croteau, R., Ketchum, R. E. B., Long, R. M., Kaspera, R. & Wildung, M. R. Taxol biosynthesis and molecular genetics. *Phytochem. Rev. Proc. Phytochem. Soc. Eur.* **5,** 75–97 (2006).

28. Kershaw, N. J. *et al.* ORF6 from the clavulanic acid gene cluster of Streptomyces clavuligerus has ornithine acetyltransferase activity. *Eur. J. Biochem. FEBS* **269,** 2052–2059 (2002).

29. Hartmann, S., Neeff, J., Heer, U. & Mecke, D. Arenaemycin (pentalenolactone): a specific inhibitor of glycolysis. *FEBS Lett.* **93,** 339–342 (1978).

30. Duszenko, M. & Mecke, D. Inhibition of glyceraldehyde-3-phosphate dehydrogenase by pentalenolactone in Trypanosoma brucei. *Mol. Biochem. Parasitol.* **19,** 223–229 (1986).

31. Duszenko, M., Balla, H. & Mecke, D. Specific inactivation of glucose metabolism from eucaryotic cells by pentalenolactone. *Biochim. Biophys. Acta* **714,** 344–350 (1982).

32. Cane, D. E. & Sohng, J. K. Inhibition of glyceraldehyde-3-phosphate dehydrogenase by pentalenolactone. 2. Identification of the site of alkylation by tetrahydropentalenolactone. *Biochemistry (Mosc.)* **33,** 6524–6530 (1994).

33. Willson, M., Lauth, N., Perie, J., Callens, M. & Opperdoes, F. R. Inhibition of glyceraldehyde-3-phosphate dehydrogenase by phosphorylated epoxides and alpha-enones. *Biochemistry (Mosc.)* **33,** 214–220 (1994).

34. Hammerschmidt, R. PHYTOALEXINS: What Have We Learned After 60 Years? *Annu. Rev. Phytopathol.* **37,** 285–306 (1999).

35. Maldonado-Bonilla, L. D., Betancourt-Jiménez, M. & Lozoya-Gloria, E. Local and systemic gene expression of sesquiterpene phytoalexin biosynthetic enzymes in plant leaves. *Eur. J. Plant Pathol.* **121,** 439–449 (2008).

36. Giannakopoulou, A. *et al.* Variation in Capsidiol Sensitivity between Phytophthora infestans and Phytophthora capsici Is Consistent with Their Host Range. *PLoS ONE* **9,** e107462 (2014).

37. Alberts, A. W. Discovery, biochemistry and biology of lovastatin. *Am. J. Cardiol.* **62,** 10J–15J (1988).

38. Alberts, A. W. *et al.* Mevinolin: a highly potent competitive inhibitor of hydroxymethylglutaryl-coenzyme A reductase and a cholesterol-lowering agent. *Proc. Natl. Acad. Sci. U. S. A.* **77,** 3957–3961 (1980).

39. Leonard, G. A., Hambley, T. W., McAuley-Hecht, K., Brown, T. & Hunter, W. N. Anthracycline—DNA interactions at unfavourable base-pair triplet-binding sites: structures of d(CGGCCG)/daunomycin and d(TGGCCA)/adriamycin complexes. *Acta Crystallogr. D Biol. Crystallogr.* **49,** 458–467 (1993).

40. Lipscomb, L. A. *et al.* Water Ring Structure at DNA Interfaces: Hydration and Dynamics of DNA-Anthracycline Complexes. *Biochemistry (Mosc.)* **33,** 3649–3659 (1994).

41. Dal Ben, D., Palumbo, M., Zagotto, G., Capranico, G. & Moro, S. DNA topoisomerase II structures and anthracycline activity: insights into ternary complex formation. *Curr. Pharm. Des.* **13,** 2766–2780 (2007).

42. Milewski, S., Chmara, H. & Borowski, E. Anticapsin: an active site directed inhibitor of glucosamine-6-phosphate synthetase from Candida albicans. *Drugs Exp. Clin. Res.* **12,** 577–583 (1986).

43. Chmara, H. Inhibition of glucosamine synthase by bacilysin and anticapsin. *J. Gen. Microbiol.* **131,** 265–271 (1985).

44. Mullaicharam, A. & Maheswaran, A. Pharmacological effects of curcumin. *Int. J. Nutr. Pharmacol. Neurol. Dis.* **2,** 92 (2012).

45. Steele, S. L. *et al.* Loss of M2 muscarinic receptor function inhibits development of hypoxic bradycardia and alters cardiac beta-adrenergic sensitivity in larval zebrafish (Danio rerio). *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **297,** R412–420 (2009).

46. Furey ML & Drevets WC. Antidepressant efficacy of the antimuscarinic drug scopolamine: A randomized, placebo-controlled clinical trial. *Arch. Gen. Psychiatry* **63,** 1121–1129 (2006).

47. Turner, J. M. & Messenger, A. J. in *Advances in Microbial Physiology* (ed. Tempest, A. H. R. and D. W.) **27,** 211–275 (Academic Press, 1986).

48. Kennedy, R. K. *et al.* 5-Methyl phenazine-1-carboxylic acid: a novel bioactive metabolite by a rhizosphere soil bacterium that exhibits potent antimicrobial and anticancer activities. *Chem. Biol. Interact.* **231,** 71–82 (2015).

49. Ma, H.-Y. *et al.* Endophytic fungal metabolite fumigaclavine C causes relaxation of isolated rat aortic rings. *Planta Med.* **72,** 387–392 (2006).

50. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42,** D1091–1097 (2014).

51. Wirmer, J. & Westhof, E. in (ed. Enzymology, B.-M. in) **415,** 180–202 (Academic Press, 2006).

52. Drawz, S. M. & Bonomo, R. A. Three Decades of β-Lactamase Inhibitors. *Clin. Microbiol. Rev.* **23,** 160–201 (2010).

53. Shimokawa, T., Kulmacz, R. J., DeWitt, D. L. & Smith, W. L. Tyrosine 385 of prostaglandin endoperoxide synthase is required for cyclooxygenase catalysis. *J. Biol. Chem.* **265,** 20073–20076 (1990).

54. Preisig-Müller, R., Schwekendiek, A., Brehm, I., Reif, H. J. & Kindl, H. Characterization of a pine multigene family containing elicitor-responsive stilbene synthase genes. *Plant Mol. Biol.* **39,** 221–229 (1999).

55. Torrance, J. W., Bartlett, G. J., Porter, C. T. & Thornton, J. M. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *J. Mol. Biol.* **347,** 565–581 (2005).

56. Crawford, J. M. & Townsend, C. A. New insights into the formation of fungal aromatic polyketides. *Nat. Rev. Microbiol.* **8,** 879–889 (2010).

57. Crawford, J. M. *et al.* Deconstruction of iterative multidomain polyketide synthase function. *Science* **320,** 243–246 (2008).

58. Yabe, K. & Nakajima, H. Enzyme reactions and genes in aflatoxin biosynthesis. *Appl. Microbiol. Biotechnol.* **64,** 745–755 (2004).

59. Korman, T. P. *et al.* Structure and function of an iterative polyketide synthase thioesterase domain catalyzing Claisen cyclization in aflatoxin biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 6246–6251 (2010).

60. Crawford, J. M. *et al.* Structural basis for biosynthetic programming of fungal aromatic polyketide cyclization. *Nature* **461,** 1139–1143 (2009).

61. Kirshner, D. A., Nilmeier, J. P. & Lightstone, F. C. Catalytic site identification—a web server to identify catalytic site structural matches throughout PDB. *Nucleic Acids Res.* **41,** W256–W265 (2013).

ANNEX 4

(1) http://metacyc.org/getxml?id=META:SECONDARY-METABOLITE-BIOSYNTHESIS

(2) http://metacyc.org/getxml?id=META:[PWY-ID]

(3) http://metacyc.org/META/pathway-genes?object=[PWY-ID]

(4) http://metacyc.org/META/NEW-IMAGE?type=GENE-IN-PWY&object=[GENE-ID]

(5) http://websvc.biocyc.org/getxml?id=META:[RXN-ID]

(6) http://websvc.biocyc.org/ getxml?META:[CMPD-ID]

Where

PWY-ID	is a pathway identifier
GENE-ID	is a gene identifier
RXN-ID	is a reaction identifier
CMPD-ID	is a compound identifier

The search for secondary metabolites biosynthetic parent pathways starts by querying (1).

Then, all children pathways are being searched recursively using the query url (2).

Once the list of all pathways is set, genes identifiers taking place in the selected pathways are searched using the query url (3).

The gene description html pages are loaded from MetaCyc website using the query url (4).

Reaction identifiers found by (3) are used to obtain an xml reaction description page using the query url (5).

Finally, smiles structures of all the compounds present in the reaction description are being searched using the query url (6).

S1. Generic query urls used for biosynthetic enzymes and compounds structures collection from MetaCyc

Table 1. Natural products biosynthesis pathways involving biosynthetic enzymes with known structures.

Pathway class	Pathway instance
Alkaloid biosynthesis	(S)-scoulerine biosynthesis
Alkaloid biosynthesis	3alpha(S)-strictosidine biosynthesis
Alkaloid biosynthesis	ajmaline biosynthesis
Alkaloid biosynthesis	ergot alkaloid biosynthesis
Alkaloid biosynthesis	taxol biosynthesis
Alkaloid biosynthesis	tropane alkaloid biosynthesis
Antibiotic biosynthesis	aclacinomycin biosynthesis
Antibiotic biosynthesis	actinorhodin biosynthesis
Antibiotic biosynthesis	bacillaene biosynthesis
Antibiotic biosynthesis	bacilysin biosynthesis
Antibiotic biosynthesis	butirosin biosynthesis
Antibiotic biosynthesis	calcium-dependent antibiotic biosynthesis
Antibiotic biosynthesis	carbapenem biosynthesis
Antibiotic biosynthesis	carminomycin biosynthesis
Antibiotic biosynthesis	cephalosporin C biosynthesis
Antibiotic biosynthesis	clavulanate biosynthesis
Antibiotic biosynthesis	daunorubicin biosynthesis
Antibiotic biosynthesis	erythromycin biosynthesis
Antibiotic biosynthesis	gramicidin S biosynthesis
Antibiotic biosynthesis	kanamycin biosynthesis
Antibiotic biosynthesis	kanosamine biosynthesis
Antibiotic biosynthesis	mersacidin biosynthesis
Antibiotic biosynthesis	mycinamicin biosynthesis
Antibiotic biosynthesis	neopentalenolactone biosynthesis
Antibiotic biosynthesis	nisin biosynthesis
Antibiotic biosynthesis	novobiocin biosynthesis
Antibiotic biosynthesis	oxytetracycline biosynthesis
Antibiotic biosynthesis	penicillin G biosynthesis
Antibiotic biosynthesis	pentalenolactone biosynthesis
Antibiotic biosynthesis	phenazine biosynthesis
Antibiotic biosynthesis	phosphinothricin biosynthesis
Antibiotic biosynthesis	rhodomycin biosynthesis
Antibiotic biosynthesis	rifamycin B biosynthesis
Antibiotic biosynthesis	streptomycin biosynthesis
Antibiotic biosynthesis	surfactin biosynthesis
Antibiotic biosynthesis	tetracenomycin C biosynthesis
Antibiotic biosynthesis	tobramycin biosynthesis
Antibiotic biosynthesis	tylosin biosynthesis
Antibiotic biosynthesis	tyrocidine biosynthesis
Antibiotic biosynthesis	vancomycin biosynthesis

Pathway class	Pathway instance
Aromatic compound metabolism	phenylpropanoid biosynthesis
Carotenoid biosynthesis	lycopene biosynthesis
Carotenoid biosynthesis	staphyloxanthin biosynthesis
Isoprenoid biosynthesis	dimethylallyl diphosphate biosynthesis
Isoprenoid biosynthesis	farnesyl diphosphate biosynthesis
Isoprenoid biosynthesis	geranyl diphosphate biosynthesis
Isoprenoid biosynthesis	geranylgeranyl diphosphate biosynthesis
Isoprenoid biosynthesis	isopentenyl diphosphate biosynthesis via DXP pathway
Isoprenoid biosynthesis	isopentenyl diphosphate biosynthesis via mevalonate pathway
Lipid metabolism	mycolic acid biosynthesis
Lipid metabolism	rhamnolipid biosynthesis
Lipid metabolism	steroid biosynthesis
Mycotoxin biosynthesis	aflatoxin biosynthesis
Phenylpropanoid metabolism	trans-cinnamate biosynthesis
Phytoalexin biosynthesis	hydropinosylvin biosynthesis
Phytoalexin biosynthesis	medicarpin biosynthesis
Phytoalexin biosynthesis	pterocarpan phytoalexin biosynthesis
Pigment biosynthesis	anthocyanin biosynthesis
Pigment biosynthesis	violacein biosynthesis
Plant hormone biosynthesis	gibberellin biosynthesis
Polyketide biosynthesis	lovastatin biosynthesis
Secondary metabolite biosynthesis	2,4-dihydroxy-1,4-benzoxazin-3-one biosynthesis
Secondary metabolite biosynthesis	epothilone biosynthesis
Secondary metabolite biosynthesis	flavonoid biosynthesis
Secondary metabolite biosynthesis	hopanoid biosynthesis
Secondary metabolite biosynthesis	terpenoid biosynthesis
Secondary metabolite metabolism	quinolate metabolism
Sesquiterpene biosynthesis	aristolochene biosynthesis
Sesquiterpene biosynthesis	epi-isozizaene biosynthesis
Sesquiterpene biosynthesis	pentalenene biosynthesis
Sesquiterpene biosynthesis	trichothecene biosynthesis
Steroid biosynthesis	cholesterol biosynthesis
Steroid biosynthesis	estrogen biosynthesis
Steroid biosynthesis	zymosterol biosynthesis
Terpene metabolism	(R)-camphor biosynthesis

Pathway class	Pathway instance
Terpene metabolism	lanosterol biosynthesis
Terpene metabolism	oleoresin biosynthesis

Table 1. Natural products biosynthesis pathways involving biosynthetic enzymes with known structures.

Table2. Miscellaneous pathways involving enzymes with known structures.

Pathway class	Pathway instance
Alcohol metabolism	ethanol degradation
Alkaloid degradation	cocaine degradation
Alkaloid degradation	nicotine degradation
Alkene biosynthesis	ethylene biosynthesis via S-adenosyl-L-methionine
Alkene metabolism	propylene degradation
Amine and polyamine biosynthesis	agmatine biosynthesis
Amine and polyamine biosynthesis	betaine biosynthesis via choline pathway
Amine and polyamine biosynthesis	carnitine biosynthesis
Amine and polyamine biosynthesis	creatine biosynthesis
Amine and polyamine biosynthesis	ectoine biosynthesis
Amine and polyamine biosynthesis	histamine biosynthesis
Amine and polyamine biosynthesis	putrescine biosynthesis via agmatine pathway
Amine and polyamine biosynthesis	putrescine biosynthesis via L-ornithine pathway
Amine and polyamine biosynthesis	S-adenosylmethioninamine biosynthesis
Amine and polyamine biosynthesis	spermidine biosynthesis
Amine and polyamine biosynthesis	spermine biosynthesis
Amine and polyamine degradation	betaine degradation
Amine and polyamine degradation	creatinine degradation
Amine and polyamine degradation	ethanolamine degradation
Amine and polyamine degradation	putrescine degradation
Amine and polyamine metabolism	carnitine metabolism
Amine and polyamine metabolism	spermidine metabolism
Amino-acid biosynthesis	beta-alanine biosynthesis
Amino-acid biosynthesis	D-alanine biosynthesis
Amino-acid biosynthesis	ergothioneine biosynthesis
Amino-acid biosynthesis	glycine biosynthesis
Amino-acid biosynthesis	L-arginine biosynthesis
Amino-acid biosynthesis	L-arginine biosynthesis [regulation].
Amino-acid biosynthesis	L-asparagine biosynthesis
Amino-acid biosynthesis	L-cysteine biosynthesis
Amino-acid biosynthesis	L-glutamate biosynthesis via GLT pathway
Amino-acid biosynthesis	L-histidine biosynthesis
Amino-acid biosynthesis	L-homocysteine biosynthesis
Amino-acid biosynthesis	L-isoleucine biosynthesis

Pathway class	Pathway instance
Amino-acid biosynthesis	L-leucine biosynthesis
Amino-acid biosynthesis	L-lysine biosynthesis via AAA pathway
Amino-acid biosynthesis	L-lysine biosynthesis via DAP pathway
Amino-acid biosynthesis	L-methionine biosynthesis via de novo pathway
Amino-acid biosynthesis	L-methionine biosynthesis via salvage pathway
Amino-acid biosynthesis	L-phenylalanine biosynthesis
Amino-acid biosynthesis	L-proline biosynthesis
Amino-acid biosynthesis	L-pyrrolysine biosynthesis
Amino-acid biosynthesis	L-serine biosynthesis
Amino-acid biosynthesis	L-threonine biosynthesis
Amino-acid biosynthesis	L-tryptophan biosynthesis
Amino-acid biosynthesis	L-tyrosine biosynthesis
Amino-acid biosynthesis	L-valine biosynthesis
Amino-acid biosynthesis	S-adenosyl-L-methionine biosynthesis
Amino-acid degradation	4-aminobutanoate degradation
Amino-acid degradation	Ehrlich pathway
Amino-acid degradation	L-alanine degradation via dehydrogenase pathway
Amino-acid degradation	L-alanine degradation via transaminase pathway
Amino-acid degradation	L-arginine degradation via ADI pathway
Amino-acid degradation	L-arginine degradation via AST pathway
Amino-acid degradation	L-glutamate degradation via hydroxyglutarate pathway
Amino-acid degradation	L-glutamate degradation via mesaconate pathway
Amino-acid degradation	L-histidine degradation into L-glutamate
Amino-acid degradation	L-kynurenine degradation
Amino-acid degradation	L-leucine degradation
Amino-acid degradation	L-lysine degradation via acetate pathway
Amino-acid degradation	L-lysine degradation via saccharopine pathway
Amino-acid degradation	L-phenylalanine degradation
Amino-acid degradation	L-proline degradation into L-glutamate
Amino-acid degradation	L-threonine degradation via oxydo-reductase pathway
Amino-acid degradation	L-threonine degradation via propanoate pathway
Amino-acid degradation	L-tryptophan degradation via kynurenine pathway
Amino-acid degradation	L-tryptophan degradation via pyruvate pathway
Amino-acid degradation	L-valine degradation
Amino-acid metabolism	lysine degradation
Amino-acid metabolism	tryptophan metabolism
Aminoacyl-tRNA biosynthesis	selenocysteinyl-tRNA(Sec) biosynthesis
Amino-sugar metabolism	1,6-anhydro-N-acetylmuramate degradation

Pathway class	Pathway instance
Amino-sugar metabolism	N-acetylmuramate degradation
Amino-sugar metabolism	N-acetylneuraminate biosynthesis
Amino-sugar metabolism	N-acetylneuraminate degradation
Amino-sugar metabolism	N-acetylneuraminate metabolism
Aromatic compound metabolism	(R)-mandelate degradation
Aromatic compound metabolism	3,4-dihydroxybenzoate biosynthesis
Aromatic compound metabolism	3-chlorocatechol degradation
Aromatic compound metabolism	3-phenylpropanoate degradation
Aromatic compound metabolism	4-hydroxyphenylacetate degradation
Aromatic compound metabolism	benzene degradation
Aromatic compound metabolism	benzoate degradation via hydroxylation
Aromatic compound metabolism	benzoate degradation via hydroxylation.
Aromatic compound metabolism	benzoyl-CoA degradation
Aromatic compound metabolism	beta-ketoadipate pathway
Aromatic compound metabolism	melatonin biosynthesis
Aromatic compound metabolism	naphthalene degradation
Aromatic compound metabolism	p-cresol degradation
Aromatic compound metabolism	phenol degradation
Aromatic compound metabolism	phenylacetate degradation
Aromatic compound metabolism	serotonin biosynthesis
Bacterial outer membrane biogenesis	enterobacterial common antigen biosynthesis
Bacterial outer membrane biogenesis	lipopolysaccharide biosynthesis
Bacterial outer membrane biogenesis	LOS core biosynthesis
Bacterial outer membrane biogenesis	LPS core biosynthesis
Bacterial outer membrane biogenesis	LPS lipid A biosynthesis
Bacterial outer membrane biogenesis	LPS O-antigen biosynthesis
Biopolymer metabolism	poly-(R)-3-hydroxybutanoate biosynthesis
Capsule biogenesis	capsule polysaccharide biosynthesis
Carbohydrate acid metabolism	2-dehydro-3-deoxy-D-gluconate degradation
Carbohydrate acid metabolism	D-galactonate degradation
Carbohydrate acid metabolism	D-galacturonate degradation via prokaryotic oxidative pathway
Carbohydrate acid metabolism	D-glucarate degradation
Carbohydrate acid metabolism	D-gluconate degradation
Carbohydrate acid metabolism	galactarate degradation
Carbohydrate acid metabolism	tartrate degradation
Carbohydrate biosynthesis	2-(alpha-D-mannosyl)-D-glycerate biosynthesis
Carbohydrate biosynthesis	3-deoxy-D-manno-octulosonate biosynthesis
Carbohydrate biosynthesis	Calvin cycle.
Carbohydrate biosynthesis	D-glycero-D-manno-heptose 7-phosphate biosynthesis
Carbohydrate biosynthesis	D-ribose 5-phosphate biosynthesis
Carbohydrate biosynthesis	dTDP-L-rhamnose biosynthesis

Pathway class	Pathway instance
Carbohydrate biosynthesis	gluconeogenesis.
Carbohydrate degradation	2-deoxy-D-ribose 1-phosphate degradation
Carbohydrate degradation	D-allose degradation
Carbohydrate degradation	glycolysis
Carbohydrate degradation	L-arabinose degradation via L-arabinitol
Carbohydrate degradation	L-arabinose degradation via L-ribulose
Carbohydrate degradation	L-fucose degradation
Carbohydrate degradation	L-rhamnose degradation
Carbohydrate degradation	pentose phosphate pathway
Carbohydrate metabolism	1,5-anhydro-D-fructose degradation
Carbohydrate metabolism	D-ribose degradation
Carbohydrate metabolism	D-sorbitol biosynthesis
Carbohydrate metabolism	D-tagatose 6-phosphate degradation
Carbohydrate metabolism	D-xylose degradation
Carbohydrate metabolism	fructose metabolism
Carbohydrate metabolism	galactose metabolism
Carbohydrate metabolism	glyoxylate and dicarboxylate metabolism
Carbohydrate metabolism	glyoxylate cycle
Carbohydrate metabolism	hexose metabolism
Carbohydrate metabolism	lactose degradation
Carbohydrate metabolism	L-fucose metabolism
Carbohydrate metabolism	L-rhamnose metabolism
Carbohydrate metabolism	pentose and glucuronate interconversion.
Carbohydrate metabolism	pyruvate metabolism
Carbohydrate metabolism	tricarboxylic acid cycle
Catecholamine biosynthesis	(R)-adrenaline biosynthesis
Catecholamine biosynthesis	(R)-noradrenaline biosynthesis
Catecholamine biosynthesis	dopamine biosynthesis
Cell wall biogenesis	cell wall polysaccharide biosynthesis
Cell wall biogenesis	lipoteichoic acid biosynthesis
Cell wall biogenesis	peptidoglycan biosynthesis
Cell wall biogenesis	peptidoglycan recycling.
Cell wall biogenesis	poly(glucopyranosyl N-acetylgalactosamine 1-phosphate) teichoic acid biosynthesis
Cell wall biogenesis	poly(glycerol phosphate) teichoic acid biosynthesis
Cell wall biogenesis	poly(ribitol phosphate) teichoic acid biosynthesis
Cell wall degradation	peptidoglycan degradation
Cofactor biosynthesis	(R)-pantothenate biosynthesis
Cofactor biosynthesis	5,6,7,8-tetrahydromethanopterin biosynthesis
Cofactor biosynthesis	7,8-dihydroneopterin triphosphate biosynthesis
Cofactor biosynthesis	adenosylcobalamin biosynthesis

Pathway class	Pathway instance
Cofactor biosynthesis	B6 vitamer interconversion
Cofactor biosynthesis	biotin biosynthesis
Cofactor biosynthesis	coenzyme A biosynthesis
Cofactor biosynthesis	coenzyme F420 biosynthesis
Cofactor biosynthesis	coenzyme M biosynthesis
Cofactor biosynthesis	FAD biosynthesis
Cofactor biosynthesis	FMN biosynthesis
Cofactor biosynthesis	iron-sulfur cluster biosynthesis
Cofactor biosynthesis	L-ascorbate biosynthesis via UDP-alpha-D-glucuronate pathway
Cofactor biosynthesis	methanofuran biosynthesis
Cofactor biosynthesis	molybdopterin biosynthesis
Cofactor biosynthesis	NAD(+) biosynthesis
Cofactor biosynthesis	NAD(+) biosynthesis [regulation].
Cofactor biosynthesis	nicotinate biosynthesis
Cofactor biosynthesis	phylloquinone biosynthesis
Cofactor biosynthesis	prenylquinone biosynthesis
Cofactor biosynthesis	pyridoxal 5'-phosphate biosynthesis
Cofactor biosynthesis	pyridoxine 5'-phosphate biosynthesis
Cofactor biosynthesis	pyrroloquinoline quinone biosynthesis
Cofactor biosynthesis	riboflavin biosynthesis
Cofactor biosynthesis	tetrahydrobiopterin biosynthesis
Cofactor biosynthesis	tetrahydrofolate biosynthesis
Cofactor biosynthesis	tetrahydrofolylpolyglutamate biosynthesis
Cofactor biosynthesis	thiamine diphosphate biosynthesis
Cofactor biosynthesis	ubiquinone biosynthesis
Cofactor degradation	B6 vitamer degradation
Cofactor degradation	L-ascorbate degradation
Cofactor degradation	nicotinate degradation
Cofactor metabolism	retinol metabolism
Energy metabolism	electron transfer.
Energy metabolism	nitrogen metabolism
Energy metabolism	oxidative phosphorylation.
Energy metabolism	photosynthesis.
Energy metabolism	sulfur metabolism
Exopolysaccharide biosynthesis	colanic acid biosynthesis
Fermentation	ethanol fermentation.
Fermentation	pyruvate fermentation
Fermentation	pyruvate fermentation to lactate
Flavonoid metabolism	quercetin degradation
Genetic information processing	DNA modification.
Genetic information processing	DNA replication.
Glucan metabolism	xyloglucan degradation

Pathway class	Pathway instance
Glycan biosynthesis	alginate biosynthesis
Glycan biosynthesis	glycogen biosynthesis
Glycan biosynthesis	starch biosynthesis
Glycan biosynthesis	trehalose biosynthesis
Glycan biosynthesis	xanthan biosynthesis
Glycan degradation	chitin degradation
Glycan degradation	glycogen degradation
Glycan degradation	starch degradation
Glycan degradation	xylan degradation
Glycan metabolism	bacterial cellulose biosynthesis
Glycan metabolism	beta-D-glucan degradation
Glycan metabolism	cellulose degradation
Glycan metabolism	exopolysaccharide biosynthesis
Glycan metabolism	heparan sulfate biosynthesis
Glycan metabolism	heparin biosynthesis
Glycan metabolism	L-arabinan degradation
Glycan metabolism	N-glycan degradation
Glycan metabolism	osmoregulated periplasmic glucan (OPG) biosynthesis
Glycan metabolism	pectin degradation
Glycan metabolism	plant cellulose biosynthesis
Glycerolipid metabolism	ether lipid biosynthesis
Glycerolipid metabolism	triacylglycerol degradation
Glycolipid biosynthesis	lipid IV(A) biosynthesis
Hydrocarbon metabolism	alkane degradation
Ketone metabolism	succinyl-CoA degradation
Lipid metabolism	arachidonate metabolism
Lipid metabolism	bile acid biosynthesis
Lipid metabolism	bile acid degradation
Lipid metabolism	C21-steroid hormone metabolism
Lipid metabolism	fatty acid beta-oxidation.
Lipid metabolism	fatty acid biosynthesis
Lipid metabolism	fatty acid metabolism
Lipid metabolism	fatty acid reduction for biolumincescence.
Lipid metabolism	hydroperoxy eicosatetraenoic acid biosynthesis
Lipid metabolism	leukotriene A4 biosynthesis
Lipid metabolism	leukotriene B4 biosynthesis
Lipid metabolism	malonyl-CoA biosynthesis
Lipid metabolism	mitochondrial fatty acid beta-oxidation.
Lipid metabolism	oxylipin biosynthesis
Lipid metabolism	peroxisomal fatty acid beta-oxidation.
Lipid metabolism	phospholipid metabolism
Lipid metabolism	prostaglandin biosynthesis

Pathway class	Pathway instance
Lipid metabolism	short-chain fatty acid metabolism
Lipid metabolism	sphingolipid metabolism
Membrane lipid metabolism	glycerophospholipid metabolism
Metabolic intermediate biosynthesis	(R)-mevalonate biosynthesis
Metabolic intermediate biosynthesis	1-deoxy-D-xylulose 5-phosphate biosynthesis
Metabolic intermediate biosynthesis	2-deoxystreptamine biosynthesis
Metabolic intermediate biosynthesis	5-phospho-alpha-D-ribose 1-diphosphate biosynthesis
Metabolic intermediate biosynthesis	acetyl-CoA biosynthesis
Metabolic intermediate biosynthesis	chorismate biosynthesis
Metabolic intermediate biosynthesis	prephenate biosynthesis
Metabolic intermediate degradation	oxalate degradation
Metabolic intermediate metabolism	(R)-mevalonate degradation
Metabolic intermediate metabolism	(S)-3-hydroxy-3-methylglutaryl-CoA degradation
Metabolic intermediate metabolism	carbamoyl phosphate degradation
Metabolic intermediate metabolism	lactate oxidation.
Metabolic intermediate metabolism	propanoyl-CoA degradation
mRNA processing	mRNA capping.
Nitrogen metabolism	(S)-allantoin degradation
Nitrogen metabolism	nitrate reduction (assimilation).
Nitrogen metabolism	nitrate reduction (denitrification)
Nitrogen metabolism	nitric oxide reduction.
Nitrogen metabolism	urea cycle
Nitrogen metabolism	urea degradation
Nucleoside biosynthesis	alpha-ribazole biosynthesis
Nucleotide metabolism	nucleotide salvage pathway
Nucleotide-sugar biosynthesis	ADP-L-glycero-beta-D-manno-heptose biosynthesis
Nucleotide-sugar biosynthesis	CDP-3,6-dideoxy-D-mannose biosynthesis
Nucleotide-sugar biosynthesis	CMP-3-deoxy-D-manno-octulosonate biosynthesis
Nucleotide-sugar biosynthesis	dTDP-4-acetamido-4,6-dideoxygalactose biosynthesis
Nucleotide-sugar biosynthesis	GDP-alpha-D-mannose biosynthesis
Nucleotide-sugar biosynthesis	GDP-L-fucose biosynthesis via de novo pathway
Nucleotide-sugar biosynthesis	UDP-4-deoxy-4-formamido-beta-L-arabinose biosynthesis
Nucleotide-sugar biosynthesis	UDP-alpha-D-glucuronate biosynthesis
Nucleotide-sugar biosynthesis	UDP-alpha-D-xylose biosynthesis
Nucleotide-sugar biosynthesis	UDP-N-acetyl-alpha-D-glucosamine biosynthesis
One-carbon metabolism	formaldehyde assimilation via RuMP pathway
One-carbon metabolism	formaldehyde assimilation via serine pathway
One-carbon metabolism	formaldehyde degradation

Pathway class	Pathway instance
One-carbon metabolism	methanogenesis from CO(2)
One-carbon metabolism	methanogenesis from methylamine.
One-carbon metabolism	methylamine degradation
One-carbon metabolism	methyl-coenzyme M reduction
One-carbon metabolism	tetrahydrofolate interconversion.
Organic acid metabolism	2-oxosuberate biosynthesis
Organic acid metabolism	glycolate biosynthesis
Organic acid metabolism	glycolate degradation
Organic acid metabolism	propanoate degradation
Organosulfur biosynthesis	taurine biosynthesis
Organosulfur degradation	taurine degradation via aerobic pathway
Organosulfur degradation	thiocyanate degradation
Phospholipid metabolism	CDP-diacylglycerol biosynthesis
Phospholipid metabolism	CDP-diacylglycerol degradation
Phospholipid metabolism	phosphatidylcholine biosynthesis
Phospholipid metabolism	phosphatidylethanolamine biosynthesis
Phospholipid metabolism	phosphatidylglycerol biosynthesis
Phospholipid metabolism	phosphatidylinositol metabolism
Phospholipid metabolism	phosphatidylinositol phosphate biosynthesis
Phosphorus metabolism	phosphonate biosynthesis
Photosynthesis	C3 acid pathway
Photosynthesis	C4 acid pathway
Photosynthesis	photorespiration
Pigment biosynthesis	melanin biosynthesis
Pigment biosynthesis	ommochrome biosynthesis
Plant hormone metabolism	auxin biosynthesis
Polyol metabolism	(R,R)-butane-2,3-diol biosynthesis
Polyol metabolism	glycerol degradation
Polyol metabolism	glycerol degradation via glycerol kinase pathway
Polyol metabolism	glycerol fermentation
Polyol metabolism	myo-inositol biosynthesis
Polyol metabolism	myo-inositol degradation into acetyl-CoA
Polyol metabolism	myo-inositol degradation into acetyl-CoA.
Polyol metabolism	myo-inositol degradation into D-glucuronate
Porphyrin-containing compound metabolism	bacteriochlorophyll biosynthesis
Porphyrin-containing compound metabolism	bacteriochlorophyll biosynthesis (light-independent).
Porphyrin-containing compound metabolism	chlorophyll biosynthesis
Porphyrin-containing compound metabolism	chlorophyll biosynthesis (light-independent).
Porphyrin-containing compound metabolism	chlorophyll degradation

Pathway class	Pathway instance
Porphyrin-containing compound metabolism	protoheme biosynthesis
Porphyrin-containing compound metabolism	protoheme degradation
Porphyrin-containing compound metabolism	protoporphyrin-IX biosynthesis
Porphyrin-containing compound metabolism	siroheme biosynthesis
Protein biosynthesis	polypeptide chain elongation.
Protein degradation	proteasomal Pup-dependent pathway
Protein degradation	proteasomal ubiquitin-dependent pathway
Protein modification	[NiFe] hydrogenase maturation.
Protein modification	cytochrome c assembly.
Protein modification	eIF5A hypusination.
Protein modification	peptidyl-diphthamide biosynthesis
Protein modification	protein glycosylation.
Protein modification	protein lipoylation via endogenous pathway
Protein modification	protein lipoylation via exogenous pathway
Protein modification	protein neddylation.
Protein modification	protein pupylation.
Protein modification	protein sumoylation.
Protein modification	protein ubiquitination.
Protein modification	sulfatase oxidation.
Purine metabolism	3',5'-cyclic AMP degradation
Purine metabolism	3',5'-cyclic di-GMP biosynthesis
Purine metabolism	3',5'-cyclic GMP degradation
Purine metabolism	7-cyano-7-deazaguanine biosynthesis
Purine metabolism	AMP biosynthesis via de novo pathway
Purine metabolism	AMP biosynthesis via salvage pathway
Purine metabolism	GMP biosynthesis
Purine metabolism	GMP biosynthesis via salvage pathway
Purine metabolism	guanine degradation
Purine metabolism	IMP biosynthesis via de novo pathway
Purine metabolism	IMP biosynthesis via salvage pathway
Purine metabolism	ppGpp biosynthesis
Purine metabolism	purine nucleoside salvage.
Purine metabolism	purine nucleotide biosynthesis [regulation].
Purine metabolism	urate degradation
Purine metabolism	xanthosine degradation
Purine metabolism	XMP biosynthesis via de novo pathway
Purine metabolism	XMP biosynthesis via salvage pathway
Pyrimidine metabolism	CTP biosynthesis via de novo pathway
Pyrimidine metabolism	CTP biosynthesis via salvage pathway

Pathway class	Pathway instance
Pyrimidine metabolism	dTMP biosynthesis via salvage pathway
Pyrimidine metabolism	dTTP biosynthesis
Pyrimidine metabolism	dUMP biosynthesis
Pyrimidine metabolism	UMP biosynthesis via salvage pathway
Quinol/quinone metabolism	1,4-dihydroxy-2-naphthoate biosynthesis
Quinol/quinone metabolism	menaquinone biosynthesis
Secondary metabolite metabolism	lignin degradation
Secondary metabolite metabolism	methylglyoxal degradation
Siderophore biosynthesis	bacillibactin biosynthesis
Siderophore biosynthesis	enterobactin biosynthesis
Siderophore biosynthesis	ferrichrome biosynthesis
Siderophore biosynthesis	mycobactin biosynthesis
Siderophore biosynthesis	petrobactin biosynthesis
Siderophore biosynthesis	pyoverdin biosynthesis
Siderophore biosynthesis	salicylate biosynthesis
Siderophore biosynthesis	vibriobactin biosynthesis
Signal transduction	phosphatidylinositol signaling pathway
Spore coat biogenesis	spore coat polysaccharide biosynthesis
Steroid metabolism	cholesterol metabolism
Sulfur metabolism	dibenzothiophene degradation
Sulfur metabolism	glutathione biosynthesis
Sulfur metabolism	glutathione metabolism
Sulfur metabolism	hydrogen sulfide biosynthesis
Sulfur metabolism	sulfate assimilation.
Sulfur metabolism	sulfite reduction.
Terpene metabolism	(4R)-limonene degradation
Terpene metabolism	(R)-camphor degradation
Terpene metabolism	1,8-cineol degradation
tRNA modification	5-methoxycarbonylmethyl-2-thiouridine-tRNA biosynthesis
tRNA modification	archaeosine-tRNA biosynthesis
tRNA modification	N(7)-methylguanine-tRNA biosynthesis
tRNA modification	tRNA-queuosine biosynthesis
tRNA modification	wybutosine-tRNA(Phe) biosynthesis
Xenobiotic degradation	1,2-dichloroethane degradation
Xenobiotic degradation	4-chlorobenzoate degradation
Xenobiotic degradation	4-chloronitrobenzene degradation
Xenobiotic degradation	4-nitrophenol degradation
Xenobiotic degradation	acetylacetone degradation
Xenobiotic degradation	atrazine degradation
Xenobiotic degradation	biphenyl degradation
Xenobiotic degradation	gamma-hexachlorocyclohexane degradation
Xenobiotic degradation	haloalkane degradation

Pathway class	Pathway instance
Xenobiotic degradation	nitrobenzene degradation
Xenobiotic degradation	nylon-6 oligomer degradation
Xenobiotic degradation	toluene degradation
Xenobiotic degradation	xylene degradation

Table2. Miscellaneous pathways involving enzymes with known structures.

Table 3. MetaCyc secondary metabolites biosynthesis pathways involving enzymes with known structures.

Pathway Class	Pathway instances
Alcohol-Biosynthesis	butanol and isobutanol biosynthesis (engineered)
Alcohol-Biosynthesis	pyruvate fermentation to isobutanol (engineered)
Alkaloids Biosynthesis	acetylaszonalenin biosynthesis
Alkaloids Biosynthesis	ajmaline and sarpagine biosynthesis
Alkaloids Biosynthesis	calystegine biosynthesis
Alkaloids Biosynthesis	chanoclavine I aldehyde biosynthesis
Alkaloids Biosynthesis	dehydroscoulerine biosynthesis
Alkaloids Biosynthesis	fumigaclavine biosynthesis
Alkaloids Biosynthesis	hyoscyamine and scopolamine biosynthesis
Alkaloids Biosynthesis	morphine biosynthesis
Alkaloids Biosynthesis	sanguinarine and macarpine biosynthesis
Alkaloids Biosynthesis	superpathway of hyoscyamine and scopolamine biosynthesis
Antibiotic Biosynthesis	(5*R*)-carbapenem carboxylate biosynthesis
Antibiotic Biosynthesis	aclacinomycin biosynthesis
Antibiotic Biosynthesis	actinorhodin biosynthesis
Antibiotic Biosynthesis	albaflavenone biosynthesis
Antibiotic Biosynthesis	aurachin RE biosynthesis
Antibiotic Biosynthesis	bacilysin biosynthesis
Antibiotic Biosynthesis	bacimethrin and bacimethrin pyrophosphate biosynthesis
Antibiotic Biosynthesis	cephalosporin C biosynthesis
Antibiotic Biosynthesis	clavulanate biosynthesis
Antibiotic Biosynthesis	daunorubicin biosynthesis
Antibiotic Biosynthesis	D-cycloserine biosynthesis
Antibiotic Biosynthesis	deacetylcephalosporin C biosynthesis
Antibiotic Biosynthesis	dehydrophos biosynthesis
Antibiotic Biosynthesis	erythromycin A biosynthesis
Antibiotic Biosynthesis	erythromycin D biosynthesis
Antibiotic Biosynthesis	fosfomycin biosynthesis
Antibiotic Biosynthesis	FR-900098 and FR-33289 antibiotics biosynthesis
Antibiotic Biosynthesis	gramicidin S biosynthesis
Antibiotic Biosynthesis	kanosamine biosynthesis I

Pathway Class	Pathway instances
Antibiotic Biosynthesis	kanosamine biosynthesis II
Antibiotic Biosynthesis	methymycin, neomethymycin and novamethymycin biosynthesis
Antibiotic Biosynthesis	mithramycin biosynthesis
Antibiotic Biosynthesis	mycinamicin biosynthesis
Antibiotic Biosynthesis	narbomycin, pikromycin and novapikromycin biosynthesis
Antibiotic Biosynthesis	neopentalenoketolactone and pentalenate biosynthesis
Antibiotic Biosynthesis	novobiocin biosynthesis
Antibiotic Biosynthesis	paromamine biosynthesis II
Antibiotic Biosynthesis	penicillin K biosynthesis
Antibiotic Biosynthesis	pentalenolactone biosynthesis
Antibiotic Biosynthesis	phenazine-1-carboxylate biosynthesis
Antibiotic Biosynthesis	phosphinothricin tripeptide biosynthesis
Antibiotic Biosynthesis	pyocyanin biosynthesis
Antibiotic Biosynthesis	rebeccamycin biosynthesis
Antibiotic Biosynthesis	rifamycin B biosynthesis
Antibiotic Biosynthesis	staurosporine biosynthesis
Antibiotic Biosynthesis	streptomycin biosynthesis
Antibiotic Biosynthesis	superpathway of butirocin biosynthesis
Antibiotic Biosynthesis	superpathway of erythromycin biosynthesis
Antibiotic Biosynthesis	superpathway of erythromycin biosynthesis (without sugar biosynthesis)
Antibiotic Biosynthesis	superpathway of penicillin, cephalosporin and cephamycin biosynthesis
Antibiotic Biosynthesis	superpathway of rifamycin B biosynthesis
Antibiotic Biosynthesis	tetracenomycin C biosynthesis
Antibiotic Biosynthesis	tylosin biosynthesis
Antibiotic Biosynthesis	validamycin A biosynthesis
Autoinducer Biosynthesis	autoinducer AI-1 biosynthesis
Autoinducer Biosynthesis	autoinducer AI-2 biosynthesis I
Autoinducer Biosynthesis	autoinducer AI-2 biosynthesis II (*Vibrio*)
Autoinducer Biosynthesis	autoinducer CAI-1 biosynthesis
Fatty acid derivarives	jasmonic acid biosynthesis
Fatty acid derivarives	superpathway of lipoxygenase
Fatty acid derivarives	traumatin and (*Z*)-3-hexen-1-yl acetate biosynthesis
Insecticides Biosynthesis	spinosyn A biosynthesis
Nitrogen-Containing Secondary Compounds Biosynthesis	fumitremorgin C biosynthesis
Phenylpropanoid Derivatives Biosynthesis	coniferyl alcohol 9-methyl ester biosynthesis
Phenylpropanoid Derivatives Biosynthesis	eugenol and isoeugenol biosynthesis
Phenylpropanoid Derivatives Biosynthesis	flavonoid biosynthesis
Phenylpropanoid Derivatives Biosynthesis	hypericin biosynthesis
Phenylpropanoid Derivatives Biosynthesis	medicarpin biosynthesis
Phenylpropanoid Derivatives Biosynthesis	naringenin biosynthesis (engineered)
Phenylpropanoid Derivatives Biosynthesis	phenylpropanoid biosynthesis
Phenylpropanoid Derivatives Biosynthesis	phenylpropanoids methylation (ice plant)

Pathway Class	Pathway instances
Phenylpropanoid Derivatives Biosynthesis	salicylate biosynthesis I
Phenylpropanoid Derivatives Biosynthesis	superpathway of pterocarpan biosynthesis (via formononetin)
Phytoalexins Biosynthesis	capsidiol biosynthesis
Phytoalexins Biosynthesis	medicarpin biosynthesis
Polyketides Biosynthesis	curcuminoid biosynthesis
Polyketides Biosynthesis	flaviolin dimer and mompain biosynthesis
Polyketides Biosynthesis	raspberry ketone biosynthesis
Secondary Metabolite Biosynthesis	2-heptyl-3-hydroxy-4(1*H*)-quinolone biosynthesis
Secondary Metabolite Biosynthesis	2-methylketone biosynthesis
Secondary Metabolite Biosynthesis	3-amino-5-hydroxybenzoate biosynthesis
Secondary Metabolite Biosynthesis	4-hydroxy-2(1*H*)-quinolone biosynthesis
Secondary Metabolite Biosynthesis	6-gingerol analog biosynthesis
Secondary Metabolite Biosynthesis	DIBOA-glucoside biosynthesis
Secondary Metabolite Biosynthesis	ergothioneine biosynthesis I (bacteria)
Secondary Metabolite Biosynthesis	fluoroacetate and fluorothreonine biosynthesis
Secondary Metabolite Biosynthesis	gliotoxin biosynthesis
Secondary Metabolite Biosynthesis	mycocyclosin biosynthesis
Secondary Metabolite Biosynthesis	$preQ_0$ biosynthesis
Secondary Metabolite Biosynthesis	pulcherrimin biosynthesis
Secondary Metabolite Biosynthesis	pyrrolnitrin biosynthesis
Secondary Metabolite Biosynthesis	superpathway of benzoxazinoid glucosides biosynthesis
Secondary Metabolite Biosynthesis	superpathway of quinolone and alkylquinolone biosynthesis
Secondary Metabolite Biosynthesis	violacein biosynthesis
Sugar Derivatives Biosynthesis	1D-*myo* -inositol hexakisphosphate biosynthesis II (mammalian)
Sugar Derivatives Biosynthesis	1D-*myo* -inositol hexakisphosphate biosynthesis V (from Ins(1,3,4)P3)
Sugar Derivatives Biosynthesis	1D-*myo*-inositol hexakisphosphate biosynthesis I (from Ins(1,4,5)P3)
Sugar Derivatives Biosynthesis	D-*myo*-inositol (1,3,4)-trisphosphate biosynthesis
Sugar Derivatives Biosynthesis	D-*myo*-inositol (1,4,5)-trisphosphate biosynthesis
Sugar Derivatives Biosynthesis	D-*myo*-inositol (1,4,5)-trisphosphate degradation
Sugar Derivatives Biosynthesis	D-*myo*-inositol (1,4,5,6)-tetrakisphosphate biosynthesis
Sugar Derivatives Biosynthesis	D-*myo*-inositol (3,4,5,6)-tetrakisphosphate biosynthesis
Sugar Derivatives Biosynthesis	D-*myo*-inositol-5-phosphate metabolism
Sugar Derivatives Biosynthesis	*myo*-inositol biosynthesis
Sugar Derivatives Biosynthesis	superpathway of 1D-*myo* -inositol hexakisphosphate biosynthesis (plants)
Sugar Derivatives Biosynthesis	superpathway of D-*myo*-inositol (1,4,5)-trisphosphate metabolism
Sugar Derivatives Biosynthesis	superpathway of inositol phosphate compounds
Sulfur-Containing Secondary Compounds Biosynthesis	3-methylthiopropanoate biosynthesis
Sulfur-Containing Secondary Compounds Biosynthesis	superpathway of Allium flavor precursors

Pathway Class	Pathway instances
Sulfur-Containing Secondary Compounds Biosynthesis	taurine biosynthesis
Terpenoids Biosyntheis	*trans*-lycopene biosynthesis I (bacteria)
Terpenoids Biosynthesis	(+)-camphor biosynthesis
Terpenoids Biosynthesis	(4*R*)-carvone biosynthesis
Terpenoids Biosynthesis	2-methylisoborneol biosynthesis
Terpenoids Biosynthesis	abietic acid biosynthesis
Terpenoids Biosynthesis	bornyl diphosphate biosynthesis
Terpenoids Biosynthesis	capsidiol biosynthesis
Terpenoids Biosynthesis	dehydroabietic acid biosynthesis
Terpenoids Biosynthesis	isopimaric acid biosynthesis
Terpenoids Biosynthesis	isoprene biosynthesis II (engineered)
Terpenoids Biosynthesis	levopimaric acid biosynthesis
Terpenoids Biosynthesis	menthol biosynthesis
Terpenoids Biosynthesis	methylerythritol phosphate pathway I
Terpenoids Biosynthesis	methylerythritol phosphate pathway II
Terpenoids Biosynthesis	mevalonate pathway I
Terpenoids Biosynthesis	mevalonate pathway II (archaea)
Terpenoids Biosynthesis	mevalonate pathway III (archaea)
Terpenoids Biosynthesis	neoabietic acid biosynthesis
Terpenoids Biosynthesis	palustric acid biosynthesis
Terpenoids Biosynthesis	perillyl aldehyde biosynthesis
Terpenoids Biosynthesis	superpathway of carotenoid biosynthesis
Terpenoids Biosynthesis	superpathway of diterpene resin acids biosynthesis
Terpenoids Biosynthesis	zeaxanthin, antheraxanthin and violaxanthin interconversion
Terpenoids Biosynthesis	B-carotene biosynthesis (engineered)
Terpenoids Biosynthesis & Alkaloid Biosynthesis	ajmaline and sarpagine biosynthesis
Terpenoids Biosynthesis & Horomone Biosynthesis	*ent*-kaurene biosynthesis I
Terpenoids Biosynthesis & Horomone Biosynthesis	superpathway of gibberellin biosynthesis
Terpenoids Biosynthesis & Horomone Biosynthesis	superpathway of gibberellin GA_{12} biosynthesis
Terpenophenolics Biosynthesis	cannabinoid biosynthesis
Terpnoids Biosynthesis	cyclooctatin biosynthesis
Terpnoids Biosynthesis	superpathway of geranylgeranyldiphosphate biosynthesis I (via mevalonate)
Terpnoids Biosynthesis	taxadiene biosynthesis (engineered)
Terpnoids Biosynthesis	taxol biosynthesis
Terpnoids Biosynthesis	epoxysqualene biosynthesis
Terpnoids Biosynthesis	superpathway of geranylgeranyl diphosphate biosynthesis II (via MEP)
Terpnoids Biosynthesis	zeaxanthin, antheraxanthin and violaxanthin interconversion

Table 3. MetaCyc secondary metabolites biosynthesis pathways involving enzymes with known structures.

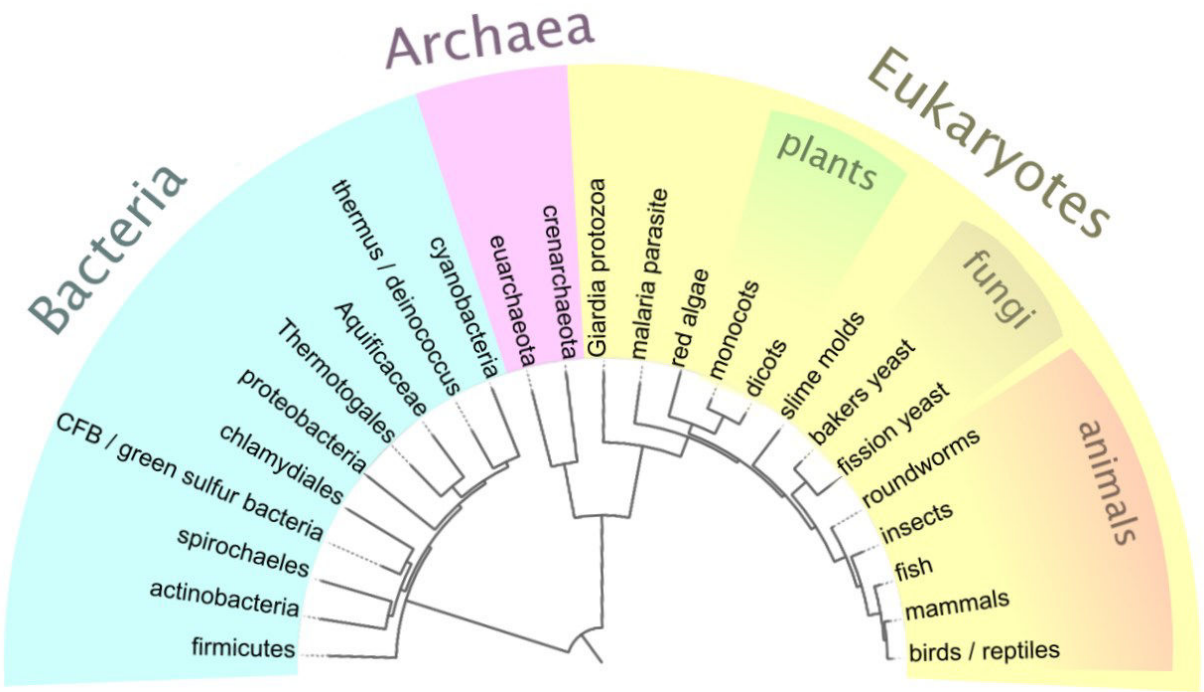


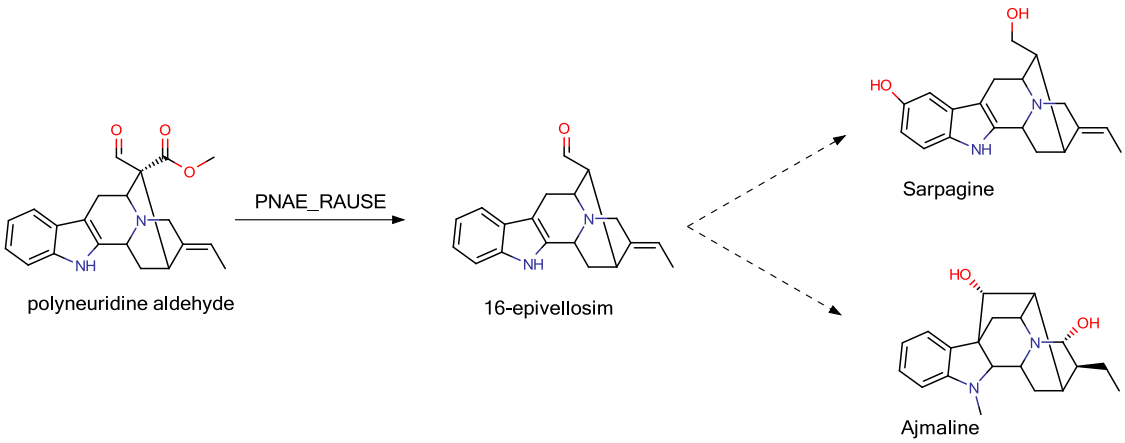
Figure 1. Simplified representation of phylogenic tree. Two species were assumed to share evolutionary relationship if they both felt into the same branch of this tree.
Source : http://upload.wikimedia.org/wikipedia/commons/e/e6/Simplified_tree.png

Enzymatic activities investigation

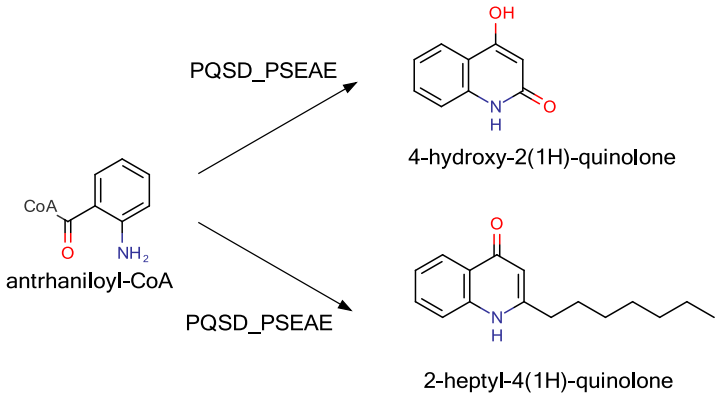
The following schemes illustrate the investigated enzymatic activities. Reactions schemes show natural product under construction only and are inspired from MetaCyc pathways, whenever there is a pathway described. Co-factors and other molecules contribution to the biosytnhetic reactions are ignored. Biosynthetic enzymes within our dataset and annontated with UniProt Identifier. Italic identifiers specify a a biosytnhetic reaction with known biosytnehtic enzyme but there is no structure available yet. In general, these enzymes were considered in the schemes when they link to biosynthetic steps for whose our dataset contains the enzyme structure. Dashed arrow represent many biosynthetic steps. When possible, final products are shown.

NIROGEN-CONTAING COMPOUNDS

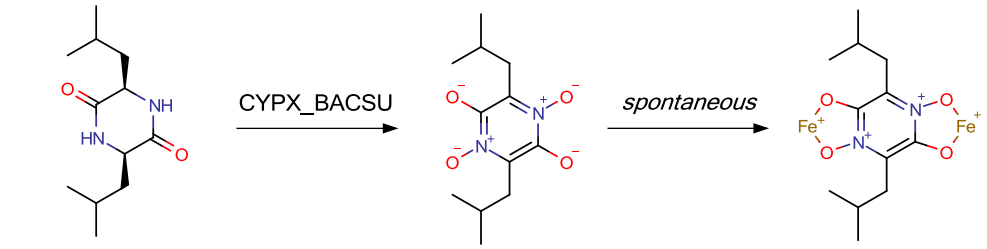
Ajmaline precursors



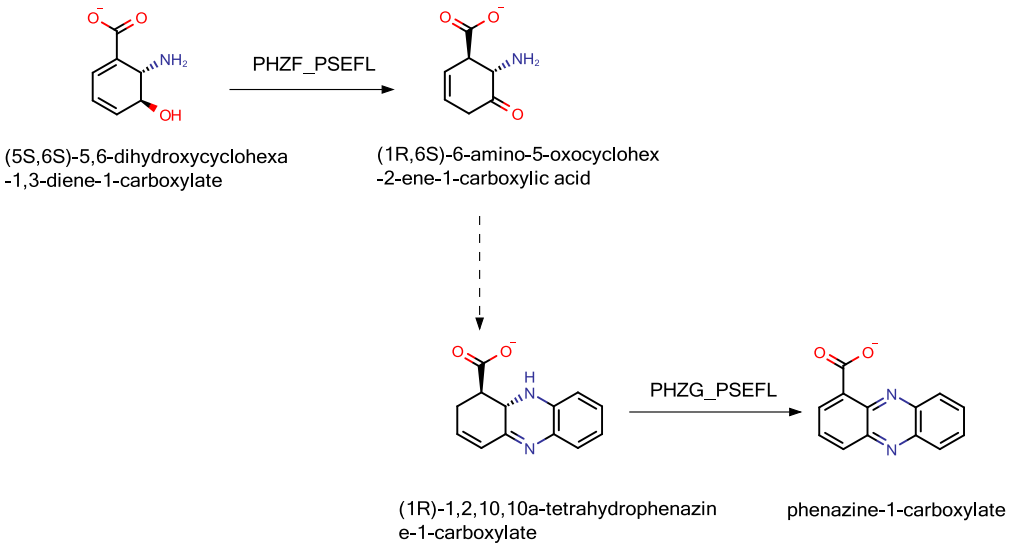
Quinolone precursor



Pulcherrimin precursors



Phenazine precursors



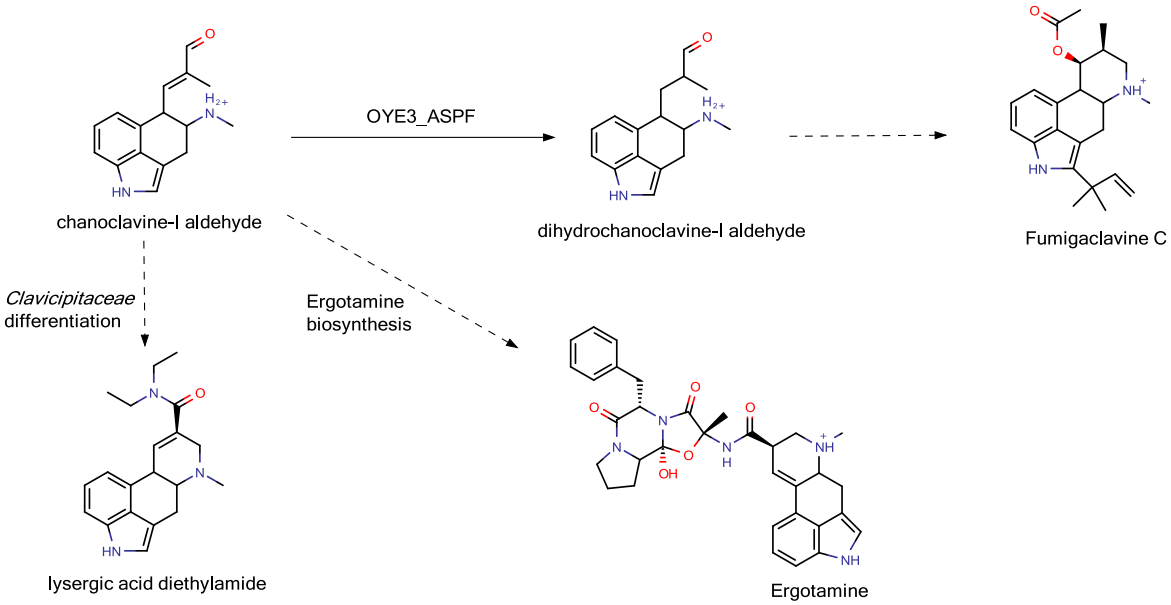
(5S,6S)-5,6-dihydroxycyclohexa
-1,3-diene-1-carboxylate

(1R,6S)-6-amino-5-oxocyclohex
-2-ene-1-carboxylic acid

(1R)-1,2,10,10a-tetrahydrophenazin
e-1-carboxylate

phenazine-1-carboxylate

Fumigaclavine C / Ergotamine precursors



chanoclavine-I aldehyde

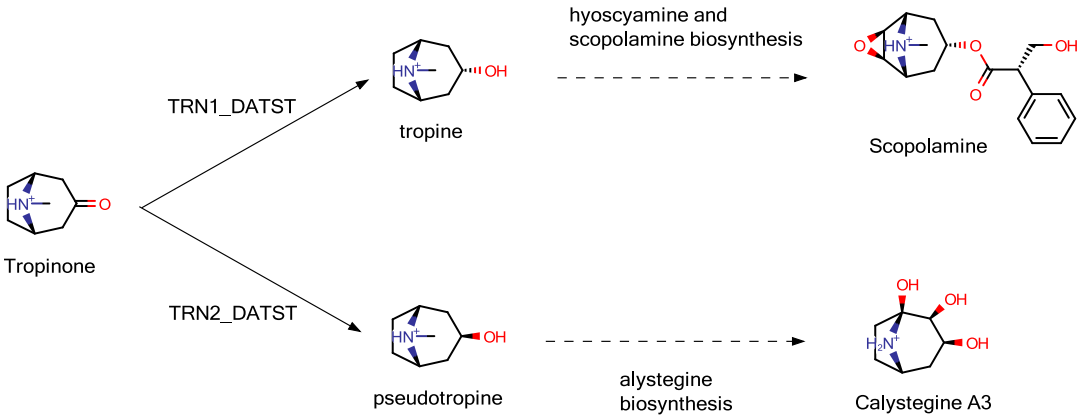
dihydrochanoclavine-I aldehyde

Fumigaclavine C

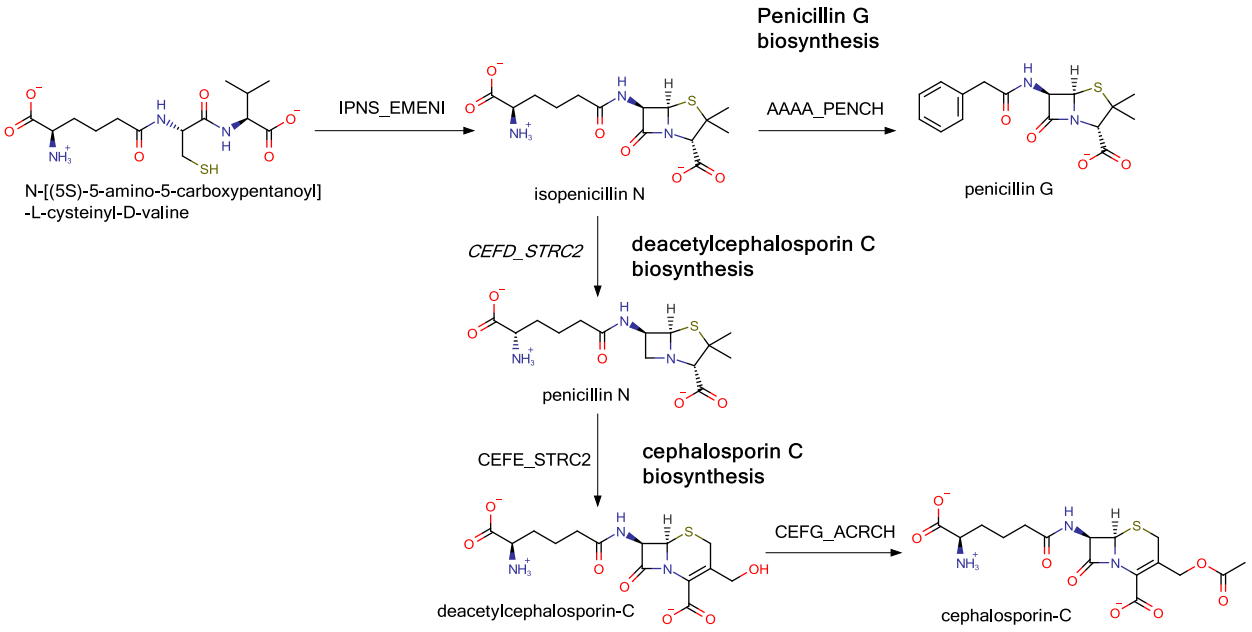
lysergic acid diethylamide

Ergotamine

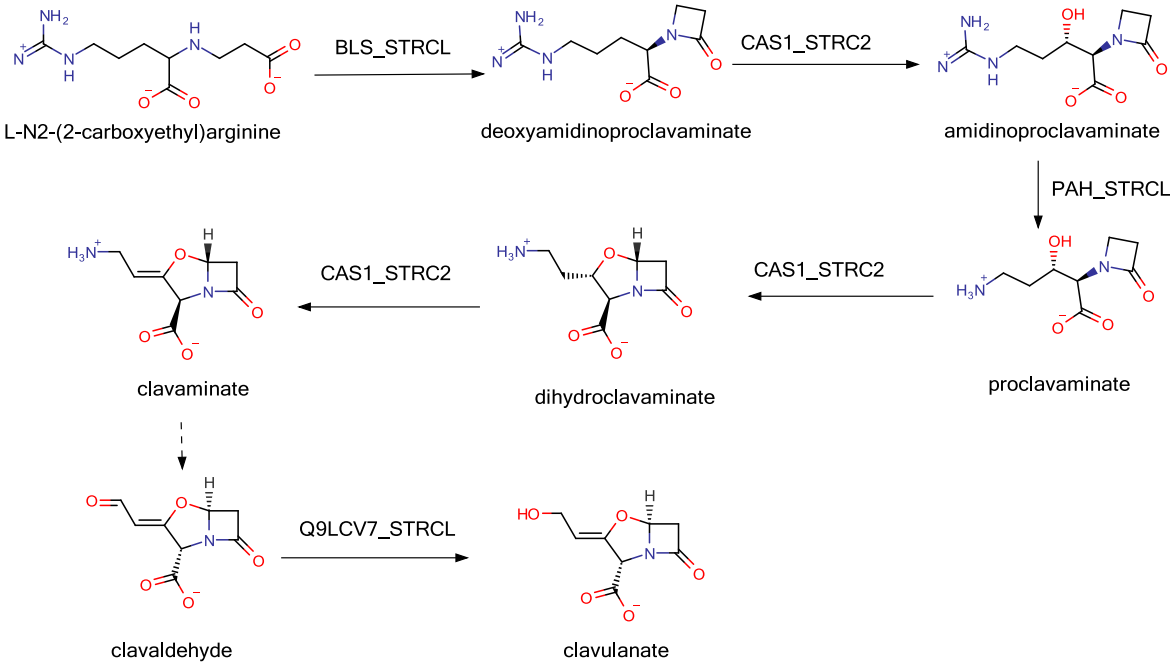
Atropine precurors



Penicillin precursors

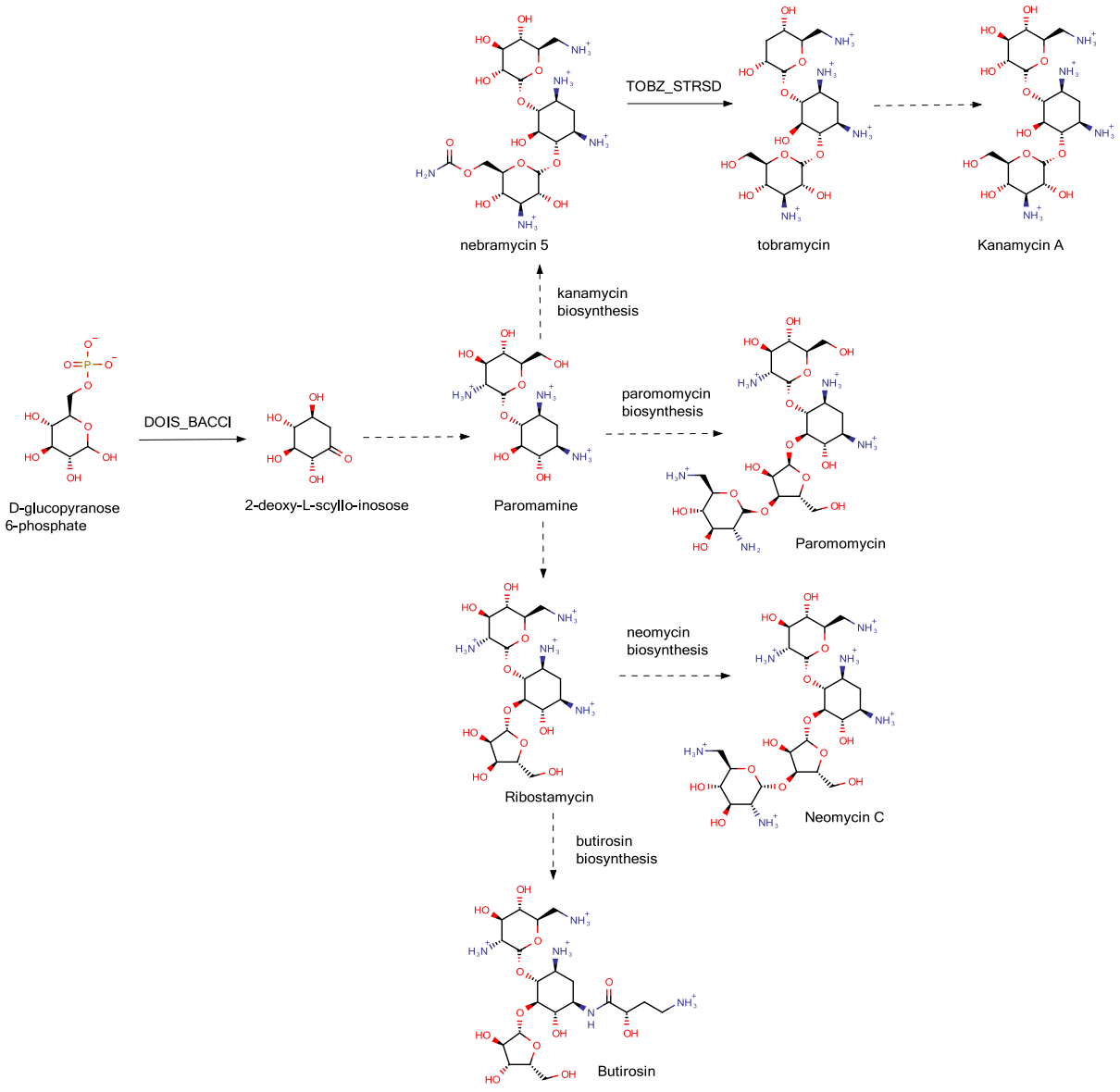


Clavulanic acid precursors

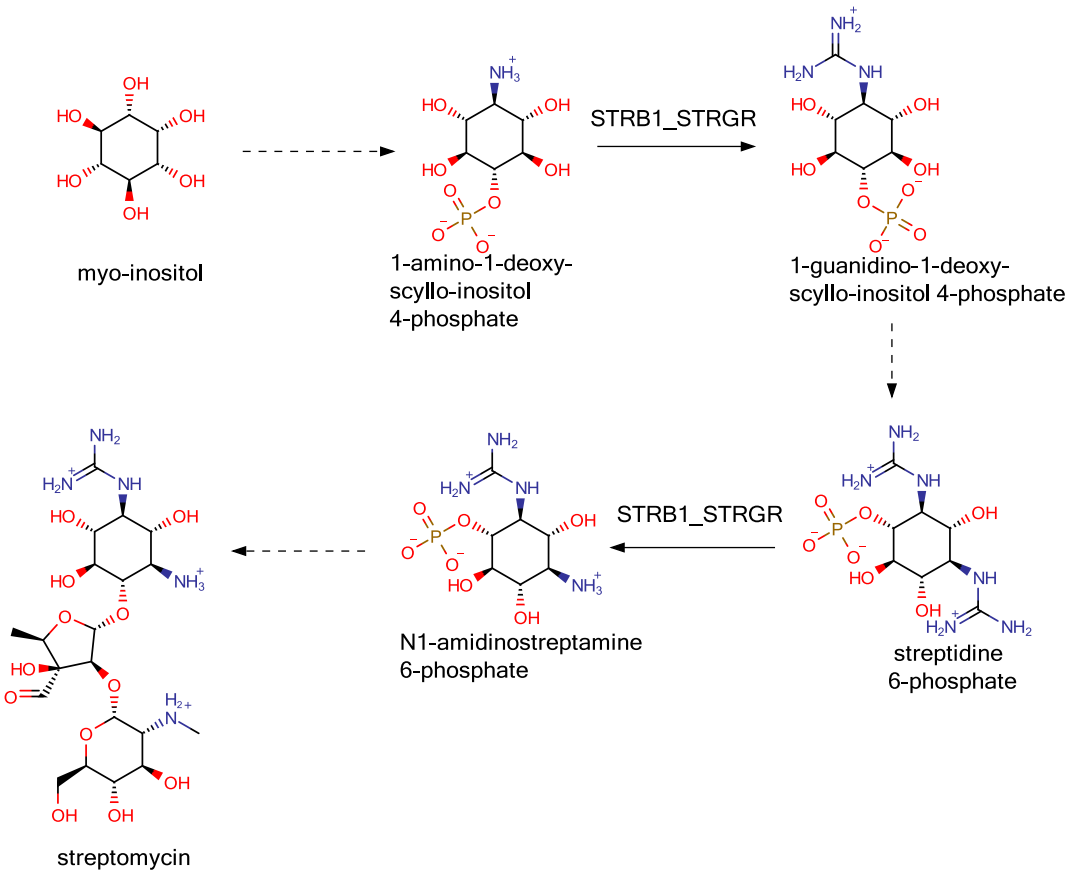


AMINOGLYCOSIDES AND RELATIVES

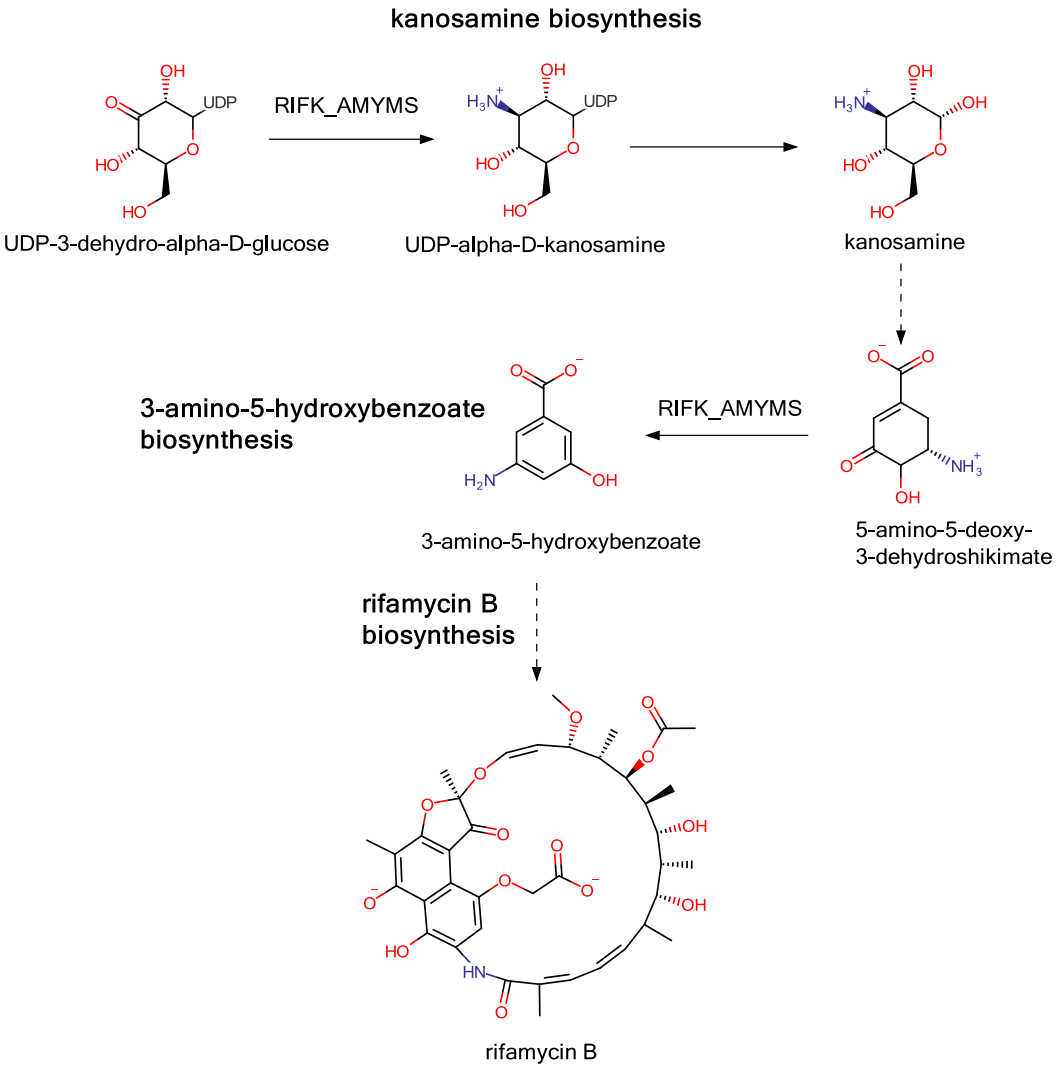
Aminoglycosides precursors



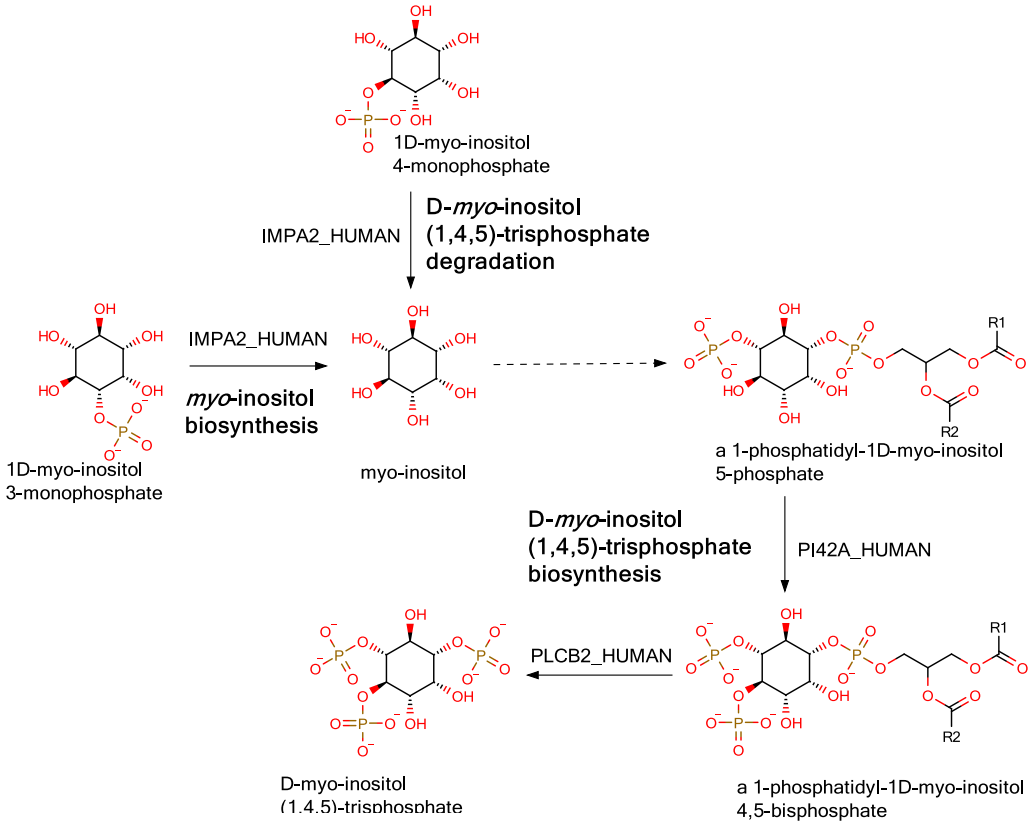
Streptomcycin ealry precursors



Rifamycin B early precursors

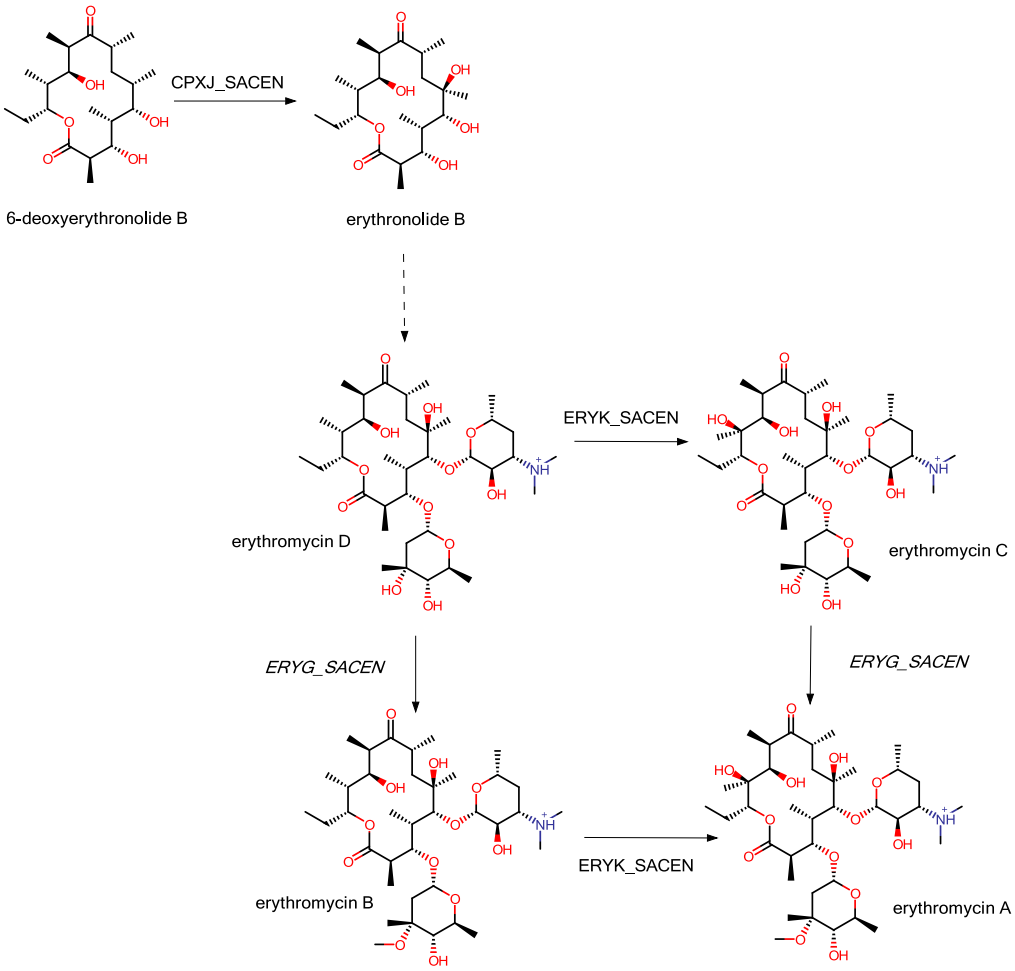


Myo-inositol an derivatives precursors

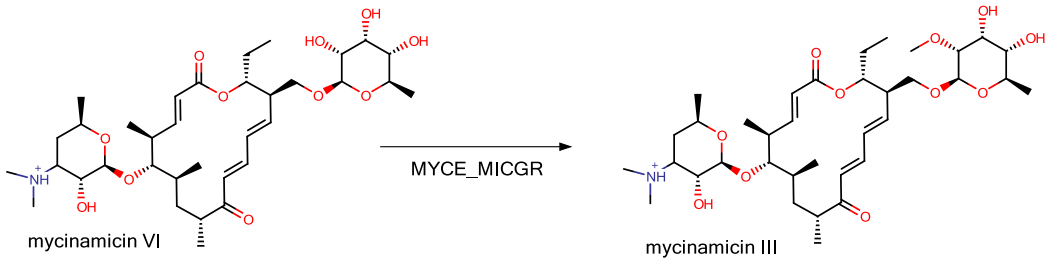


MACROLIDES

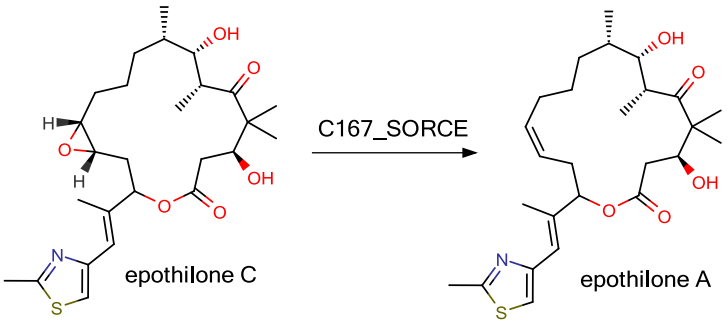
Erythromycin A precursors



Mycinamicin biosynthesis

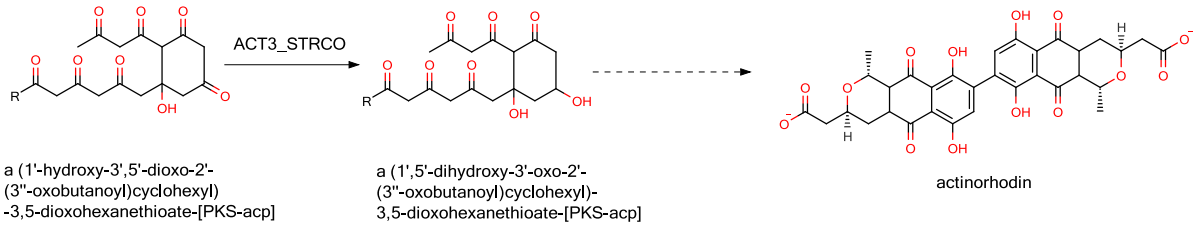


Epothilone precursors



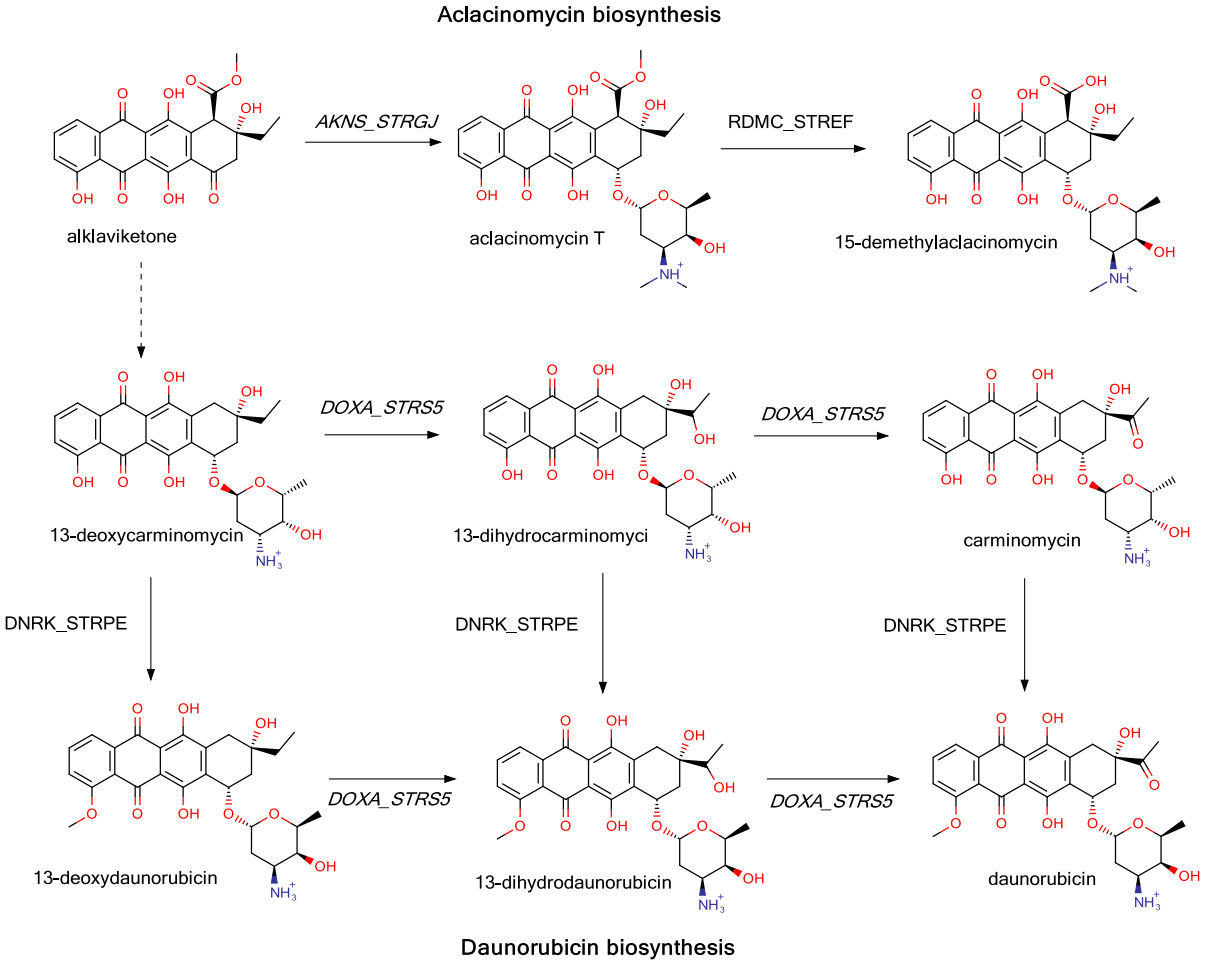
POLYKETIDES

Actinorhodin precursors

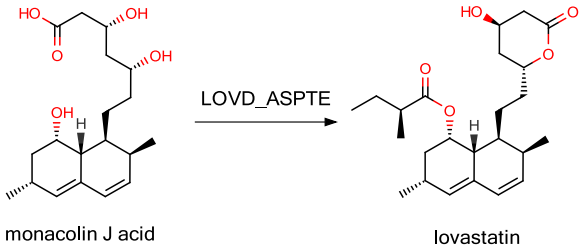


R1 = a polyketide synthase with ACP domain

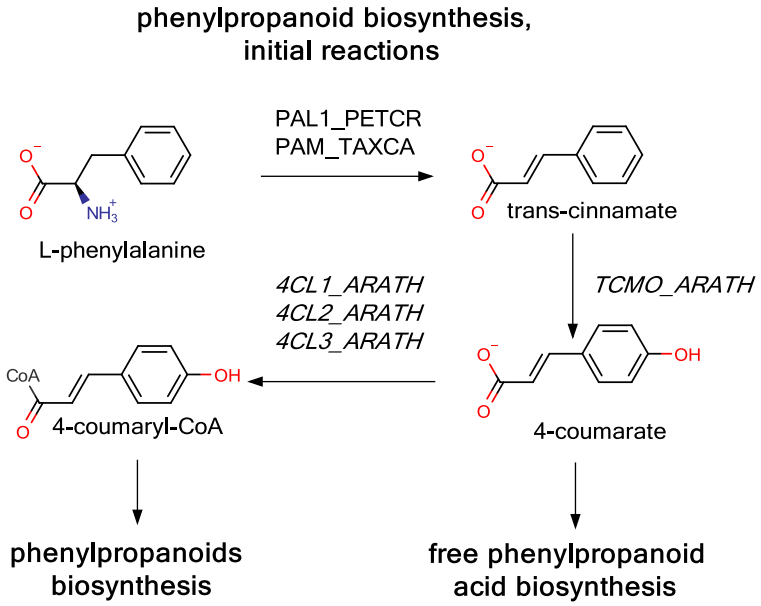
Anthracycline precursors



Lovastatin precursors

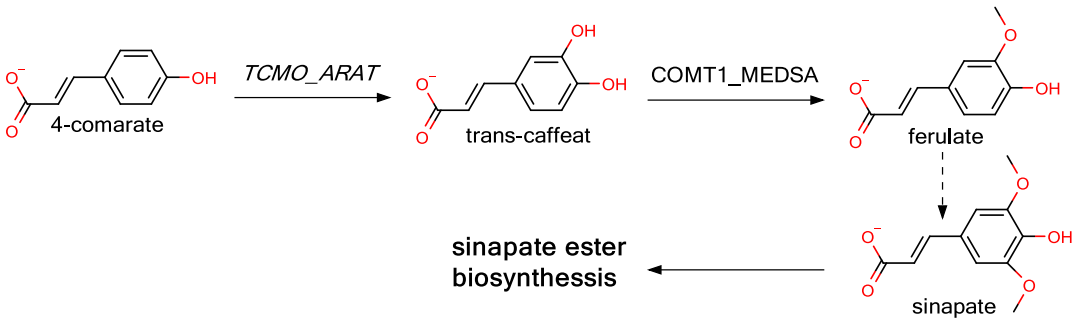


PHENYLPROPANOIDS



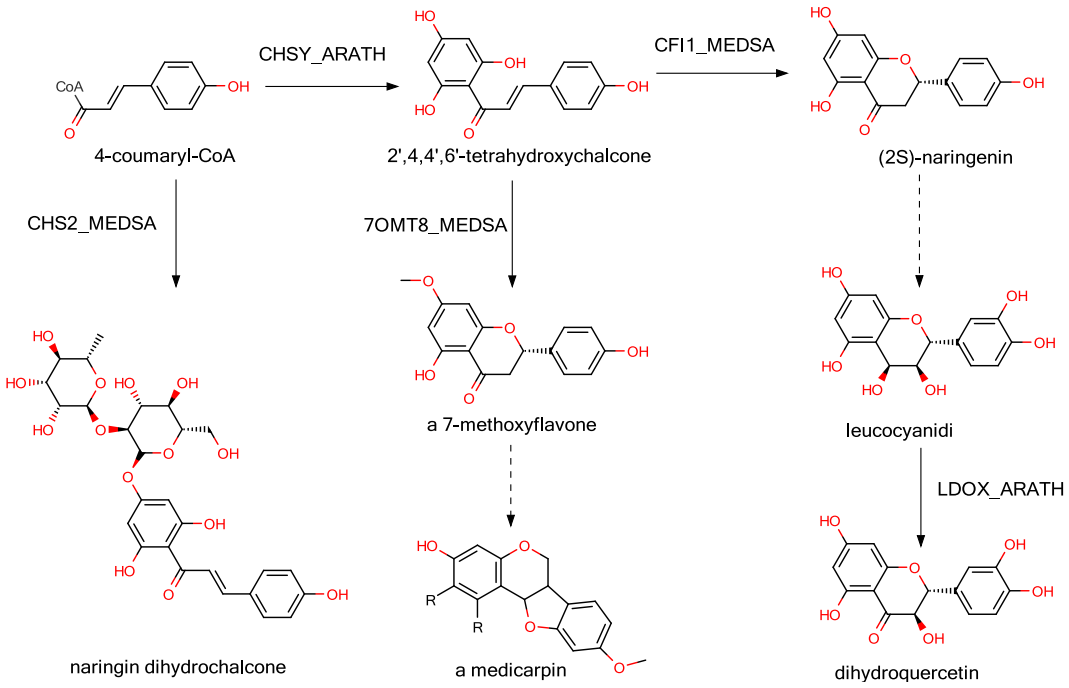
Phenylpropanoids biosynthesis leads to many biosynthetic pathways with ligandable biosynthetic enzyme structures such as curcuminoids, raspberry ketones, flavonoids, hydropinosylvin, and medicarpin.

Free phenylpropanoids precursors

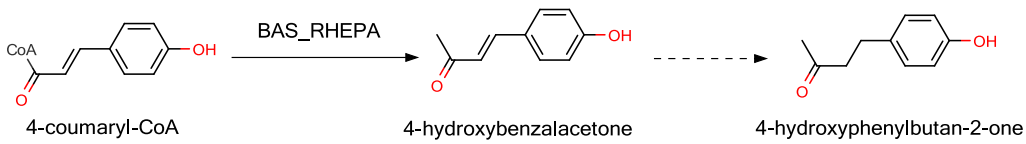


Dihydropinosylvin precursors

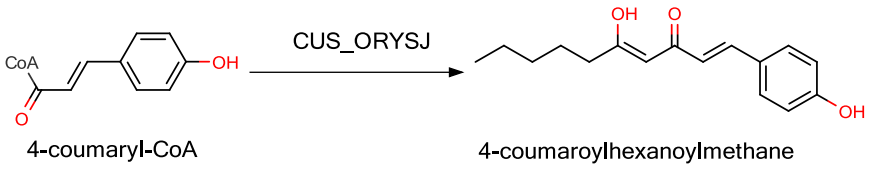
Flavonoids and medicarpin precursors



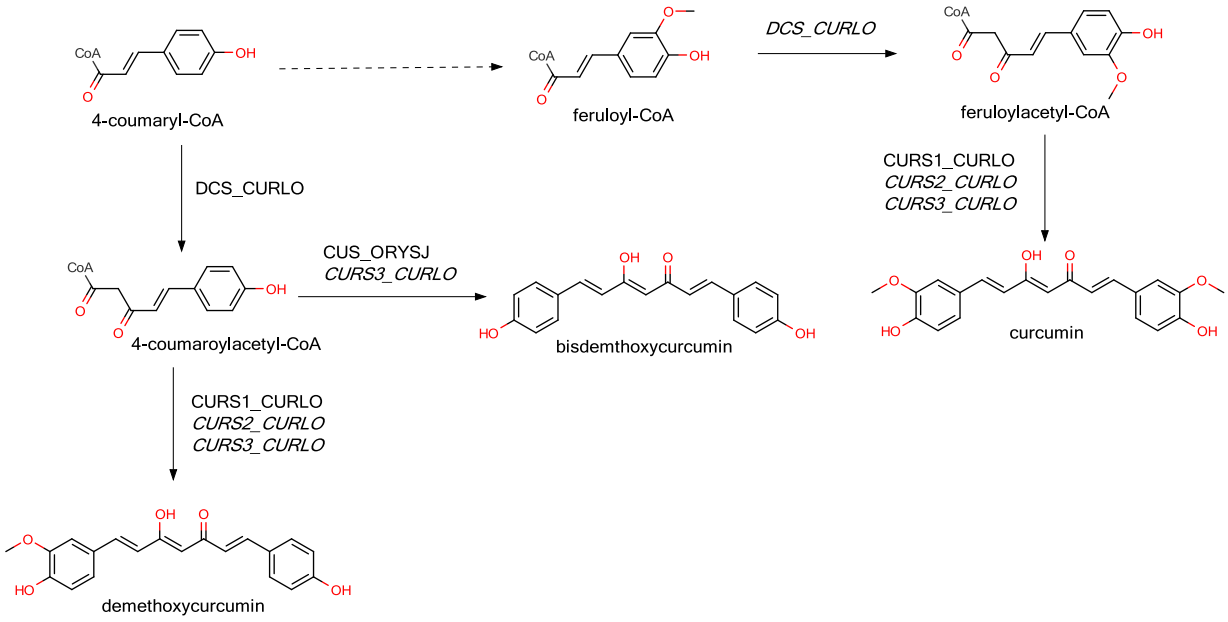
Rasberry ketones precursors



6-gingerol analog

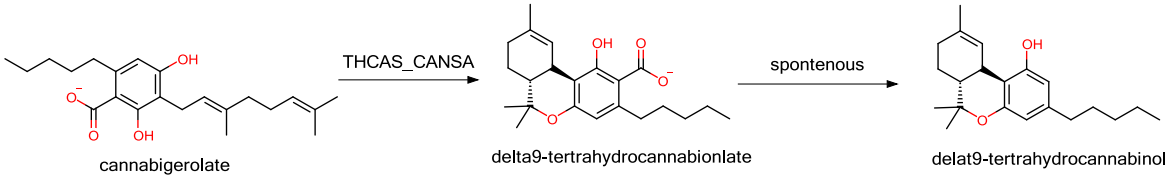


Curcumminoids biosynthesis

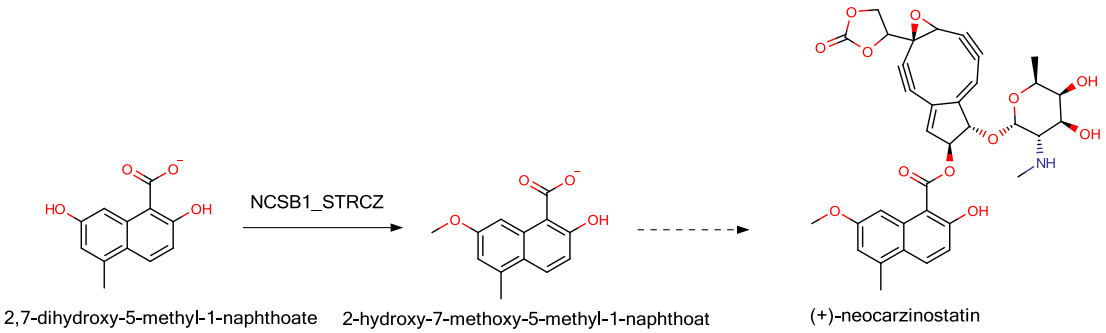


TERPENOPHENOLICS

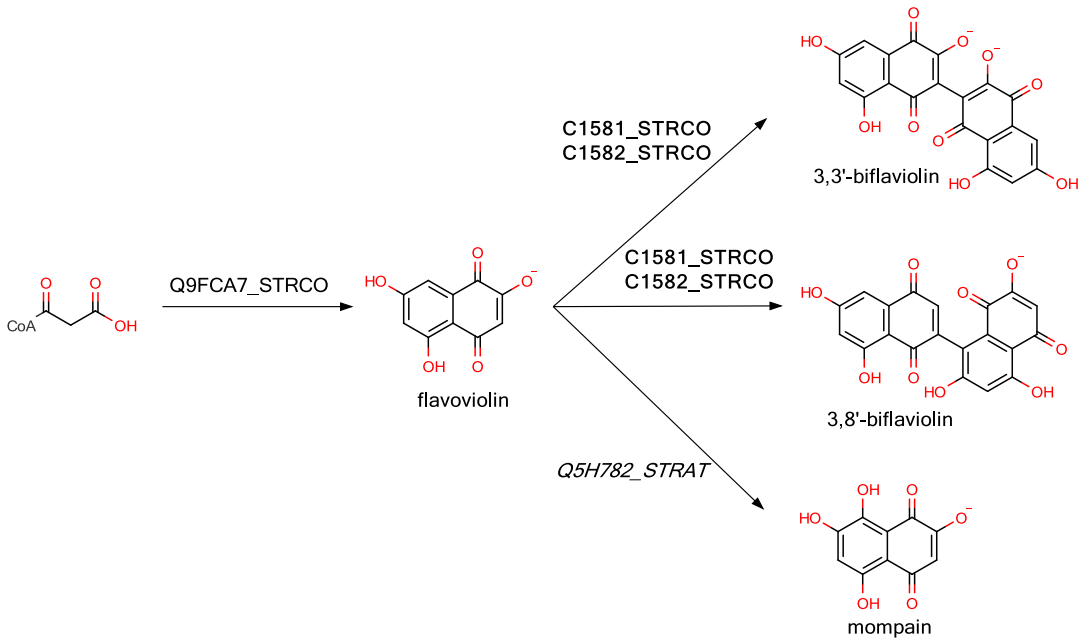
Dronabinol precursors (cannabionoid biosynthesis)



Neocarzinostatin precurors

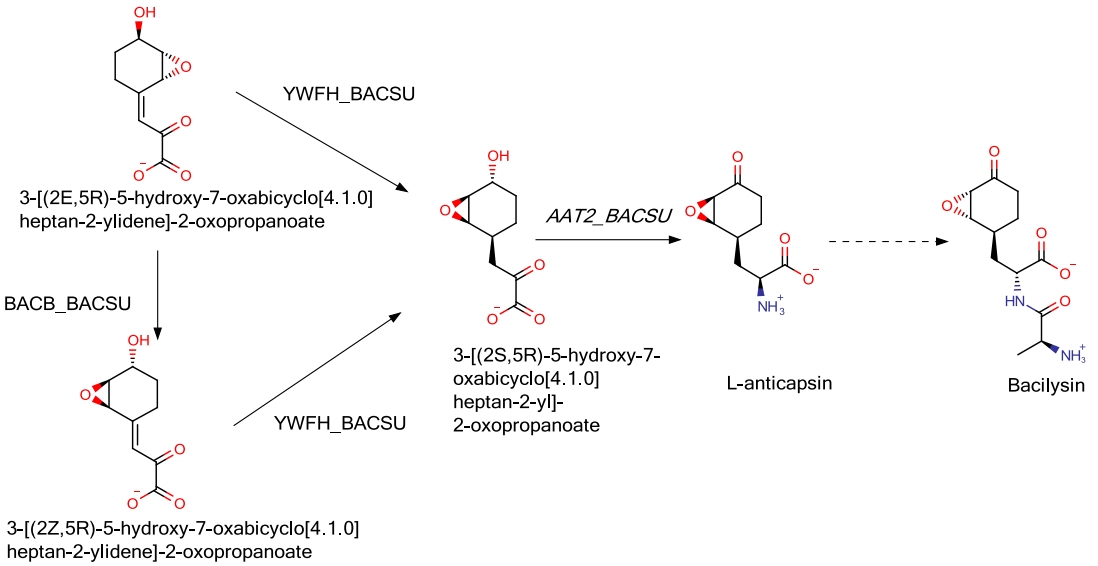


Flaviolin and mompain precursors



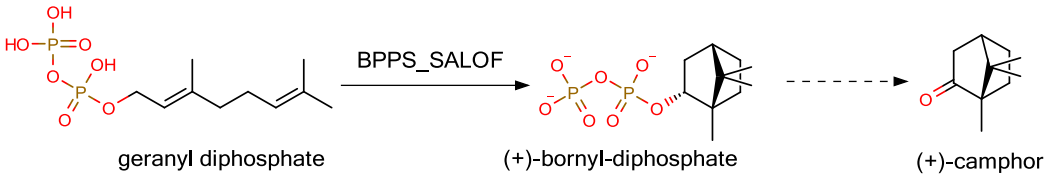
C1582_STRCO and C1582_STRCO have known structures but are not ligandable.

Bacilysin precursors

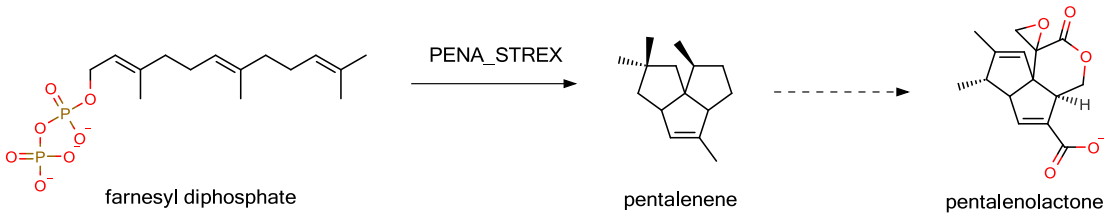


HYDROCARBONS

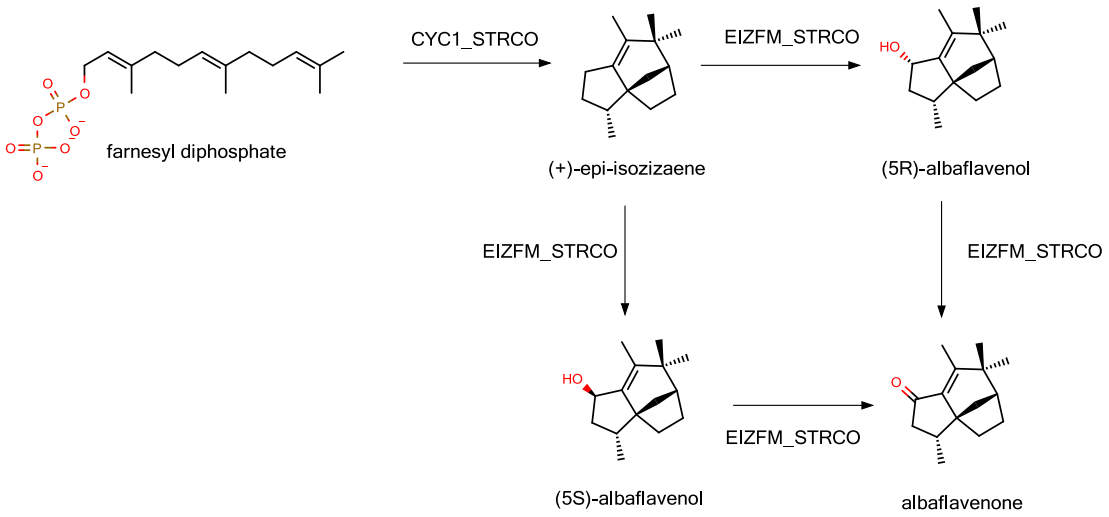
(+)-camphor precursors



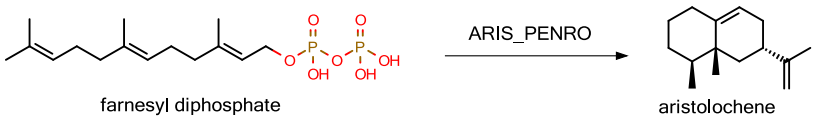
Pentalenone precursor



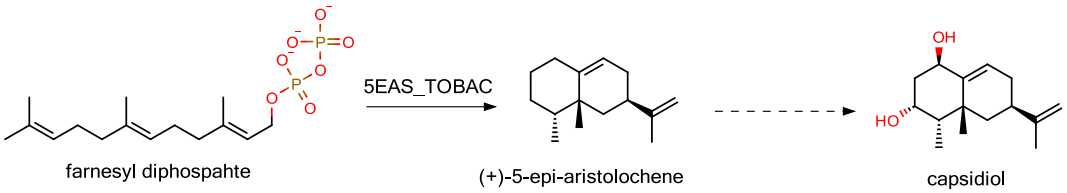
Albaflavenone precursors



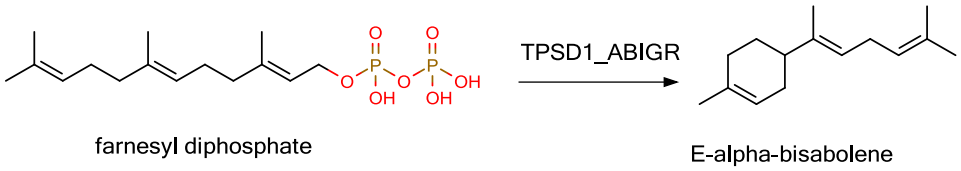
Aristolochene biosynthesis



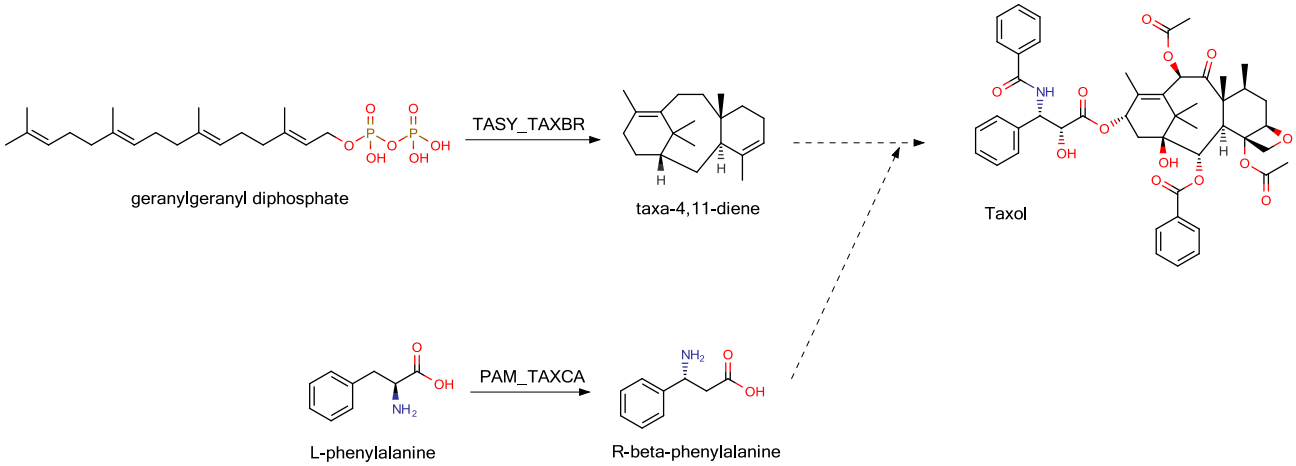
Capsidiol precursor



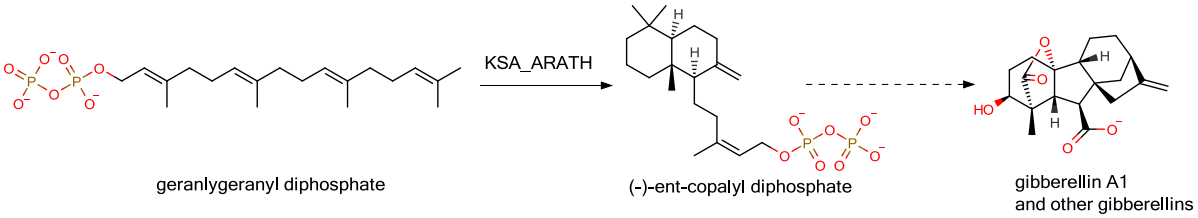
Oleoresins biosynthesis (sesquiterpene)



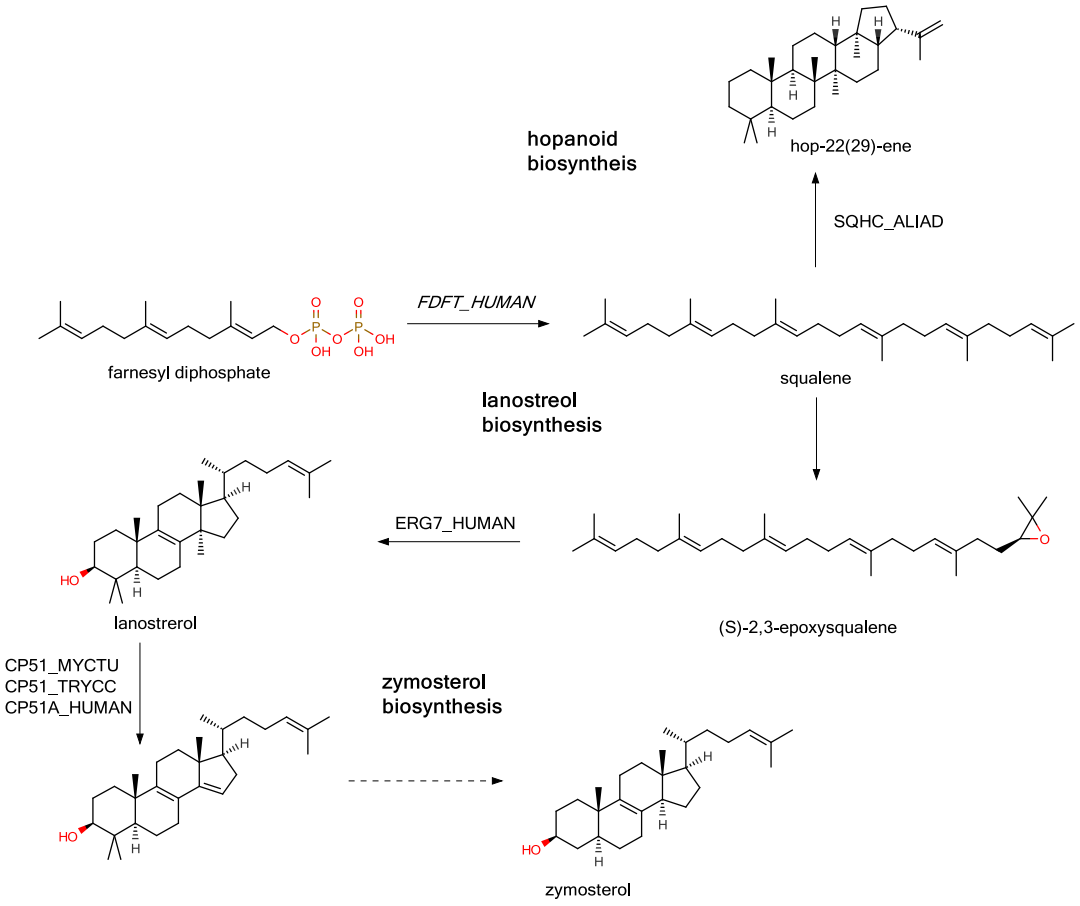
Taxol precursors



Ent-kaurene precursors

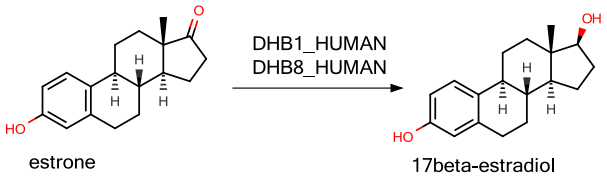


1. STEROIDS

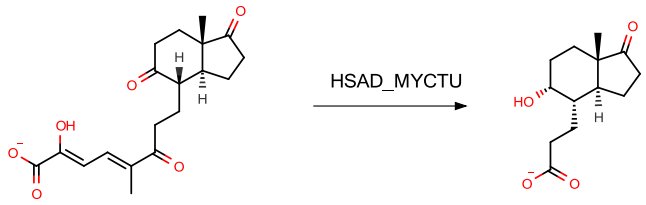


There is a structure of FDFT_HUMAN but was not found ligandable.

Estrogen precursor

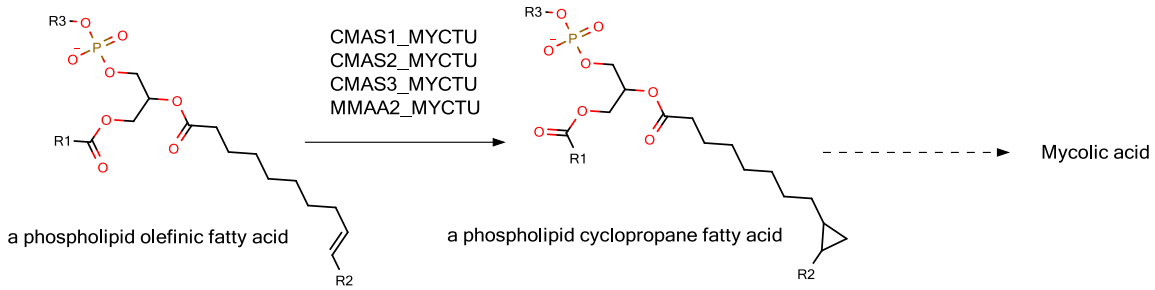


Androstenedione degradation

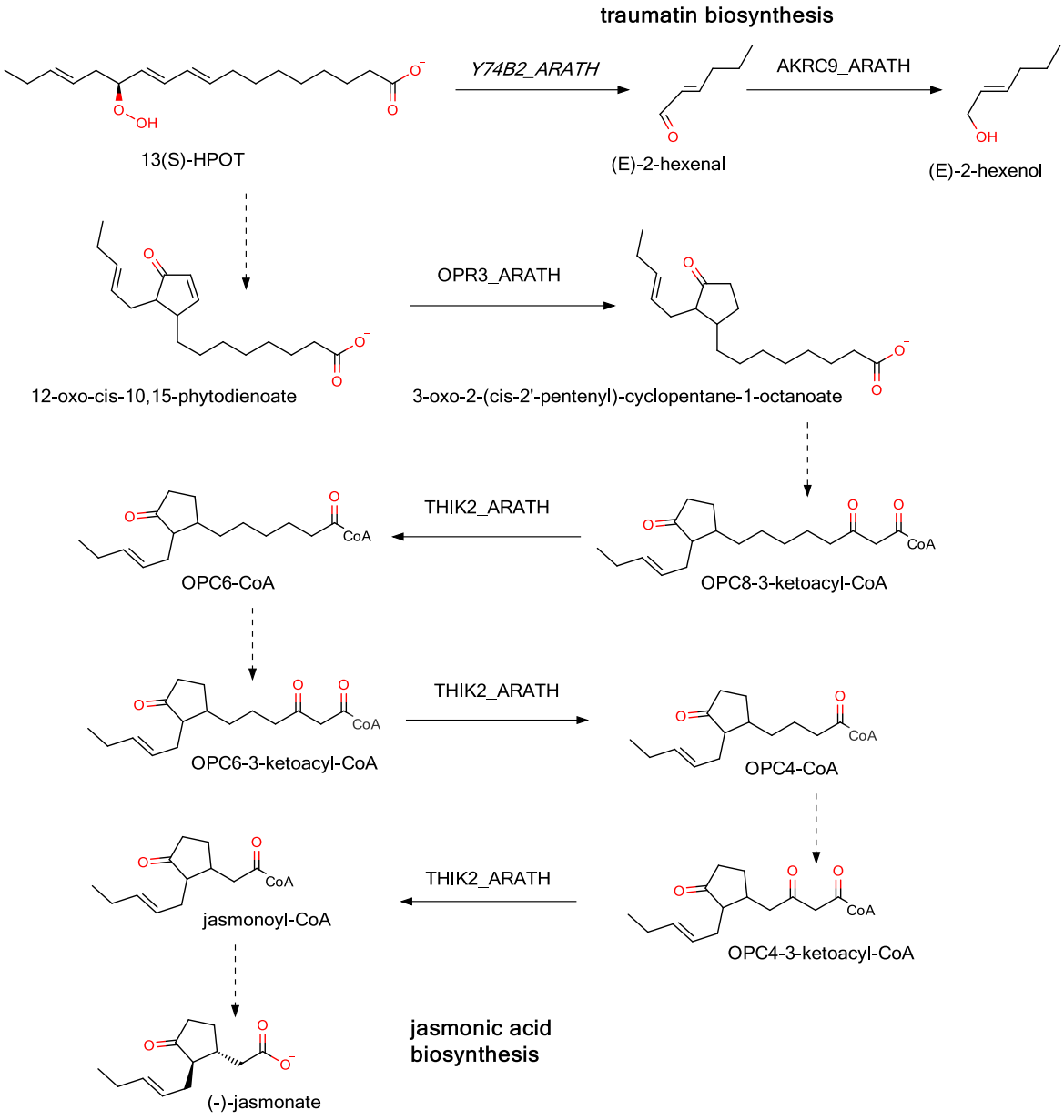


LIPIDS & FATTY ACIDS

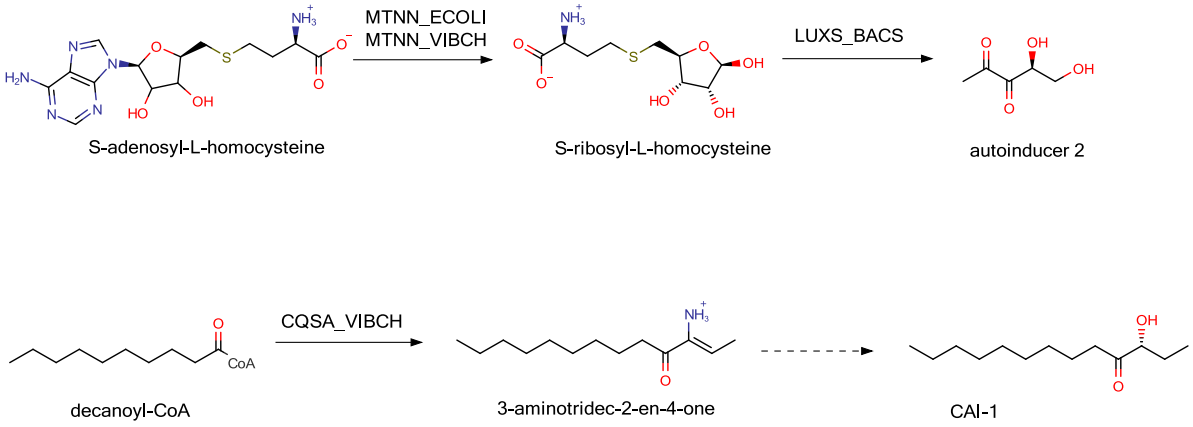
Mycolic acid precursors



Jasmonic acid precursors

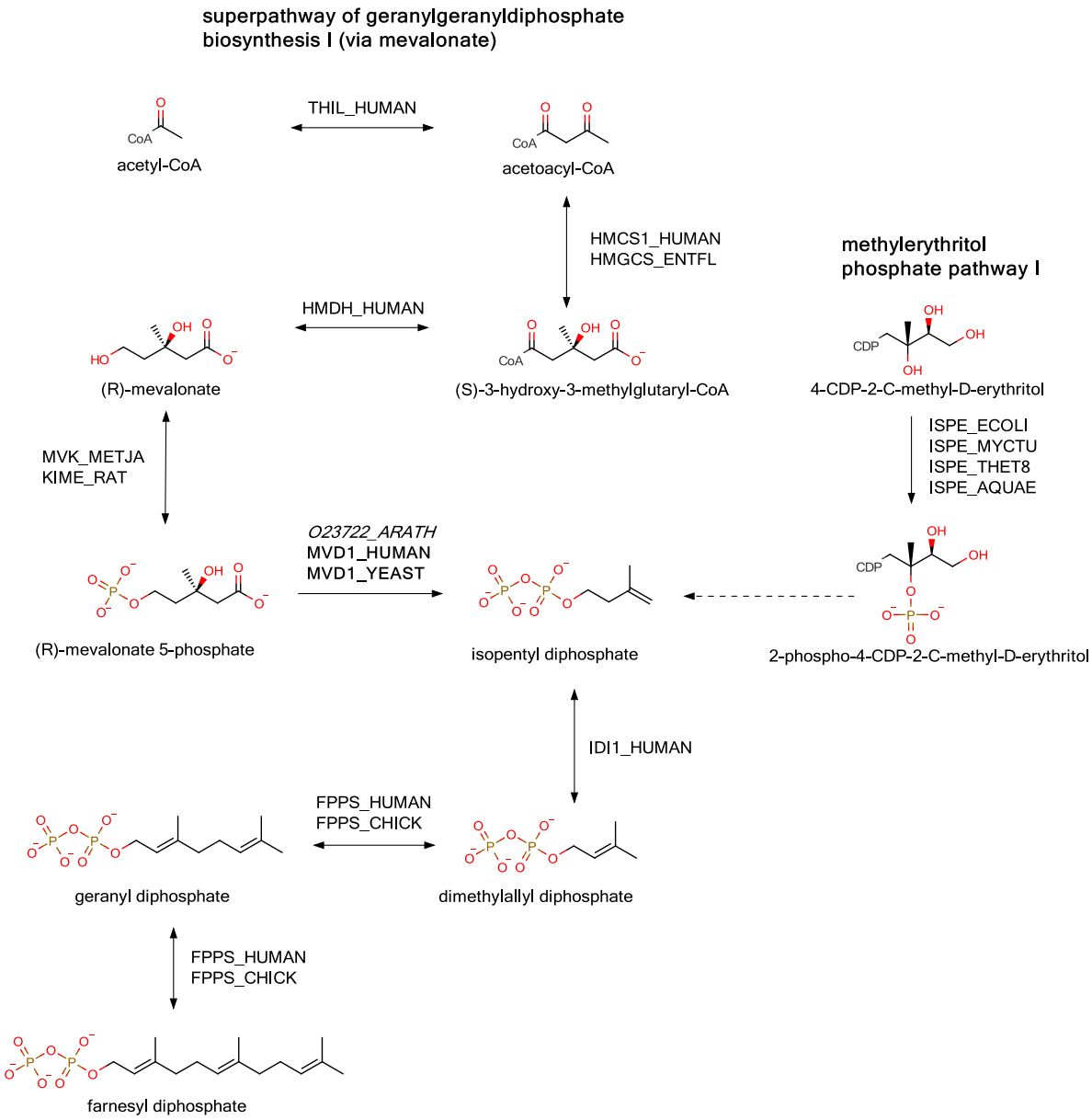


Autoinducers precursors



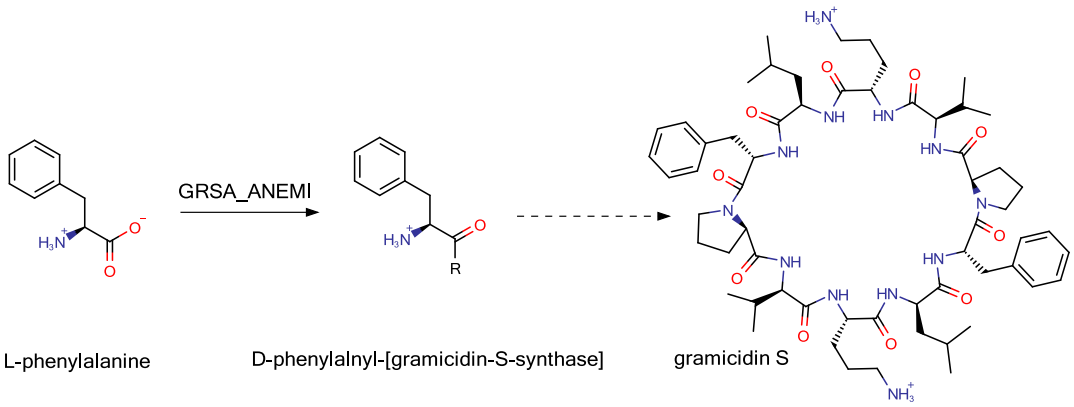
PRECURSORS AND SMALL MOLECULES

Isoprenoids



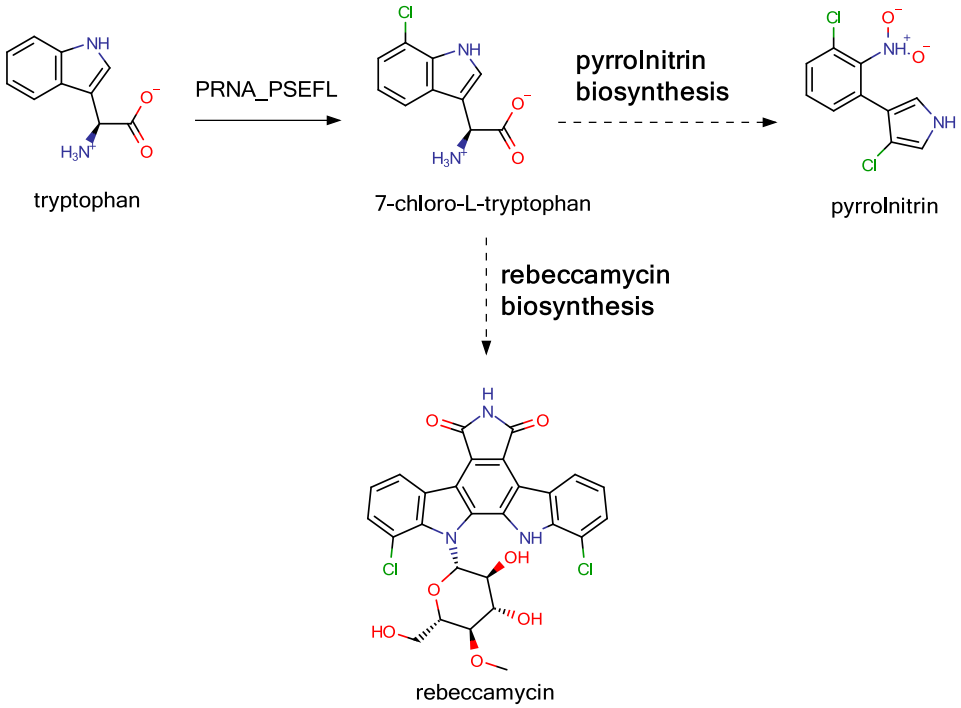
MVD1_YEAST, MVD1_HUMAN have structures but were not found ligandable.

Gramicidin S early precursors

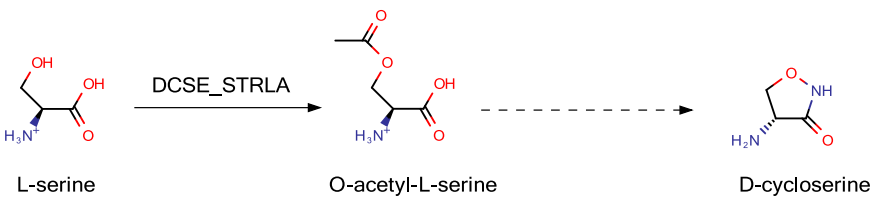


R = non-ribosomal peptide synthetase

Pyrrolnitrin & rebeccamycin precursors



D-cycloserine precursors



L-serine O-acetyl-L-serine D-cycloserine

Chapter 5.

Structural Investigations of Natural Product Biological Imprints for Binding Site Comparison

INTRODUCTION

X-ray crystallography and Nuclear Magnetic resonance have delivered and continue to deliver a rapidly increasing number of protein structures. Technical advances has opened the door to new challenges in structural biology (e.g. membrane proteins, large macromolecular complexes …).[1] A particular class of proteins are biosynthetic enzymes. They are nature's chemists responsible for the synthesis of natural products. In the domain of biosynthesis, many structural biology studies have focused in understanding the functions of enzymes due to their interesting aspects for pharmaceutical, energy and food industries. The difficult nature of the work prompts biologists, chemist and computational scientists to combine their efforts around biosynthetic instances in order to characterize enzymatic reaction mechanism. As a consequence, knowledge in the field has dramatically increased, yielding in vast and sometimes versatile data spread over thousands of scientific articles. Thankfully, some structural biologist have synthetized their knowledge and thus, facilitate the work of others by publishing articles reviewing many structural biology studies at once. However, these reviews tend to focus on the exploration of structural data within a family of biosynthetic enzymes such as for example methyltransferases[2] or terpene synthases.[3,4] In addition, these studies sometimes investigate structural relationships between related enzymes, for example in order to infer an enzyme's function, but hardly investigate structural similarity with (potential) target proteins of the produced compounds.

In order to find structural relationships between natural product biosynthetic enzymes and natural product target proteins, we focus on binding site similarity. To compare binding site, we used, namely SiteAlign[5] and Shaper,[6] two programs developed on the underlying "key-lock" principle postulating that similar binding sites bind the same

ligands. In our earlier study focusing on the structural basis of ligand promiscuity, we showed that different targets of a small molecular weight compound can share common structural patterns.[7] Following this study, we examined the ability of binding site similarity to capture biological imprint of flavonoids in flavonoid target proteins. We could demonstrate that flavonoid target proteins share binding site features with flavonoid biosynthetic enzymes, thereby proposing that biological imprints are embedded within catalytic sites of biosynthetic enzymes.[8] However, virtual screening results has shown very versatile results and thus, questioned the reliability of our binding site representation to describe this particular biological imprints. In order to further study how biosynthetic enzymes can be related to natural product target proteins, it was required to carry on with structural investigations of the molecular recognitions made by biosynthetic enzymes to their natural substrates on a broader scale.

We have collected, a set of 117 natural product biosynthetic enzymes. We investigated if their binding sites are prone to share binding site features with targets and more particularly, we examined if their binding-modes with natural product substrates suit the “key-lock” principle onto which our binding site similarity methods rely. Examinations were carried out on two different levels. First, we did a chemical analysis of enzymatic activities. From there, we could already propose that not all biosynthetic enzymes interact with a substrate that is representative of a natural product. Then, we examined binding-modes of relevant substrates within crystallographic structures. We raised several points characterizing how biological imprints relate to our representation of binding sites.

Lastly, a focus is made on one biosynthetic of penicillin G. Screening results of the enzyme versus the sc-PDB[9] will be presented and we will discuss the ability of site comparison methods to capture biological imprints similarity in known targets binding sites.

1. METHOD AND MATERIALS

Ligandable biosynthetic enzyme collection

Ligandable active site of 117 biosynthetic enzymes were identified in the PDB as described in chapter 4. Each enzyme was annotated with appropriate substrate and product structures. The list of selected enzymes is given in **table 2** of **annex 5**.

1.1. Virtual screening

We searched for similar binding sites in the sc-PDB following the method described in chapter 3. Briefly, ligandable biosynthetic enzymes active sites, were compared to sc-PDB binding sites using two different methods, SiteAlign and Shaper. For each screen, hit lists was obtained using a distribution-based similarity score cutoff. The method described in chapter 3, was used for similarity cutoffs definition of SiteAlign screenings with mean and standard deviation of the complete distribution of D2 similarity score and thus, before D1 filtering. Protein hits had Z-score higher than 2.5.

1.2. Chemical structure of substrates and products

Chemical structures of substrates and products in the enzymatic reactions were tagged with their biosynthetic pathways. Pathways were analyzed based on metabolic networks provided on MetaCyc[10] website (www.metacyc.org/), and if available, from UniPathway[11] website (www.unipathway.org/).

1.3. Molecular recognition

Molecular recognition of enzymatic substrates and products by the enzyme active site was analyzed by visual inspection of crystallographic complexes using the molecular viewing software Chimera.[12] If necessary, identification of catalytic sites was supported by literature reports.

RESULT & DISCUSSION

In the rest of this chapter, we assumed that binding sites embedding the biological imprint of natural products are most likely to give high similarity scores to target proteins. We examined substrates structures and their three-dimensional molecular recognitions to decipher the quality of the biological imprint. From thereon, enzymes embedding a natural product biological imprint will be called “good baits” as opposed to other enzymes, which will be called “bad baits”.

Part 1: chemical structures of substrates and products

An enzyme is prone to carry the biological imprint of a natural product if it acts on a substrate closely related to that natural product. However, it is not the case that all biosynthetic enzymes in our dataset have substrates closely related to the natural product. As shown in **figure 1**, substrate (and product) of a biosynthetic reaction can be structurally different to final natural products. Flavin-dependent tryptophan halogenases RebH and PrnA (UniProt ID: Q8KHZ8, P95480 respectively) *b*oth catalyze the chlorination of tryptophan into 7-chloro-L-tryptophan[13,14] in the first step of rebeccamycin and pyrolnitrin biosynthesis respectively. As a consequence, the enzymes should be considered as "bad baits". Following this observation, we adopted an overall approach, which assumed that "good baits" are more likely to take place in the downstream part of pathways, near the end products. Further on, phenylalanine aminomutase (UniProt ID: Q6CZ04) is responsible for the rearrangement of L-phenylalanine to R-β-phenylalanine[15] in the first step of taxol's 13C-side chain biosynthesis. The biosynthetic step is far from the overall end product taxol and, even if it is close to the final step of taxol's 13C side chain biosynthesis (a sub-pathway in taxol biosynthesis), we can reasonably affirm that the enzyme is a "bad bait" (**figure 2)**.

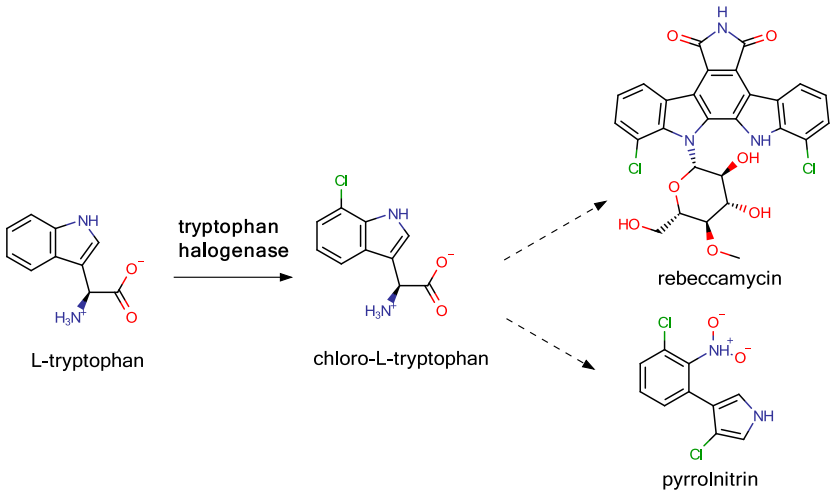


Figure 1. A common precursor of two different antibiotics: L-tryptophan

Farnesyl diphosphate and geranyl diphosphate are both key metabolites at the origin of diverse terpene natural products. The construction of these precursor metabolites is described in multi-step biosynthetic pathways (via mevalonate and via methylerythritol) involving not less than 10 enzymes from our dataset. Even if the enzymes are taking place in downstream part of their respective pathways, almost all of their reaction products are not characteristic of mature terpenes. Therefore, enzymes in the biosynthesis of farnesyl diphosphate and geranyl diphosphate should be considered as "bad baits" (**figure 3**).

Furthermore, after construction, these isoprenoid metabolites can undergo a cyclization reaction that yields in the core of cyclic terpenes. These reactions are often described as the first committed step in their respective biosynthetic pathway. For example, aristolochene synthase (UniProt ID: Q03471) catalyzes the cyclization of farnesyl

diphosphate into aristolochene[16] in what is described as the first step of aristolochene biosynthesis.

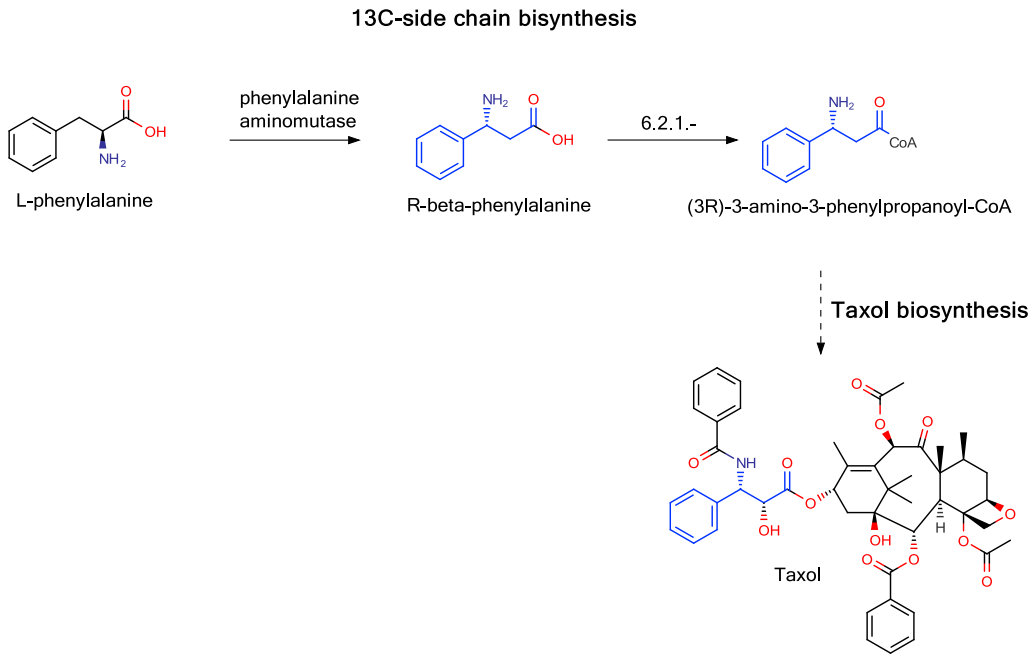


Figure 2. Synthesis of 13C-side chain of taxol.
Structures shown with blue bonds represent the common scaffold and it presence in the end products

Other similar examples from our dataset are illustrated in **figure 4**. Because these enzymes are forming the core of cyclic terpenes, they should be considered as "good baits", even though they take place in the first step of their respective pathways. Due to the complexity of biosynthetic pathways compartmentation, neither the position of enzymatic reactions in the pathways nor the scaffolds of end products in pathways is sufficient to predict if an enzyme is a "good bait". Hence, we have opted for a manual investigation rather than an automated procedure to discriminate "good baits" from "bad baits". Nevertheless, a hand-made list of natural products resulting from the pathways could be used as reference molecules, and in conjunction with an appropriate

similarity method, it could be feasible to define a cutoff beyond which biosynthetic enzyme can be considered “good baits”. If such a method is being used, ubiquitous transporter groups such as coenzyme A attached should be ignored when comparing metabolites to the end product of the pathways as they may interfere with the similarity measure.

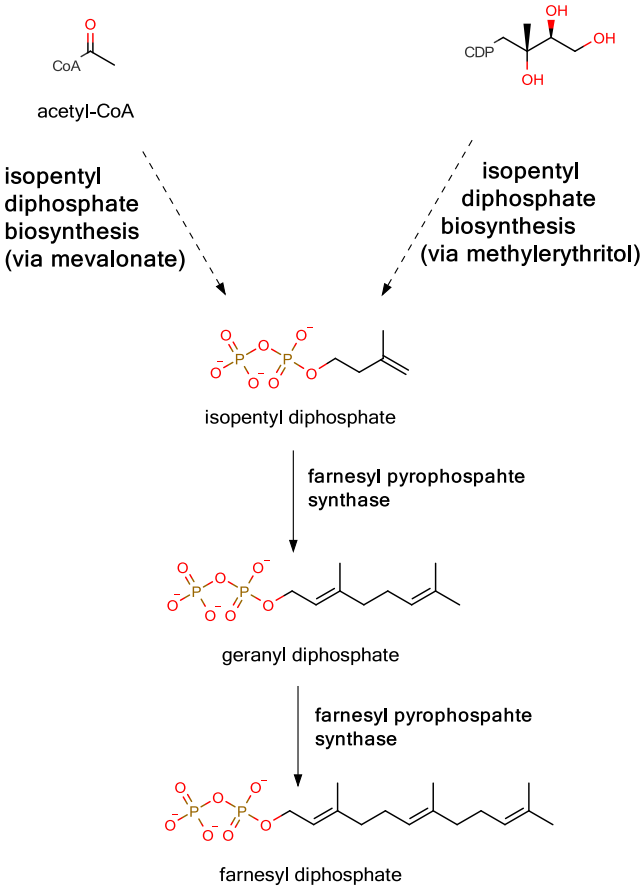


Figure 3. Building blocks of sesquiterpenes: isoprenoids.
The top of the figure illustrates acetyl-CoA (left) and methylerythritol (right), the two possible origins of isopentyl diphosphate.

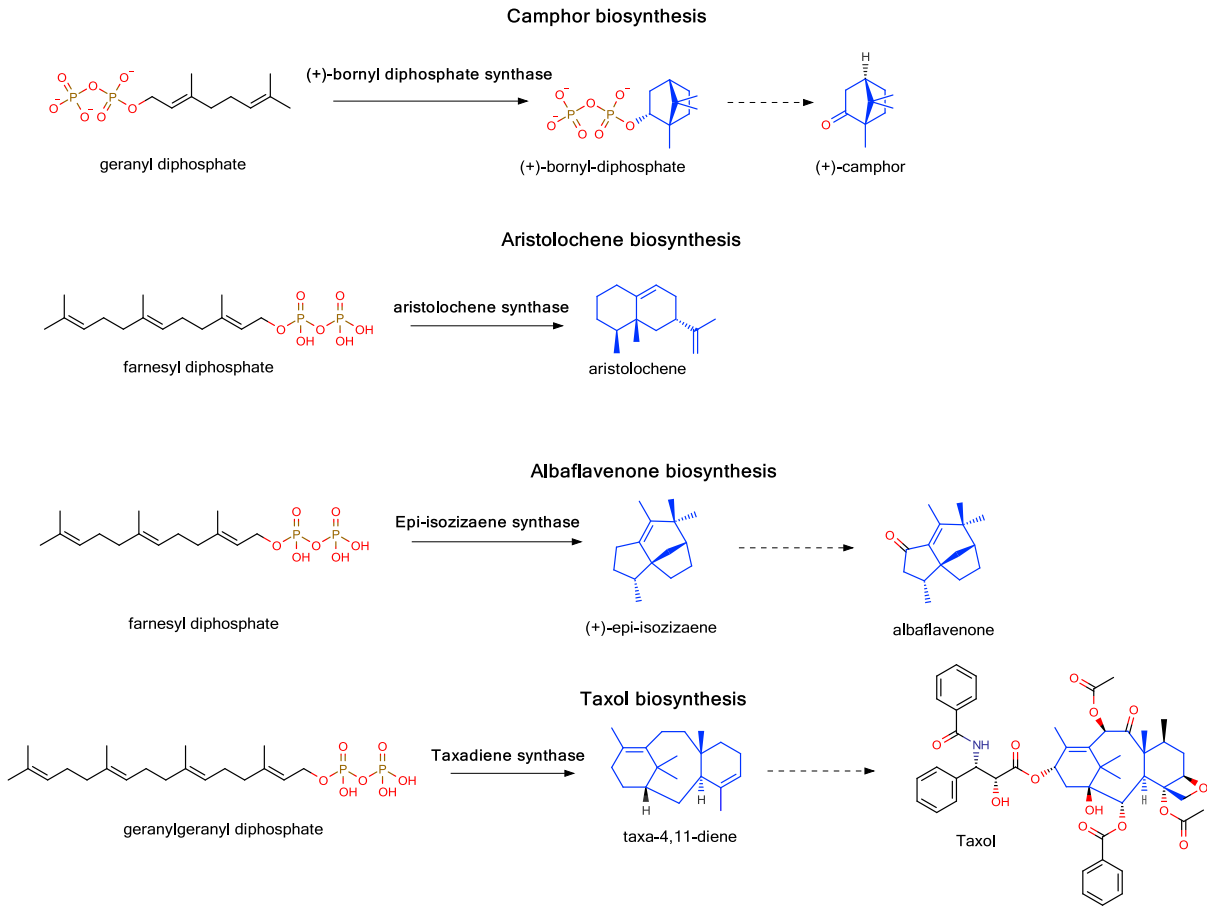


Figure 4. Illustration of some isoprenoid cyclization examples.
Structures shown with blue bonds represent the common scaffold and it presence in the end product.

Part 2: substrates molecular recognition

In this section, we examined crystallographic structures to find rules predicting if an enzyme is a “good bait”. We visually inspected substrates and their derivatives binding sites and evaluated their binding mode. Throughout case-by-case analysis we could raise four scenarios. Typical examples will be discussed.

1. The enzyme sequesters tightly the natural substrate

A biological imprint gives shape and properties to a natural product. In other words, the enzymatic environment in contact with a substrate is the structural basis that steers the formation of a natural product thanks to shape and property complementarity. Therefore, full specific biological imprints are embedded in enzymes that encapsulate natural product ligands tightly. A family of biosynthetic enzymes showing this peculiar property is the terpenoid cyclase family. These enzymes catalyze the first step in the biosynthesis of a vast variety of terpenes, including cyclic terpenes.[3] (+)-bornyl diphosphate synthase (UniProt ID: O81192) is one of them, it catalyzes the formation of (+)-camphor precursor, (+)-bornyl diphosphate, from geranyl diphosphate.[17] The **Figure 5** shows how the reaction product is sequestered in the active site with an almost perfect complementarity. In this particular example, the active site has been suggested to serve as a template to chaperone substrate conformation in order to initiate the reaction mechanism.[18] A number of hydrophobic residues, including Trp323, Ile334, Val452 and Phe578, all located around the cyclic terpene moiety, steer the flexible substrate conformation towards what will become the reaction product. It is tempting to suggest that these particular residues are playing a key role in the biological imprint that shapes the molecular core of (+)-camphor. Hence, we should consider this scenario as a “good bait” marker for further virtual screenings. Similar observations can be made in more terpene cyclases responsible for the synthesis of monoterpenes, sesquiterpenes, or triterpenes. Most commonly, isoprenoid substrates are sequestered within hydrophobic pockets. Nevertheless, these pockets do have different amino-acids arrangements and thus steer reaction mechanisms towards various terpene products.[4]

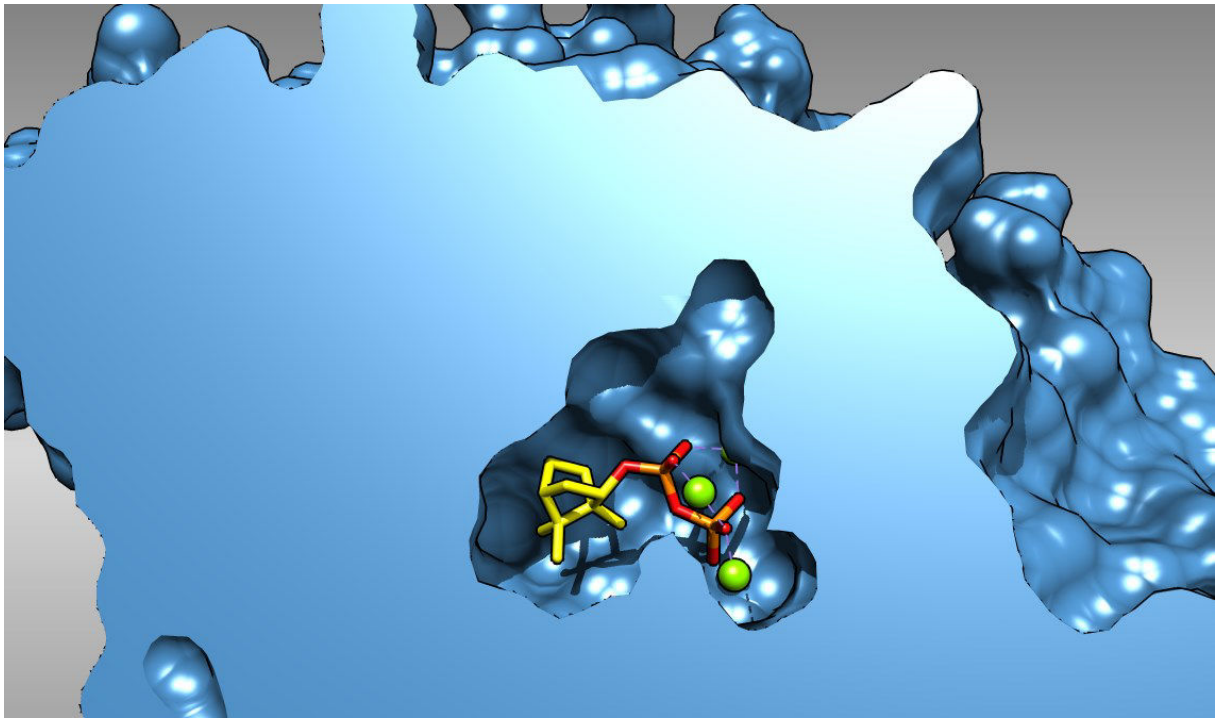


Figure 5. Capped view of the active site of (+)-bornyl diphosphate synthase (PDB ID: 1N24).
The enzyme surface is represented in steel blue. The reaction product, (+)-bornyl diphosphate is represented with yellow sticks. Green spheres represent magnesium atoms responsible for the fixation of pyrophosphate group (orange).

2. The enzyme partially recognizes the natural substrate

We assumed that an enzyme carries the biological imprint of a natural product if it recognizes the complete structure of the natural product. However, in many cases natural products achieve their functions throughout series of enzymatic reactions attaching or modifying substituent components to the core of the molecule under construction. For example hydroxyl groups are commonly methylated in order to modulate natural products bioavailability, bioactivity or reactivity.[2] These reactions are most commonly catalyzed by methyltransferases through transferring a methyl group from S-adenosyl-L-methionine co-factors to a methyl accepting atom of a natural substrate. Methyl accepting atoms are often part of the natural product decoration and thus they are not central to the core of the molecule. Therefore, catalytic cores of

methyltransferases do not necessarily need to encapsulate entire substrate structures during their activities. Mycinamicin III 3"-O-methyltransferase MycF (UniProt ID: Q49492), involved in the biosynthesis of the antibiotic mycinamicin[19], illustrates an extreme scenario (**figure 6**).

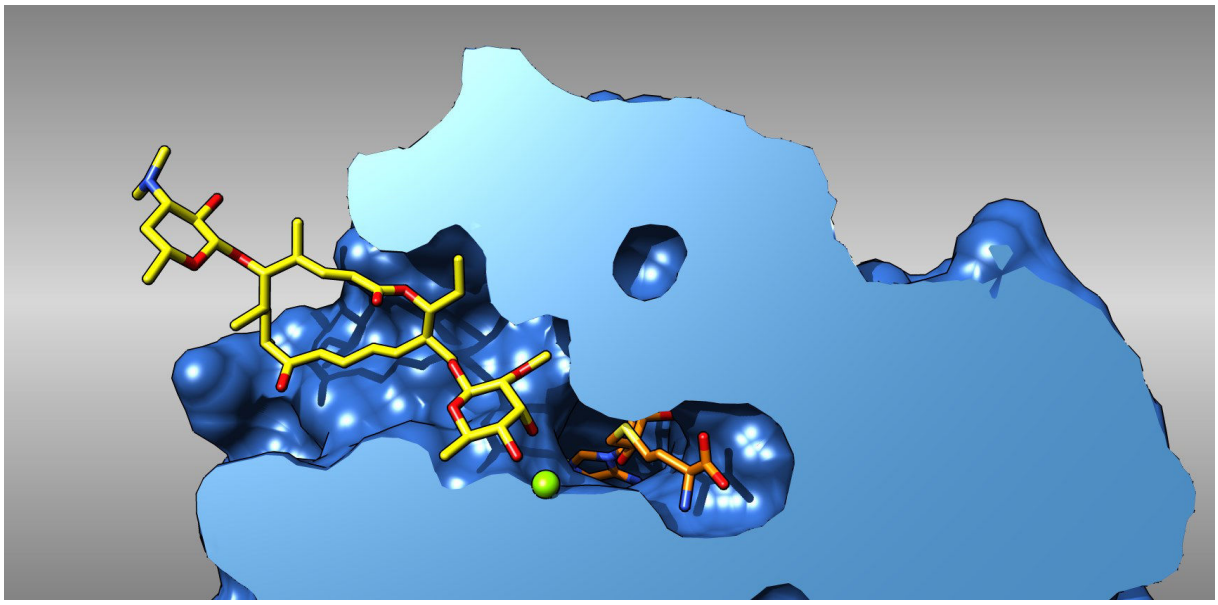


Figure 6. Capped view of mycinamicin III 3"-O-methyltransferase in complex with its natural substrate mycinamicin III (PDB ID: 4X7U). The enzyme surface is represented with steel blue color. Natural substrate is represented with yellow sticks. The co-substrate, S-adenosyl-L-homocysteine is represented with orange sticks and is presently mimicking the natural co-substrate S-adenosyl-methionine. The green sphere represents a magnesium ion.

MycF catalyzes the 3'-O-methylation of the javose moiety of mycinamicin III to form the mycinose moiety of mycinamicin IV[20] (**annex 4, enzymatic activity investigation**). The **figure 6** shows the natural substrate mycinamicin III in the active site of MycF as described in PDB ID: 4X7U. On one hand, in the buried part of the cavity, the javose moiety of the substrate makes specific contacts with the catalytic core of the enzyme. The hydroxyl groups at the position 3 and 4 of the sugar both coordinate a magnesium

ion and interact through H-bonds with the residues Asn191 and Gln246. On the other hand, the macrolactone ring of the substrate does not show any specific contacts with MycF as it is located in a less buried (and hydrophobic) region. Furthermore, the desosamine sugar is totally exposed to the solvent, it makes no contact with MycF at all, which suggests that the enzyme could tolerate javose substrates bearing different macrolactone rings. In fact, the recent study of substrate specificity in MycF has reported that an alternate substrate containing javose and the macrolactone ring but with an additional sugar attached to desosamine was not affecting MycF enzymatic activity.[20] Thus we can affirm that the enzyme specifically recognizes the javose moiety but that macrolactone and desosamine rings are not specifically recognized. Nevertheless, this example does not stand for all methyltransferases as suggests the binding mode of mycinamicin VI in the active site of MycE. The enzyme catalyzes the 2'-O-methylation of 6'-deoxyallose right before MycF biosynthetic step and shows more specificity to the macrolactone ring.[20] Furthermore, it is not a scenario specific to methyltransferases. For example isopenicillin N synthase is involved in the formation of the β-lactam ring of penicillin but does not make specific contacts with the side chain. Lastly, partial recognition of natural should not be associated to "bad baits" always. Especially if the recognized fragment is responsible for pharmacological activity. In a certain extent, such cases could be considered "good baits" for virtual screening.

3. The molecular recognition is not specific of the natural product

Biosynthetic enzymes are most likely carrying the biological imprint of a natural product when they specifically recognize substrates. However, as discussed in the previous point, biological imprint in an enzyme binding site sometimes accounts for fragments of the natural product. Beyond partial substrate recognition, we have observed complexes of enzymes without specific binding-modes to their putative substrate. It is the case of some enzymes exhibiting monooxygenase activity within cytochrome p450 domains. *E*pi-isozizaene 5-monooxygenase (UniProt ID: Q9K498) is one of them, it catalyzes a two-step allylic oxidation.[21] First, it carries out an oxidation of epi-isozizaene and then it performs another oxidation to yield albaflavenone (**annex 4 section enzymatic investigation**). As visible in the structure of the enzyme complexed with the reaction substrate (**figure 7**), an epi-isozizaene molecule positions its reactive carbon over the heme group responsible for monooxygenase activity. However, the study of albaflavenone biosynthesis reported that the enzyme product was a mixture of roughly equivalent amounts of (5R)-albaflavenol and (5S)-albaflavenol, demonstrating the lack of sterospecificity. Further on, structural studies of the enzyme revealed the presence of a second epi-isozizaene molecule, bound at the entrance of the active site. The endo and exo orientations of the two ligands lead the team of structural biologist to suggest the existence of two substrate binding-modes, each of them yielding in a different stereoisomer.[22] In that particular case, minor specific contacts with the substrate and an overwhelming proportion of the cavity render the identification of a proper biological imprint difficult, if it is present. This suggestion might stand for other cytochromes p450 as their binding sites are usually formed by a large hydrophobic cage holding the heme group and thus not favoring specific contacts with the substrate. Besides cytochromes

p450, other enzymes have been characterized with non-specific reactions. For example, many terpene synthases produce multiple compounds,[2–4] which should warn us in the interpretation of their biological imprints. Lastly, as visible in **figure 7**, the cavity exceeds largely the coverage of the substrate molecule, showing that the example is particularly inconsistent with our aim. The cavity generated by VolSite, and thus the cavity that would be used to search for similar binding sites, is a very loose representative of albaflavenone biological imprint.

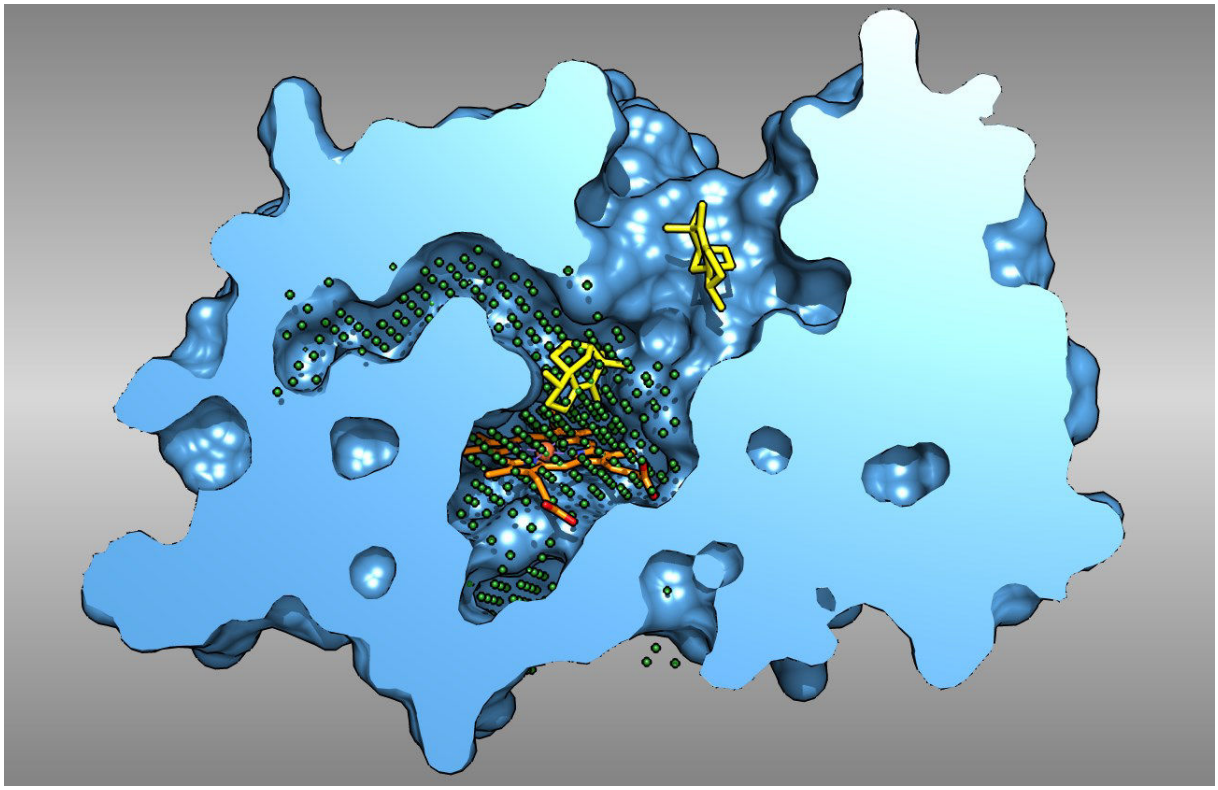


Figure 7. Capped view of epi-isozizaene 5-monooxygenase in complex with the natural substrate (+)-epi-isozizaene (PDB ID: 3EL3). The enzyme surface is colored in steel blue. Natural substrates are represented with yellow sticks whereas the heme group responsible for monooxygenase activity is represented with orange sticks. Green points represent VolSite cavity points.

4. The active site is not representative of the enzyme active state

Biosynthetic enzymes can undergo structural changes upon binding of substrates or co-factors. It is the case of terpene synthases for example. In terpene synthase apo-structures, the active site cleft is exposed to the solvent. Upon binding of the pyrophosphate group of isoprenoid substrates a loop closes the active site, shielding the reacting chamber from solvent.[4] However, in these examples, structural rearrangement of the active site entrance has only a minor impact on substrate molecular core recognition, as this rearrangement recognizes mainly the pyrophosphate group. Nevertheless, these structural changes affect our binding site definition and thus, could affect virtual screening results. In other cases, natural product biosynthesis was explored through series of mutational studies. For example, epi-isozizaene synthase *(*accession code = Q9K499), which is the enzyme responsible for the formation of the molecular core of albaflavenone,[23] has been extensively mutated in order to study the catalytic mechanism[24] and in order to explore chemodiversity of possible "unnatural" reaction products.[25] As a result, a number of enzyme structures with mutated residues are available and labeled under the uniprot accession code of the wild type enzyme. Therefore, a moderate credit should be given to the biological imprint as they might be bad representative of natural products if the mutations affect atomic coordinate too much. Other enzymes undergo more dramatic conformational changes. For example, an Ntn-hydrolases involved in the biosynthesis of penicillin undergoes a considerable conformational change between precursor and mature states. Acyl coenzyme A:isopenicillin N acyltransferase (AT) (UniProt ID: P15802) is responsible for the conversion of isopenicillin N to penicillin G by exchanging the hydrophilic side chain with a phenyl group.[26] AT is produced as an inactive precursor enzyme and undergoes a

posttranslational modification that cleaves its peptide chain. As shown in **figure 8**, the entrance of the active site in its inactive form is blocked by an α-helix composed of 10 residues (yellow helix). However, after cleavage, a whole segment of the of the chain folds outwards, exposing the cavity to the solvent. This major structural change affects the binding site definition and thus, it would be highly inadvisable to use the enzyme precursor structures as bait for virtual screening.

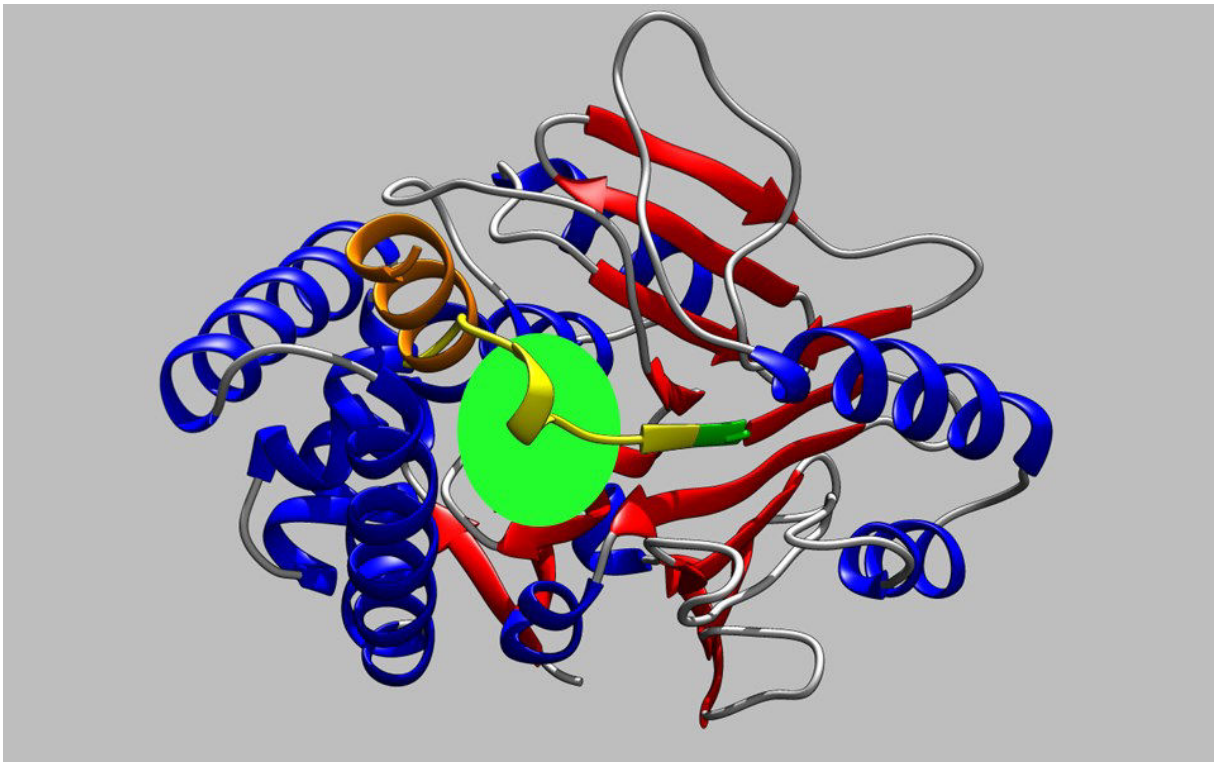


Figure 8. Active and inactive state of acyl coenzyme A:isopenicillin N acyltransferase.
The enzyme is represented with ribbons. The enzyme is oriented so that the reader looks down into active site. The yellow segment represents the entrance of the active site in the inactive form of the enzyme (PDB ID: 2X1C). When activated, the yellow segment is cleaved at Cys103 (represented in green) and folds outwards, exposing the active site cleft (green patch) to the solvent. Orange helix represents the cleaved segment in the active form of the enzyme (PDB ID: 2X1E).

Part 3: A new example of Protein Fold Topology

Considering the series of observations discussed in the previous sections, we selected "good baits" for experiments aiming at finding structural relationships with natural products target proteins. Focus was given to enzymes synthetizing a natural product with known targets, thus allowing us to investigate retrospective examples. In order to identify these enzymes, we designed an automated pipeline able to search compounds in the bioaffinity database ChEMBL,[27] the database of approved drugs DrugBank[28] and the protein data bank[29] (**annex 5**, **figure 1**). The pipeline returns reported target proteins of natural products in our dataset of biosynthetic enzymes. Adjustments in the search method are still needed. Notably, we plan to set a similarity search (instead of the exact match already in place) that would enable us to find target proteins of natural product using a closely related metabolite. This type of search was carried out manually until now. Last but not least, we ensured of the presence of known targets in the sc-PDB, the screened dataset.

In our dataset, we dispose of enzymes involved in penicillin G biosynthesis. Amongst them is isopenicillin N synthase (IPNS) (UniProt ID: P05326). The enzyme catalyzes the formation of the lactam core, which is the pharmacological principle of β-lactam antibiotics.[30] The biosynthetic step prepares the β-lactam moiety before addition of a benzyl in the last biosynthetic reaction. Thereby the enzyme catalyzes a reaction close to the final step. Structures of IPNS are in complex with analogues of product and substrate of the natural enzymatic reaction and all have highly conserved binding modes, indicating the consistency of the "good baits".

Penicillin G is an antibiotic that inhibits the formation of peptidoglycan cross-links in bacterial cell wall, thus favoring cell membrane degradation.[31] Unfortunately, target

proteins affected by the pharmacologic activity (penicillin binding proteins) are not present in our screened dataset. However, the sc-PDB contains eight different β-lactamases (28 structures) involved in bacterial resistance against β-lactam antibiotics. This enzyme is known to hydrolyze penicillin lactam cores and therefore disable their antibiotic activity.[31] The fact that β-lactamases hydrolyze lactam cores of antibiotics indicates us that they constitute interesting true positives, as they recognize the lactam core in penicillins.

Virtual screening experiments using IPNS as the bait was successful in finding β-lactamases with SiteAlign only. When using the structure PDB ID: 1W05 as the bait, we could identify a New-Dehli Metallo-β-lactamase (NDM-1, PDB ID: 4HL2) with significant similarity compared to the rest of the comparisons. In this screening, distribution of similarity scores is characteristic of a normal distribution, allowing us to assign a Z-score of 2.55 to NDM-1 (**annex 5, figure 2**). The bait we used for virtual screening is in complex with an analogue of isopenicillin N precursor, a tripeptide onto which the valine carboxylic end was truncated to an alanine carboxylic end.[30] The substrate anchors its thiol group to the catalytic iron atom, indicating that the lactam ring will be formed here. Actually, two structures of IPNS (PDB ID: 2JB4 and 1ODN) support this indication as they are positioning the lactam ring of isopenicillin N analogues onto this exact same spot.[32,33] In NDM-1 structure, the ligand represents the hydrolyzed form of penicillin G, and thereby represents a picture of the system after hydrolysis. At the difference to IPNS, the lactam sulfur atom does not interact with the catalytic center but is located at the opposite side of the ring, indicating that the lactam ring opens from the C-N bond. Indeed, nitrogen atom and carboxylic group of what is left of the lactam ring are anchored to the catalytic zincs.

SiteAlign was able to superimpose a number of features successfully. Mainly, two patches superimposed well, aligning IPNS cavity to one part of NDM-1 active site cleft. As shown on **figure 9**, a hydrophobic region in contact with the benzyl moiety of hydrolyzed penicillin G superimposed partially with IPNS cavity lining to isopenicillin N precursor. More precisely, Ile35, Leu65 and Val73 in NDM-1 occupy a spatial location that is equivalent to Leu317, Leu321 and Val217. More importantly, catalytic cores anchoring the lactam cores are well superimposed. In IPNS, residues of the catalytic triad coordinating the iron atom, His120-His122-Asp124, matched residues coordinating a zinc atom in NDM-1, His214-His270-Asp216 respectively. Other than that, punctual matches are spread over the cavities. A total of 19 residues overlapped onto the superimposed polyhedrons of SiteAlign out of which, 11 overlaps have a specific distance score value less of equal to 0.2 (**annex 5, table 1**). Since SiteAlign method represents binding sites with a degree of fuzziness, one does not expect an optimized atomic superimposing. Differently to what we could observe in the flavonoid-kinase example (chapter 3), no conserved secondary structure elements could be visualized in the alignment. Here, secondary structure elements of the two enzymes join or cross punctually in order to form a conserved pattern. Hence, the example demonstrates that two proteins do not need to present similar fold arrangements in the vicinity of the binding sites to recognize a same molecular core.

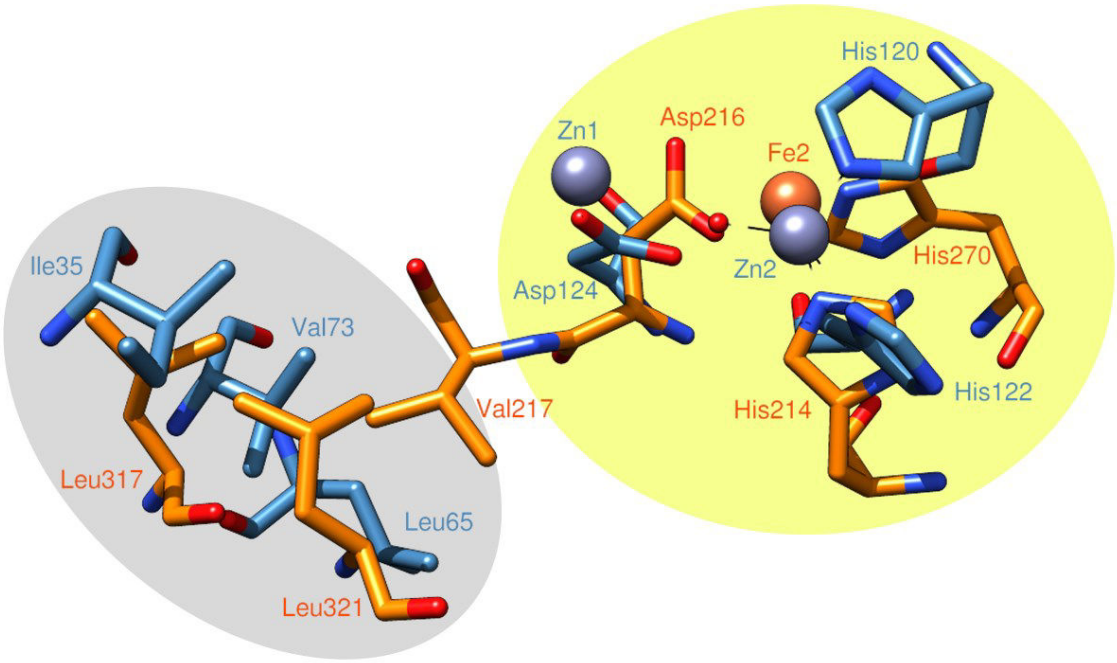


Figure 9. Conserved patterns in IPNS and NDM-1.
This figure represents an alignment made with SiteAlign, superimposing active sites of IPNS (PDB ID: 1W05, orange sticks) and NDM-1 (PDB ID: 4HL2, steel blue sticks). Grey spheres represent catalytic zinc ions in NDM-1 whereas the orange sphere represents the catalytic iron in IPNS. A patch of hydrophobic residues (grey ellipse) and the catalytic center are conserved in the two active sites (grey ellipse) and the catalytic center is also equivalently located.

When compared together, the two binding sites have an overall different shape (**figure 10**), explaining why Shaper was not able to identify the pair of enzymes as similar. The active site of IPNS squeezes isopenicillin N precursor into a closed cavity whereas the active site of NDM-1 is a wide cleft exposed to the solvent. As indicated by the position of isopenicillin N precursor in **figure 10**, the cavity of IPNS aligns to roughly one half of the active site cleft in NDM-1, including the catalytic center. Consequently, IPNS binding site does not align with the hydrolyzed form of penicillin G in NDM-1. The fact that the two enzymatic products have a common β-lactam core but bear different substitutions, indirectly explains why the alignment did not superimpose the two ligand bioactive poses. In IPNS, the cavity provides apolar contacts (squeezing the ligand into the cavity) partly matching those of NDM-1 (grey surface) whereas at the two extremities of IPNS

cavity, polar residues lock the substrate into the bioactive position. On one side (yellow surface) the polar residues form the catalytic core whereas on the other side, polar residues in IPNS are superimposed to a region of NDM-1 active site that is different to the environment of the benzyl in hydrolyzed penicillin G. Nonetheless, this mismatch does not discredit our alignment because corresponding sub-pockets do not recognize the same substrate moiety. Besides, one can reasonably affirm that the biological imprint of penicillin G is not contained within the mismatched regions of IPNS and NMD-1 but is mostly located in the catalytic cores. Although the mismatched part of IPNS binding site is rather specific to its substrate, it recognizes a moiety of the ligand that is not present in penicillin G. Alternatively, the benzyl moiety of hydrolyzed penicillin G is not specifically recognized since NDM-1 is known for its ability to accommodate β-lactam antibiotics with substitution variants at this precise location.[34,35]

This result highlights the fact that metallo-β-lactamase, and more generally natural product targets, do not need to reproduce complete biological imprint in order to interact with a natural product. Furthermore on this line, since biosynthetic enzymes construct natural products throughout series of reactions, (much like building steps add pieces to a final work) they do not need to recognize the complete natural product but rather the specific parts that are being assembled. Thereby, one can say that the biological imprint of a natural product may be fragmented into several biosynthetic enzymes each containing sub-parts of a global natural product biological imprint. Lastly, our binding site representation suits the “key-lock” model but biosynthetic enzymes interact with natural product by induced fit. Even if or screening result suggests that our binding site representations essentially contains the biological imprint of a natural product, the representation does not specifically represent the biological imprints. In

many cases, binding sites overcast the set of residues that is relevant to the synthesis of the natural product, or the residues in the binding site are not configured in active state. Regarding biological imprint, these cases contains non-representative data that can mislead binding site comparison experiments.

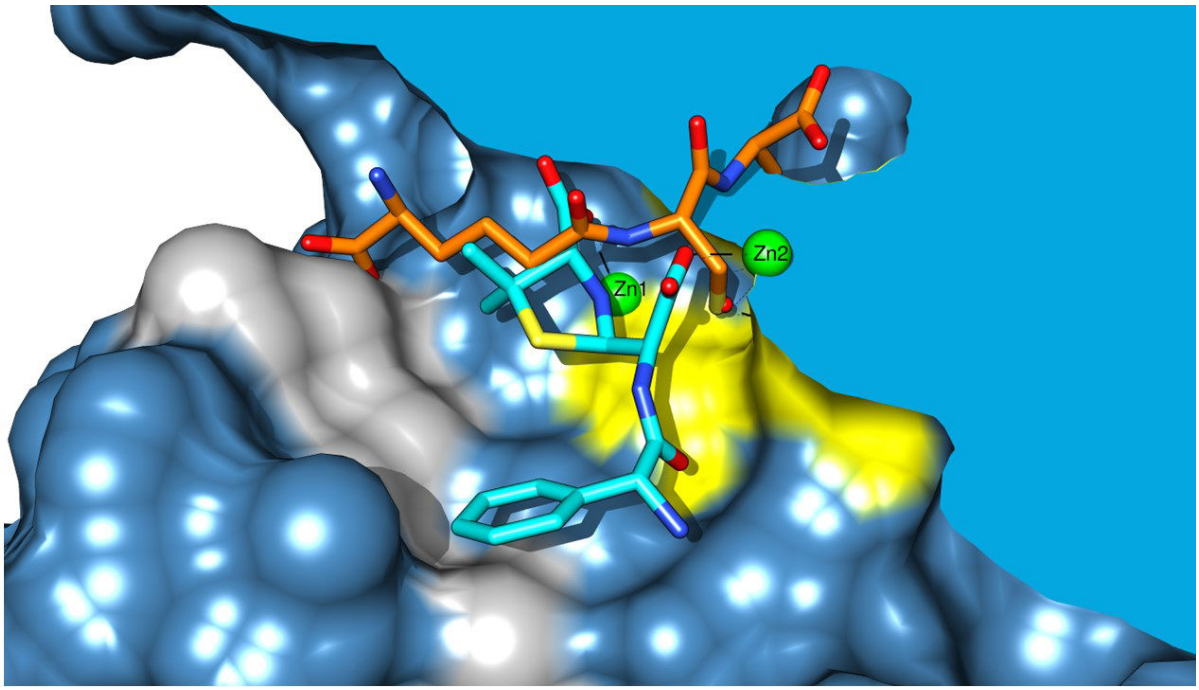


Figure 10. Capped view of NDM-1 active site with isopenicillin N precursor after SiteAlign superimposition.
The surface of the protein is represented with steel blue. The grey patch represents the conserved hydrophobic residues whereas the yellow patch represents the conserved catalytic core. Green spheres represent zinc ions from NDM-1. Orange sticks represent isopenicillin N precursor, after superimposition according to SiteAlign, whereas the cyan sticks represent the hydrolyzed form of penicillin G in its binding site.

CONCLUSION

In this chapter, we have seen that biosynthetic enzymes embed biological imprint of natural products with various levels of representability. On one extreme, enzymes located in upstream parts of biosynthetic pathways may not be representative of the biological imprint at all. On the other extreme, an enzyme that sequesters tightly a natural product may embed a large portion of the biological imprint. Alternatively, enzymes contributing in the assembly of certain parts of natural products embed a portion of the biological imprint that is specific to the assembled part of the produced compound. Considering our structural observations, we can propose a new definition of biological imprints. A biological imprint is fragmented into the sequence of enzymes involved in the biosynthetic pathway of a natural product. The imprint is formed by the set structural patterns within residues directly contributing to the assemblage of chemical features present in the final natural product.

In addition, our binding site similarity screening experiment has demonstrated that some components of biological imprint may be more related to pharmacological activities than other. In our case, the enzyme that catalyzes the formation of the β-lactam core in penicillin antibiotics has shown structural relationships with β-lactamases. Moreover, this experiment has proved once again that binding site similarity could capture “portions” of biological imprints reproduced within natural product targets. However, the remote similarity that we found in the penicillin example has emphasized even more that fact that binding site representations are maybe not the most suited to find this kind of similarities. In fact, our representation relies on the “key-lock” model which takes most sense when a molecule adapts to a binding cavity as often observed with inhibitors binding a protein. However, in biosynthesis, recognition of the

substrates often happen by induced fit, it is the enzymes that adapts to the substrate upon binding. Hence, our findings with the penicillin example are more coincidental than verifying the “key-lock” principle. Therefore, capturing biological imprint of natural products in target protein binding sites requires more appropriate binding site representation.

REFERENCES

1. Zheng, H. *et al.* X-ray crystallography over the past decade for novel drug discovery - where are we heading next? *Expert Opin. Drug Discov.* **10,** 975–989 (2015).
2. Liscombe, D. K., Louie, G. V. & Noel, J. P. Architectures, mechanisms and molecular evolution of natural product methyltransferases. *Nat. Prod. Rep.* **29,** 1238–1250 (2012).
3. Degenhardt, J., Köllner, T. G. & Gershenzon, J. Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry* **70,** 1621–1637 (2009).
4. Christianson, D. W. Structural Biology and Chemistry of the Terpenoid Cyclases. *Chem. Rev.* **106,** 3412–3442 (2006).
5. Schalon, C., Surgand, J.-S., Kellenberger, E. & Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins Struct. Funct. Bioinforma.* **71,** 1755–1778 (2008).
6. Desaphy, J., Azdimousa, K., Kellenberger, E. & Rognan, D. Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes. *J. Chem. Inf. Model.* **52,** 2287–2299 (2012).
7. Sturm, N., Desaphy, J., Quinn, R. J., Rognan, D. & Kellenberger, E. Structural insights into the molecular basis of the ligand promiscuity. *J. Chem. Inf. Model.* **52,** 2410–2421 (2012).
8. Sturm, N., Quinn, R. J. & Kellenberger, E. Similarity between Flavonoid Biosynthetic Enzymes and Flavonoid Protein Targets Captured by Three-Dimensional Computing Approach. *Planta Med.* **81,** 467–473 (2015).
9. Desaphy, J., Bret, G., Rognan, D. & Kellenberger, E. sc-PDB: a 3D-database of ligandable binding sites--10 years on. *Nucleic Acids Res.* **43,** D399–404 (2015).
10. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **42,** D459–D471 (2014).
11. Morgat, A. *et al.* UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.* **40,** D761–D769 (2012).
12. Pettersen, E. F. *et al.* UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
13. Kirner, S. *et al.* Functions encoded by pyrrolnitrin biosynthetic genes from Pseudomonas fluorescens. *J. Bacteriol.* **180,** 1939–1943 (1998).
14. Yeh, E., Garneau, S. & Walsh, C. T. Robust in vitro activity of RebF and RebH, a two-component reductase/halogenase, generating 7-chlorotryptophan during rebeccamycin biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 3960–3965 (2005).
15. Feng, L., Wanninayake, U., Strom, S., Geiger, J. & Walker, K. D. Mechanistic, mutational, and structural evaluation of a Taxus phenylalanine aminomutase. *Biochemistry (Mosc.)* **50,** 2919–2930 (2011).

16. Proctor, R. H. & Hohn, T. M. Aristolochene synthase. Isolation, characterization, and bacterial expression of a sesquiterpenoid biosynthetic gene (Ari1) from Penicillium roqueforti. *J. Biol. Chem.* **268,** 4543–4548 (1993).

17. Wise, M. L., Savage, T. J., Katahira, E. & Croteau, R. Monoterpene synthases from common sage (Salvia officinalis). cDNA isolation, characterization, and functional expression of (+)-sabinene synthase, 1,8-cineole synthase, and (+)-bornyl diphosphate synthase. *J. Biol. Chem.* **273,** 14891–14899 (1998).

18. Whittington, D. A. *et al.* Bornyl diphosphate synthase: Structure and strategy for carbocation manipulation by a terpenoid cyclase. *Proc. Natl. Acad. Sci.* **99,** 15375–15380 (2002).

19. Li, S., Anzai, Y., Kinoshita, K., Kato, F. & Sherman, D. H. Functional analysis of MycE and MycF, two O-methyltransferases involved in the biosynthesis of mycinamicin macrolide antibiotics. *Chembiochem Eur. J. Chem. Biol.* **10,** 1297–1301 (2009).

20. Bernard, S. M. *et al.* Structural Basis of Substrate Specificity and Regiochemistry in the MycF/TylF Family of Sugar O-Methyltransferases. *ACS Chem. Biol.* **10,** 1340–1351 (2015).

21. Zhao, B. *et al.* Biosynthesis of the sesquiterpene antibiotic albaflavenone in Streptomyces coelicolor A3(2). *J. Biol. Chem.* **283,** 8183–8189 (2008).

22. Zhao, B. *et al.* Crystal Structure of Albaflavenone Monooxygenase Containing a Moonlighting Terpene Synthase Active Site. *J. Biol. Chem.* **284,** 36711–36719 (2009).

23. Lin, X., Hopson, R. & Cane, D. E. Genome mining in Streptomyces coelicolor: molecular cloning and characterization of a new sesquiterpene synthase. *J. Am. Chem. Soc.* **128,** 6022–6023 (2006).

24. Aaron, J. A., Lin, X., Cane, D. E. & Christianson, D. W. Structure of epi-isozizaene synthase from Streptomyces coelicolor A3(2), a platform for new terpenoid cyclization templates. *Biochemistry (Mosc.)* **49,** 1787–1797 (2010).

25. Li, R. *et al.* Reprogramming the chemodiversity of terpenoid cyclization by remolding the active site contour of epi-isozizaene synthase. *Biochemistry (Mosc.)* **53,** 1155–1168 (2014).

26. Bokhove, M. *et al.* Structures of an Isopenicillin N Converting Ntn-Hydrolase Reveal Different Catalytic Roles for the Active Site Residues of Precursor and Mature Enzyme. *Structure* **18,** 301–308 (2010).

27. Bento, A. P. *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **42,** D1083–D1090 (2014).

28. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* **42,** D1091–1097 (2014).

29. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

30. Long, A. J. *et al.* Structural Studies on the Reaction of Isopenicillin N Synthase with the Truncated Substrate Analogues δ-(l-α-aminoadipoyl)-l-cysteinyl-glycine and δ-(l-α-aminoadipoyl)-l-cysteinyl-d-alanine†,‡. *Biochemistry (Mosc.)* **44,** 6619–6628 (2005).

31. Fisher, J. F., Meroueh, S. O. & Mobashery, S. Bacterial Resistance to β-Lactam Antibiotics: Compelling Opportunism, Compelling Opportunity. *Chem. Rev.* **105,** 395–424 (2005).

32. Stewart, A. C., Clifton, I. J., Adlington, R. M., Baldwin, J. E. & Rutledge, P. J. A Cyclobutanone Analogue Mimics Penicillin in Binding to Isopenicillin N Synthase. *ChemBioChem* **8,** 2003–2007 (2007).

33. Elkins, J. M. *et al.* Crystallographic studies on the reaction of isopenicillin N synthase with an unsaturated substrate analogue. *Org. Biomol. Chem.* **1,** 1455–1460 (2003).

34. Kumarasamy, K. K. *et al.* Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: a molecular, biological, and epidemiological study. *Lancet Infect. Dis.* **10,** 597–602 (2010).

35. King, D. T., Worrall, L. J., Gruninger, R. & Strynadka, N. C. J. New Delhi Metallo-β-Lactamase: Structural Insights into β-Lactam Recognition and Inhibition. *J. Am. Chem. Soc.* **134,** 11362–11365 (2012).

ANNEX 5

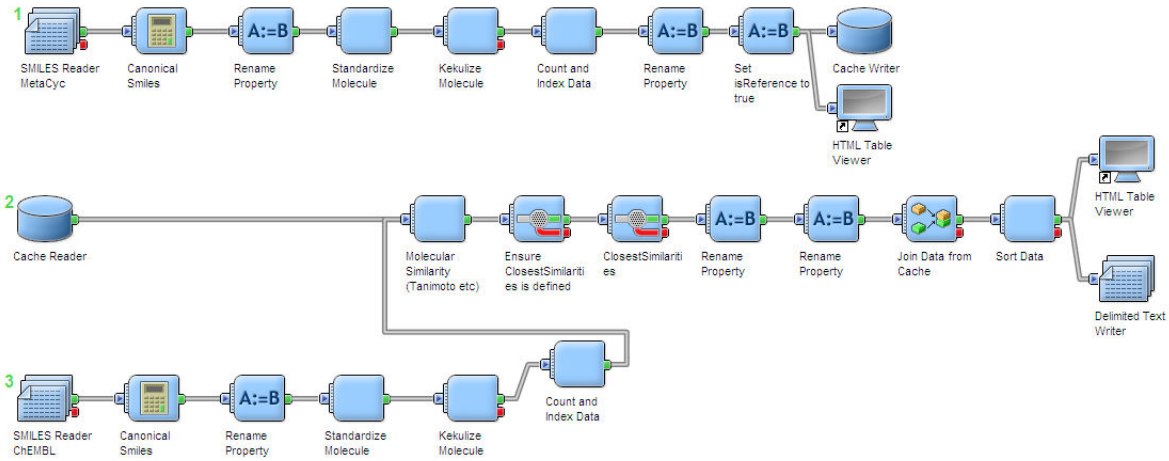


Figure 10. Pipeline pilot protocol for natural products known target search.

1/ The pipeline starts by reading natural product substrates. The structures are processed for standardization and assigned as references. 2/ References are stored in cache memory and given as input of the molecular similarity component. 3/ Structures previously extracted from ChEMBL are read and processed in the same way as the reference molecules. Known targets of ChEMBL molecules are assigned a compounds. The two sets of molecules are compared to each other. A filter discards all molecules that did not match any reference molecule. Before output, duplicate the results are merged together by keeping all targets.

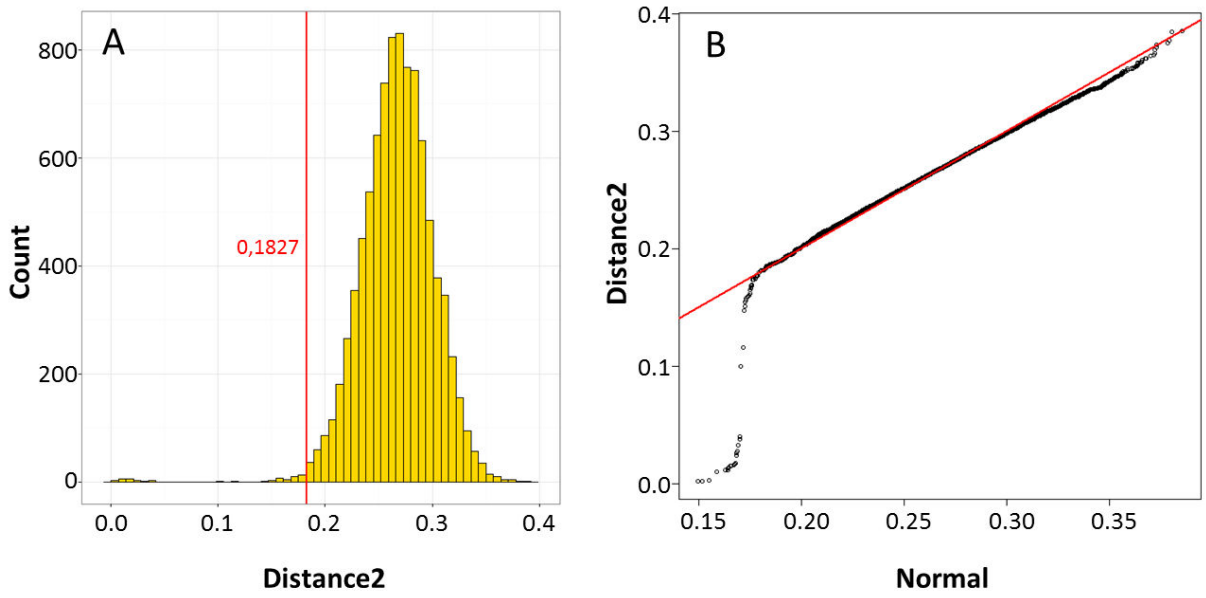


Figure 11. A: distribution of D2 score. Red line represents the cutoff over which binding sites are considered similar. B: quantile-quantile plot of D2 score distribution. The x-axis represents percentiles of a normal distribution (generated by rnorm in R) scaled to D2 distribution. The y-axis represents percentiles of D2 distribution.

1W05			4HL2	
Distance to center	Residue	Specific score	Residue	Distance to center
17	**His214**	**0.125**	**His122**	**14**
16	**Ser183**	**0.2042**	**His250**	**12**
18	Phe211	0.3750	Asp223	18
20	Leu223	0.3208	Gly207	18
22	Val272	0.1208	Thr190	18
22	Asn287	0.4167	Ala74	22
27	**Gln225**	**0.2000**	**Cys208**	**9**
14	**Val217**	**0.0917**	**Leu65**	**21**
15	**Asp216**	**0.0083**	**Asp124**	**13**
13	Thr331	0.3083	Leu221	17
19	Tyr91	0.5208	Gly219	9
15	**Leu321**	**0.0458**	**Met67**	**21**
26	Thr123	0.2083	Ile35	16
22	**Leu317**	**0.0833**	**Val73**	**17**
22	**Gln330**	**0.0583**	**Asn220**	**8**
21	**Gly329**	**0.0125**	**Gly222**	**18**
24	**His270**	**0.0167**	**His120**	**20**
26	**Tyr189**	**0.1167**	**His189**	**13**
11	Val185	0.3167	Lys211	17

Table 1. Residue overlaps onto SiteAlign polyhedrons.
Distance to center are expressed by the number of bins (interval of 0.5Å) separating their Cβ to the center of the compared polyhedrons. Residues from the superimposed catalytic centers are highlighted in yellow. Residues from the superimposed hydrophobic region are highlighted in grey. Bold residues represent overlapped triangles of the superimposed polyhedrons with a local distance score value less or equal to 0.2.

Table 2. List of biosynthetic enzymes from the considered dataset (next page).
Protein names are UniProt recommended names. If multiple names were recommended, we selected the first one only.

UniProt ID	Protein Name	Species
TPSD1_ABIGR	Alpha-bisabolene synthase	*Abies grandis*
CEFG_ACRCH	Acetyl-CoA--deacetylcephalosporin C acetyltransferase	*Acremonium chrysogenum*
SQHC_ALIAD	Squalene--hopene cyclase	*Alicyclobacillus acidocaldarius subsp. acidocaldarius*
RIFK_AMYMS	3-amino-5-hydroxybenzoate synthase	*Amycolatopsis mediterranei*
C5B3_AMYOR	Cytochrome P450 165B3	*Amycolatopsis orientalis*
C5C4_AMYOR	Cytochrome P450 165C4	*Amycolatopsis orientalis*
GRSA_ANEMI	Gramicidin S synthase 1	*Aneurinibacillus migulanus (Bacillus migulanus).*
ISPE_AQUAE	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	*Aquifex aeolicus*
KSA_ARATH	Ent-copalyl diphosphate synthase, chloroplastic	*Arabidopsis thaliana*
LDOX_ARATH	Leucoanthocyanidin dioxygenase	*Arabidopsis thaliana*
OPR3_ARATH	12-oxophytodienoate reductase 3	*Arabidopsis thaliana*
AKRC9_ARATH	Aldo-keto reductase family 4 member C9	*Arabidopsis thaliana*
THIK2_ARATH	3-ketoacyl-CoA thiolase 2, peroxisomal	*Arabidopsis thaliana*
LOVD_ASPTE	Acyltransferase LovD	*Aspergillus terreus.*
BTRK_BACCI	L-glutamyl-[BtrI acyl-carrier protein] decarboxylase	*Bacillus circulans*
GLDSA_BACCI	L-glutamine:2-deoxy-scyllo-inososе aminotransferase	*Bacillus circulans*
DOIS_BACCI	2-deoxy-scyllo-inosose synthase	*Bacillus circulans.*
MRSD_BACSY	Mersacidin decarboxylase	*Bacillus sp.*
BACB_BACSU	Bacilysin biosynthesis protein BacB	*Bacillus subtilis*
YWFH_BACSU	Bacilysin biosynthesis oxidoreductase YwfH	*Bacillus subtilis*
CYPX_BACSU	Pulcherriminic acid synthase	*Bacillus subtilis (strain 168).*
THCAS_CANSA	Tetrahydrocannabinolic acid synthase	*Cannabis sativa*
IEMT_CLABR	(Iso)eugenol O-methyltransferase	*Clarkia breweri*
CURS1_CURLO	Curcumin synthase 1	*Curcuma longa*
TRN1_DATST	Tropinone reductase 1	*Datura stramonium*
TRN2_DATST	Tropinone reductase 2	*Datura stramonium*
5BPOR_DIGLA	3-oxo-Delta(4,5)-steroid 5-beta-reductase	*Digitalis lanata*

UniProt ID	Protein Name	Species
IPNS_EMENI	Isopenicillin N synthase	*Emericella nidulans*
HMGCS_ENTFL	Hydroxymethylglutaryl-CoA synthase	*Enterococcus faecalis*
ILVC_ECOLI	Ketol-acid reductoisomerase	*Escherichia coli*
ISPE_ECOLI	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	*Escherichia coli*
IDI_ECOLI	Isopentenyl-diphosphate Delta-isomerase	*Escherichia coli*
QUED_ECOLI	6-carboxy-5,6,7,8-tetrahydropterin synthase	*Escherichia coli*
MTNN_ECOLI	5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase	*Escherichia coli*
FPPS_CHICK	Farnesyl pyrophosphate synthase	*Gallus gallus*
DCS1_GOSAR	(+)-delta-cadinene synthase isozyme XC1	*Gossypium arboreum*
CP51A_HUMAN	Lanosterol 14-alpha demethylase	*Homo sapiens*
DHB1_HUMAN	Estradiol 17-beta-dehydrogenase 1	*Homo sapiens*
DHB8_HUMAN	Estradiol 17-beta-dehydrogenase 8	*Homo sapiens*
FPPS_HUMAN	Farnesyl pyrophosphate synthase	*Homo sapiens*
HMCS1_HUMAN	Hydroxymethylglutaryl-CoA synthase, cytoplasmic	*Homo sapiens*
HMDH_HUMAN	3-hydroxy-3-methylglutaryl-coenzyme A reductase	*Homo sapiens*
IMPA2_HUMAN	Inositol monophosphatase 2	*Homo sapiens*
THIL_HUMAN	Acetyl-CoA acetyltransferase, mitochondrial	*Homo sapiens*
IMPA1_HUMAN	Inositol monophosphatase 1	*Homo sapiens*
PTEN_HUMAN	Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN	*Homo sapiens*
IDI1_HUMAN	Isopentenyl-diphosphate Delta-isomerase 1	*Homo sapiens*
PI42A_HUMAN	Phosphatidylinositol 5-phosphate 4-kinase type-2 alpha	*Homo sapiens*
PLCB2_HUMAN	1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase beta-2	*Homo sapiens*
ERG7_HUMAN	Lanosterol synthase	*Homo sapiens*
NISP_LACLL	Nisin leader peptide-processing serine protease NisP	*Lactococcus lactis subsp. lactis*
REBH_NOCAE	Flavin-dependent tryptophan halogenase RebH	*Lechevalieria aerocolonigenes*
CFI1_MEDSA	Chalcone--flavonone isomerase 1	*Medicago sativa*

UniProt ID	Protein Name	Species
CHS2_MEDSA	Chalcone synthase 2	*Medicago sativa*
COMT1_MEDSA	Caffeic acid 3-O-methyltransferase	*Medicago sativa*
7OMT8_MEDSA	Isoflavone-7-O-methyltransferase 8	*Medicago sativa*
CHOMT_MEDSA	Isoliquiritigenin 2'-O-methyltransferase	*Medicago sativa*
I4OMT_MEDTR	Isoflavone 4'-O-methyltransferase	*Medicago truncatula*
BSUHB_METJA	Fructose-1,6-bisphosphatase/inositol-1-monophosphatase	*Methanocaldococcus jannaschii*
MVK_METJA	Mevalonate kinase	*Methanocaldococcus jannaschii*
MYCF_MICGR	Mycinamicin III 3''-O-methyltransferase	*Micromonospora griseorubida*
MYCE_MICGR	Mycinamicin VI 2''-O-methyltransferase	*Micromonospora griseorubida.*
CMAS1_MYCTU	Cyclopropane mycolic acid synthase 1	*Mycobacterium tuberculosis*
CMAS2_MYCTU	Cyclopropane mycolic acid synthase 2	*Mycobacterium tuberculosis*
CMAS3_MYCTU	Cyclopropane mycolic acid synthase 3	*Mycobacterium tuberculosis*
ISPE_MYCTU	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	*Mycobacterium tuberculosis*
CP51_MYCTU	Lanosterol 14-alpha demethylase	*Mycobacterium tuberculosis*
HSAD_MYCTU	4,5:9,10-diseco-3-hydroxy-5,9,17-trioxoandrosta-1(10),2-diene-4-oate hydrolase	*Mycobacterium tuberculosis*
MMAA2_MYCTU	Cyclopropane mycolic acid synthase MmaA2	*Mycobacterium tuberculosis*
MMAA4_MYCTU	Hydroxymycolate synthase MmaA4	*Mycobacterium tuberculosis*
OYE3_ASPFU	Chanoclavine-I aldehyde reductase	*Neosartorya fumigata*
5EAS_TOBAC	5-epi-aristolochene synthase	*Nicotiana tabacum*
CUS_ORYSJ	Bisdemethoxycurcumin synthase	*Oryza sativa subsp. japonica*
AAAA_PENCH	Acyl-coenzyme A:6-aminopenicillanic-acid-acyltransferase 40 kDa form	*Penicillium chrysogenum*
ARIS_PENRO	Aristolochene synthase	*Penicillium roqueforti.*
PAL1_PETCR	Phenylalanine ammonia-lyase 1	*Petroselinum crispum*
DPSS_PINSY	Dihydropinosylvin synthase	*Pinus sylvestris*
SILD_PODPE	Secoisolariciresinol dehydrogenase	*Podophyllum peltatum*
PQSD_PSEAE	2-heptyl-4(1H)-quinolone synthase PqsD	*Pseudomonas aeruginosa*

UniProt ID	Protein Name	Species
RHLG_PSEAE	Rhamnolipids biosynthesis 3-oxoacyl-[acyl-carrier-protein] reductase	*Pseudomonas aeruginosa*
PHZD_PSEFL	Probable isochorismatase	*Pseudomonas fluorescens*
PHZF_PSEFL	Trans-2,3-dihydro-3-hydroxyanthranilate isomerase	*Pseudomonas fluorescens.*
PHZG_PSEFL	Phenazine biosynthesis protein PhzG	*Pseudomonas fluorescens.*
PRNA_PSEFL	Flavin-dependent tryptophan halogenase PrnA	*Pseudomonas fluorescens.*
KIME_RAT	Mevalonate kinase	*Rattus norvegicus*
RG1_RAUSE	Raucaffricine-O-beta-D-glucosidase	*Rauvolfia serpentina*
PERR_RAUSE	Perakine reductase	*Rauvolfia serpentina*
SG1_RAUSE	Strictosidine-O-beta-D-glucosidase	*Rauvolfia serpentina*
PNAE_RAUSE	Polyneuridine-aldehyde esterase	*Rauvolfia serpentina*
BAS_RHEPA	Polyketide synthase BAS	*Rheum palmatum*
PDC1_YEAST	Pyruvate decarboxylase isozyme 1	*Saccharomyces cerevisiae*
CPXJ_SACEN	6-deoxyerythronolide B hydroxylase	*Saccharopolyspora erythraea*
ERYK_SACEN	Erythromycin C-12 hydroxylase	*Saccharopolyspora erythraea*
BPPS_SALOF	(+)-bornyl diphosphate synthase, chloroplastic	*Salvia officinalis*
C167_SORCE	Cytochrome P450 167A1	*Sorangium cellulosum*
TOBZ_STRSD	nebramycin 5' synthase	*Streptoalloteichus tenebrarius*
CHMJ_STRBI	dTDP-4-dehydro-6-deoxyglucose 3-epimerase	*Streptomyces bikiniensis*
NCSB1_STRCZ	2,7-dihydroxy-5-methyl-1-naphthoate 7-O-methyltransferase	*Streptomyces carzinostaticus.*
CAS1_STRC2	Clavaminate synthase 1	*Streptomyces clavuligerus*
CEFE_STRC2	Deacetoxycephalosporin C synthase	*Streptomyces clavuligerus*
BLS_STRCL	Carboxyethyl-arginine beta-lactam-synthase	*Streptomyces clavuligerus.*
GNAT2_STRCL	Glutamate N-acetyltransferase 2	*Streptomyces clavuligerus.*
PAH_STRCL	Proclavaminate amidinohydrolase	*Streptomyces clavuligerus.*
ACT3_STRCO	Putative ketoacyl reductase	*Streptomyces coelicolor*
CYC1_STRCO	Epi-isozizaene synthase	*Streptomyces coelicolor*
RPPA_STRCO	1,3,6,8-tetrahydroxynaphthalene synthase	*Streptomyces coelicolor*

UniProt ID	Protein Name	Species
Q7DC80_PSEAE	1,3,6,8-tetrahydroxynaphthalene synthase	*Streptomyces coelicolor*
EIZFM_STRCO	Epi-isozizaene 5-monooxygenase/(E)-beta-farnesene synthase	*Streptomyces coelicolor.*
PENA_STREX	Pentalenene synthase	*Streptomyces exfoliatus*
STRB1_STRGR	Inosamine-phosphate amidinotransferase 1	*Streptomyces griseus.*
DCSE_STRLA	L-serine/homoserine O-acetyltransferase	*Streptomyces lavendulae.*
DNRK_STRPE	Carminomycin 4-O-methyltransferase DnrK	*Streptomyces peucetius.*
RDMC_STREF	Aclacinomycin methylesterase RdmC	*Streptomyces purpurascens.*
TASY_TAXBR	Taxadiene synthase	*Taxus brevifolia*
PAM_TAXCA	Phenylalanine aminomutase (L-beta-phenylalanine forming)	*Taxus canadensis*
NCS_THLFG	S-norcoclaurine synthase	*Thalictrum flavum subsp. glaucum*
BSUHB_THEMA	Fructose-1,6-bisphosphatase/inositol-1-monophosphatase	*Thermotoga maritima*
ISPE_THET8	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	*Thermus thermophilus*
PILR1_THUPL	Bifunctional pinoresinol-lariciresinol reductase 1	*Thuja plicata*
CP51_TRYCC	Sterol 14-alpha demethylase	*Trypanosoma cruzi*
LUXS_BACSU	S-ribosylhomocysteine lyase	*Undef_OS*
CQSA_VIBCH	CAI-1 autoinducer synthase	*Vibrio cholerae serotype O1*
MTNN_VIBCH	5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase	*Vibrio cholerae serotype O1*

Conclusions

In this thesis, we have investigated if binding site similarity could help pharmacognosy using biological imprints of natural products as baits to find target proteins. Through the whole project, we were able to address several questions.

We demonstrated that the modelling of ligand binding sites was a factor of most importance in binding site similarity screenings. We gambled on the thin line between highly detailed representations and simplistic representations of binding sites. The inherent problem to protein structures, their resolution, has told us that throwing off sensitivity to small conformational changes was the winning strategy.

From thereon, we tackled the “key-lock” model, underlying principle of binding site similarity approaches. We identified a significant number of drug-like molecules, including natural products that are very promiscuous, thereby raising some of the limits in solving the PFT hypothesis by binding site similarity.

Despite the difficulties, we were able to show that binding site similarity was not a hopeless strategy to prove the PFT hypothesis. We gave a proof of concept showing that even if flavonoids are promiscuous molecules, robust methods such as SiteAlign are able to identify structural features shared between flavonoid biosynthetic enzymes and kinase proteins. Nevertheless, more evidence was needed.

Therefore, we developed a method capable of searching all possible natural product biosynthetic enzymes in order to study structural relations between biosynthetic origins of natural products and their potent biological activities on a wider extend. We created a dataset of 117 natural product biosynthetic enzymes.

Last we made a diagnostic of biological imprints in our dataset of enzymes. We have seen that natural products rather interact with proteins by induced fit than following the “key-lock” model, thus pointing at a possible use of other methods to validate the PFT hypothesis. Nevertheless there is still a fraction of biosynthetic enzyme structures recognizing specifically at least pharmacological principles of natural products. For example, we could raise a structural resemblance in the biosynthesis of penicillins and the bacterial resistance against penicillins.

The content of this thesis is a strong ground for the discovery of new PFTs and adds a stone to the bigger question:

“*Why does Evolution create Natural Products?*”

Thanks for reading,

Noé Sturm.

Noé STURM

Caractérisation de l'empreinte biologique des produits naturels pour la conception rationnelle de médicament assistée par ordinateur

Résumé

La comparaison de site peut-elle vérifier l'hypothèse: «*Les origines biosynthétiques des produits naturels leurs confèrent des activités biologiques*»? Pour répondre à cette question, nous avons développé un outil modélisant les propriétés accessibles au solvant des sites de liaison. La méthode a montré des aspects intéressants, mais elle souffre d'une sensibilité aux coordonnées atomiques. Cependant, des méthodes existantes nous ont permis de prouver que l'hypothèse est valide pour la famille des flavonoïdes. Afin d'étendre l'étude, nous avons développé un procédé automatique capable de rechercher des structures d'enzymes de biosynthèse de produits naturels disposant de sites actifs capables de lier une molécule de petite taille. Nous avons trouvé les structures de 117 enzymes.

Les structures nous ont permis de caractériser divers modes de liaison substrat-enzyme, nous indiquant l'empreinte biologique des produits naturels ne correspond pas toujours au modèle « clé-serrure ».

Résumé en anglais

Can computational binding site similarity tools verify the hypothesis: *"Biosynthetic moldings give potent biological activities to natural products"*? To answer this question, we designed a tool modeling binding site properties according to solvent exposure. The method showed interesting characteristics but suffers from sensitivity to atomic coordinates.

However, existing methods have delivered evidence that the hypothesis was valid for the flavonoid chemical class. In order to extend the study, we designed an automated pipeline capable of searching natural products biosynthetic enzyme structures embedding ligandable catalytic sites. We collected structures of 117 biosynthetic enzymes. Finally, according to structural investigations of biosynthetic enzymes, we characterized diverse substrate-enzyme binding-modes, suggesting that natural product biological imprints usually do not agree with the "key-lock" model.