Thieme

Structural Searching of Biosynthetic Enzymes to Predict Protein Targets of Natural Products

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

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

Key words
structural similarity, protein binding site, biosynthetic enzymes, virtual screening, natural products

received July 6, 2017
revised October 18, 2017
accepted October 20, 2017

Bibliography
DOI https://doi.org/10.1055/s-0043-121992
Published online November 3, 2017 | Planta Med 2018; 84: 304–310 © Georg Thieme Verlag KG Stuttgart · New York | ISSN 0032-0943

Correspondence
Prof. 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 J. Quinn
Griffith Institute for Drug Discovery, Griffith University, Nathan Campus
Don Young Road, Brisbane, QLD 4111, Australia
Phone: + 61 7 37 35 60 06, Fax: + 61 7 37 35 60 01
r.quinn@griffith.edu.au

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

ABSTRACT

Recently, we have demonstrated that site comparison methodology using flavonoid biosynthetic enzymes as the query could automatically identify structural features common to different flavonoid-binding proteins, allowing for the identification of flavonoid targets such as protein kinases. With the aim of further validating the hypothesis that biosynthetic enzymes and therapeutic targets can contain a similar natural product imprint, we collected a set of 159 crystallographic structures representing 38 natural product biosynthetic enzymes by searching the Protein Databank. Each enzyme structure was used as a query to screen a repository of approximately 10 000 ligandable sites by active site similarity. We report a full analysis of the screening results and highlight three retrospective examples where the natural product validates the method, thereby revealing novel structural relationships between natural product biosynthetic enzymes and putative protein targets of the natural product. From a prospective perspective, our work provides a list of up to 64 potential novel targets for 25 well-characterized natural products.

Introduction

Natural products (NPs) constitute an important source of drugs and inspiration for drug discovery. Many marketed drugs have natural origins [1]. NP intrinsic properties result in privileged interaction with the biological world because each NP has met at least one biosynthetic enzyme, providing an embedded biological imprint. NPs often have many stereocenters giving them three-dimensional (3D) shapes, which are well recognized by protein active sites [2].

Considering that the biological imprint endows NPs with specific binding properties, the structure-based computing methods represent an attractive alternate to current strategies for the identification of NP targets [3]. The recent literature provides examples of NP target fishing by molecular docking, a computing method which is commonly used in a virtual screening approach to identify novel bioactive molecules from natural resources [4–

6]. Large-scale docking of approximately 197 k NPs in the Universal Natural Products Database (UNPD) to 332 FDA-approved drug targets hence contributed to the construction of an extensive NP-target network [6]. Recently, we suggested that true and potential targets of an NP could be identified by direct comparison of the NP biosynthetic enzyme 3D structure to the 3D structure of proteins in the Protein Databank (PDB). We established the proof of concept on the flavonoid example, showing that a virtual screening of 2379 PDB proteins by site similarity to the active site of five flavonoid biosynthetic enzymes successfully retrieved known flavonoid targets, such as kinases [7].

In this study, we collected and annotated publicly available 3D structures of NP biosynthetic enzymes. We compared them to protein structures in the sc-PDB, a database describing ligand-binding sites in the PDB [8]. We assessed similarities between the NP biosynthetic enzyme active site and the sc-PDB protein ligand-binding site. We further validated our method on three new retrospective examples, and suggested new potential targets for 25 NPs.

Results and Discussion

We explored the PDB to identify NP biosynthetic enzyme structures. Our strategy involved integration of data from the PDB [9], the Dictionary of Natural Products (http://dnp.chemnetbase.com), and the metabolic databases MetaCyc [10] or UniPathway [11]. These databases provide complementary information: the PDB is a repository of the experimental 3D structures of biomolecules; the Dictionary of Natural Products is the most comprehensive source of NPs with chemical and biological annotation from the literature on over 290 k NPs; MetaCyc and UniPathway describe biosynthetic pathways, including both the NP structure and name of genes coding for biosynthetic enzymes.

We employed two methods to identify biosynthetic enzymes in the PDB. Firstly, we searched for cross-references to the PDB in NP biosynthetic pathways of MetaCyc and UniPathway (see the selection of pathways in the Material and Methods section and in **Table S1**, Supporting Information). Secondly, we directly looked into the PDB and searched for keywords such as biosynthesis (see Materials and Methods for the entire list).

The two searches yielded a total of 9550 structure files, with 80% of the structures selected by the keyword search. We parsed and validated structure files using functional and structural annotation according to a six-step protocol (▶ **Table 1**). In the first step, we assigned a protein name, source organism, and catalytic reaction to 9140 structures. We used the annotation protocol of the sc-PDB [8], and added catalytic reactions found in UniProt [12]. In the second step, we identified the catalytic residues of half of the entries using data in UniProt or in the Catalytic Site Atlas [13]. In the two steps, we discarded entries with inconsistent or missing information, e.g., if no catalytic residue was reported in UniProt or in the Catalytic Site Atlas. Noteworthy, the catalytic residues reported in UniProt and the Catalytic Site Atlas were inferred from published experimental evidences. In the third step, we validated PDB entries selected by the keyword search whether the corresponding protein was encoded by a gene referenced in the NP biosynthetic pathways considered in this study, or by an ortho-

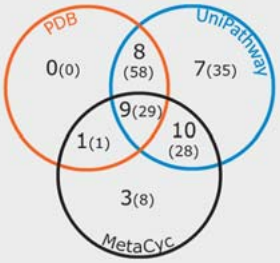
▶ **Table 1** Count of structures in the collection workflow.

	Number of PDB files		
	Using keywords	**From UniPathway**	**From MetaCyc**
Selection	7603	1035	907
Analysis and filtering			
1-Annotation using UniProt	7209	1026	905
2-Identification of catalytic residues	3384	449	318
3- Pathways filtering	1642	449	318
4- Active site detection	1118	335	238
5- Ligandability filter	760	280	169
6- Manual curation	88	150	66

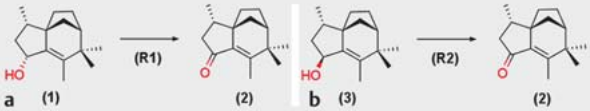
logue of a gene referenced in the NP biosynthetic pathways considered in this study (definition of related species is given in **Fig. S1**, Supporting Information). We found that 1642 PDB entries actually describe a NP biosynthetic enzyme. In the fourth step, we detected all cavities in every protein structure based on the geometrical description of the protein surface [14], and we checked whether one of the cavities was the actual enzyme active site. We verified that, in 70% of the tested structures, there is one cavity containing at least one of the residues tagged as catalytic. We discarded many entries because none of their cavities contained a catalytic residue. In the fifth step, we selected ligandable active sites, i.e., sites able to bind a drug-like ligand with high affinity, according to a prediction made from geometrical and pharmacophoric properties [14]. Ligandability is a prerequisite for accurate site comparison. About 70% of the tested structures described a ligandable active site. In the last step, we excluded the enzymes which metabolize early precursors structurally unrelated to the biosynthetic pathway end product. For example, isopentenyl diphosphate is an early precursor in the biosynthesis of many complex terpenes, such as albaflavenone, capsidiol, pentalenolactone, and lanosterol. Biosynthetic enzymes metabolizing isopentenyl diphosphate are not expected to share structural similarity with the protein targets of the complex end products. This manual filtering caused a large cut in the data set, discarding 90% of the structures selected by the keyword search.

▶ **Table 1** summarizes the number of PDB structures selected at each step. Searching cross-references to the PDB in MetaCyc and UniPathway finally yielded 66 and 150 PDB structures, respectively, 57 of them in common (▶ **Fig. 1**). The resource-specific structures represent 19 different enzymes, thus indicating that the two metabolic databases provide complementary information. The direct search in the PDB yielded 88 structures, but all were already referenced in MetaCyc and UniPathway.

In total, we identified 38 different biosynthetic enzymes, including 18 bacterial enzymes, 17 plant enzymes, and three fungal enzymes. Several enzymes have more than one structure in the PDB, and therefore our data set contains more than 38 crystallographic structures, 159 in total. The number of structures of each



▶ **Fig. 1** Number of NP biosynthetic enzymes found in the Protein DataBank using the different search approaches. Counts of the PDB structures are given below counts of proteins.



▶ **Fig. 2** Two different reactions for the synthesis of the same product. In our data set, one reaction product represents multiple enzymatic reactions. Reactions **R1** and **R2**, catalyzed by one monooxygenase from *Streptomyces coelicolor*, are represented by albaflavenone (**2**). **a** Last step of the biosynthesis of albaflavenone from (5R)-albaflavenol. **b** Last step of the biosynthesis of albaflavenone from (5S)-albaflavenol.

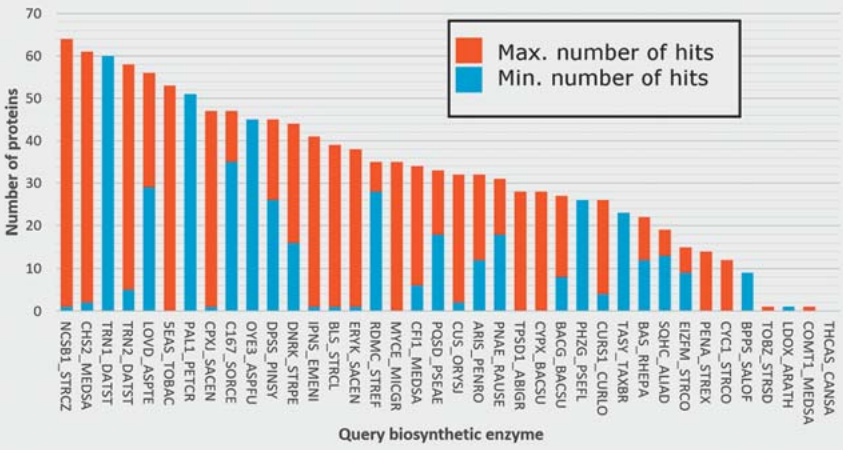
enzyme is given in the Supporting Information (see the excel file named 10-1055-s-0043-121992-BiosyntheticEnzyme_Shaper_hitlits.xls). Three enzymes have 10 or more structures. The most represented enzyme is isopenicillin N synthase (IPNS) from *Emericella nidulans* with 37 crystallographic structures. In the PDB structures, IPNS is always in a complex with either the natural substrate, a substrate analogue, the NP penicillin, or analogues of penicillin [15]. The second most represented enzyme is 5-epi-aristolochene synthase from *Nicotiana tobacum* with 11 crystallographic structures. Two PDB structures describe the enzyme alone, and the nine others show a complex between the enzyme and an analogue of the NP [16]. The third most represented enzyme is nebramycin synthase from *Streptoalloteichus tenebrarius* with 10 crystallographic structures of the enzyme in a complex with either the substrate, a cofactor, or an analogue of the cofactor. Considering the 38 enzymes in the entire data set, only 5% of 159 structures describe the enzyme alone. In most of the cases, a non-natural ligand is bound in the enzyme active site, often for the purpose of studying the enzymatic mechanism.

The 38 selected enzymes cover diverse chemical reactions. The six EC classes are represented: transferases (34%), oxydo-reductases (28%), lyases (14%), hydrolases (19%), isomerases (11%), and ligases (4%). The enzymes in the data set are involved in the biosynthesis of antibiotics, terpenes, steroids, phenylpropanoids, alkaloids, polyketides, and pigments. The complete list of biosynthetic enzymes and pathways is given in **Table S1**, Supporting Information. From the UniProt annotation, we assigned 49 generic chemical reactions to the 38 enzymes. Several enzymes recognize multiple substrates and, thus, were assigned multiple reactions. The 49 reactions described the formation of only 25 natural products (see their name and SMILES code in **Table S2**, Supporting Information). An example of an enzyme able to metabolize two substrates to form a single product is shown in ▶ **Fig. 2**: the monooxygenase (UniProt ID: EIZFM_STRCO) is able to oxidize both (5R)-albaflavenol (compound **1**) and its stereoisomer (5S)-albaflavenol (compound **3**) to produce albaflavenone (compound **2**). As a consequence, it was associated with two reactions (reaction labeled R1 and R2 in ▶ **Fig. 2**) and a single NP (compound **2**).

To summarize, we gathered and annotated 159 crystallographic structures of 38 NP biosynthetic enzymes. Every structure contains a ligandable and well-delimited cavity encompassing part or all of the active site, and thus constitutes a suitable structural reference for site comparison.

We screened the sc-PDB, which contains 9283 ligandable 3D-binding sites, by site similarity to each of the 159 NP biosynthetic enzyme structures. An all-against-all comparison, which represents about 1.47 million 3D structure pairs, was performed using the program Shaper [14]. Shaper 3D aligns two sites by optimizing the shape overlap of grid points filling the cavity defined by site residues. Shaper scores the similarity by comparing the number, position, and pharmacophoric type (as defined from the properties of surrounding protein atoms) of the aligned grid points. In the virtual screening experiments, hits were selected if their similarity score differs by more than 2.5 standard deviations from the mean value of the distribution of scores. We previously demonstrated that this scoring scheme efficiently recovers flavonoid target proteins when sc-PDB is compared to flavonoid biosynthetic enzymes [7]. All of the 38 NP biosynthetic enzymes, except one, shared structural similarity with up to 64 sc-PDB proteins (▶ **Fig. 3**). If several crystallographic structures of the same enzyme were used as a query, the number of hits can vary largely. For instance, acetyltransferase LovD (Uniprot ID: LOVD_ASPTE) is represented by six crystallographic structures (PDB IDs: 3HLC, 3HLD, 3HLE, 3HLF, 3HLG, 4LCL) and an sc-PDB screening using these six structures retrieved from 29 to 56 proteins. The largest and smallest hit lists correspond to query active sites of different sizes. Typically, more hits were found if a smaller definition of the active site was used: hit lists of 29 and 56 proteins were obtained using query sites comprising 31 and 25 residues, respectively. Virtual screenings of the sc-PDB showed no hits for tetrahydrocannabinolic acid synthase (THC) as a query because its active cavity is much larger than the average sc-PDB sites.

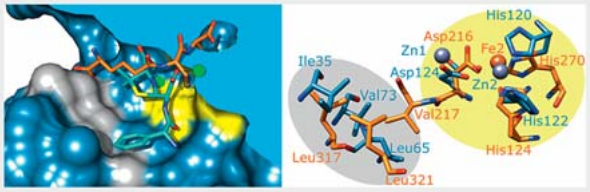
Virtual screenings using flavonoid biosynthetic enzymes revealed structural relationships with the flavonoid targets in the kinase family, as published previously [7]. We then searched ChEMBL [17], DrugBank [18], and the PDB for affinity data between the 25 identified NPs (exact match of structure) and all the sc-PDB proteins. Unfortunately, none of the NPs were already described as a ligand for any of the hit proteins. We thus analyzed the hit lists manually to propose new pairs of NP biosynthetic enzymes and ligandable proteins. All the screening results are available in Supporting Information (one spreadsheet per screening is given in 10-1055-s-0043-121992-BiosyntheticEnzyme_Shaper_hitlits.xls).



▶ **Fig. 3** Comparing ligandable protein sites with the active site of NP biosynthetic enzymes. The number of hits in sc-PDB is given for all 38 enzymes in the data set, with maximum and minimum values obtained for the different structure files of an enzyme. The x-axis shows UniProt identifiers of the query enzymes. Enzymes Uniprot identifiers and names are given in **Table S1**, Supporting Information.

We repeated site comparisons on several interesting cases using SiteAlign [19], which is computationally more expensive than Shaper, but gives more effective scores if the query and compared sites have very different sizes. Virtual screening of the sc-PDB by site similarity to NP biosynthetic enzyme paired three NPs to one of their target proteins: penicillin G to the New-Delhi Metallo β-lactamase (NDM-1), Δ^9-tetrahydrocannabinol to acetylcholinesterase (AChE), and lovastatin to histone deacetylases (HDAC). In the three cases, the 3D alignment by site similarity programs of the biosynthetic enzyme active site with the ligand-binding site of the NP target enzyme overlaid key determinants for the NP binding. Noteworthy, the three considered NP biosynthetic enzymes, IPNS, thetrahydrocannabinolic acid synthase (THCA) and acyltransferase lovD, are involved in a late stage of the NP biosynthetic pathway and recognize a large part of the substrate.

IPNS is the most represented enzyme in our data set. Its 3D structure has been solved in a complex with many ligands mimicking the endogenous substrate, N-((5S)-5-amino-5-carboxypentanoyl)-L-cysteinyl-D-valine, or the product, penicillin G. The enzymatic mechanism is well known, and the catalytic site is unambiguously defined. The virtual screening of the sc-PDB using IPNS as a query retrieved from 0 to 40 hits, including the NDM-1. Like all β-lactamases, NDM-1 confers its bacterial source organism resistance to β-lactam antibiotics such as penicillin. Both IPNS and NDM-1 act on a lactam ring by catalyzing ring closure and opening, respectively. The 3D-alignment by site similarity of the two active sites superimposed the metallic cofactors and also the catalytic residues. The His214-His270-Asp216 triad in IPNS matched the His122-His120-Asp124 triad in NDM-1, as highlighted in yellow in ▶ **Fig. 4**. The 3D-alignment also revealed a common hydrophobic patch required for the correct positioning of the substrate in the vicinity of the catalytic center. The residues Leu317, Leu321, and Val217 in IPNS were found to be equivalent to Ile35, Val73, and Leu65 in NDM-1, as highlighted in grey in ▶ **Fig. 4**. Noteworthy, the two proteins do not share sequence similarity (13% identity over 376 residues). Moreover, they have different



▶ **Fig. 4** Local similarity between isopenicillin N synthase and β-lactamases. The isopenicillin N synthase of *Emericella nidulans* in a complex with a substrate analogue is shown in blue (PDB ID: 1W05). The New-Dehli Metallo β-lactamase of *Klebsiella pneumoniae* in a complex with hydrolyzed ampicillin is shown in orange (PDB ID: 4HL2).

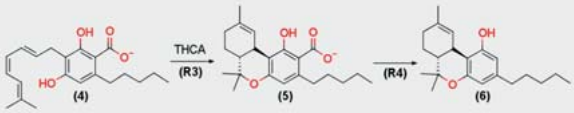
folds. Nevertheless, the 3D-alignment of the two catalytic sites superimposed the recognition sites of the lactam ring. Lastly, virtual screening did not identify any of the eight other β-lactamases present in the sc-PDB because either they differ in type (metallo lactamases vs. serine lactamases) [20] or their sc-PDB site did not encompass the catalytic cavity. Noteworthy, site similarity methods have been designed to capture common ligand-binding features and, consequently, can only pair two proteins that bind the same ligand if the ligand-binding mode is conserved [21].

THCA catalyzes ring closure in cannabigerolate (compound **4**) to produce the Δ^9-tetrahydrocannabinolate (compound **5**), which then undergoes a spontaneous decarboxylation leading to the end product of the biosynthetic pathway, Δ^9-tetrahydrocannabinol (THC or compound **6**; ▶ **Fig. 5**). THC is a psychoactive drug known to bind to synaptic receptors, primarily CB1 and CB2 receptors [22], but also other hippocampal proteins involved in signal transduction mediated by acetylcholine, such as AChE [23]. The single copy of THCA in our data set represents a ligand-free enzyme, yet bound to the flavine-adenine dinucleotide cofactor. As previously mentioned, virtual screening of the sc-PDB using

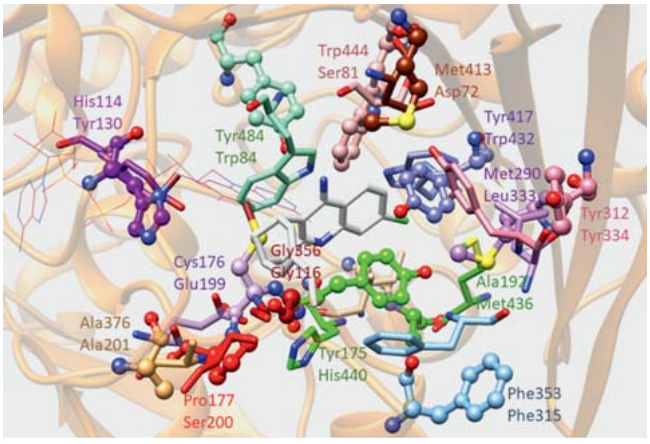
the program Shaper and THCA as a query retrieved no hit. A second screening using the program SiteAlign yielded 41 hits, including AChE of *Homo sapiens* and of *Tetronarce californica*, and a second enzyme binding acetylcholine, namely, human choline esterase. The 3D-alignment of THCA and AChE superimposed 20 residues, in particular the aromatic cage recognizing the ligand. More precisely, Tyr175-Tyr312-Phe353-Tyr417-His114-Tyr484 in THCA matched His440-Tyr334-Phe315-Trp432-Tyr130-Trp84 in AChE (▶ **Fig. 6**).

Acyltransferase LovD is involved in the synthesis of lovastatin. It catalyzes the formation of lovastatin from monacolin J and lovF-bound 2-methylbutyrate [24]. Lovastatin belongs to the therapeutic class of statin drugs that are indicated to lower circulating cholesterol level [25]. Lovastatin inhibits cholesterol biosynthesis by competing with the substrate of 3-hydroxy-3-methylglutaryl-coenzyme A reductase (HMG-CoA reductase). Lovastatin also inhibits HDAC2, which, like other HDACs, plays an important role in the regulation of DNA transcription [26, 27]. HDACs constitute an important class of therapeutic targets for neurodegenerative diseases such as Alzheimer's disease and schizophrenia [28, 29]. Virtual screening of the sc-PDB using the program Shaper and acyltransferase LovD as a query retrieved several dozen hits, including HDAC8 but not HMG-CoA reductase. A second round of screening using the program SiteAlign confirmed this result, ranking human HDAC8 as first in the hit list obtained using the PDB entry 3HLC as a query. HDAC8 is a homolog of HDAC2, which is not present in the sc-PDB. The two enzymes have the same biological activity and similar sequences (overall identity = 30 %). Their active site is highly conserved, with 16 identical residues among the 24 lining the catalytic cavity, as observed in the PDB files 4LXZ for HDAC2 and 3SFF for HDAC8. In addition, the 3D-structure of the active site is well conserved: the RMSD computed on 20 pairs of equivalent Cα atoms is equal to 1 Å (as computed using the Chimera matchmaker [30] sequence-based alignment of 4LXZ and 3SFF PDB proteins). The 3D-alignment by site similarity using SiteAlign of acyltransferase LovD and HADC8 overlaid the catalytic centers of the two enzymes (▶ **Fig. 7**) and the networks of polar side chains involved in their catalytic mechanism. Interestingly, both acyltransferase LovD and HDAC2/HDCA8 catalyze the transfer of an acyl group. Acyltransferase LovD adds an acetyl group to the substrate, monacolin J acid, whereas HDAC2 and HDAC8 remove an acetyl group from an acetyl-lysine residue of a histone protein.

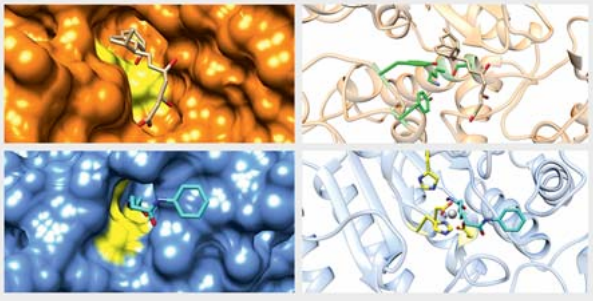
These three retrospective examples constitute additional validation that a NP biosynthetic enzyme structure is highly valuable data to identify potential targets of NPs using the site comparison method. Moreover, we constituted a set of 38 biosynthetic enzymes that fully met the requirements to apply the method. Virtual screening of all ligandable sites in the PDB yielded 1 to 64 protein hits per query enzyme. Thus, from a prospective perspective, our work provides a list of up to 64 potential novel targets for 25 well-characterized NPs. All of the results are available in Supporting Information. Validation of a new hypothesis could stir up the structural study of more NP biosynthetic enzymes, and thus increase the scope, which is still limited, of our approach for the identification of NP protein targets.



▶ **Fig. 5** Synthesis of Δ^9-tetrahydrocannabinol.



▶ **Fig. 6** Local similarity between tetrahydrocannabinolic acid synthase and acetylcholinesterase. The backbone of tetrahydrocannabinolic acid synthase of *Cannabis sativa* is shown in orange (PDB ID: 3VTE). Residues in the active site are represented with colored ball and sticks. The FAD cofactor is represented with wires. Residues in the binding site of acetylcholinesterase of *Tetronarce californica* are represented as colored sticks (PDB ID: 1E66). The bound ligand, (−)-huprine, is represented with CPK-colored sticks.



▶ **Fig. 7** Local similarity between acyltransferase LovD and histone deacetylase 2. Acyltransferase LovD of *Aspergillus terrus* in a complex with monacolin J (PDB ID: 3HLD, top) and histone deacetylase 2 of *Homo sapiens* in a complex with the inhibitor vorinostat (PDB ID: 4LXZ, bottom) are shown in the orientation after SiteAlign superposition. Pictures on the left side show the protein surface, with catalytic residues highlighted in yellow. Pictures on the right side show the proteins backbone as ribbons and catalytic residues as sticks.

Material and Methods

Collecting the NP biosynthetic enzyme structures in the PDB. Since 2015, the UniProt database [12] has provided metabolic information by integrating the Chemical Entities of Biological Interest ontology known as ChEBI [31], a description of expert-curated biochemical reactions, namely Rhea [32], and a hierarchy of pathways as given in UniPathway [11]. In UniPathway, pathways (e.g., erythromycin biosynthesis) are described by a set of reactions annotated with the enzyme name, and grouped into families (e.g., antibiotic biosynthesis). The UniProt pathway index file (release 2015_08) establishes association links between enzymes and pathway families. Pathway families representing the biosynthesis of NPs were selected manually (**Table S1**, Supporting Information).

The MetaCyc database [10] describes metabolic pathways with a detailed cascade of enzymatic reactions, and also the gene and source organism of metabolic enzymes. Arrows in reaction schemes connect the metabolite structures from the initial substrates to the final products. Pathways are organized according to a hierarchical classification. All the pathways in the "secondary metabolite biosynthesis" branch of the tree were considered in this study. The gene names in selected pathways were obtained using the BioCyc REST-based web service. They were then related to their corresponding enzyme names using the MetaCyc dictionary. The link between the enzyme name and 3D-structure was established using the PDB-UniProt dictionary provided by the SIFT initiative [33].

The full PDB [9] was also directly scanned searching for the keywords "biosynth*", "natural product*", "secondary metabol*", and "plant defense*" in the KEYWDS, TITLE, and JRNL lines of the header section of the PDB files, and also in the abstract of primary articles cited. Selected files were filtered if neither the described enzyme nor a related enzyme (with identical gene name and related species as defined in **Fig. S1**, Supporting Information) could be found in NP biosynthetic pathways selected in either UniProt or in MetaCyc.

Annotating NP biosynthetic enzyme structures. The 3D-structures were annotated with the enzyme protein and gene names, as recommended in UniProt, following the sc-PDB annotation protocol [8]. We verified that EC was consistent with the protein name. We searched UniProt files for the *ACT_SITE* field to define a list of catalytic residue numbers. If the *ACT_SITE* field was not present, catalytic residues were searched in the Catalytic Site Atlas [13]. If no catalytic residue was identified at all, then the entry was discarded. Alignment between PDB and UniProt sequences allowed for the mapping of catalytic residues to the enzyme structure. The chemical structure of substrates and products were retrieved from ChEBI.

Processing NP biosynthetic enzyme structures. Structure files were downloaded from the PDB website (www.rcsb.org) and converted in MOL2 format using the UCSF Chimera package [34]. Cavities at the enzyme surface were detected using the program VolSite [14]. The catalytic site was defined as the set of residues surrounding a ligandable cavity (within a distance cutoff of 4 Å) and containing the highest number of catalytic residues. If the PDB file contained multiple copies of the enzyme, we considered the catalytic site with the lowest average temperature factor.

Comparing protein sites. Protein site similarity was computed using SiteAlign [19] and Shaper [14].

SiteAlign represents a binding site with an 80-triangle polyhedron centered on the protein cavity. Physicochemical properties of amino acids in the site are projected onto triangles of the polyhedron (cofactor, metal ions, and water molecules are ignored). Null property is assigned to triangles not hit by any projection. Sites are aligned by optimizing the superimposition of two polyhedrons for the best match of physicochemical properties. SiteAlign quantifies site similarity using two distances: *D1*, which scores all triangles pairs, and *D2*, which only scores pairs of non-null triangles in the reference and the compared polyhedron. Shaper considers amino acids, cofactor(s), and water molecule(s) as part of the protein. It represents sites with a cubic grid filling the cavity. Each grid point is annotated with the pharmacophoric property of the nearest protein atoms. Sites are aligned by maximizing the geometric overlap of grids. *RefTversky* coefficient scores similarity.

The screened dataset: the sc-PDB, a collection of ligandable sites in the PDB. Active sites of NP biosynthetic enzymes were compared to the sc-PDB release 2013 [8]. 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) in the vicinity of the ligand. A ligand-binding site was defined as all protein residues delimiting the cavity detected using VolSite (considering the ligand position) and with at least one heavy atom distant from less than 6.5 Å from any ligand heavy atom. The dataset contains 9283 structures representing 3678 different proteins.

Virtual screening of the sc-PDB by site comparison. Active sites of NP biosynthetic enzymes were compared to the sc-PDB using SiteAlign and Shaper. Each screening yielded a list of 9283 sites sorted by decreasing similarity to the query site. Ranking of the 3678 corresponding proteins was obtained by considering only the highest similarity score of multi-copy proteins. Similarity score significance was assessed by computing expectation from the score distribution provided that these are bell-shaped (Gaussian-like). An expectation threshold of less than 0.06 was used to select hit proteins (Z-scores ≥ 2.5 using Shaper, Z-scores ≤ -2.5 using SiteAlign).

Supporting information

A simplified phylogenic tree, selection of biosynthetic pathways, and a list of natural products and their SMILES structures are available as Supporting Information. A separate spreadsheet reports the potential novel targets of the 25 NPs produced by the 38 biosynthetic enzymes described in this study.

Acknowledgements

Financial support by Griffith University, Brisbane, and LabEx "Drug research center Medalis" is acknowledged. The authors would like to thank the calculation center of the IN2P3 (CNRS, Villeurbanne, France) for allocation of computing time, as well as collaborators G. Beck for the search of NPs in CHEMBL, and M. Campitelli and G. Bret for technical support. Molecular graphics and analyses were performed with the UCSF

Chimera package. Chimera was developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311).

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Newman DJ, Cragg GM. Natural products as sources of new drugs from 1981 to 2014. J Nat Prod 2016; 79: 629–661

[2] Tajabadi FM, Campitelli MR, Quinn RJ. Scaffold flatness: reversing the trend. Springer Sci Rev 2013; 1: 141–151

[3] Romo D, Liu JO. Editorial: Strategies for cellular target identification of natural products. Nat Prod Rep 2016; 33: 592–594

[4] Olğaç A, Orhan IE, Banoglu E. The potential role of *in silico* approaches to identify novel bioactive molecules from natural resources. Future Med Chem 2017; 9: 1663–1684

[5] Ma ST, Feng CT, Dai GL, Song Y, Zhou GL, Zhang XL, Miao CG, Yu H, Ju WZ. *In silico* target fishing for the potential bioactive components contained in Huanglian Jiedu Tang (HLJDD) and elucidating molecular mechanisms for the treatment of sepsis. Chin J Nat Med 2015; 13: 30–40

[6] Gu J, Gui Y, Chen L, Yuan G, Lu HZ, Xu X. Use of natural products as chemical library for drug discovery and network pharmacology. PLoS One 2013; 8: e62839

[7] Sturm N, Quinn RJ, Kellenberger E. Similarity between flavonoid biosynthetic enzymes and flavonoid protein targets captured by three-dimensional computing approach. Planta Med 2015; 81: 467–473

[8] 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

[9] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein Data Bank. Nucleic Acids Res 2000; 28: 235–242

[10] Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, 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 2014; 42: D459–D471

[11] Morgat A, Coissac E, Coudert E, Axelsen KB, Keller G, Bairoch A, Bridge A, Bougueleret L, Xenarios I, Viari A. UniPathway: a resource for the exploration and annotation of metabolic pathways. Nucleic Acids Res 2012; 40: D761–D769

[12] The UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res 2015; 43: D204–D212

[13] Furnham N, Holliday GL, de Beer TAP, Jacobsen JOB, Pearson WR, Thornton JM. The catalytic site atlas 2.0: cataloging catalytic sites and residues identified in enzymes. Nucleic Acids Res 2014; 42: D485–D489

[14] 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

[15] Roach P, Clifton I, Hensgens C, Shibata N, Schofield C, Hajdu J, Baldwin J. Structure of isopenicillin N synthase complexed with substrate and the mechanism of penicillin formation. Nature 1997; 387: 827–830

[16] Starks C, Back K, Chappell J, Noel J. Structural basis for cyclic terpene biosynthesis by tobacco 5-epi-aristolochene synthase. Science 1997; 277: 1815–1820

[17] Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR. The ChEMBL database in 2017. Nucleic Acids Res 2017; 45: D945–D954

[18] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2014; 42: D1091–1097

[19] Schalon C, Surgand JS, Kellenberger E, Rognan D. A simple and fuzzy method to align and compare druggable ligand-binding sites. Proteins Struct Funct Bioinforma 2008; 71: 1755–1778

[20] Bush K, Jacoby GA. Updated functional classification of β-lactamases. Antimicrob Agents Chemother 2010; 54: 969–976

[21] Sturm N, Rognan D, Quinn RJ, Kellenberger E. Comparing atom-based with residue-based descriptors in predicting binding site similarity: do backbone atoms matter? Future Med Chem 2016; 8: 1871–1885

[22] Hoffman AF, Lupica CR. Synaptic targets of Δ9-tetrahydrocannabinol in the central nervous system. Cold Spring Harb Perspect Med 2013; 3: a012237

[23] Moss DE, Peck PL, Salome R. Tetrahydrocannabinol and acetylcholinesterase. Pharmacol Biochem Behav 1978; 8: 763–765

[24] Xie X, Meehan MJ, Xu W, Dorrestein PC, Tang Y. Acyltransferase mediated polyketide release from a fungal megasynthase. J Am Chem Soc 2009; 131: 8388–8389

[25] Stender S, Nordestgaard B. [Cholesterol – when is preventive treatment with statin indicated?]. Ugeskr Laeger 2011; 173: 1747; author reply 1747

[26] Lin YC, Lin JH, Chou CW, Chang YF, Yeh SH, Chen CC. Statins increase p21 through inhibition of histone deacetylase activity and release of promoter-associated HDAC1/2. Cancer Res 2008; 68: 2375–2383

[27] Shahbazian MD, Grunstein M. Functions of site-specific histone acetylation and deacetylation. Annu Rev Biochem 2007; 76: 75–100

[28] Guidotti A, Auta J, Chen Y, Davis JM, Dong E, Gavin DP, Grayson DR, Matrisciano F, Pinna G, Satta R, Sharma RP, Tremolizzo L, Tueting P. Epigenetic GABAergic targets in schizophrenia and bipolar disorder. Neuropharmacology 2011; 60: 1007–1016

[29] Chuang DM, Leng Y, Marinova Z, Kim HJ, Chiu CT. Multiple roles of HDAC inhibition in neurodegenerative conditions. Trends Neurosci 2009; 32: 591–601

[30] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera – a visualization system for exploratory research and analysis. J Comput Chem 2004; 25: 1605–1612

[31] Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M, Steinbeck C. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res 2013; 41: D456–D463

[32] Morgat A, Axelsen KB, Lombardot T, Alcántara R, Aimo L, Zerara M, Niknejad A, Belda E, Hyka-Nouspikel N, Coudert E, Redaschi N, Bougueleret L, Steinbeck C, Xenarios I, Bridge A. Updates in Rhea – a manually curated resource of biochemical reactions. Nucleic Acids Res 2015; 43: D459–D464

[33] Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, Oldfield TJ, O'Donovan C, Martin MJ, Kleywegt GJ. SIFTS: Structure integration with function, taxonomy and sequences resource. Nucleic Acids Res 2013; 41: D483–D489

[34] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970; 48: 443–453