JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

PROLIX: Rapid Mining of Protein−Ligand Interactions in Large Crystal Structure Databases

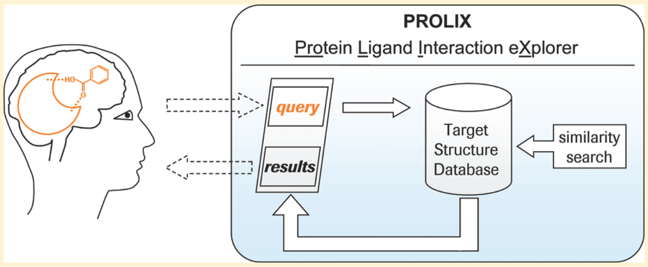
Martin Weisel,*,[†] Hans-Marcus Bitter,[‡] François Diederich,[§] W. Venus So,[∥] and Rama Kondru[†]

[†]Discovery Chemistry, Hoffmann-La Roche, Inc., 340 Kingsland Street, Nutley, New Jersey 07110, United States

[‡]Translational Research Sciences, Hoffmann-La Roche, Inc., 340 Kingsland Street, Nutley, New Jersey 07110, United States

[§]Laboratorium für Organische Chemie, ETH Zürich, Wolfgang-Pauli-Strasse 10, HCI, 8093 Zürich, Switzerland

[∥]Pharma Research & Early Development Informatics, Hoffmann-La Roche, Inc., 340 Kingsland Street, Nutley, New Jersey 07110, United States


PROLIX
Protein Ligand Interaction eXplorer
query
results
Target Structure Database
similarity search
HO
O

ABSTRACT: A central problem in structure-based drug design is understanding protein−ligand interactions quantitatively and qualitatively. Several recent studies have highlighted from a qualitative perspective the nature of these interactions and their utility in drug discovery. However, a common limitation is a lack of adequate tools to mine these interactions comprehensively, since exhaustive searches of the protein data bank are time-consuming and difficult to perform. Consequently, fundamental questions remain unanswered: How unique or how common are the protein−ligand interactions observed in a given drug design project when compared to all complexed structures in the protein data bank? Which interaction patterns might explain the affinity of a tool compound toward unwanted targets? To answer these questions and to enable the systematic and comprehensive study of protein−ligand interactions, we introduce PROLIX (Protein Ligand Interaction Explorer), a tool that uses sophisticated fingerprint representations of protein−ligand interaction patterns for rapid data mining in large crystal structure databases. Our implementation strategy pursues a branch-and-bound technique that enables mining against thousands of complexes within a few seconds. Key elements of PROLIX include (i) an intuitive interface that enables users to formulate complex queries easily, (ii) exceptional speed for results retrieval, and (iii) a sophisticated results summarization. Herein we describe the algorithms developed to enable complex queries and fast retrieval of search results, as well as the intuitive aspects of the user interface and summarization viewer.

■ INTRODUCTION

High-throughput X-ray crystallography[1] and NMR[2] have led to a dramatic increase in the number of the protein structures publicly available in the Protein Data Bank (PDB).[3] Analyses of protein−ligand complex structures have contributed significantly to an improved understanding of the protein−ligand interactions[4,5] that drive ligand binding. This improved understanding about protein−ligand interactions has provided valuable information for receptor-based drug design projects.[6]

At the time this manuscript was being prepared, more than 77 000 structures had been deposited in the PDB.[3] This collection includes a significant number of validated protein-small molecule complexes as reported in the PDBbind database.[7−10] The primary challenge for efficient protein−ligand interaction analysis and mining in terms of search retrieval speed is the large number of available complexes. To address this challenge, bit string fingerprinting is widely utilized in various computer programs that extract information from protein−ligand interactions in large protein complex databases.[11−23] Bit string and integer-based fingerprinting can be used to store large amounts of data in a relatively small space, enabling rapid comparison of complex data.[24]

When analyzing protein−ligand interactions, established fingerprint methods commonly annotate information about the nature of an interaction (e.g., hydrogen bond, polar contact, $\pi-\pi$ interaction), the considered interaction distances (e.g., Euclidean distance between the heavy atoms of a hydrogen bond), as well as the types of the interacting ligand atoms and receptor residues.

Protein−ligand interaction fingerprints described in the literature are designed to address different problems in early drug design. A number of methods use protein−ligand interaction fingerprints to improve the performance of protein−ligand docking programs.[11−14] Herein, interaction fingerprints are calculated for all docking solutions generated by the docking software annotating favorable interactions to the receptor. The docking poses are then reranked, marking the pose with the most favorable interactions set to the receptor as the optimal docking solution. These methods have been proven to offer increased success rates in identifying crystallographic ligand binding modes.[11−14]

Received: January 18, 2012

Several methods implement protein–ligand interaction fingerprints for the virtual screening of small molecules that feature a desired ligand interaction pattern to their respective receptors.[11,15−18,21] The interacting fragment fingerprints (IF-FP)[15−17] introduced by Tan and co-workers incorporate interaction information into conventional 2D fingerprints by emphasizing the interacting parts of a ligand during similarity searching. This approach is realized by encoding only those ligand parts in the IF-FP which feature interactions with the receptor while ignoring noninteracting parts. The method achieves improved active compound recall rates in virtual screening searches in comparison to complete ligand fingerprints.[15] While IF-FPs highlight the interacting parts of the ligand, information concerning the interaction counts, types or distances is not considered in the fingerprints. Furthermore, the types of the interacting receptor residues are not regarded in this approach. These issues are in part addressed by the structural interaction fingerprints (SIFt)[11,18−20] proposed by Deng and co-workers. SIFts are 1D binary fingerprint representations of protein–ligand interactions in three-dimensional protein-inhibitor complexes. In contrast to the IF-FPs that emphasize the interacting ligand fragments, the SIFts focus on the binding site residues that form interactions with the ligand. Each interacting residue is represented by a 7-bit-long bit string that annotates information about the interaction type and whether the interaction involves the residue main-chain or side-chain. The complete SIFt for a protein–ligand complex is generated by sequentially concatenating the bit string of each interacting residue. SIFts have been applied to a wide range of applications including virtual screening, ligand docking pose analysis and interaction profile identification within protein target families.[18] However, since the fingerprints are dependent on the residue sequence of the respective receptor they were derived from, SIFts are limited to perform analyses within the same protein family.

Therefore, existing fingerprint-based tools to analyze protein–ligand interactions focus on encoding either the receptor or the ligand side of the complex rather than recording the chemical information of the respective interaction. Moreover, interaction fingerprints derived from the amino acid sequence of one receptor cannot be applied to receptors in different protein families.

To resolve these drawbacks of existing approaches, we implemented a new fingerprinting method that features a sequence-independent encoding of interaction types and distances. These fingerprints were integrated into our new database mining tool PROLIX (Protein Ligand Interaction Explorer) which enables rapid matching of protein–ligand interaction patterns in large crystal structure databases. The comprehensive mining of protein–ligand interactions found in the PDB poses several challenges: first, complex queries must be easy to formulate; second, retrieval of search results must be fast (on the order of seconds); third, the results must be summarized in an informative and interpretable manner. To address the ease of query definition, we developed an intuitive interface that allows users to easily draw a small molecule, introduce pocket residues and then specify various noncovalent interactions and additional geometric constraints. For rapid results retrieval, we implemented a novel branch-and-bound fingerprinting technique that enables rapid database mining of user queries. This method performs database searches against thousands of complexes typically achieving subsecond runtimes. Finally, to facilitate the interpretation of results, PROLIX features a sophisticated results display that summarizes database hits matching the user query and highlights preferred protein–ligand interactions across different protein families.

We designed PROLIX to answer questions occurring in daily drug design work: What other receptors feature protein–ligand interaction networks similar to my project? What functional groups or fragments bind to a similar receptor environment and can be used as tool compounds? A query in PROLIX can therefore be defined as a specific set of different interaction types with user-defined distances between discrete binding site residues and a ligand molecule. We implemented a database mining routine that is capable of matching complex query layouts in large target databases holding millions of interactions achieving subsecond runtimes. PROLIX specifically operates on fingerprint representations of binding site residues, interaction types and interaction distances. A structural alignment or superposition of binding sites is not required in PROLIX. This approach distinguishes our tool from other methods that use structural alignments of either binding site residues or pharmacophore-like pseudocenters for binding site comparisons.[25−31]

PROLIX can be used to gain further knowledge about the propensities of certain pocket environments to participate in specific protein–ligand interactions. In this manuscript, we describe some challenges we faced while developing PROLIX and present the algorithms implemented to address these challenges.

■ METHODS

PROLIX offers a graphical user interface (GUI) that enables the user to define complex protein–ligand interaction queries to be matched against large databases of complexed structures. User queries are translated into a series of fingerprints that are employed to find matching protein–ligand interaction patterns as well as distance constraints between binding site residues in the target database. In this section, we describe (i) how user queries are defined in PROLIX, (ii) our multistep matching strategy, (iii) the databases utilized in the matching process, (iv) the protein–ligand interaction fingerprints design and matching procedure, (v) the residue distance matching process, (vi) the scoring and ranking of the matched database hits, (vii) the design of the PROLIX GUI, and (viii) the generation of the protein family and subfamily names used to classify the database hits.

Query Definition in PROLIX. A query in PROLIX comprises a ligand molecule, binding site residues, as well as interaction types and distances defined toward the ligand (Figure 1A). Of note, we differentiate between “interacting residues” and what we term “shell residues”. We define the shell as all residues that lie within 4.5 Å of the ligand molecule. While interacting residues feature at least one noncovalent interaction to the ligand molecule, a shell residue (or nonbonding residue) is defined as a residue that is present in the binding site (or shell) of the complex and does not feature a direct interaction with the ligand.

When defining the protein–ligand interactions, the user can designate different interaction types and their distances. The interaction distances are measured between the respective interacting heavy atoms on the ligand molecule and the specified residue. Differences between the query interaction distances and the respective matched target interactions results in penalty scores that affect the hit ranking.

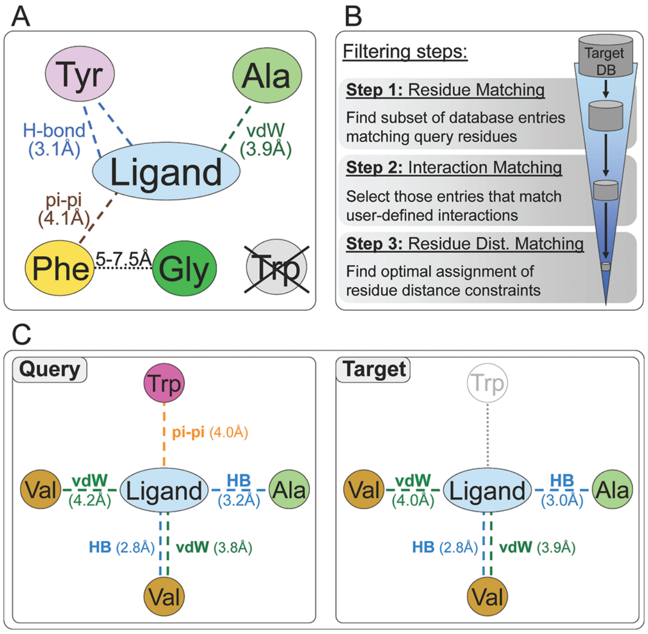


Figure 1. (A) Simplified protein–ligand interaction user query, interactions are shown as dashed lines, distance constraints between residues are shown as dotted lines. This exemplary query features three "interacting residues" (Tyr, Ala, Phe) that interact with the ligand through different interaction types (H-bond = hydrogen bond, *vdW* = van der Waals interaction, *pi-pi* = pi-pi interaction). The ligand molecule is shown as a blue oval to simplify the depiction. Besides the three interacting residues, this example query features one "shell residue" (Gly) with a distance constraint (5–7.5 Å) toward another interacting residue (Phe). The crossed out residue (Trp) represents an "excluded" (or unwanted) residue. All target entries featuring at least one tryptophan in their shell will be discarded. (B) Overview of the branch-and-bound filtering approach implemented in PROLIX. Entries of the target database are discarded if they do not feature the residues (step 1), interactions (step 2) or potential distance constraints (step 3) defined by the user. The user can influence the filtering process by specifying partial matching cutoffs for the requested residues. Penalty scores will be assigned for database hits that do not match all requested residues or distance constraints. Database entries that do not meet a user-defined partial matching cutoff will be discarded. (C) Concept of partial residue matching as implemented in PROLIX. The example query (left panel) features four binding site residues interacting with the ligand through various interactions. One exemplary entry of the target database (right panel) features a protein–ligand interaction pattern that is similar to the user query. In this example three of the four requested residues and their interactions are matched, yielding a partial matching of the residues of 75%. This target will only be considered as a hit if the user defined a partial residue matching cutoff of <75%. A penalty score will be assigned for the differences in the interaction distances for the matched interactions of the three matched residues.

PROLIX enables distance constraint definitions between receptor residues to express their spatial arrangement in the binding site. Distance constraints can be defined between pairs of interacting or nonbonding residues or mixtures of both. Each residue can have distance constraints to multiple other residues. A lower and an upper cutoff are defined for each residue distance constraint. A corresponding distance in a target entry is only considered a match if it lies within the user-defined cutoff range.

PROLIX additionally supports the definition of "excluded" residues to further characterize a specific binding site environment in more detail. For example, if a tryptophan is defined as an excluded (or unwanted) residue, any database entry featuring this residue in its shell will be removed from the hit list. The excluded residue definition applies to all residues within the 4.5 Å shell around the ligand and does not differentiate between interacting and nonbonding residues.

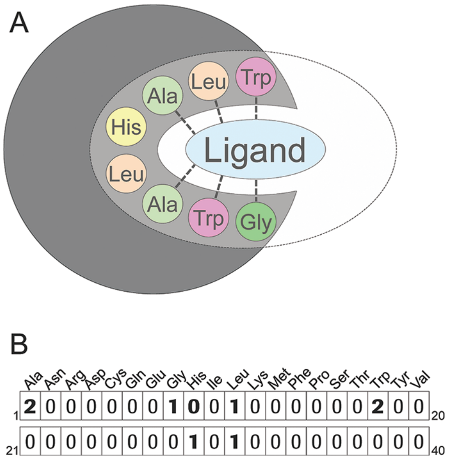
All residues, interactions and distances defined in the user query are handled as fingerprints in PROLIX enabling the rapid comparison of the query against the target database. The general matching strategy is summarized in the following section.

General Database Mining Strategy. Matching a single protein–ligand interaction pattern defined in a user query against a large target database requires comparisons to hundreds of thousands of binding site residues and millions of interaction types and distances. A powerful matching algorithm is needed to handle the vast number of required comparisons. Therefore, we implemented an approach that is known as the *branch-and-bound* strategy in informatics. This approach pursues a systematic enumeration of all possible solutions where large subsets of unpromising candidates are discarded en masse at an early stage of the matching process. Employing this approach, numerous unnecessary comparisons can be avoided thus resulting in a dramatic reduction in overall matching runtime. For efficient database mining, a three-step filtering approach was implemented in PROLIX comprising (i) a coarse-grained fingerprint filtering step focusing on the binding site residues, followed by (ii) a detailed fingerprint filtering of interaction types and distances and (iii) a distance constraints analysis if defined by the user. Thus, the PROLIX fingerprint matching algorithm examines in the following order whether a database entry features (i) the interacting and nonbonding binding site residues, (ii) the interaction types, and (iii) the pairwise residue distance constraints that the user defined in the query. Figure 1B illustrates this process.

The PROLIX fingerprint-matching algorithm supports partial residue matching in the first filtering step. For example, if the user defines 10 binding site residues and specifies a 75% partial matching cutoff, a database entry that matches 8 of 10 residues is still considered a hit, whereas an entry matching only 7 residues is discarded. The partial residue matching is depicted in Figure 1C.

Databases Used for the Calculation of PROLIX Fingerprints. The target database comprises the protein–ligand complexes available in both the RCSB Protein Data Bank[3] (PDB, www.pdb.org) and the Roche in-house X-ray structure database. These two data sets were accessed through Proasis software.[32] Information explicitly stored in the raw PDB text file format, such as atom coordinates, b-factors and sequence information, is accessible in Proasis through a relational database. Crystal structures complexed with small molecules imply information about noncovalent protein–ligand interactions. Proasis extracts these noncovalent interactions from complexed crystal structures by following a detailed set of rules defining several different protein–ligand interaction types. This information about noncovalent interaction types, their distances and the atoms on the protein and ligand side is also stored in the relational database mentioned above. A detailed list about the different Proasis interaction types and how they are defined was recently published by Kuhn and co-workers.[33]

Generation and Matching of "Shell Residue Fingerprints". The first filtering step focuses on the residues defined by the user in the input query. The frequencies of the 20 standard amino acids of both the interacting and nonbonding shell residues are separately annotated in what we term the "query shell residues fingerprint" (qSRFP). The resulting 40-position-fingerprint (Figure 2) is subdivided into two sections: the frequencies of the interacting residues are annotated in the first 20 positions, whereas the occurrences of the nonbonding shell residues are listed in the second 20 positions with each position representing one of the 20 standard amino acids. The qSRFP is calculated on the fly from the input user query at the time the matching is launched. The corresponding "target shell residues fingerprint" (tSRFP) has the same format as the qSRFP. Unlike the qSRFP, the tSRFPs are precalculated for the entire target database, significantly reducing total matching runtimes. The standard amino acid frequencies in the interacting residues part of the tSRFP (first 20 positions) are derived from Proasis. We implemented a protocol in PipelinePilot[34] to determine the frequencies of the interacting residues directly from the Proasis relational database. Residues that feature multiple interactions to the ligand are only counted once in the fingerprint. The frequencies of the 20 nonbonding shell residues were directly derived from the PDB X-ray structure files using a Java program we implemented. The program counts the frequencies of the different residue types in the shell, not counting the residues that interact with the ligand. The frequencies of the nonbonding shell residues are then annotated in the last 20 positions of the tSRFP. The tSRFP generation process from a protein−ligand complex is summarized in Figure 2.



Ala	Asn	Arg	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
2	0	0	0	0	0	0	**1**	**0**	0	**1**	0	0	0	0	0	0	**2**	0	0
0	0	0	0	0	0	0	0	**1**	0	**1**	0	0	0	0	0	0	0	0	0

Figure 2. Generation of the target shell residues fingerprint (tSRFP) (A) The tSRFPs are precalculated for all PDB files stored in the database of target protein−ligand complex structures. All interacting and nonbonding residues that are binding site "shell" are annotated in the tSRFP. The shell is defined as all interacting and nonbonding residues within 4.5 Å (dashed outline oval) around the ligand. The example receptor shell features six interacting residues (dashed interaction lines) and two nonbonding residues. (B) Format of the shell residues fingerprint (SRFP). Both the SRFPs for the query (**q**SRFP) and the targets (**t**SRFP) have the same format: the frequencies of the interacting residues are annotated in the first 20 positions (upper row) and the frequencies of the nonbonding residues are annotated in the last 20 positions of the fingerprint (lower row). The fingerprint in this example is the tSRFP derived from the complex in Figure 2A. The 40 position fingerprint is shown in two rows of 20 positions for formatting reasons.

Shell residue fingerprints are used for the first filtering step of PROLIX where matching target complexes are identified with regard to the residues in their binding site. The actual qSRFP fingerprint matching against the precalculated tSRFPs database was realized through the extensive use of matrix operations. Specifically, the precalculated 40 position tSRFPs for all target complexes in the database are read as row vectors t_i into a single target fingerprint matrix T. We used the JAMA package[35] to create these matrices in our PROLIX Java code. The qSRFP that is generated on the fly from the user query is handled as a 40 position row vector q that is iteratively matched against all i tSRFP vectors t_i stored in target matrix T. We used iterative subtractions of the qSRFP vector q from all tSRFP vectors t_i to identify matching targets. A result vector r_i is calculated as the difference of each subtraction (eq 1).

$$r_i = t_i - q \tag{1}$$

The result vector r_i features the same format as the query and target vectors q and t_i with all vectors having a length of $j = 40$ positions. Negative positions in r_i indicate query residues that were not fully matched by the respective target fingerprint vector t_i. The number u of unmatched residues (i.e., residues defined in the query vector q that were not matched by a respective target fingerprint vector t_i) can be calculated as the absolute value of the sum of all negative entries in r_i (eq 2).

$$u_i = |\sum_{j=1}^{40} r_{ij}|, \quad \text{for all } r_{ij} < 0 \tag{2}$$

The ratio m_i of matched residues can be calculated with regard to the sum of all query residues n (eq 3).

$$m_i = \frac{n - u_i}{n} \tag{3}$$

If, for example, a partial residue matching cutoff p of 0.75 (75%) was defined by the user, a target entry would only be considered as a potential hit if the SRFP matching against the input query yields a matching ratio $m_i \geq 0.75$ in this first filtering step. All targets not achieving the preset partial residue cutoff p will be ruled out as potential hits and therefore be excluded from subsequent database matching efforts.

As mentioned previously, the user can define "excluded residues" (see Figure 1A) in the query. An excluded (or unwanted) residue is marked by a −1 in both the respective interacting and nonbonding qSRFP positions. For example, if the user selected alanine as an excluded residue, positions 1 and 21 would be set to −1. A target will be excluded as a potential hit when its corresponding tSRFP features one or more unwanted residues at the respective position in the fingerprint. This search for excluded residues is performed prior to the actual fingerprint matching process described in this section. In case no excluded residues are found in the tSRFP, the positions highlighting unwanted residues in the qSRFP as −1 will be set to 0, ensuring that the negative −1 marker is not misinterpreted as an unmatched residue in the SRFP matching.

Generation and Matching of Interaction Fingerprints. All entries remaining as hit candidates after the first residue filtering step are submitted to a second filtering step that focuses on the interactions defined between the binding site residues and the ligand in the user query. The user can define multiple interactions of different types and distances between the ligand and each residue. The interaction types defined in the user query are translated into what we term the "query interaction fingerprint" (qIFP). Similar to the frequencies of the different residue types annotated in the qSRFP (Figure 2), the qIFP annotates the frequencies of the user-defined interaction types. In contrast to the qSRFP, which describes the frequencies of all binding site residues in one single fingerprint, a separate qIFP is created for each residue of the user query. For example, a user query featuring five interacting residues would result in five individual qIFPs annotating the respective frequencies of all interaction types for each query residue. While the qIFPs are directly generated from the user query, the corresponding "target interaction fingerprints" (tIFPs) are precalculated for the entire target database. A PipelinePilot[34] script was implemented to create the tIFPs directly from the Proasis relational database. The tIFPs are stored in a JAMA[35] matrix in PROLIX.

The matching of the qIFPs against the precalculated tIFPs database is similar to the SRFP matching process. A result vector r_i is again calculated as the difference between a target vector t_i and a query vector q_i (eq 1). Unmatched interactions of a target can again be identified as negative positions in the result vector r_i. In contrast to the first residue matching filtering step, partial matching is not considered for interactions in the second filtering step. Therefore, a residue that was matched in the residue filtering step will no longer be considered as a potential match if it does not feature all interaction types defined in the corresponding query residue (Figure 3). Consequently, the ratio of matched residues is reduced. Hence, if the user-defined partial residue matching cutoff is no longer achieved, a respective target will be instantly discarded as a potential database hit (Figure 3, bottom panel).

Since one qIFP is calculated for each interacting query residue, multiple comparisons have to be performed per target to find corresponding residues with matching interactions. Moreover, if a given target features multiple interacting residues of the same type, multiple fingerprint matchings have to be performed for a query residue. For example, a qIFP that is generated for an alanine in the user query featuring one hydrogen bond to the ligand (see Figure 3, top panel) has to be matched against the corresponding tIFPs generated for all alanine residues in each target shell. To execute this concept, we implemented a procedure in the PROLIX interaction matching that performs a complete enumeration of all possible pairs of interacting query and target residues of the same type.

Regarding the previous query example of an alanine residue with a single hydrogen bond, a list holding all matching target alanines that feature the requested hydrogen bond is generated. We refer to this list as mriList (matching residues with matching interactions list). Alanines not featuring the user-defined hydrogen bond are not annotated in the mriList. The mriList will subsequently be completed to list all matching target residues for all query residues and their interactions. The residues listed in the mriList hence feature the same residue type and interactions as their corresponding query residue.

The distances for the interactions matched in the target might differ from those defined for the query interactions. Furthermore, the matched target residues may feature more interactions of the same type than specified in the query. Moreover, each target residue annotated in the mriList could potentially match multiple different query residues of the same type. A method is consequently required that performs an accurate matching between the requested query residues and their interactions with the corresponding target residues. Therefore, we implemented an algorithm in PROLIX that guarantees the optimal assignment of the matching target residues and interactions stored in the mriList to best match the user query. The method is optimal such that it maximizes the number of matching target residues (as a first priority) and minimizes the summed difference D of the distances matched between the query and the target.

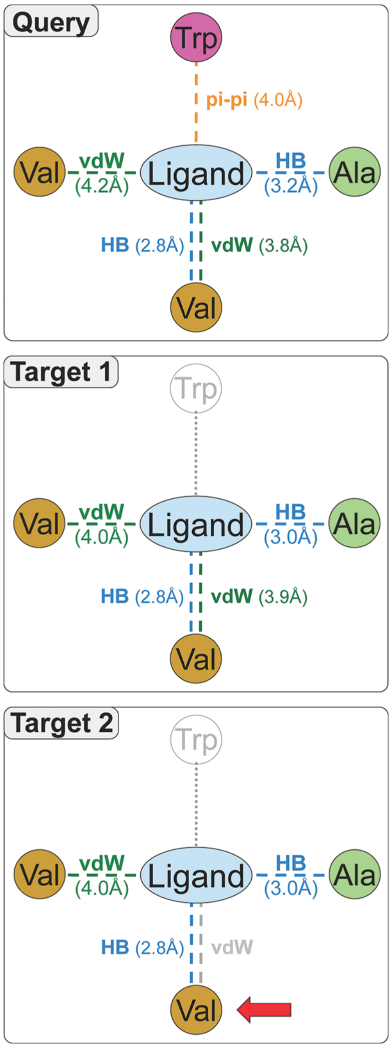


Figure 3. Interaction matching in PROLIX. A user query is defined with four residues and their interactions to the ligand (top panel). Target 1 (middle panel) is considered a hit, if a partial residue matching cutoff of 75% (or smaller) is defined, since 3 out of 4 (75%) query residues and **all** their interactions are matched. Target 2 (bottom panel) is not considered a hit, if a 75% partial residue matching cutoff is selected, since only two residues (50%) feature **all** interactions defined in the query. The red arrow marks a valine residue in target 2 that is missing the van der Waals interaction requested in the user query. Target 2 will be removed from the hit list of matching targets.

To illustrate the optimal assignment procedure, the interaction matching can be explained using a graph model (Figure 4).

Let G be a graph representing the interacting query residues R and the interacting atoms of the ligand L as vertices V in the graph $(R,V \in V)$. User-defined interactions are represented as edges E in G, connecting residue nodes R and ligand atom nodes L. Distances defined for the query interactions are saved as edge properties μ_E. The optimal assignment method performs the complete enumeration of all possible matching scenarios between the query residues and the target residues

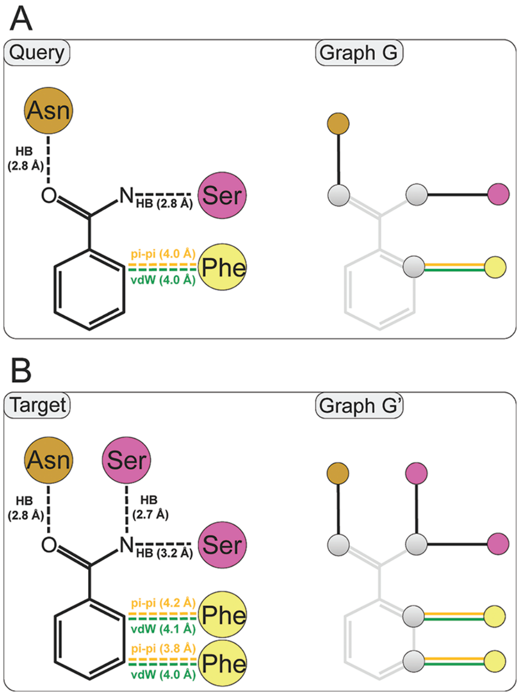


Figure 4. Optimal assignment of residues and their interactions in PROLIX. (A) A query is defined with three residues interacting with a ligand molecule. The query can be expressed as graph G that represents residues and ligand atoms as nodes and interactions as connecting edges. Information about interaction types and distances are stored as edge properties. Graph G offers an abstracted representation of the protein−ligand interaction pattern defined in the Query. (B) A matching target might feature more residues and interactions than a respective query. Target graph G′ illustrates that there are multiple possible matches between query and receptor residues. PROLIX enumerates all possible matching scenarios. The optimal assignment method determines which scenario features the maximum number of matched residues while having the smallest sum of differences between query and interaction distances. Matched target residues have to be congruent with the corresponding query residue types and also have to feature all respective interactions defined for the query residue.

annotated in the mriList. The comparison of a matching protein−ligand interaction pattern between the query and a target scenario can be expressed by the corresponding query graph $G = (V, E, \mu_E)$ and a target graph $G' = (V', E', \mu'_E)$. Interaction distances defined in the user query might differ from the distances of the matched interactions in each matching scenario of a given target. Let edge $v \in V$ be the edge in G representing an interaction in the query that is matched by edge $v' \in V'$ in G' representing a corresponding interaction in the target. The interaction distances are represented as edge properties μ_v and $\mu_{v'}$ in the graphs G and G'. The difference d of the interaction distances can therefore be calculated as the absolute value of the difference between μ_v and $\mu_{v'}$ (eq 4).

$$d(\mu_v, \mu_{v'}) = |\mu_v - \mu_{v'}| \tag{4}$$

The total interaction distance difference D is defined as the sum of all interaction difference distances d of the i interactions matched by a target scenario.

$$D = \sum_i d_i \tag{5}$$

The total interaction distance difference D is used as a penalty score to rank the different matching scenarios for one target. If multiple matching scenarios achieve the same number $n_{\max}$ of matched residues, the total interaction distance difference D is used as a secondary scoring function. A scenario where all n query residues are matched ($n = n_{\max}$), and all interaction distances are identical in query and target scenario ($D = 0$) is considered a perfect match. In general, the scenario achieving $n_{\max}$ and featuring the lowest accumulated D represents the optimal solution for a respective target. The two scoring parameters $n_{\max}$ and D are not used solely to determine the optimal matching scenario for a single target entry. They are also used for the global ranking of the optimal matches of all targets that were not discarded by the residue and interaction matching. The number of matched residue distance constraints is used as a third parameter for the scoring of matching targets. The residue distance matching is summarized in the next section.

Residue Distance Matching. In addition to the definition of different residue types, interaction types, and interaction distances, distance constraints between binding site residues can be defined in PROLIX user queries (Figure 1A). Residue distances can be defined between interacting and nonbonding residues. Multiple distances to different residues can be defined for each residue in the query. However, only one distance constraint can be defined for a pair of residues. Minimum and maximum values are defined as hard cutoffs for residue distance constraints. A distance constraint in the target is only considered a match if (i) it is defined between the same pair of residue types as defined in the query and (ii) it meets the specified distance cutoffs. The user can specify whether the distance constraints are measured between alpha-carbons or the closest atoms between the residues. For all target complexes, pairwise residue distances were precalculated for all binding site residues. The pairwise residue distances were saved in two databases, one using closest atoms as reference points and the other using distances calculated between alpha-carbons.

Similar to the interaction matching, multiple solutions potentially exist in one target when matching the user-defined residue distance constraints pattern. An elaborate matching process was implemented accordingly to enumerate all possible matching residue distance scenarios for each target entry not yet discarded in the preceding filtering steps. This process enumerates all possible matching scenarios and maximizes the number of residue distance constraints matched in the target (Figure 5). The matching process has to ensure that (i) the matched residue distance lies within the user-defined minimum and maximum distance cutoffs, (ii) the distance constraint is defined between the same residue types as specified in the query, and (iii) the matched constrained residues feature the same interactions as the corresponding query residues.

The residue distance constraint matching procedure identifies the optimal matching scenario for each target. The matching algorithm considers not only the residue distance cutoffs but also the types of the constrained residues and their interactions to the ligand. The scenario featuring the largest number of matched distance constraints is determined as the optimal match for each target.

Scoring and Ranking of Database Hits. A scoring function is implemented in PROLIX that ranks the database hits as best matching the user query. The ranking procedure considers the number of residues matched, the number of distance constraints matched and the total interaction distance

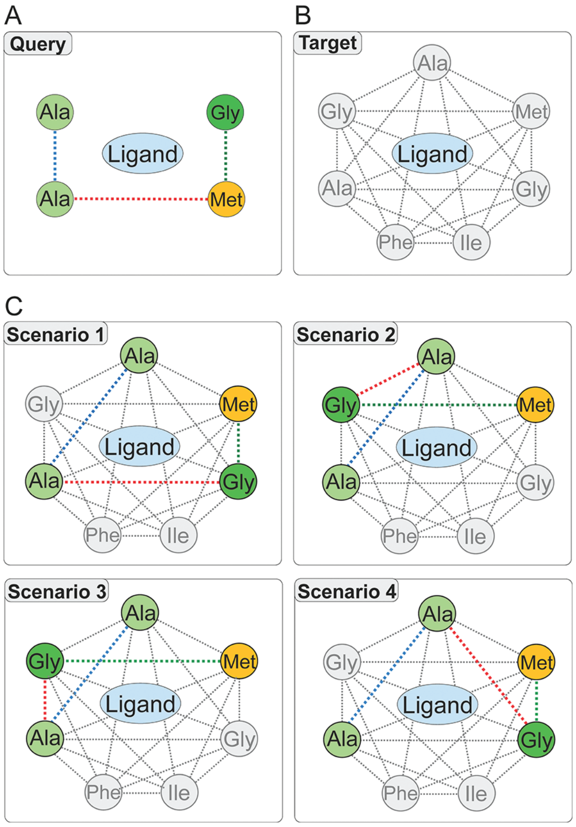


Figure 5. Residue distance constraint matching in PROLIX. (A) A query is defined with four residues and three distance constraints (dotted lines). Protein−ligand interactions are not shown to simplify the depiction. The residue distance constraints are color-coded for a better illustration of the distance constraints matching in panel C. (B) The query will be matched against a target complex shown here. Twenty-one pairwise distances for the seven target residues are highlighted here. (C) The residue distance constraints defined in the query (panel A) can be matched in multiple ways by the target (panel B). Shown here are the four different scenarios that represent a complete matching of all distance constraints defined in the query. The matched distance constraints are defined between the same residue types as given in the query. Furthermore, the matching method has to guarantee that the user-defined minimum and maximum distance cutoffs are met and that the matched residues also feature potential interactions to the ligand (not shown here) that the user might have defined.

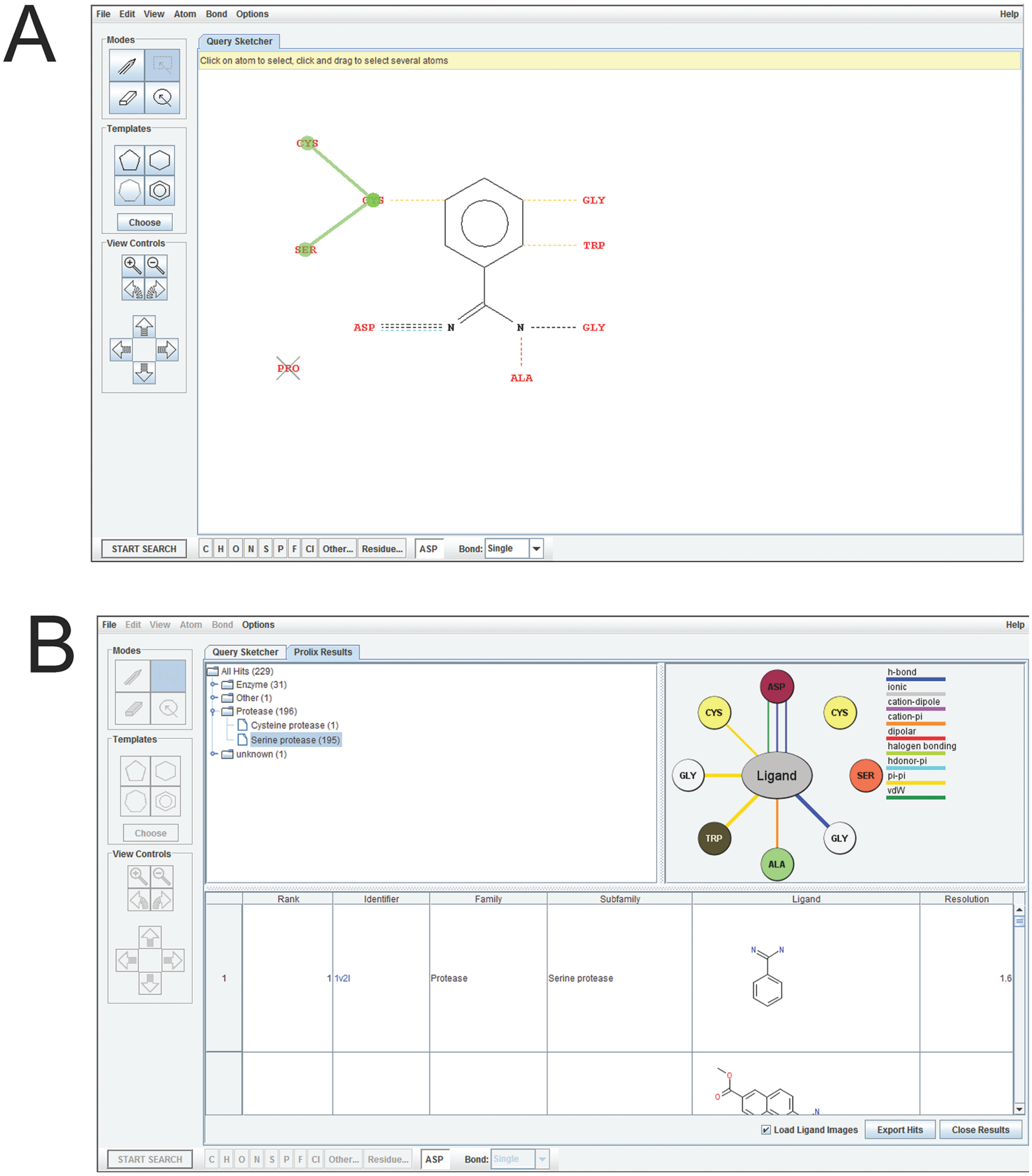
difference *D*, in that order. Consequently, the target hit featuring the largest number of matched residues, the largest number of matched distance constraints and the smallest total interaction distance *D* is ranked as the top hit. In the unlikely case that two hits achieve the same score, the hit first found in the database is ranked one position higher than the second hit. A shared rank for hits achieving the same score is not supported in PROLIX. The final ranked hit list comprises only those hits that satisfy the user-defined partial residue matching cutoff where the matched residues feature the interaction types defined in the query.

Design of the PROLIX Graphical User Interface. A graphical user interface (GUI) was developed in collaboration with the Cambridge Crystallographic Data Centre[36] (CCDC). The GUI is available as a web application to users at Roche. It offers a molecular sketcher panel where the user can define a protein−ligand interaction as a query. The GUI query sketcher is described in Figure 6A.

The user query comprises the ligand molecule, the interacting and nonbonding residues, the protein−ligand interactions and the residue distance constraints. The query defined in the GUI is translated to a query XML file. Subsequently, the query fingerprints are generated from the query XML file and compared against the precalculated target fingerprints during the matching process. The ranked list of database hits is written to a results XML file that serves as input for the PROLIX GUI results display. A screenshot of the results panel is shown in Figure 6B. The database hits are summarized in a scrollable ranked list at the bottom of the results display. The list features additional information for each hit, such as which protein family and subfamily it belongs to, 2D depictions of the respective reference ligands or the resolution of the respective X-ray structures. The list can be filtered by a navigation tree located in the upper left panel of the results display. Clicking a certain protein family or subfamily in the tree will filter the results list to show only those hits belonging to the current selection. The graphic display in the top right corner summarizes the conservation of individual protein−ligand interactions (as defined in the user query) across different protein families. The protein−ligand interaction picture depicts the occurrence of the interactions defined in the query for the protein family selected in the browsing tree. The picture highlights strongly conserved interactions as thicker lines, whereas less frequently observed interactions are depicted as thinner lines. Hovering over an interaction line will display a tooltip summarizing all PDB identifiers of the current selection in the browsing tree featuring the respective interaction. The tooltip also shows the percentage of the hits in the current selection that feature the respective interaction.

In addition to revealing which interactions are prevalent for a selected protein family, the protein−ligand interaction graphic is interactive in the sense that the user can select one or multiple interactions or residues of interest. The ranked hit list at the bottom of the results panel will be filtered accordingly to show only those hits featuring the respective interaction or residue. Clicking a residue in the protein−ligand interaction graphic highlights the residue and *all* interactions defined to the ligand, as partial matching is not considered for interactions in PROLIX. Selecting multiple residues and interactions will typically reduce the number of hits shown in the result hit list.

Definition of the Protein Family Nomenclature. The classification of protein family and subfamily names is based on annotations from individual public protein family databases. Each structure record (PDB code or in-house structure ID) is mapped to both Entrez[37] Gene ID and SwissProt[38] Accession Numbers using an in-house gene database as well as UniProt ID mapping.[39,40] The Entrez Gene IDs and SwissProt Accession Numbers are used to map the structure records to protein families (as each individual protein family database contains Entrez Gene ID or SwissProt Accession Number as identification for each protein). The structure records are mapped to protein family names according to a predetermined priority order as listed here (starting with higher priority): kinases, GPCRs, proteases, nuclear hormone receptors, ion channels, cytochrome P450 and "other" enzyme classes. The public sources for these individual protein families are: KinBase[41] for kinases, IUPHAR[42] for GPCRs, nuclear hormone receptors and ion channels, MEROPS[43] for proteases and EBI Enzyme Database[44] for various other enzyme classes. Since some of the protein family databases contain only human proteins, the nonhuman Entrez Gene ID for each structure



	Rank	Identifier	Family	Subfamily	Ligand	Resolution
1	1	1v2l	Protease	Serine protease		1.6

Figure 6. Introduction of the PROLIX graphical user interface (GUI). (A) Screenshot of the query sketcher panel. The depicted user query comprises a benzamidine ligand molecule, six interacting residues and their interactions (dashed lines) to the ligand, two nonbonding shell residues, one excluded residue (crossed out) and two residue distance constraints (solid green lines). The ligand is sketched using the standard molecule drawing tools on the left panel. A list of the 20 standard amino acids is available clicking the "Residue..." button on the bottom. Interactions can be defined by dragging a line from the residue to the ligand. The type and the distance of the interaction can be defined in a pop-up menu. The interactions appear as color-coded dashed lines in the query panel. Distances between a pair of residues can be defined by dragging a line between them. The minimum and maximum distance cutoff values can be defined in a pop-up menu. Residue distance constraints appear as green lines between the constrained pair of residues. A residue can be defined as an "excluded residue" through a menu that opens when right-clicking on a residue. The database search is launched by pressing the "Start Search" button. (B) The results display panel is subdivided into three sections: (i) A navigation tree (top left) enables browsing through the interaction networks of database hits by protein family and subfamily; (ii) the interactive graphic display (top right) illustrates the conservation of individual protein−ligand interactions across different protein families (in comparison to the user query). The thickness of an interaction line represents the conservation of the respective interaction for the protein family selected in the navigation tree. (iii) A tabular ranked list provides a summary of the database hits (including protein family, 2D ligand images, X-ray-resolution) as a ranked list (bottom). The list is updated according to the protein family selected in the navigation tree. Clicking on residues or interactions in the graphic display will filter the list to those hits featuring the respective interaction or residue. Multiple interactions or residues can be selected at the same time. The hit list can be exported as a tab-separated text file using the "Export Hits" button.

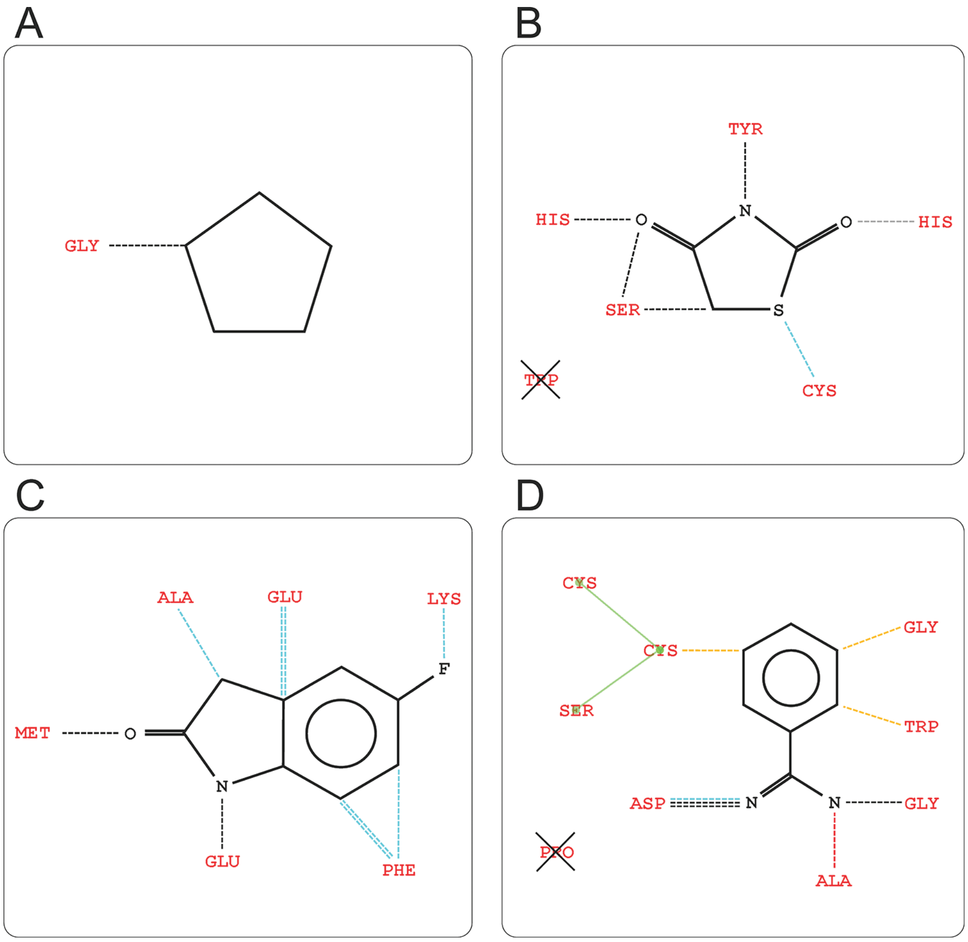


Figure 7. Example user queries used for runtime analyses of the PROLIX web application. The runtime performances of these queries are given in Table 1. (A) Simplistic query featuring only one prevalent residue (glycine) with just one common interaction (hydrogen bond). (B) More specific query with six interacting residues featuring different interaction types with some residues having multiple interactions to the ligand. (C) More restrictive query with five interacting residues and one excluded residue. (D) Very complex query with six interacting residues, two nonbonding residues, one excluded residue, four different interaction types and two residue distance constraints (green lines).

record is also converted to the corresponding human Entrez Gene ID using NCBI HomoloGene.[45] This conversion increases protein family assignment to the PDB structure entries. The process, starting from downloading public protein family data to the subsequent mappings, is automated as much as possible. Some missing information in the current public downloads is manually added using both public and in-house data sources and incorporated in the automated process.

■ RESULTS AND DISCUSSION

PROLIX is a novel tool for efficient mining of protein−ligand interactions in large crystal structure databases. The graphical user interface features a query sketcher that enables the definition of very specific interaction patterns between a ligand molecule and receptor residues (Figure 6). In addition to basic query features like defining different interaction types and distances to different receptor residues, the interaction fingerprints implemented in PROLIX are capable of handling more complex search functionality. The user can specify exact interaction counts (e.g., exactly 2 hydrogen bonds), define noninteracting residues or exclude residues from the query (e.g., no cysteine in the pocket). Furthermore, the tool allows the definition of geometric constraints to describe a spatial arrangement of residues in the receptor binding site. While the translation of complex user queries into the PROLIX fingerprints is straightforward, it proved to be challenging to develop a query database matching algorithm to handle several query residues of the same type featuring multiple constraints and potentially match multiple different residues in a target structure. The matching procedure described in this manuscript features a sophisticated procedure to enumerate all possible matching scenarios and determine the optimal matching with the respective target.

The abstraction of complex interaction patterns and distance constraints realized in PROLIX fingerprints allows an ultrafast matching against all public and in-house protein−ligand complexes available to Roche. The efficient mining is realized by our multistep branch-and-bound approach that identifies and discards nonmatching targets en masse at an early mining stage thus dramatically reducing the total number of matching operations to be performed. Additionally, the translation of query constraints into fingerprints enables the implementation as efficient matrix operations against precalculated target database fingerprints. We analyzed the database mining runtime and the total query runtime to evaluate the speed performance of the PROLIX web tool. The total query runtime describes the time span between the launch of a query in the query sketcher panel (Figure 6A) and the time the results with all data and images are displayed in the results panel of the GUI (Figure 6B). Total query runtimes vary depending on the complexity of the query and the amounts of data to be displayed for the identified hits in the results panel. Contrary to intuition, the most complex query definitions achieve the fastest runtimes. This is due to the fact that large parts of the target database can be discarded in the first filtering steps of the matching routine for very specific queries. We performed runtime analyses for four example queries as shown in Figure 7.

These query examples were chosen to highlight the relationship between the complexity of a user query and the corresponding runtime performance. A 65% partial residue matching cutoff was chosen for the runtime analyses of all queries, adding additional complexity to the database mining process. The total query runtimes and the fingerprint matching runtimes of the four user queries (Figure 7) are given in Table 1. Each database query was performed on a 2.4 GHz Xeon quad core processor accessing up to 64 gigabytes of memory. The results show that simplistic queries prove to be computationally more expensive than more complex queries: A query with only one prevalent interacting residue and just one interaction (Figure 7A) shows significantly increased runtimes compared to more complex queries (Figures 7B−D, Table 1). This is due to the fact that a comparably simplistic query will match a large number of target database entries and therefore create immense amounts of data to be mined. Furthermore, simplistic queries ultimately generate an increased number of matching database targets. Consequently, this results in extended total query runtimes, since more data and images have to be loaded into the GUI's results display. In contrast to simplistic queries, the total runtime for an average query (Figures 7B−D) is typically about two seconds (Table 1), whereas the actual runtime for the database mining process is considerably faster, achieving runtimes of about 150 ms or less (Table 1). However, even simplistic queries featuring only one query residue with just one interaction require only about five seconds of total query runtime. We do not consider these simplistic queries to be very meaningful to answer questions about conserved protein−ligand interaction patterns or binding site similarity. However, since these queries represent the computationally most expensive input, they are useful to demonstrate the speed of the database mining algorithms implemented in PROLIX. The runtime performance of only a few seconds (as a worst case scenario) for simplistic queries underscores the database mining capabilities of PROLIX, especially considering the fact that each query requires comparisons to hundreds of thousands of binding site residues, millions of protein−ligand interactions and large numbers of possible matching scenarios in every target complex. We consider this extremely fast data retrieval a key feature that is critical for the success of this mining tool that is frequently used by scientists in drug design.

Table 1. Runtimes for Example Queries in PROLIX[a]

query	database mining runtime	total query runtime
Figure 7A	1.9 s	5.2 s
Figure 7B	146 ms	2.1 s
Figure 7C	102 ms	1.9 s
Figure 7D	85 ms	1.9 s

[a] The input queries used for the runtime analysis are depicted in Figure 7. The table indicates the runtime for (i) the database mining process (including scoring and ranking of results) and (ii) the total query runtime. The total query runtime describes the complete runtime from launching the query until the time all data and images are loaded into the GUI results panel.

We used the protein−ligand complex definitions in Figure 7B−D as example queries to demonstrate the search capabilities of PROLIX. The three queries represent protein−ligand interactions found in complexes of different target classes. All interactions were defined between a fragment of the respective bound ligand and a number of residues in the binding site. The selected queries comprise: (i) an IL-2-inducible T cell kinase in complex with a fragment of the inhibitor sunitinib (Figure 7B, PDB 3miy[46]), (ii) a peroxisome proliferator activated receptor (PPAR) in complex with a fragment of rosiglitazone (Figure 7C, PDB 1zgy[47]), and (iii) a thrombin serine protease complexed with a benzamidine-fragment of the bound inhibitor (Figure 7D, PDB 2pks[48]). A partial residue matching cutoff of 65% was used in the database matching. A detailed summary of the retrieved database hits with focus on their protein family and subfamily classification is given in Table 2. The data shows that all queries were found as top ranked PDB hits. This is noteworthy since only a part of all binding site residues and interactions (defined by the respective ligand fragment of the actual bound ligand) was used in each query. In addition, a partial residue matching cutoff of 65% was applied in the matching process. For each of the three examples the respective protein family and subfamily classification of the query was found as the largest group in the hit list. This result was most pronounced in the thrombin query (Figure 7D) where 263 out of 276 hits were serine proteases. The high enrichment of proteases (273 out of 276 hits) can be explained by the fact that the thrombin query is the most restrictive of all

Table 2. Hit Statistics for Three Example Queries Illustrated in Figures 7B−D[a]

	3miy	1zgy	2pks
query classification	IL-2-inducible T cell kinase (tyrosine kinase)	LRH-1 nuclear receptor (PPAR)	thrombin (serine protease)
query ligand	sunitinib fragment (Figure 7B)	rosiglitazone fragment (Figure 7C)	benzamidine (fragment of 2pks ligand, Figure 7D)
top ranked PDB hit	3miy	1zgy	2pks
total hits	52	54	276
largest **protein family** (and subfamily) in hit list	**44 kinases** (38 tyrosine kinases, 4 CMGC kinases, 2 AGC kinases)	**33 nuclear receptors** (all PPAR)	**273 proteases** (263 serine proteases, 10 cysteine kinases)
other **protein families** (and subfamilies) in hit list	**3 ion channels** (all Ca_V ion channels), **2 enzymes** (all transferases), **2 binding proteins** (all calmodulin) **1 protease** (cysteine protease)	**21 enzymes** (11 oxidoreductases, 5 transferases, 3 lyases, 2 hydrolases)	**3 enzymes** (2 transferases, 1 epimerase)

[a] The table summarizes the **protein family** (and subfamily) classification of the identified target hits. The hit set comprises public as well as Roche in-house structures. Kinases and proteases are promoted from the level of subfamilies in the enzyme family to the protein family level.

three example queries. The protein family and subfamily classification of the matching database hits are indicated in the results table and the browsing tree in the results panel of the PROLIX GUI (Figure 6B). Since proteases and kinases represent large protein families with extensive subfamily branches, we decided to promote the two groups from the level of protein subfamilies within the enzyme family to the top level of protein family. This solution provides the user with more information about kinase or protease subfamilies and represents additional information that would be lost otherwise.

We designed the PROLIX tool to be easily used by both experts, as well as users with limited computer skills. Therefore, we decided to offer PROLIX as a web application that features an easy-to-use 2D query sketcher (an applet provided by CCDC[36]). The frontend query sketcher communicates the user query input and matching result output to and from the backend matching algorithms through XML files. This component-based application design facilitates independent development and easier modification of each component. For example, in future versions, we plan to offer a 3D molecular viewer as an input interface to allow expert users to define query interaction patterns directly from an X-ray structure. The integration of this new 3D input dialogue is straightforward, since this frontend query component can be modified independently without affecting the backend matching routine.

We considered introducing the concept of residue groups (e.g., acidic, aromatic, basic, hydrophobic, positively charged, negatively charged) defined by similar physicochemical properties. These residue groups can be easily annotated in the fingerprints as super residues. However, this approach would dramatically increase the number of valid matchings in the target receptor (consider the various possibilities to match a query of five "hydrophobic" super residues in a target binding site). Additionally, the results would be difficult to interpret and visualize. Consequently, we chose not to implement these functionalities.

CONCLUSIONS AND OUTLOOK

In this manuscript, we present PROLIX, a new method to rapidly mine protein−ligand interaction patterns in large crystal structure databases. Protein−ligand complex structures are a treasure trove that can be mined to gain invaluable insight into structure-based drug design. We have highlighted some of the key elements and challenges to develop an easy-to-use protein−ligand interaction mining tool for medicinal chemists. This tool provides knowledge about the preferred protein−ligand interactions for specific chemo-types. In addressing some of these challenges, we have taken a pragmatic approach of developing a software tool that balances simplicity (ease-of-use) and functionality. These include the design of a graphical user interface that enables the definition of complex queries and an advanced summarization of results. Furthermore, we developed a sophisticated multistep fingerprint-matching algorithm that enables fast search retrieval against large databases of precalculated target fingerprints. PROLIX was designed to answer questions that arise in daily drug design research: Which complexes in the PDB feature a protein−ligand interaction network similar to my query? How are these interaction patterns conserved in different protein families? PROLIX answers these questions, gives insight into interaction networks that explain activity and selectivity, and highlights early safety issues. Furthermore, the tool can be used as an idea generator as the ligands of the identified target matches present new structural features that can be exploited for drug design.

Future work will focus on adding enhanced functionality to the PROLIX mining capabilities. We plan to extend the fingerprint design and matching algorithms to also handle interactions bridged by water molecules or metals. A priority for a future version is to implement the differentiation between backbone and side chain interactions, which is currently not supported. As a further enhancement, we intend to enable the structural alignment of database hits that match the user query. The user would therefore select several hits displayed in the results table of the GUI for structural alignment which would be presented in a molecular viewer such as PyMOL.[49] Another valuable enhancement comprises the addition of a method to filter the database hits in the results table by X-ray resolution. We intend to incorporate these features in future versions of the software.

PROLIX is presently available to Roche scientists and is being beta tested, however in the future we do not rule out a possibility of making the tool available to the public.

AUTHOR INFORMATION

Corresponding Author

*Phone: +1 (973) 235-6809. Fax: +1-(973)-235-6084. E-mail: Martin.Weisel@roche.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Roche Postdoc Fellowship (RPF) program. The authors thank Dr. Simon Cottrell, Dr. Simon Bowden, Dr. Ian Bruno and Dr. John Liebeschuetz for the development of the graphical user interface of PROLIX. We thank Dr. Neil Taylor from Desert Scientific Software for help using the Proasis database. We thank Dr. Luis Tari, Dr. Bernd Kuhn and Dr. Martin Stahl for helpful discussions. Also, we would like to thank Donovan Makund for integrating the GUI applet with the PROLIX search engine on the web server and Eric Scott for his help on implementing the protocols to generate the protein−ligand fingerprints used in PROLIX.

REFERENCES

(1) Sharff, A.; Jhoti, H. High-throughput crystallography to enhance drug discovery. *Curr. Opin. Chem. Biol.* **2003**, *7*, 340−345.

(2) Hajduk, P. J.; Gerfin, T.; Boehlen, J. M.; Haberli, M.; Marek, D.; Fesik, S. W. High-throughput nuclear magnetic resonance-based screening. *J. Med. Chem.* **1999**, *42*, 2315−2317.

(3) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(4) Böhm, H.-J.; Schneider, G., *Protein−Ligand Interactions from Molecular Recognition to Drug Design*; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2003.

(5) Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein−ligand interaction. *Proteins* **2002**, *49*, 457−471.

(6) Babine, R. E.; Bender, S. L. Molecular recognition of protein−ligand complexes: Applications to drug design. *Chem. Rev.* **1997**, *97*, 1359−1472.

(7) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.* **2009**, *49*, 1079−1093.

(8) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding affinities for protein-ligand complexes with

known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.

(9) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.

(10) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114−2125.

(11) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): A novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337−344.

(12) Kelly, M. D.; Mancera, R. L. Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1942−1951.

(13) Mpamhanga, C. P.; Chen, B.; McLay, I. M.; Willett, P. Knowledge-based interaction fingerprint scoring: a simple method for improving the effectiveness of fast scoring functions. *J. Chem. Inf. Model.* **2006**, *46*, 686−698.

(14) Perez-Nueno, V. I.; Rabal, O.; Borrell, J. I.; Teixido, J. APIF: A new interaction fingerprint based on atom pairs and its application to virtual screening. *J. Chem. Inf. Model.* **2009**, *49*, 1245−1260.

(15) Tan, L.; Lounkine, E.; Bajorath, J. Similarity searching using fingerprints of molecular fragments involved in protein-ligand interactions. *J. Chem. Inf. Model.* **2008**, *48*, 2308−2312.

(16) Tan, L.; Bajorath, J. Utilizing target−ligand interaction information in fingerprint searching for ligands of related targets. *Chem. Biol. Drug Des.* **2009**, *74*, 25−32.

(17) Tan, L.; Vogt, M.; Bajorath, J. Three-dimensional protein-ligand interaction scaling of two-dimensional fingerprints. *Chem. Biol. Drug Des.* **2009**, *74*, 449−456.

(18) Singh, J.; Deng, Z.; Narale, G.; Chuaqui, C. Structural interaction fingerprints: a new approach to organizing, mining, analyzing, and designing protein-small molecule complexes. *Chem. Biol. Drug Des.* **2006**, *67*, 5−12.

(19) Chuaqui, C.; Deng, Z.; Singh, J. Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening. *J. Med. Chem.* **2005**, *48*, 121−133.

(20) Deng, Z.; Chuaqui, C.; Singh, J. Knowledge-based design of target-focused libraries using protein-ligand interaction constraints. *J. Med. Chem.* **2006**, *49*, 490−500.

(21) Sato, T.; Honma, T.; Yokoyama, S. Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening. *J. Chem. Inf. Model.* **2010**, *50*, 170−185.

(22) Schreyer, A.; Blundell, T. CREDO: A protein-ligand interaction database for drug discovery. *Chem. Biol. Drug Des.* **2009**, *73*, 157−167.

(23) Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50*, 123−135.

(24) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(25) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387−406.

(26) Shulman-Peleg, A.; Shatsky, M.; Nussinov, R.; Wolfson, H. J. Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC Biol.* **2007**, *5*, 43.

(27) Konc, J.; Janezic, D. Protein−protein binding-sites prediction by protein surface structure conservation. *J. Chem. Inf. Model.* **2007**, *47*, 940−944.

(28) Carl, N.; Konc, J.; Janezic, D. Protein surface conservation in binding sites. *J. Chem. Inf. Model.* **2008**, *48*, 1279−1286.

(29) Shulman-Peleg, A.; Shatsky, M.; Nussinov, R.; Wolfson, H. J. MultiBind and MAPPIS: Webservers for multiple alignment of protein 3D-binding sites and their interactions. *Nucleic Acids Res.* **2008**, *36*, W260−W264.

(30) Angaran, S.; Bock, M. E.; Garutti, C.; Guerra, C. MolLoc: A web tool for the local structural alignment of molecular surfaces. *Nucleic Acids Res.* **2009**, *37*, W565−W570.

(31) Konc, J.; Janezic, D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **2010**, *26*, 1160−1168.

(32) *Proasis*, version 3; Desert Scientific Software: Castle Hill, New South Wales, Australia, 2011.

(33) Kuhn, B.; Fuchs, J. E.; Reutlinger, M.; Stahl, M.; Taylor, N. R. Rationalizing tight ligand binding through cooperative interaction networks. *J. Chem. Inf. Model.* **2011**, *51*, 3180−3198.

(34) *Pipeline Pilot*, version 8.0; Accelrys: San Diego, CA, 2011.

(35) Hicklin, J.; Moler, C.; Webb, P.; Boisvert, R. F.; Miller, B.; Pozo, R. *JAMA: A Java matrix package*, version 1.0.2; National Institute of Standards and Technology: Gaithersburg, MD, 2005.

(36) Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, United Kingdom.

(37) Maglott, D.; Ostell, J.; Pruitt, K. D.; Tatusova, T. Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Res.* **2005**, *33*, D54−D58.

(38) Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **2003**, *31*, 365−370.

(39) Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. UniProt: The Universal Protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, D115−D119.

(40) Bairoch, A.; Apweiler, R.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2005**, *33*, D154−D159.

(41) The kinase database at Sugen/Salk. http://www.kinase.com/kinbase (accessed November 2011).

(42) Sharman, J. L.; Mpamhanga, C. P.; Spedding, M.; Germain, P.; Staels, B.; Dacquet, C.; Laudet, V.; Harmar, A. J. IUPHAR-DB: New receptors and tools for easy searching and visualization of pharmacological data. *Nucleic Acids Res.* **2011**, *39*, D534−D538.

(43) Rawlings, N. D.; Barrett, A. J.; Bateman, A. MEROPS: The database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **2011**.

(44) Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **2000**, *28*, 304−305.

(45) Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; DiCuccio, M.; Edgar, R.; Federhen, S.; Geer, L. Y.; Kapustin, Y.; Khovayko, O.; Landsman, D.; Lipman, D. J.; Madden, T. L.; Maglott, D. R.; Ostell, J.; Miller, V.; Pruitt, K. D.; Schuler, G. D.; Sequeira, E.; Sherry, S. T.; Sirotkin, K.; Souvorov, A.; Starchenko, G.; Tatusov, R. L.; Tatusova, T. A.; Wagner, L.; Yaschenko, E. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **2007**, *35*, D5−D12.

(46) Kutach, A. K.; Villasenor, A. G.; Lam, D.; Belunis, C.; Janson, C.; Lok, S.; Hong, L. N.; Liu, C. M.; Deval, J.; Novak, T. J.; Barnett, J. W.; Chu, W.; Shaw, D.; Kuglstatter, A. Crystal structures of IL-2-inducible T cell kinase complexed with inhibitors: insights into rational drug design and activity regulation. *Chem. Biol. Drug Des.* **2010**, *76*, 154−163.

(47) Li, Y.; Choi, M.; Suino, K.; Kovach, A.; Daugherty, J.; Kliewer, S. A.; Xu, H. E. Structural and biochemical basis for selective repression of the orphan nuclear receptor liver receptor homolog 1 by small heterodimer partner. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 9505−9510.

(48) Blomberg, D.; Fex, T.; Xue, Y.; Brickmann, K.; Kihlberg, J. Design, synthesis and biological evaluation of thrombin inhibitors based on a pyridine scaffold. *Org. Biomol. Chem.* **2007**, *5*, 2599−2605.

(49) *The PyMOL Molecular Graphics System*, version 1.3; Schrödinger, LLC: New York, NY, 2011.