Identification of Potential Small Molecule Peptidomimetics Similar to Motifs in Proteins

Ivan Baran,[†,§] Radka Svobodova Varekova,[‡] Laavanya Parthasarathi,[§] Simon Suchomel,[‡]
Fergal Casey,[§] and Denis C. Shields*[,§]

Siemens Research Ireland, Dublin, Ireland, ANF Data, a Siemens Company, Brno, Czech Republic, and UCD
Conway Institute of Biomolecular and Biomedical Sciences, University College Dublin, Belfield,
Dublin 4, Ireland

Received September 18, 2006

Protein−protein interactions are central to most biological processes and represent a large and important class of targets for human therapeutics. Small molecules containing peptide substituents may mimic regions of interacting proteins and inhibit their interactions. We set out to develop efficient methods to screen for similarities between known peptide structures within proteins and small molecules. We developed a method to rank peptide-compound similarities, that is restricted to small linear motifs in proteins, and to compounds containing amino acid substituents. Application to a search of the PubChem database (5.4 million compounds) using all short motifs on accessible surface areas in a nonredundant set of 11 488 peptides from the protein structure database PDB demonstrated the feasibility of the method for high throughput comparisons and the availability of compounds with comparable substituents: over 6 million compound−peptide pairs shared at least three amino acid substituents, ∼100 000 of which had an rmsd score of less than 1 Å. A Z-score function was developed that compares matches of a compound to different instances of the peptide motif in PDB, providing an appropriate scoring function for comparison among peptide-compound similarities involving different numbers of atoms (while simultaneously enriching for similarities that are likely to be more specific for the protein of interest). We applied the method to searches of known short protein motifs against the National Cancer Institute Developmental Therapeutic Program compound database, identifying a known true positive.

INTRODUCTION

Protein−protein interactions play a key role in most biological processes and therefore represent a large and important class of targets for human therapeutics,[1] including not only those interactions involved in mediating signals in pathways but also those required for protein modifications (such as protein kinases, protein phosphatases, glycosyl transferases, acyl transferases, proteases). These interactions regulate fundamental processes such as cell growth, cell cycle, metabolic pathways, and signal transduction. Inhibition or activation of protein−protein binding interactions have been widely studied but represents a challenging problem. One of the goals is the discovery of new compounds, which control these processes more effectively.

Natural inhibitors or activators of protein−protein interactions are native biologically active short peptides structurally similar to the ligand's interacting motif or region. Although biologically active short peptides have a great potential for medical applications, they often need to be modified to overcome their poor pharmacological properties, such as poor bioavailability, high susceptibility to enzymatic degradation, and limited cellular penetration. Accordingly, replacement of these peptides by effective peptidomimetic compounds is highly desirable but for an individual peptide can represent a very considerable effort.[2] Over past decades various computational as well as experimental screening methods for rational peptidomimetics design have been developed.[3]

Despite many achievements in the development of peptidomimetics from bioactive peptides,[4,5] it remains a nontrivial problem. In practice, peptidomimetics come either from rational design to modify a starting peptide sequence or else are obtained from high throughput compound screens.[6] In this paper, we are interested in whether computational screening for peptidomimetic compounds (both peptidelike and those only sharing the side chain substituents) may be feasibly performed on a wide variety of protein regions, to define more limited compound sets for screening. Rather than invest a large effort into the biological identification of oligopeptides that inhibit the interaction between proteins, and then subsequently invest a large effort in the identification of peptidomimetics for the individual peptides of interest, an alternative strategy may be to identify and screen a large set of compounds that mimic potential interaction regions of proteins. Screening this subset of compounds for activity (in place of screening peptides) has potential to accelerate the discovery of drugs that target protein−protein interactions.

Here, we identify an efficient workflow for such a large-scale screening. This relies on identifying small molecules with substituents that topologically and structurally resemble key amino acid side chains. We demonstrate the utility of this workflow, by comparing a large compound database to a large database of peptide regions from the surfaces of protein structures, and illustrate its application to known short linear motifs that mediate protein−protein interactions.

* Corresponding author phone: +353-1-716 6831; fax: +353-1-716 6701; e-mail: denis.shields@ucd.ie.
[†] Siemens Research Ireland.
[‡] ANF Data.
[§] University College Dublin.

10.1021/ci600404q CCC: $37.00 © 2007 American Chemical Society
Published on Web 02/20/2007

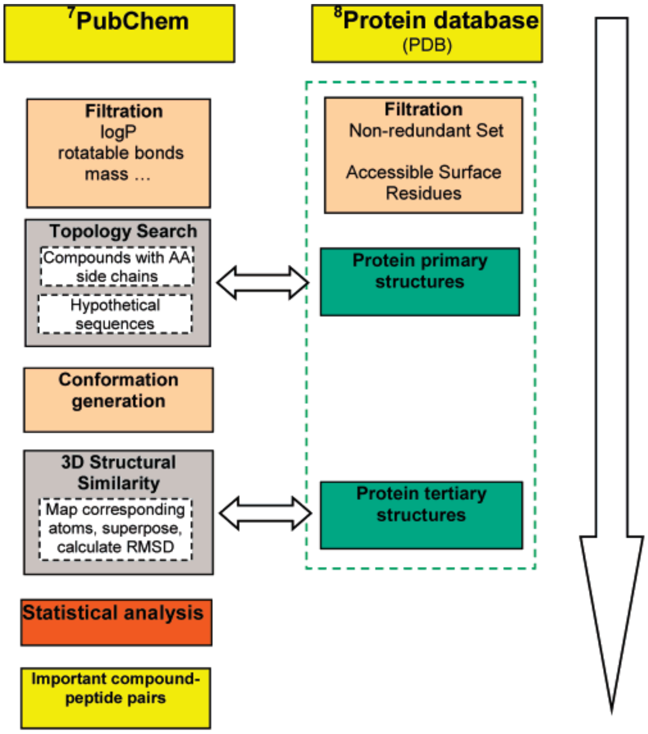


Figure 1. Simplified graphical representation of the workflow. Double headed arrows indicate the two points where comparisons are made between compound libraries and the peptide structure database. The algorithm identifies in a filtered library of compounds those with substituents similar to amino acids side chains and generates from them hypothetical amino acid sequences. Each substituent found in a compound is represented by a single letter abbreviation and each hypothetical sequence by a regular expression. Next, primary structures of peptide chains in FASTA format are scanned by simple regular expression matching to identify compound−peptide pairs with shared motifs. Thereafter, the workflow calls an external program to generate conformers only for compounds with detected shared motifs. Finally, corresponding atoms are mapped and structures are superposed. Compounds with outstanding structural similarity are potential peptidomimetic candidates of the corresponding peptide.

METHODOLOGY

The method has four main steps: (1) identification of compound substituents that are similar to amino acid side chains, (2) generation of hypothetical "sequences" from all possible combinations of these substituents, (3) a regular expression search of the hypothetical sequences in a database of primary amino acid sequences, and (4) superposition and estimation of the root-mean-square deviation (rmsd) between the known structures of the amino acid sequence taken from the PDB protein structure database and the modeled conformations of the compound. The first three steps correspond to a search for 2D topological similarity, while the final step then determines the 3D structural similarity.

Input, Output, and Workflow. The input is a file containing multiple compounds in MDL SD file format, a text file listing the filenames of PDB-formatted protein structure files, and a directory containing the PDB files. The output is a list of all small molecule−peptide pairs with their associated RMSDs, the number of atoms superposed, and the amino acid sequence contributing to the comparison, along with ancillary information such as the number of structural conformations provided for that compound. The workflow that enables this is shown in Figure 1, and each step of this is detailed below in the particular application to comparing PubChem[7] compounds to the PDB database.[8]

Identifying Amino Acid Substituents in Compounds. Despite the existence of numerous methods of fragment mining[9] we decided for reasons of efficiency to develop our own algorithm for searching subfragments similar to amino acid side chains. The algorithm first breaks molecular graphs into subgraphs. Every subgraph connected by a single nonring bond to the rest of the molecule is considered as a "substituent". The connection atom between the substituent and the scaffold is named "substituent C_α" (later "sC_α"). Only substituents of up to 10 non-hydrogen atoms (the size of tryptophan's side chain) are stored. Overlapping subgraphs are merged and represented by the largest one. Those substituents of compounds that resemble amino acids are then identified, using an approach similar to binary fingerprints. Amino acid side chains in various tautomeric forms and protonation states are considered. An example of a compound with substituents similar to amino acid side chains is given in Figure 2. Since similarities to substituents of amino acid side chains are considered, it is possible to detect small active substituents that are responsible for biochemical activity of amino acids. For example, finding the entire glutamine side chain in small molecules is exceptional, but occurrences of the amide groups are frequent. The current version of the algorithm recognizes 17 amino acids side chains and their subparts. All mined fragments are shown in Table 1. Considering tautomeric states, this comprises a total of 42 possible substituents. Glycine is not considered, since it has no side chain, alanine is excluded, since its simple side chain is very common, and proline is omitted since the current implementation of the algorithm does not handle its iminoacid structure.

Definition of Amino Acid Like Substituent Containing Compounds by Hypothetical Sequence Generation. The algorithm controls the number of created hypothetical sequences with the aid of four major parameters: (i) the minimal number of amino acid substituents (here set at 3), (ii) the hypothetical sequence length (minimal and maximal, set at 3 and 5, respectively), (iii) the maximal allowed graph distance between two sC_α atoms (set at 7), and (iv) the number of allowed gaps in the hypothetical sequence (set at 1, see below). The program calculates graph distances among subfragment's sC_α atoms by using modified Dijstra shortest path algorithm.[10] This parameter has a critical influence on the number of potential sequences identified. Additional parameters include controlling the maximal allowed number of sequences per compound. Since compounds in libraries that are derived from peptides may contain many amino acid fragments, this can result in a combinatorial explosion of an excessively large number of hypothetical sequences. We eliminated the compounds that had greater than 100 000 sequences per compound. Note that the hypothetical sequence generation step can be bypassed when we seek compound matches to known motifs (involved in protein−protein interactions, for example) or a particular subclass of motif definitions. In that case, both the computational effort is reduced, and we can relax the restrictions on motif length and numbers of gaps. We will provide examples of such applications below.

Search for Shared Motifs in Proteins. In the final phase of 2D topological comparison the program searches for hypothetical shared motifs between compounds and primary

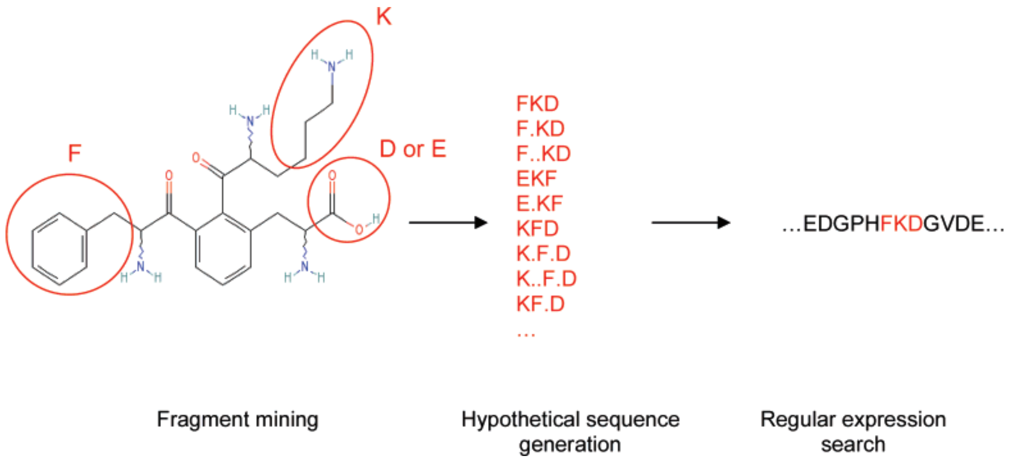


Figure 2. Graphical representation of the topology search for small molecule (CID 3017427). The process consists of three main steps: (1) identification of compound substituents that are similar to amino acid side chains, (2) generation of all possible combinations of these substituents into hypothetical "sequences", and (3) a regular expression search of the hypothetical sequences in a database of primary amino acid sequences.

structures of proteins to create corresponding pairs. Each amino acid substituent is represented by its one letter abbreviation. Subsequently the algorithm generates all possible combinations of these letters with or without gaps for each compound, represented as simple FASTA sequences. The number of potential gaps between two neighboring substituents (similar to amino acid side chains) was set at 1. Table 2 illustrates the great increase in hypothetical sequences if a larger gap is permitted: we chose a maximum gap size of one in order to reduce the false positive rate. After sequence generation, a simple regular expression search finds identical matches between the hypothetical sequences and peptide subsequences in the FASTA formatted primary sequences derived from the PDB database.

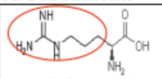
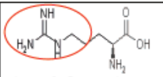
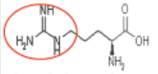
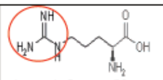
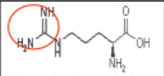
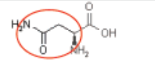
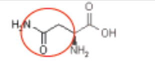
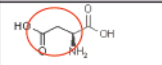
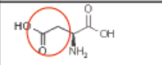
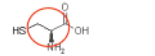
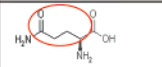
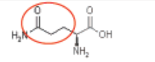
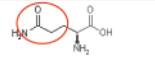
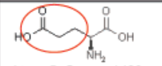
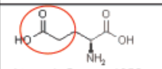
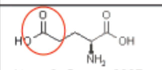
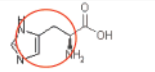
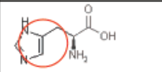
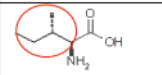
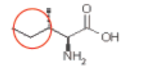
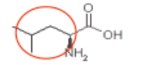
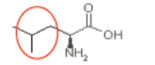
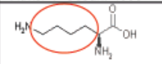
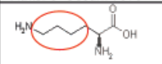
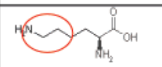
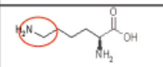
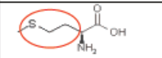
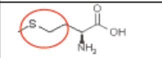
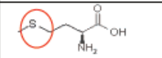
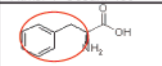
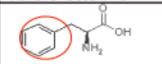
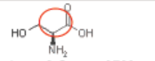
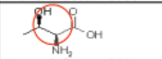
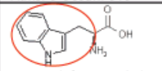
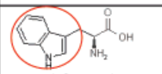
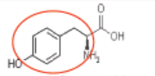
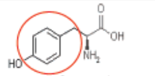
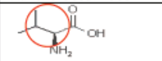
Structural Superposition. In order to take into account molecular flexibility, the workflow calls an external program to generate 3D conformations of compounds with shared patterns. The core of 3D structural comparison is superposition of compounds represented by a set of their conformers and proteins with shared motifs on their matched side chains. The algorithm first extracts shared residues from the PDB protein structure file and maps corresponding atoms from both structures to create fixed and moving sets of points for superposition. The algorithm checks for each atom its type and graph position in the substituent and residue to ensure ideal atom mapping. This approach yields an equal number of atoms in both sets of points, even if there are incomplete substituents similar to a subpart of amino acid side chain. The mapping is illustrated in Figure 3. The superposition is a pure geometrical optimization problem, being simply the process of moving and rotating two sets of points in space to minimize their geometrical distances.[11] The quality of superposition is measured by rmsd. During superposition the program automatically checks all possible combinations of mappings among atoms in specific situations, such as ambiguous substituents with similar active side-chain fragments (e.g., valine and leucine, aspartic and glutamic acid) or ambiguous atoms in specific chemical groups (carboxyl group), to choose the best hit with the lowest rmsd.

Data Analysis. The following statistical characteristics were estimated for each compound−protein pair, across all compound conformers: the number of superposed atoms, the best rmsd, the average rmsd, the number of rmsd lower than certain limit, and two Z-scores. Z-scores are calculated in the standard way as the value minus the mean, divided by the standard deviation. The first Z-score, here termed local (lZ-score), determines how significantly different the best hit is, compared to the rest of the rmsd values for all other conformers of that particular compound matching that peptide. The second Z-score, here termed global (gZ-score), determines how significantly different the best hit for a particular compound is compared to the rest of the best hits for different instances of the motif in PDB with the same hypothetical sequence. The second criterion has practical sense only during analysis of larger compound data sets, but it is one of the best for identification of significant structural similarities from the large set of compound−protein pairs. The meaning of the statistical scores is elucidated in the practical example below.

Implementation. The framework and many of the basic applications and scripts were written in Python,[12] while computationally intensive modules (such as substituents search, hypothetical sequence generation, or mapping of corresponding atoms) are written in C/C++. BioPython's Structural Bioinformatics module[13] is used for superposition. The workflow is platform independent and realized in three versions: single CPU, multi-CPU with shared memory, and cluster version. The size of job dictates which version is optimal. The analysis presented here was performed using the cluster version. The program splits input compound database into subfiles and distributes them to individual nodes. The core calculation for each subset is performed independently to the rest as new process. The master process controls the runs of the slave processes. Finally, all result files are merged and analyzed to obtain global statistical scores. In practice, the three most critical factors in determining computational intensity are the number of generated hypothetical sequences, the number of compound−protein pairs detected, and the number of allowed generated conformers per compound.

External programs are an indivisible part of the workflow. Our implementation utilizes for filtration and conformer generation the chemistry package Molecular Operating Environment (MOE) from Chemical Computing Group Inc.[14]

Table 1. Number of Identified Substituents in the PubChem Database Similar to Partial or Complete Amino Acids Side Chains[a]

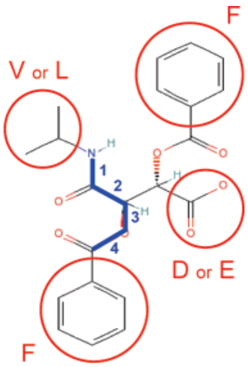
Amino Acids		Mined Substituents					Total
Arginine	R	 Atms: 7 Occur.: 239	 Atms: 6 Occur.: 1	 Atms:5 Occur.: 8	 Atms: 4 Occur.: 58	 Atms: 3 Occur.: 339	645
Asparagine	N	 Atms: 4 Occur.: 693	 Atms: 3 Occur.: 3351				4041
Aspartic acid	D	 Atms: 4 Occur.: 1952	 Atms: 3 Occur.: 6837				8789
Cysteine	C	 Atms: 2 Occur.: 139					139
Glutamine	Q	 Atms: 5 Occur.: 103	 Atms: 4 Occur.: 693	 Atms: 3 Occur.: 3351			4147
Glutamic acid	E	 Atms: 5 Occur.: 1480	 Atms: 4 Occur.: 1952	 Atms: 3 Occur.: 6837			10269
Histidine	H	 Atms: 6 Occur.: 148	 Atms: 5 Occur.: 3				151
Isoleucine	I	 Atms: 4 Occur.: 1359	 Atms: 2 Occur.: 52749				54108
Leucine	L	 Atms: 4 Occur.: 4008	 Atms: 3 Occur.: 7970				11978
Lysine	K	 Atms: 5 Occur.: 209	 Atms: 4 Occur.: 40	 Atms: 3 Occur.: 158	 Atms: 2 Occur.: 135		542
Methionine	M	 Atms: 4 Occur.: 469	 Atms: 3 Occur.: 79	 Atms: 2 Occur.: 2			550
Phenylalanine	F	 Atms: 7 Occur.: 17569	 Atms: 6 Occur.: 38551				56120
Serine	S	 Atms: 2 Occur.: 3703					3703
Threonine	T	 Atms: 3 Occur.: 815					815
Tryptophan	W	 Atms: 10 Occur.: 1258	 Atms: 9 Occur.: 247				1505
Tyrosine	Y	 Atms: 8 Occur.: 906	 Atms: 7 Occur.: 877				1783
Valine	V	 Atms: 3 Occur.: 7970					7970

[a] Occur. = number of occurrences.

Other external programs can be used without restrictions, because the communication to workflow is the standard MDL SD file import/export operation.

Compound Library Filtration and Conformation Generation. PubChem,[7] curated by U.S. National Center for Biotechnology Information (NCBI), is a large freely available database, which provides information on the structure as well as the biological activities of small molecules. We took the PubChem database version May 2006. To concentrate on small compounds with desirable chemical properties we filtered out molecules with molecular weight greater than 700, with a calculated LogP greater than 6 or less than −4, with greater than 6 hydrogen-bond donors, with greater than 11 hydrogen-bond acceptors, and with greater than 15 rotatable bonds. This is similar to standard procedures used in assembling of cheminformatics databases.[15] We used MOE's conformation import option, which constructed conformers using a parallelized fragment-based approach: molecules are subdivided into overlapping fragments each of which is subjected to a rigorous stochastic search. The fragment conformations are rapidly assembled by superposing the overlapping atoms. In order to reduce the computational requirements, and possibly to somewhat improve the signal-to-noise ratio, we limited the number of conformations

Table 2. Example of the Influence of Various Parameters Settings on the Number of Generated Hypothetical Sequences for Regular Expression Searching[a]



maximal graph distance	4	10	4	10	4	10
maximal allowed gaps	0	0	1	1	2	2
examples of sequences	FDF, FEF, FDFV, FDFL, FEFL, VFDF, DVF, FDF, ...		F.DF, F.DF.V, VFD.F, D.V.F, F.D.F, ...		F..D.F, F..DF.V, VFD..F, D..V.F, F..D..F, ...	
number of sequences[b]	172	600	1280	5376	4644	21384

[a] The possible number of substituents is between 3 and 4 for all parameter combinations. The graph distance between two fragments is highlighted in blue. [b] Total number of generated hypothetical sequences in FASTA format.

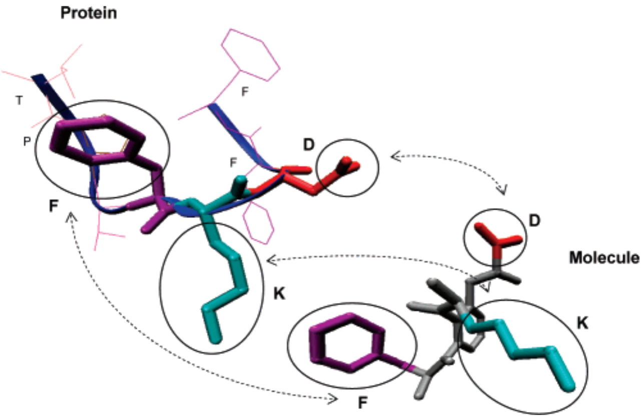


Figure 3. Mapping of corresponding atoms for substituents similar to amino acid side chains in proteins. The algorithm first extracts shared residues from the PDB file. Thereafter, the algorithm checks for each atom its type and graph position in the substituent and residue to ensure ideal atom mapping and constructs fixed sets of atoms for superposition.

to 50 per compound. Conformations were modeled in vacuum. The failure limit was increased to 100. Merck molecular force field MMFF94 was chosen for molecular mechanics calculations. MOE processed 40 465 compounds and generated 10 437 408 conformers. Twenty lowest energy conformers per compound were selected for further analysis. We note that despite the many developed methods effective over past decades, calculation of molecular conformation remains an active area of research in the field,[16] and clearly many other strategies are possible.

Protein Data Preparation. PDB[8] is the worldwide repository for processing and distribution of 3-D structure data of proteins and nucleic acids. The objective was to determine the solvent accessible regions of proteins from a nonredundant version of the PDB database. The PDB version January 2006 contained 34 348 proteins, comprising 69 952 peptide chains. These have been represented as alternative nonredundant sets, where all chains were structurally aligned and divided into 11 488 clusters for 90% similarity, 10 055 clusters for 70% similarity, and 8623 clusters for 50%. Clustering is performed by CD-HIT algorithm.[17] Each cluster is represented by the highest ranked member. Since the number of clusters was not remarkably increased by selecting those with 90% similarity, we chose this data set for further analysis; many of the regions for comparison on the surface of the proteins are likely to adopt alternative conformations, which are of interest in comparing flexible peptides to compounds. Accessible Surface Area (ASA) for each residue was calculated and normalized according to the maximal possible ASA for a particular amino acid type, using the standard DSSP program.[18] Thereafter the preparation program generated new primary structures in FASTA format, masking out as X all residues which did not have a normalized ASA above 0.5 (An example of part of a generated sequence is XXXXREDWXXXXDRAXXXXX.).

RESULTS AND DISCUSSION

Application to High-Throughput Comparison of Many Short Motifs on Protein Surfaces to a Large Compound Database. The PubChem database version May 2006 contained 5 499 083 molecules, and after filtration this was reduced to 4 607 561 small molecules. 46 495 compounds were identified with 3 and more substituents, representing 0.84% of all compounds collected in the PubChem database. 87.64% of these compounds contained 3 substituents, 11.36% compounds contained 4 substituents, and the rest of compounds (<1.0%) contained 5 or more substituents. The maximal number of substituents found in a compound was 8 in a small symmetric molecule, but all of these comprised the simplest 2-carbon substituent of isoleucine. Of greater interest are the compounds containing a number of different types of substituents per compound, since these have a greater chance of matching important hypothetical sequences of potential peptidomimetics. A summary of the identified substituents similar to amino acid side chains or their

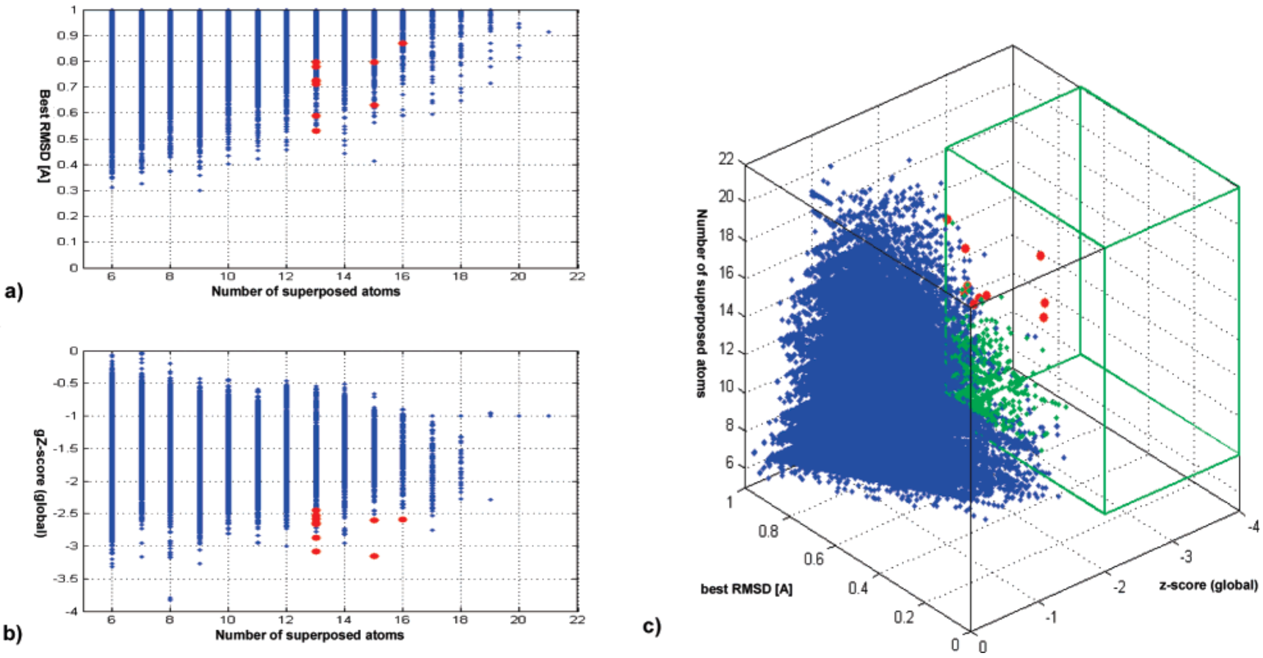


Figure 4. Statistical distributions of compound−protein pairs with shared motifs. Distribution of 98 838 pairs where rmsd of the best hit <1.0 A. All red points represent selected significant compound−protein pairs listed in Table 3. (a) Distribution of the best hit values for each compound against the number of superposed atoms. (b) Dependency between gZ-score (global) and number of superposed atoms. (c) Green box represents 334 pairs lying in the volume of main interest with the following constraints: best hit rmsd <0.7 A, gZ-score <−2.0, and number of superposed atoms >8.

functional subparts for all compounds with at least 3 fragments are shown in Table 1.

Simple statistical and visual inspections of the fragments and their occurrences, led to the following observations: 1. Compounds with a large number of substituents were typically short peptides or peptidomimetics. 2. Most of the compounds (99%) contained 3 or 4 substituents. 3. The graph distance is usually higher than the “optimal” distance of 3, which corresponds to graph distance between two nearest sC_{α} atoms in peptide backbone. 4. The distribution of chemical groups is in general agreement with the distributions described in compound libraries by other researchers.[9,19] For example, common groups include the phenyl ring (phenylalanine, 56 120 occurrences) and carboxyl group (aspartic acid, 8789 occurrences or glutamic acid, 10 269 occurrences); the simple two carbon atoms substituent of isoleucine is represented by 52 749 occurrences. The most infrequent substituents are histidine with 151 occurrences and cysteine with 139 occurrences.

The program generated 541 287 hypothetical sequences, with an average of 12 per compound. The highest number of generated hypothetical sequences was 3696 per compound. The most frequently generated hypothetical sequences were typically combinations of the most populated substituents. All variations (inclusive of gaps, ambiguous amino acids, and alternative orderings of residues) of these sequences were represented by 41 177 152 motifs in FASTA format in total. 6 813 648 compound−protein pairs with shared motifs were found by a regular expression search of the surface exposed regions of the nonredundant peptide data set.

Finally, the program superposed 6 367 896 compound−protein pairs. The number of superposed pairs are lower than the total number of pairs identified by 2-D search (with losses at about 7%). These defects are caused by three sources of errors: (i) problems in conformer computation, (ii) breaks in peptide chains (PDB annotation issues), and (iii) most commonly, absence of 3D structure for one or more of the residues.

The initial filter applied to this data set was to restrict to pairs where the rmsd of the best matching conformer was lower than 1.0 A. This preliminary filtration decreased the investigated set to 98 838 compound−peptide pairs with shared motifs and potential significant structural similarity. This subset represents 1.55% of all calculated pairs. Statistical characteristics such as gZ-score or number of conformers lower than the prescribed limit were calculated within this subdata set, making these calculations quicker, but somewhat stricter in the sense that much higher gZ-score values would have been obtained had the entire data set been used.

While the rmsd value between the two structures can be considered a criterion for judging the quality of a superposition, it does not result in a fair assessment, since the number of superposed atoms considered in the rmsd computation may vary significantly from one compound−protein pair to the next (between 6 and 21 atoms in this data set). This dependency is illustrated in Figure 4a. However, in addition to this simple dependency on the number of atoms, there are more subtle dependencies arising from the molecular constraints on substituents: those that are more flexible will tend to have better matches for a particular conformer. We therefore considered how the matches for the same substituent combinations elsewhere in the data set might provide an effective “control” to modify the scoring function appropriately. To achieve this, we calculated a gZ-score that contrasts the rmsd value for a compound compared to the peptide to the set of rmsd values for that compound compared to all other equivalent instances of that peptide in the data set. The gZ-score should correct for the dependency on both

Table 3. Superposition of 10 Representative Examples of Similar Compound−Proteins Pairs with Shared Motif on Proteins Surface[k]

motif[a]	CID[b]	PDB[c]	chain	residues	rmsd[d]	atoms[e]	average rmsd[f]	rmsd < 1.0 A (%)[g]	lZ-score[h]	gZ-score[i]	Instances[j]
ED.F	320239	1qkl	A	805 806 808	0.531	13	1.696	33.34	−2.018	−2.87	124
VEF	*5107381*	*1hw7*	*A*	*228 229 230*	*0.588*	*13*	*1.388*	*66.67*	*−1.697*	*−3.086*	*52*
DDE	4384800	1wcm	D	204 205 206	0.712	13	1.759	19.05	−2.614	−2.45	1129
FKD	*3017427*	*1xo8*	*A*	*147 148 149*	*0.629*	*15*	*1.558*	*42.86*	*−2.042*	*−3.157*	*55*
ED.F	95142	1qkl	A	805 806 808	0.721	13	1.751	47.62	−1.806	−2.664	100
IEF	5202952	1r1q	A	55 56 57	0.727	13	1.323	57.143	−1.417	−2.579	18
DDF	278311	2ajf	E	414 415 416	0.78	13	1.885	33.33	−1.798	−2.527	124
LDF	4363920	1hb6	A	47 48 49	0.78	13	1.856	35.01	−2.159	−2.635	52
LF.L	4966789	1o51	A	123 124 126	0.796	15	1.877	52.38	−1.691	−2.602	12
EFR	448593	1iqs	A	18 19 20	0.871	16	1.799	52.38	−1.319	−2.589	31

[a] Shared motif. [b] Compound ID in PubChem. [c] Protein ID in PDB. [d] Root-mean-square deviation. [e] rmsd of the 20 conformer's best alignment. [f] Average rmsd of 20 conformers' alignments. [g] Percentage of conformer's alignments lower than 1.0 A. [h] lZ-score (local). [i] gZ-score (global). [j] Number of instances in nonredundant surface segments in PDB. [k] Identified by the presented method, with a relatively high number of superposed atoms, low gZ-score, and low best hit rmsd: in addition, the number of conformers per compound with rmsd < 1.0 A was higher than 30%. Superposed structures of the gray italicized highlighted compound−protein pairs are shown in Figure 5.

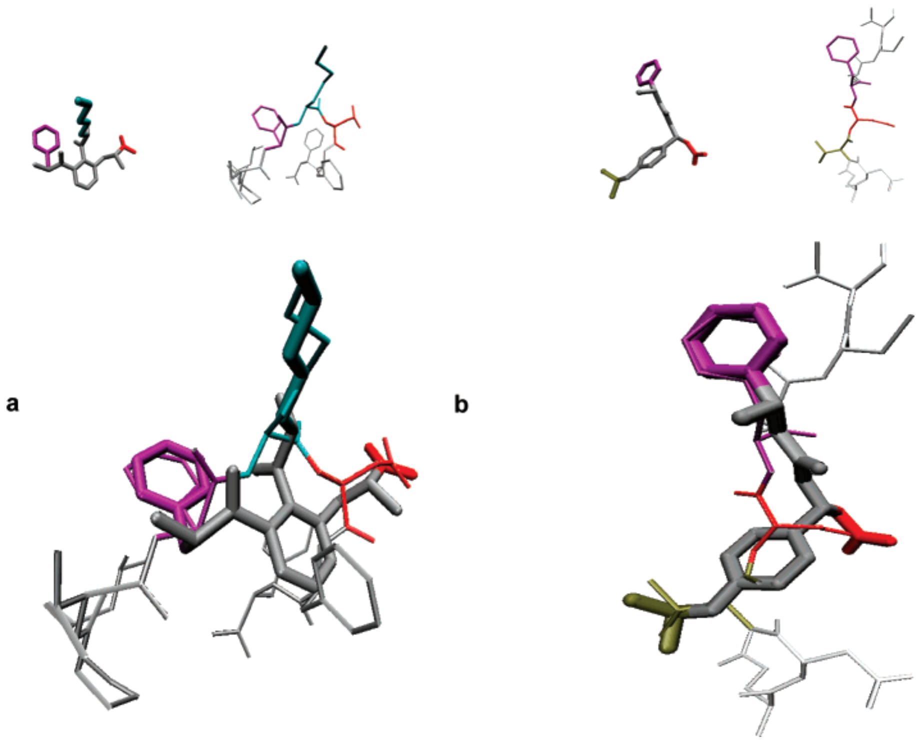


Figure 5. Examples of the superimposition of compounds and surface exposed subregions of proteins. Individual molecules (thick for compound, thin for peptide) are indicated at a smaller scale above the superimpositions. Substituents used in superimposition calculation are colored: (a) compound CID 3017427 with a segment of LEA14 protein (PDB-code: 1xo8:A) and (b) compound CID 5107381 with a segment of Hsp33 protein (PDB-code: 1hw7:A)

the number of atoms and other constraints: it can be seen from Figure 4b that gZ-score is indeed largely independent of the number of atoms.

Given the gZ-score, what are the appropriate ways of ranking the most interesting peptide-compound pairs? Figure 4c illustrates the distribution of values for three parameters, gZ-score, the rmsd for the best matching conformer, and the number of superposed atoms. We defined the following criteria for selecting significant compound−protein pairs appropriate for further structural and biological investigation: 1. *Number of superposed atoms >8*. This constraint helped by discarding very simple motifs (6 and 7 atoms) to eliminate many instances of less interesting motifs including substituent representations of peptide sequences such as III, SSS, and SIS. *2. rmsd of the best hit <0.7A. 3. 50% of conformers rmsd <1.0 A.* This constraint weights against compounds for which the majority of conformers are very unlike the peptide structure. 4. *gZ-score <−2.0 with number of instances >20*. In order to concentrate on motifs for which a relatively high gZ-score was possible, only gZ-scores calculated on more than 20 matches of the compound to different PDB instances were included.

A total of 334 compound−protein pairs satisfied these criteria and were selected for further investigation. These are colored by green marks in Figure 4c. Ten representative examples are shown in Table 3 and are highlighted in Figure 4 in red. Superposed structures of two highlighted compound−protein pairs are shown in Figure 5.

While this approach gives an indication of the feasibility of such searches, how might it be implemented in practice? Referring to a protein (1iqs) illustrates how the method might prove useful (Figure 6). An investigator examining this protein can rapidly identify from Figure 6 that there are 4 potential peptidomimetics mapping to three different re-

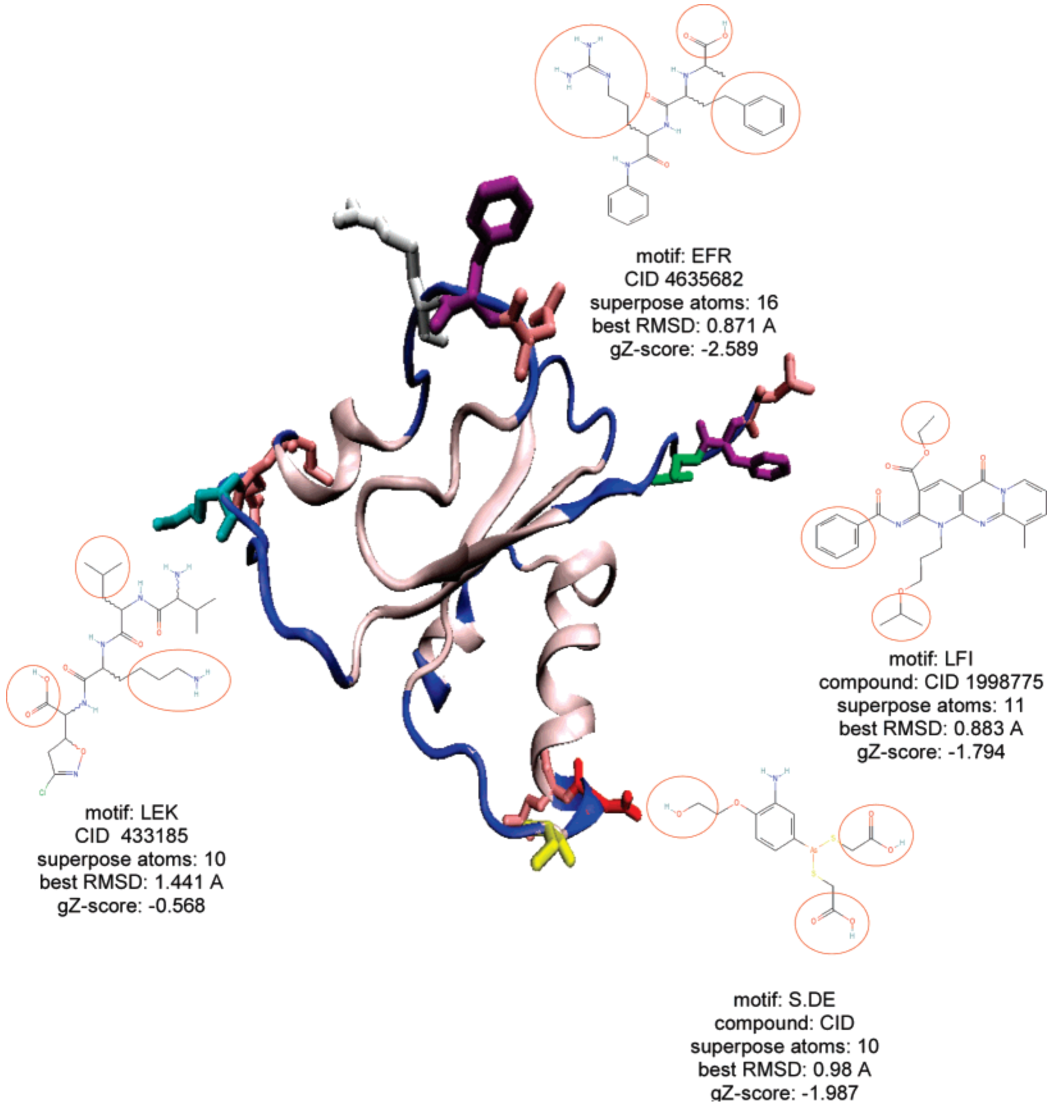


Figure 6. Examples of the protein (PDB-code: 1iqs) and its potential small compound peptidomimetics. Bright blue regions of the peptide backbone represent surface-exposed regions. The 4 compound−peptide motif pairs that scored best are indicated, along with the rmsd of the best hit, the corresponding motif, the gZ-score, and the number of atoms used for superposition.

gions: however, one of these is more strongly favored, on the grounds of having a much lower gZ-score of −2.59, suggesting a higher degree of specificity for this particular protein region. Thus, an investigator can rapidly look at a protein of interest and visually assess any potential compound leads for inhibition of surface interactions. However, given the general difficulties of inhibiting protein−protein interactions, this approach, although worth quickly investigating, is likely to have a relatively low yield of compounds with sufficient affinity to have a significant impact in typical screening assays. For this reason, the discovery of novel compounds arising from this search approach is more likely to be successful if systematically applied to a much larger number of protein targets simultaneously.

Practical Application to Known Short Linear Motifs. Short linear motifs are known to mediate many protein−protein interactions.[20,21] We investigated the occurrence of 8 short linear motifs from the ELM (Eukaryotic Linear Motif) database[21] that occurred on the surface of proteins of known structure in PDB. ELM motifs were expanded to a total of 19 instances, representing conformationally distinct states in PDB for a given ELM (see Table 4). We searched for similarity to compounds in the database of the National Cancer Institute (NCI) Developmental Therapeutic Program (DTP).[22] The advantage of this drug library is the availability of antitumor screening information for a subset of compounds. 46 644 compounds initially were selected on the basis of having 2 or more amino acid substituents and a maximal graph distance of 4. As for the Pubchem database searches (see above), there was pronounced enrichment of certain substituents in the compound database (e.g., benzene and carboxyl groups). A total of 913 drugs were matched, after filtration to eliminate those with MW > 700 and more than 15 rotatable bonds, 307 drugs were obtained, which had 340 superimposable similarities to the 9 ELM motifs. Superimposition was restricted to the terminal substituents of each amino acid.

Among these 340 matches, it is reassuring to note that a known true positive is returned. The rmsd for the known RGD mimetic Cilengitide[23] is ranked 61st, with an rmsd value of 0.64 (based on superimposition of the R and D groups), placing it in the top 20% of the best matches. RGD mimetics act like fibrinogen and vitronectin (proteins that contain the RGD motif) in binding to the extracellular regions

Table 4. Important Protein Motifs, Their General Regular Expressions Representations, Instances in PDB Database, and Extracted 3D Structures

ELM database identifier[a]	description[a]	general regular expressions	simplified regular expressions	instances in PDB[b]
LIG_AP2alpha_2	DPF/W motif binds alpha and beta subunits of AP2 adaptor complex.	DP[FW]	DW	**1EDU**, 1EYH, 1H0A
LIG_CYCLIN_1	Substrate recognition site that interacts with cyclin and thereby increases phosphorylation by cyclin/cdk complexes.	[RK].L.[FYLIVMP]	[RK]L[FYLIVM]	**1AXC, 1DT7, 1H24**, **1H25**, 1H26, 1H27, 1H28, 1JSU
LIG_NRBOX	The nuclear receptor box motif (LXXLL) confers binding to nuclear receptors.	[∧P](L)[∧P][∧P](L)(L)[∧P]	LLL	**1GWQ, 1GWR, 1T63**, 1T65
LIG_PCNA	The PCNA binding site is found in proteins involved in DNA replication, repair and cell cycle control.	(∧.{0,3}|Q).[∧FHWY][ILM] [∧P][∧FHILVWYP] [DHFM][FMY]..	L[DF]F, MD[FY], MFY	1AXC, 1UL1
LIG_RB	Interacts with the Retinoblastoma protein.	[LI].[CSF].[DE]	LCD	**1GH6, 1KNA,** 1KNE, 1Q3L
LIG_RGD	Found in extracellular matrix and recognized by members of the integrin family. In fibronectin, the RGD motif lies on a flexible loop.	RGD	RD	1FNA, **1FNF**, 1MFN, 1OC0, **1S4G**, 1SSU, **1TTF**, 1TTG
LIG_SH2_GRB2	GRB2-like Src Homology 2 (SH2) domains binding motif.	Y.N.	YN	1IVO, 1M14, 1M17, 1MOX, 1NQL, **1R0P**, 1R1W, 1SHY, 1XKK
LIG_TRAF2_1	Major TRAF2-binding consensus motif. Members of the tumor necrosis factor receptor (TNFR) superfamily initiate intracellular signaling by recruiting TRAFs through their cytoplasmic tails.	[PSAT].[QE]E	[ST]EE	**1CA9**, 1CZZ, **1D00**, 1D01, 1D0A, **1FLL**, 1JMA

[a] From the ELM database. [b] Those identifiers in bold were chosen as representative of a group of members with similar structure, and searched versus the NCI DTP database.

of cell-surface integrin proteins.[24] Comparison of the score for this compound to other instances of R.D on the surface of PDB structures confirms that the true RGD match is stronger than that for random R.D motifs (i.e., the Z-score is negative); nevertheless, the match is in fact rather weak, since many other matches of ELMs to drugs have much stronger Z-scores, reflected in the fact that out of the 340 matches, the Z-score rank of RGD and Cilengitide drops to 129. This practical application raises a number of key points regarding the method: first, the peptide backbone is disregarded (which is relevant to the RGD interaction), second, it is not clear what conformation of RGD is most likely to be bioactive. Thus, for this drug-motif comparison, there is no clear advantage to applying the Z-score in distinguishing true hits from likely false positives.

We compared the matches of the motifs to drugs based on both rmsd and Z-score. We noted that of the 4 compounds with a Z-score lower than -3 that also had cancer screening data available, one inhibited the growth of leukemic cells in the NCI screen, with a log GI50 of between -6.7 and -7.2 in the six different leukemic cell lines. This reflected a match to the RGD-like compound NSC627744 (N^5-(amino-(imino)methyl)-N^2-((10-methyl-10λ^5-acridin-1-yl)methyl)-ornithine compound with methyl hydrogen sulfate). While further inspection of the profile of this compound revealed that it shared a characteristic profile with other acridine-containing compounds (which act as DNA intercalators), it raises the question whether this compound may act through multiple mechanisms or some synergistic combination of the two. The four matches with the highest rmsd for which drug screening was available either showed inactivity or had a common effect on all cell lines at a log concentration of -4 molar (compounds NSC627892 and NSC627898) suggesting nonspecific toxicity. Thus, in this application, the Z-score and rmsd highlight different compounds as being of most interest: clearly, experimental validation will be required to prove that Z-score is superior to rmsd in screens for compounds resembling protein short linear motifs.

Searches for similarities between short peptide motifs involved in protein interactions and compounds may be greatly strengthened when the structural model for the peptide motif is obtained from its structure in complex with the interacting protein, particularly given the enrichment of short linear motifs in disordered regions of proteins.[20] To illustrate such an application, we searched for similarity of the F.D.F motif contained in a 10-mer peptide complexed with amphyphysin[25] (structure obtained from PDB entry 1ky7). Figure 7 illustrates the contrast between rmsd and Z-score. In this application the Z-score needs to be interpreted with some caution, since the motif instances used to calculate the Z-score are in uncomplexed proteins. Since the constraints are likely to be different in these two contexts, it is not a perfect comparator. An alternative approach would be to restrict the comparison to similar motifs that lie at the interface of two proteins in complex. However, in most cases the sample size of such motifs from PDB is too small to permit this. Nevertheless, visual inspection (Figure 7) aids

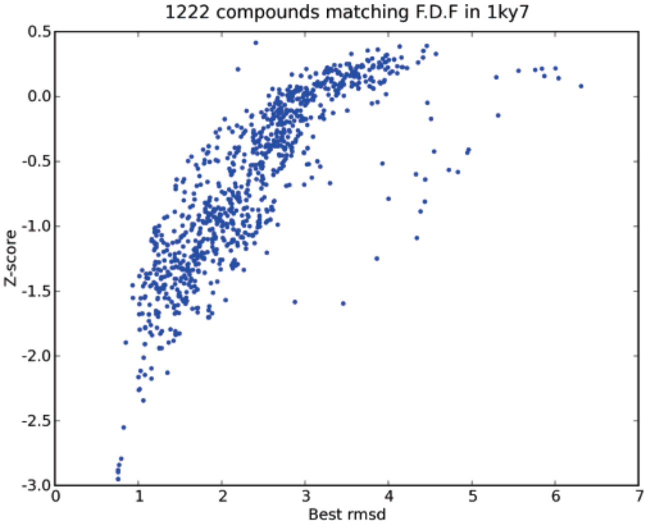


Figure 7. Z-score versus rmsd for the set of 1222 compounds matching the F.D.F motif within the 10-residue peptide derived from amphiphysin, complexed with the AP2 Clathrin Adaptor Alpha Appendage.

in making appropriate decisions on which compounds might be screened for bioactivity.

While the major focus of drug discovery sensibly concentrates on high-affinity druggable pockets, there remains a need to further develop methods for finding drugs that inhibit protein−protein interactions. One approach is to find one or more weak low affinity inhibitors and refine them in order to achieve an adequate inhibition of the protein−protein interaction. This study concentrates on the initial stage of identifying sublibraries of compounds that may be screened for their ability to inhibit protein−protein interactions. The strategy adopted is to assume that the peptide surface itself is a model for a compound that will interact with the corresponding surface of the interacting protein. On this basis, we seek compounds with similarity to subregions of the peptide surface.

What are the advantages of our method? First, it is complementary to other methods, concentrating on the reduced search space populated by exactly matching substituents. By forcing a matching of the amino acid substituents, as opposed to the general chemical properties represented in a pharmacophore search, it will be superior in those contexts where the recognition site is highly specific for a particular substituent. Second, the availability of a comparison set of biological structures from PDB for the appropriate scoring correction (Z-score calculation) is attractive. Third, it is a rapid method, which permits a set of compounds corresponding to a large interface of a protein complex, or to the entire surface of a protein, to be rapidly defined for further biological testing.

What are the drawbacks of our method? Clearly, it is in no sense intended to replace other approaches, since pharmacophore definition in terms of broad chemical properties for a given peptide region permits the interrogation of a much larger compound space. A formal comparison of the power of our substituent method versus standard pharmacophore approaches would be of great interest: however in the absence of an obvious unbiased “training set” of small molecules with peptide similarities, it is difficult to assess the relative value of different approaches through computational analysis alone. While we would not advocate our approach as the only method to identify mimetics of a peptide sequence, it provides a useful component in the cheminformatic identication of a series of potential compounds for further screening. The method as presented here has a number of limitations in relation to important amino acid modifications such as phosphorylation, but in principle extension to such alternative substituents is possible. The method ignores backbone interactions, so that for short motifs where such interactions are critical for defining binding specificity, it is clearly not ideal.

Our method is limited to contexts where the recognition epitope lies within a certain range. While there are many PPIs that involve contacts of highly dispersed residues, most of these will not be druggable by small molecule compounds or are at the very least not suitable for computational searches against libraries of small molecules. We have instead concentrated on the short linear motifs that lie typically within much less than 10 residues, that mediate many protein−protein interactions.[21,26] Even then, with motifs such as those contained in α-helices that spread a few contact points over a 10 residue region, Table 2 illustrates the explosion of potential comparisons that are required in generating potential motifs from drug libraries, as longer gaps are considered, so that in practice, only the more tightly grouped motifs are realistically addressed by this method. Allowing gaps equivalent to one turn of an α-helix is a possible extension to this method that might partially address this issue. More specifically, if the residues of interest within a protein are predefined, the search space can be narrowed down sufficiently, such that given a set of motifs with any number of gaps, a set of matching compounds can be easily provided.

The described approach works on small molecules and on modified or natural peptides. The method proposed is best suited to peptide microdomains where the oligopeptide sequence itself has biological activity: i.e. the specificity of action is a property of the short region, without dependence on features of the protein outside of the short motif. An obvious more general caveat to the interpretation of such matches is that the active conformation of a protein surface motif may adopt a very different conformation to that seen in a particular PDB structure, particularly for those structures that are not in complex. Since such signaling motifs are often in disordered regions of proteins,[20] this is a quite general problem. The best solution is to obtain the protein structure in complex with the interacting proteins. In the absence of availability of such a structure, it is possible that considering all known structures (e.g., from NMR) may include the active conformation of these highly disordered regions, but this is not guaranteed.

An alternative approach to the one taken here is that of Goede et al.,[27,28] who define compounds similar to peptide backbone of protein sequences by analyzing main chain dihedral angles. Their approach is different, since the focus is on the backbone (four “stem” atoms), with our approach more closely resembling classical peptides superposition applied to nonpeptidic molecules by omitting the backbone only superposing side chain atoms. More generally, there have been comparatively few approaches to systematically comparing surface regions of proteins to compound libraries: this in part reflects the perceived difficulty in identifying high affinity compounds acting against protein−protein interaction surfaces. While we do not underestimate these difficulties, we believe that characterizing this search space in more detail, as we have done here, contributes to an improved understanding of strategies for identifying small compounds that target protein interactions.

Identification of the bioactive conformation of short linear peptide motifs remains one of the key challenges in this area, since many linear motifs occur in highly disordered regions;[20] and synthesized oligopeptides may only rarely adopt the bioactive conformation, leading to various medicinal chemistry approaches that seek to stabilize the peptides appropriately.[29] The correct identification of the conformation of the peptide that is bioactive is not always available. Computational prediction of peptide conformations may provide little relevance to peptide activity[30] beyond what can be more simply predicted from peptide primary structure.[31]

While the native conformation of the peptide region in the uncomplexed protein may provide an adequate model (e.g., for those involving stable α-helices), the induced fit conformation represented by the conformation of the peptide complexed with its interacting partner is more appropriate for peptides from disordered regions; while for certain peptides, novel protein conformations outside of these two alternatives may be most appropriate.[6] In our analysis here we presented examples of the first two approaches. It is possible that the conformational sampling of disordered protein regions containing short linear interaction motifs provided by NMR structures of proteins may provide a library of conformers which could include the conformer selected by an interacting protein, such as a protease.[32] However, the absence of the interacting surface may in some cases prevent any of these aqueous conformers resembling the most bioactive conformation, and the increase of the conformational search space also has the downside that many more false positive compounds will be identified.

ACKNOWLEDGMENT

This work was supported by Scientific Foundation Ireland and Siemens as Industry Supplement Research Partnership program between UCD and Siemens Research Ireland. We thank the Irish Centre for High-End Computing (ICHEC) for their computational resources.

REFERENCES AND NOTES

(1) Arkin, M. R.; Wells, J. A. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat. Rev. Drug Discovery* **2004**, *3*, 301−317.

(2) Hruby, V. J.; Balse, P. M. Conformational and topographical considerations in designing agonist peptidomimetics from peptide leads. *Curr. Med. Chem.* **2000**, *7*, 945−970.

(3) Toogood, P. L. Inhibition of protein-protein association by small molecules: approaches and progress. *J. Med. Chem.* **2002**, *45*, 1543−1558.

(4) Gante, J. Peptidomimetics - Tailored Enzyme Inhibitors. *Angew. Chem., Int. Ed. Engl.* **1994**, *33*, 1699−1720.

(5) Perdih, A.; Kikelj, D. The Application of Freidinger Lactams and their Analogs in the Design of Conformationally Constrained Peptidomimetics. *Curr. Med. Chem.* **2006**, *13*, 1525−1556.

(6) Bursavich, M. G.; Rich, D. H. Designing non-peptide peptidomimetics in the 21st century: inhibitors targeting conformational ensembles. *J. Med. Chem.* **2002**, *45*, 541−548.

(7) *PubChem*. http://pubchem.ncbi.nlm.nih.gov/ (accessed May 2006).

(8) *Protein Databank (PDB*). www.pdb.org/pdb/ (accessed Jan 2006).

(9) Lameijer, E. W.; Kok, J. N.; Back, T.; Ijzerman, A. P. Mining a chemical database for fragment co-occurrence: discovery of "chemical cliches". *J. Chem. Inf. Model.* **2006**, *46*, 553−562.

(10) Thomas, H.; Cormen, C. E. L.; Rivest, R. L.; Stein, C. *Introduction to Algorithms*, 2nd ed.; MIT Press and McGraw-Hill: 2001; Vol. Section 24.3: Dijkstra's algorithm, pp 595−601.

(11) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. B: Struct. Sci.* **1976**, *B32*, 922−923.

(12) *Python*. www.python.org (accessed month year).

(13) Hamelryck, T.; Manderick, B. PDB parser and structure class implemented in Python. *Bioinformatics* **2003**, *19,* 2308−2310.

(14) *Molecular Operating Environment;* Chemical Computing Group Inc. http://www.chemcomp.com/ (accessed Jun 2005).

(15) Irwin, J. J.; Shoichet, B. K. ZINC−a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(16) Leach, A. R. *Molecular Modelling. Principles and Applications;* Pearson Education Limited: 2001; Vol. Confomational Analysis, pp 457−506.

(17) Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658−1659.

(18) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−637.

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

(20) Neduva, V.; Russell, R. B. Linear motifs: evolutionary interaction switches. *FEBS Lett.* **2005**, *579*, 3342−3345.

(21) Puntervoll, P.; Linding, R.; Gemünd, C.; Chabanis-Davidson, S.; Mattingsdal, M.; Cameron, S.; Martin, D. M. A.; Ausiello, G.; Brannetti, B.; Costantini, A.; Ferrè, F.; Maselli, V.; Via, A.; Cesareni, G.; Diella, F.; Superti-Furga, G.; Wyrwicz, L.; Ramu, C.; McGuigan, C.; Gudavalli, R.; Letunic, I.; Bork, P.; Rychlewski, L.; Küster, B.; Helmer-Citterich, M.; Hunter, W. N.; Aasland, R.; Gibson, T. J. ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.* **2003**, *31*, 3625−3630.

(22) DTP-NCI. http://dtp.nci.nih.gov/ (accessed Oct 2006).

(23) Friess, H.; Langrehr, J. M.; Oettle, H.; Raedle, J.; Niedergethmann, M.; Dittrich, C.; Hossfeld, D. K.; Stoger, H.; Neyns, B.; Herzog, P.; Piedbois, P.; Dobrowolski, F.; Scheithauer, W.; Hawkins, R.; Katz, F.; Balcke, P.; Vermorken, J.; van Belle, S.; Davidson, N.; Esteve, A. A.; Castellano, D.; Kleeff, J.; Tempia-Caliera, A. A.; Kovar, A.; Nippgen, J. A randomized multi-center phase II trial of the angiogenesis inhibitor Cilengitide (EMD 121974) and gemcitabine compared with gemcitabine alone in advanced unresectable pancreatic cancer. *BMC Cancer* **2006**, *6*, 285.

(24) Huang, C.; Cheng, J.; Stern, A.; Hsieh, J.; Liao, C.; Tseng, C. Disabled-2 is a novel IIb-integrin-binding protein that negatively regulates platelet-fibrinogen interactions and platelet aggregation. *J. Cell Sci.* **2006**, *119*, 4420−4430.

(25) Brett, T. J.; Traub, L. M.; Fremont, D. H. Accessory protein recruitment motifs in clathrin-mediated endocytosis. *Structure* **2002**, *10*, 797−809.

(26) Neduva, V.; Linding, R.; Su-Angrand, I.; Stark, A.; de Masi, F.; Gibson, T. J.; Lewis, J.; Serrano, L.; Russell, R. B. Systematic Discovery of New Recognition Peptides Mediating Protein Interaction Networks. *PLoS Biol.* **2005**, *3*, e405.

(27) Preissner, R.; Goede, A.; Rother, K.; Osterkamp, F.; Koert, U.; Froemmel, C.; Matching organic libraries with protein-substructures. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 811−817.

(28) Goede, A.; Michalsky, E.; Schmidt, U.; Preissner, R. SuperMimic−fitting peptide mimetics into protein structures. *BMC Bioinformatics* **2006**, *7*, 11.

(29) Fairlie, D. P.; West, M. L.; Wong, A. K. Towards Protein Surface Mimetics. *Curr. Med. Chem.* **1998**, *5*, 29−62.

(30) Edwards, R. J.; Moran, N.; Devocelle, M.; Kiernan, A.; Meade, G.; Signac, W.; Foy, M.; Park, S. D. E.; Dunne, E.; Kenny, D.; Shields, D. C. Bioinformatic discovery of novel bioactive peptides. *Nature Chem. Biol.* In press.

(31) Parthasarathi, L.; Devocelle, M.; Sondergaard, C.; Baran, I.; O'Dushlaine, C. T.; Davey, N. E.; Edwards, R. J.; Moran, N.; Kenny, D.; Shields, D. C. Absolute net charge and the biological activity of oligopeptides. *J. Chem. Inf. Model.* **2006**, *46*, 2183−2190.

(32) Fairlie, D. P.; Tyndall, J. D. A.; Reid, R. C.; Wong, A. K.; Abbenante, G.; Scanlon, M. J.; March, D. R.; Bergham, D. A.; Chai, C. L. L.; Burkett, B. A. Conformational Selection of Inhibitors and Substrates by Proteolytic Enzymes: Implications for Drug Design and Polypeptide Processing. *J. Med. Chem.* **2000**, *43*, 1271−1281.

CI600404Q