Published paper
Kennewell, E., Willett, P., Ducrot, P. and Luttmann, C. (2006) *Identification of target-specific bioisosteric fragments from ligand-protein crystallographic data.* Journal of Computer-Aided Molecular Design, 20 (6). pp. 385-394.

Identification of Target-Specific Bioisosteric Fragments from Ligand-Protein Crystallographic Data

Elizabeth A. Kennewell[1] and Peter Willett[2]

Krebs Institute of Biomolecular Research and Department of Information Studies, University of Sheffield, Sheffield S1 4DP, UK

Pierre Ducrot and Claude Luttmann

Chemical Sciences, Sanofi-Aventis, 94400 Vitry-sur-Seine, France

Abstract. Bioisosteres are functional groups or atoms that are structurally different but that can form similar intermolecular interactions. Potential bioisosteres were identified here from analysing the X-ray crystallographic structures for sets of different ligands complexed with a fixed protein. The protein was used to align the ligands with each other, and then pairs of ligands compared to identify substructural features with high volume overlap that occurred in approximately the same region of geometric space. The resulting pairs of substructural features can suggest potential bioisosteric replacements for use in lead-optimisation studies. Experiments with twelve sets of ligand-protein complexes from the Protein Data Bank demonstrate the effectiveness of the procedure.

Keywords. Bioisostere, Fragment substructure, Ligand-protein complex, Protein Data Bank, Shape similarity

[1] Current address: Health Informatics Service, Sheffield West Primary Care Trust, Sheffield, S10 3TG

[2] To whom all correspondence should be addressed: Email p.willett@sheffield.ac.uk; Tel. +44-114-2222633

INTRODUCTION

Bioisosteres were originally defined as molecules or functional groups that have similar chemical and physical properties and that hence exhibit similar biological activities [1, 2]. This definition has since been extended to include substructural features that are structurally different but that can form similar intermolecular interactions. The concept of bioisosterism is an important approach in the lead-optimisation stage of a drug-design programme as it provides a way of enhancing some desirable chemical or physical property, e.g. to improve solubility or metabolic stability, whilst still maintaining the biological activity of interest [3]. For example, Merderski et al. reported the use of the benzothiadiazole group as a bioisoster for the methylendioxyphenyl group in a study of endothelin receptor antagonists [4]; Uddin et al. described the use of the sulfonylazide group as a booster for the sulphonamide group in a study of celecoxib analogues [5]; and Showell and Mills advocated the use of silicon as a replacement for a fully substituted sp^3 carbon [6]. The use of bioisosterism to support lead optimisation will also be the focus of the work reported here. It is, however, worth noting that there is now also increasing interest in the use of scaffold-hopping techniques to suggest replacement ring systems that can locate functionality at the appropriate locations in 3D space whilst providing a novel patent position (see, e.g., [7-9]).

A convenient source of bioisosteres is the BIOSTER database from Accelrys Inc., which details pairs of compounds that have been reported in the literature as being biologically interchangeable [10]. The current version of the database, Version 2003.1, contains almost eleven-thousand pairs of potential bioisosteres, including drugs, agrochemicals and enzyme inhibitors [11]. Alternatively, means can be found to identify bioisosteres automatically and a range of approaches have been described that are based on calculating measures of similarity between pairs of substituents to find those that are closely related using the chosen similarity measure. Then, given an existing bioactive molecule, potential analogues are obtained by replacing one or more of the substituents on a central scaffold by those substituents that have previously been shown to be most similar. There have been several reports of such techniques, differing principally in the types of information that are used for the calculation of the inter-substituent similarities.

The simplest way of measuring the similarity between a pair of substituents uses the 2D fingerprint measures that are widely used for similarity searching in chemical databases. Such an approach has been reported recently by Wagener and Lommerse, who describe a system that has been developed at Organon for suggesting bioisosteric replacements and that is based on fingerprints encoding topological pharmacophore information about the atoms comprising a substituent [12]. In another topology-based study, Sheridan analysed pairs of molecules that belonged to the same activity class in the MDL Drug Data Report (MDDR) database and that differed in only one location [13]. Use of a maximum common substructure algorithm identified the common parts of the two molecules that were being compared; this common substructure was removed and the remaining pair of substructures stored as potential bioisosteres. Some of the replacements were generic, in that they occurred frequently throughout the MDDR, while others were identified only within specific therapeutic classes.

A focus on the identification of topological equivalences inevitably means that less account is taken of physicochemical properties that may be of particular importance in the context of bioactivity. Both Ertl et al. [14] and Holliday et al. [15] have reported work in which a substituent is characterised by computed physicochemical properties of various sorts. The first of these represents a substituent by a vector of properties that are computed for the substituent as a whole; related ideas have been reported recently by Zhu et al. in work on the measurement of superstructure similarity in the design of reaction schemes [16]. The more complex system described by Holliday *et al*. represents a substituent by a series of vectors that encode the sum of the atomic properties at increasing numbers of bonds away from the point of attachment of the substituent. The more detailed representation thus takes account of both the physicochemical characteristics of a substituent and its topology (and, implicitly, of its geometry in the case of low-flexibility substituents). The resulting similarities have been used successfully for both database searching and QSAR [15, 17].

Finally, there is IsoStar, which is a knowledge base containing information on the geometries of non-bonded interactions between specified pairs of chemical groups [18, 19]. The geometric data is used to generate scatterplots showing all the possible positions of a chosen contact group around a chosen central group, thus providing an overview of the preferred orientations that allow a particular group-to-group interaction to take place. Watson et al. have discussed geometric similarity measures based on these scatterplots, so as to identify groups that are oriented similarly with respect to a given central group, such as a key amino acid in a protein active site [20]. This study is perhaps the most closely related to the work reported here in that both approaches use X-ray crystallographic data as the basis for identifying pairs of similar substructures; however, our work identifies equivalences that are specific to a particular target, and that are hence more likely to be associated with changes in the biological activity of interest.

METHODS

Introduction

Our techniques seek to identify potential bioisosteres within a set of ligands for a particular protein target. In brief, each ligand within a dataset is chosen in turn to act as the *reference ligand*, which is then compared to all the other ligands in the dataset, each of which is referred to as a *query ligand*. The query ligand is split into a set of fragments (the *query fragments*) to identify small regions within a pair of ligands that might be bioisosteric. The potential bioisosteres are identified based on volume overlap between a query fragment and a region within the current reference ligand that occupies the same space as the query fragment. The procedure is summarised in Figure 1, and explained in greater detail in the remainder of this section. The majority of the software was written using the Scientific Vector Language (SVL) scripting language available in the Molecular Operating Environment (MOE) that has been developed by the Chemical Computing Group [21].

Alignment and splitting of the ligands

The Protein Data Bank (PDB) was searched to retrieve sets of structures that all shared a common amino acid sequence and that all contained a ligand bound to the protein [22]. One of the structures was chosen to act as a template, and the protein coordinates in the other structures fitted to the coordinates for the protein in the chosen reference structure. This fitting stage was carried out with an MOE 3D-alignment procedure that uses all of the protein backbone atoms for the superposition of the different structures of the common protein. The ligands were then aligned by extracting them from the set of fitted protein structures.

Each of the extracted ligands is broken down into a set of overlapping fragments by the breaking of appropriate bonds. For example, Figure 3 shows the results obtained from splitting the molecule shown in Figure 2. The output in Figure 3 is obtained by breaking all the single bonds within the molecule unless they are either ring bonds or bonds involving terminal atoms; the four bonds broken in the example molecule are shown by red lines in the figure. The fragments are generated by breaking the identified bonds in all possible combinations, and this set of fragments is then filtered to remove those that contain just a single atom, e.g., the single nitrogen atom fragment in Figure 3.

Identification of fragment pairs

Each query fragment is then compared with the current reference molecule to identify the reference-molecule atoms that best overlay the query fragment. The fragment and the reference molecule are already aligned, and it is hence simple to score the overlaps between the query fragment and sections of the reference molecule to determine the best mapping. The mapping is based on the degree to which the two fragments overlap in terms of the volume of their constituent atoms. Fragments with a high degree of overlap will occupy a similar position within the protein's active site and are hence assumed to have a similar role within the ligand.

Scoring reference molecule sections

As mentioned previously, the best overlap of each query fragment with the reference molecule needs to be identified, the resulting overlap being called the *reference fragment*. In order to do this, the reference molecule is split into sections, where a section is defined as being part of a molecule in which all the atoms within it are connected by ring bonds or multiple bonds. The sections within a particular molecule can hence be identified by breaking all of its non-ring single bonds, as exemplified in Figure 4. Here, the four single, non-ring bonds marked in red are split to generate the sections labelled 1-5 on the right-hand side of the figure.

The sectioned reference molecule is then compared with the query fragment. Sections consisting of only one atom are kept because these sections may overlap with the query fragment and therefore need to be retained as part of the reference fragment. If these sections do not overlap with the query molecule then they are excluded from the reference fragment, thus ensuring that the smallest reference fragment is identified.

Computing volume overlaps is time-consuming and so it was decided to measure the overlap between a reference section and a query fragment using an equation based on the distances between pairs of their constituent atoms, specifically, a simplified version of the SEAL scoring function developed by Kearsley and Smith [23]. The volume overlap for each atom within a specific section is computed with each atom in the query fragment, and these overlaps summed. The sums for all the atoms within the section are then added together to create the overall section score.

$$\sum_{j=1}^{m}\sum_{i=1}^{n} e^{-d_{ij}^{2}}$$

where *m* and *n* are the numbers of atoms in the reference section and the query fragment, respectively. An average score is calculated by dividing this value by the number of atoms within the section, and the resulting mean score used to determine whether this section should be included within the reference fragment: this is done if the mean score is at least 0.5. This procedure is illustrated in Figure 5, comparing the same reference molecule as in Figure 4 with a query fragment shown in orange. Here, two of the five sections of the reference molecule score highly enough, compared to the query fragment, to be included in the reference fragment (as indicated by the ticks), but the other three are omitted (as indicated by crosses). The two selected sections in Figure 5 hence comprise the *fragment pair* shown in Figure 6, with the two substructural moieties making equivalent hydrogen bonding interactions with the protein structure.

Calculating the average overall score

Once the fragment pair has been identified then an average overall score for the pair needs to be calculated: this score is used to determine whether the fragment pair should be saved and to rank the pairs in the results database. The distances between each possible pairing of one atom from the query fragment and one atom from the reference fragment are calculated, and this information used to score the match. The score computed is

$$\frac{2}{m+n}\sum_{j=1}^{m}\sum_{i=1}^{n} e^{-d_{ij}^{2}}$$

where *m* and *n* are the numbers of atoms in the reference fragment and the query fragment, respectively.

Fragment pairs scoring less than a cut-off value 0.7 are excluded from further consideration, thus removing poorly aligned fragment pairs. In addition, a series of filters was applied so as to remove fragment pairs that could not meet one or more of several criteria that are necessary for a fragment pair to represent a potential bioisosteric pair [23]. Examples of such criteria include the following. First, only query fragments containing 20 atoms or less and reference fragments containing more than one atom are considered. Second, it is possible for the query fragment and the reference fragment to be identical (especially if the ligands in the dataset have structurally similar regions to each other); these pairs are obviously not bioisosteric and are hence also removed. Third, disjointed reference fragments were removed: these arise when the sections of the reference molecule that scored highly enough to be part of the reference fragment were not all connected together within the reference molecule.

RESULTS AND DISCUSSION

The procedure described above was run on several sets of ligands drawn from the PDB, as listed in Table 1. It will be seen from this table that there are major variations in the numbers of distinct fragment pairs identified. The number depends on several factors, including: the number of ligands within the dataset; whether the ligands all bind to the same active site within the protein (and hence occupy a similar space); and the structural diversity of the ligands (as structurally homogeneous ligands are likely to produce multiple non-unique fragment pairs). Even so, the procedure is sufficiently rapid in execution to enable datasets of the sort shown in Table 1 to be processed in 5-10 minutes on a Linux PC.

Examples of the fragment pairs identified are shown in Figure 7. In this figure, the fragment pair is shown on the left with the ligands the pair were derived from shown to the right of the fragment pair; and the reference ligand is always shown in purple with the query ligand in orange. The results in Figure 7 demonstrate that our procedures are able to identify pairs of fragment substructures that occupy the same space within a protein's active site, and that may function as target-specific bioisosteres. Such pairs may be involved in the same molecular interactions (as is the case with the bioisosteres identified using IsoStar) but may instead have other roles, such as being part of the scaffold region or a linker. Once the fragment pairs have been identified they can be made available for consideration by medicinal chemists working on that target as potential aids for lead optimisation. Further examples of fragment pairs are shown in Figure 8, which illustrates the wide range of types of structural equivalence identified by our procedure.

Several of the fragment pairs from Figure 8 are illustrated in Figure 9, which demonstrates the types of interaction identified by our procedure. Each of these figures shows the aligned proteins together with the associated fragment pairs, illustrating the ways in which different substructures are able to make the same interactions with the protein. Thus, Figure 9a shows a hydrogen bonding interaction in CDK4, and there is also a stacking of the aromatic rings above the amide in the protein; Figure 9b shows a hydrogen-bonding interaction in Factor Xa; Figure 9c shows a hydrophobic interaction in Factor Xa between the amide and the centres of the bicyclic rings; Figure 9d shows a hydrophobic interaction in tyrosine phosphatise between the phenylaniline and the rings; Figure 9e shows a polar interaction in tyrosine phosphatase between an arginine and carboxylate or phosphate (there are also two interactions with backbond amides); Figure 9f shows a zinc binding interaction in MMP3 involving carboxylate and hydroxymate, a well-known pairing.

One of the principal applications of bioisosteres during lead optimisation is to enhance a molecule's ADMET profile. There is hence a need to link the substructural equivalences identified here with locally-generated physicochemical data. Specifically, the data is scanned to find pairs of molecules that differ from one another just by that particular fragment pair. The property data associated with such a molecule-pair is then used to compute ΔP, where P is the altered property. The procedure is repeated for all molecule-pairs with the chosen fragment-pair and the mean ΔP computed, so as to identify substituent replacements that are expected to improve the chosen property P. A prototype system based on these ideas is now under development at Sanofi-Aventis.

CONCLUSIONS

There is a rapidly increasing number of ligand-protein complexes for which X-ray crystallographic data are available, with many important biological targets for which there are complexes with a range of different ligands. The availability of such data provides a basis for the identification of bioisosteres that are target-specific. The resulting bioisosteres might be expected to provide more reliable information when modifying an existing lead compound than do existing approaches, which are based either on empirical measures of inter-substituent similarity or on non-target-specific crystallographic data. In this paper, we have described one such approach, in which ligands extracted from PDB ligand-protein complexes are aligned in 3D space to identify substructural features with high volume overlap that occur in approximately the same regions of geometric space. Experiments with twelve sets of ligand-protein complexes demonstrate that our approach is both effective and efficient in operation in identifying potential substructural replacements. These replacements may be used to provide a knowledge-based approach to the enhancement of the ADMET profile of a lead compound.

REFERENCES

1. Burger, A. Prog. Drug Res. 37 (1991) 287.
2. Patani, G. and LaVoie E., Chem. Rev., 96 (1996) 3147.
3. Olesen, P.H., Curr. Opin. Drug Discov. Develop., 4 (2001) 471.
4. Mederski, W., Osswald, M., Dorsch, D., Anzali, S., Christadler, M., Schmitges, C. and Wilm, C., Bioorg. Med. Chem. Lett., 8 (1998) 17.
5. Uddin, M., Praveen Rao, P. and Knaus, E., Bioorg. Med. Chem., 11 (2003) 5273.
6. Showell, G. and Mills, J., Drug Discov. Today, 8 (2003) 551.

7. Schneider, G., Neidhart, W., Giller, T. and Schmid, G. Angew. Chem. Int., 38 (1999) 2894.
8. Bohl, M., Dunbar, J.B., Gifford, E., Heritage, T., Wild, D.J., Willett, P. and Wilton, D.J., Quant. Struct.-Act. Relat. 21 (2002) 590.
9. Böhm, H.-J., Flohr, A. and Stahl, M., Drug Discov. Today: Technol., 1 (2004) 217.
10. Ujváry, I., Pest. Sci., 51 (1997) 92.
11. The BIOSTER database is available from Accelrys Inc. at http://www.accelrys.com/products/chem_databases/databases/bioster.html
12. Wagener, M. and Lommerse, J.P.M., J. Chem. Inf. Model., 46 (2006) 677.
13. Sheridan, R., J. Chem. Inf. Comp. Sci., 42 (2002) 103.
14. Ertl, P., J. Mol. Graph. Model., 16 (1998) 11.
15. Holliday, J.D., Jelfs, S.P., Willett, P. and Gedeck, P., J. Chem. Inf. Comp. Sci., 43 (2003) 406.
16. Zhu, Q., Yao, J., Yuan, S., Li, F., Chen, H., Cai, W. and Liao, Q. J. Chem. Inf. Model. 45 (2005) 1214.
17. Hirons, L., Holliday, J.D., Jelfs, S.P., Willett, P. and Gedeck, P., QSAR Combin. Sci., 24 **(**2005) 611.
18. Bruno I.J., Cole J., Lommerse, R., Rowland, R., Taylor, R. and Verdonk, M., J. Comp.-Aid. Mol. Des., 11 (1997) 425.
19. The IsoStar database is available from the Cambridge Crystallographic Data Centre at http://www.ccdc.cam.ac.uk/products/knowledge_bases/isostar/
20. Watson, P., Willett, P., Gillet, V.J. and Verdonk, M., J. Comp.-Aid. Mol. Des. 15 (2001) 835.
21. Chemical Computing Group is at http://www.chemcomp.com
22. Berman, H.M., Battistuz, T., Bhat, T.N., Blum, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D. and Zardecki, C., Acta Crystallographica, D58 (2002) 899.
23. Kearsley, S.K. and Smith, G.M. Tetrahed. Comput. Methodol., 3 (1990) 615.
24. Kennewell, L. Identifying Bioisosteric Fragments from Databases of Protein-Ligand Complex X-Ray Crystallographic Structures. MSc dissertation, University of Sheffield, 2004.

Protein target	Number of ligands	Number of non-unique fragment pairs identified (scoring over 0.7)
ACHE	63	125
Beta-glucosidase	23	28
CDK2	32	702
CDK	12	136
Factor XA	20	347
HIV-1 protease	78	5837
MAO	16	16
MMP13	6	15
MMP3	5	19
PDE4	12	68
Tyrosine kinase	3	0
Tyrosine phosphatase 1b	33	585

Table 1. Results of the procedure using PDB datasets

```
Extract a set of ligands from the PDB that are all complexed with the same protein.
Align the ligands on the basis of the common protein structure.
FOR each ligand DO
    Make it the reference ligand, RL
    FOR each of the remaining ligands DO
        Make it the query ligand, QL
        Split QL into fragments
            FOR each query fragment DO
                Score its volume overlap with RL
                Identify the best matching region in RL
            ENDDO
    ENDDO
ENDDO
```

Figure 1. Overview of the bioisostere identification procedure

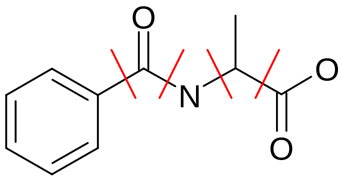


Figure 2. Splitting of a ligand

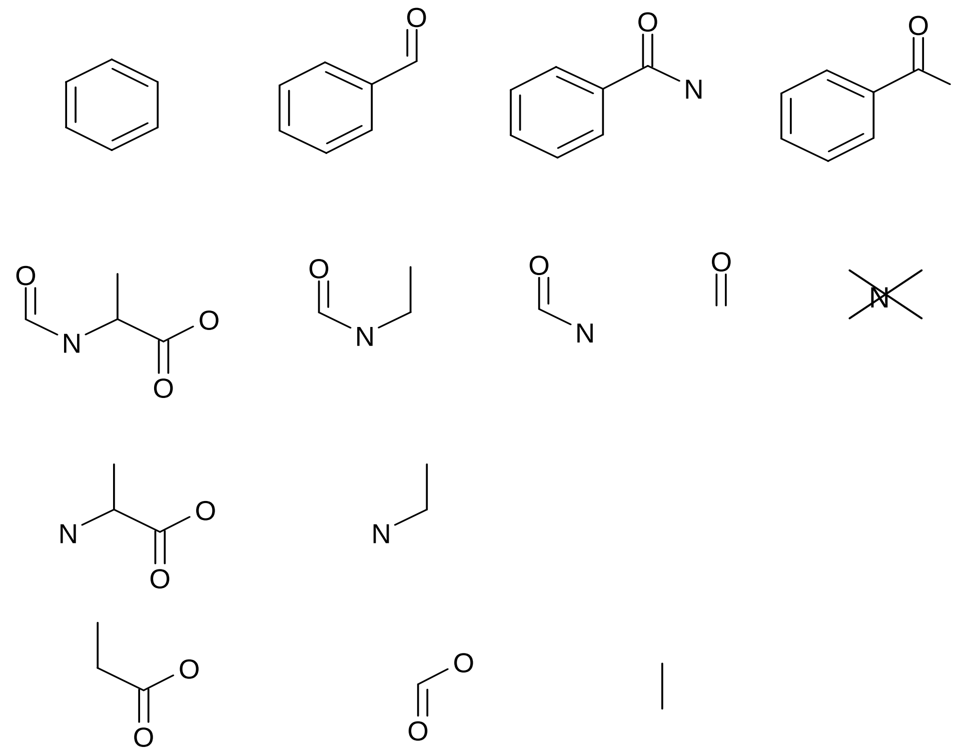


Figure 3. Fragments generated from the molecule shown in Figure 2

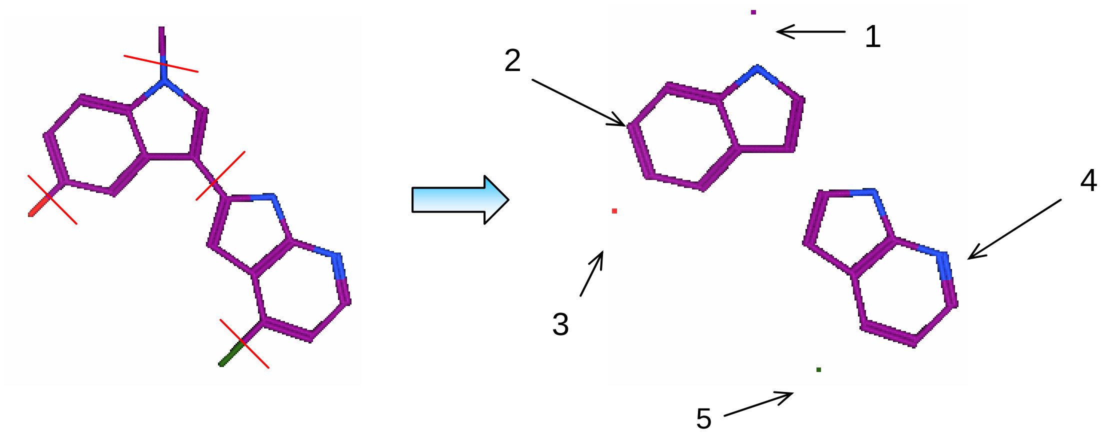


Figure 4. A reference molecule broken down into five sections

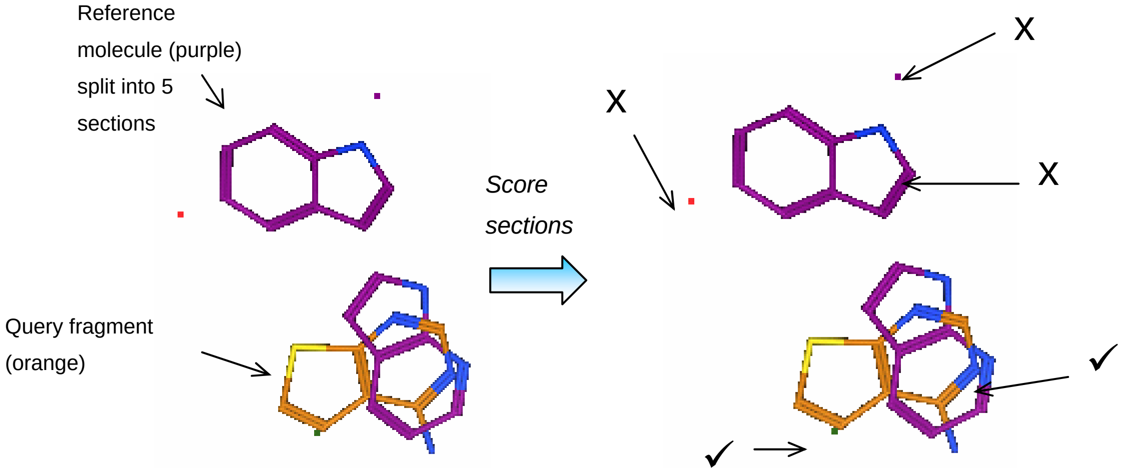


Figure 5. Selection of sections for inclusion in the reference fragment

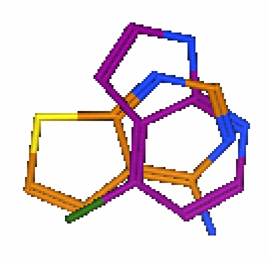
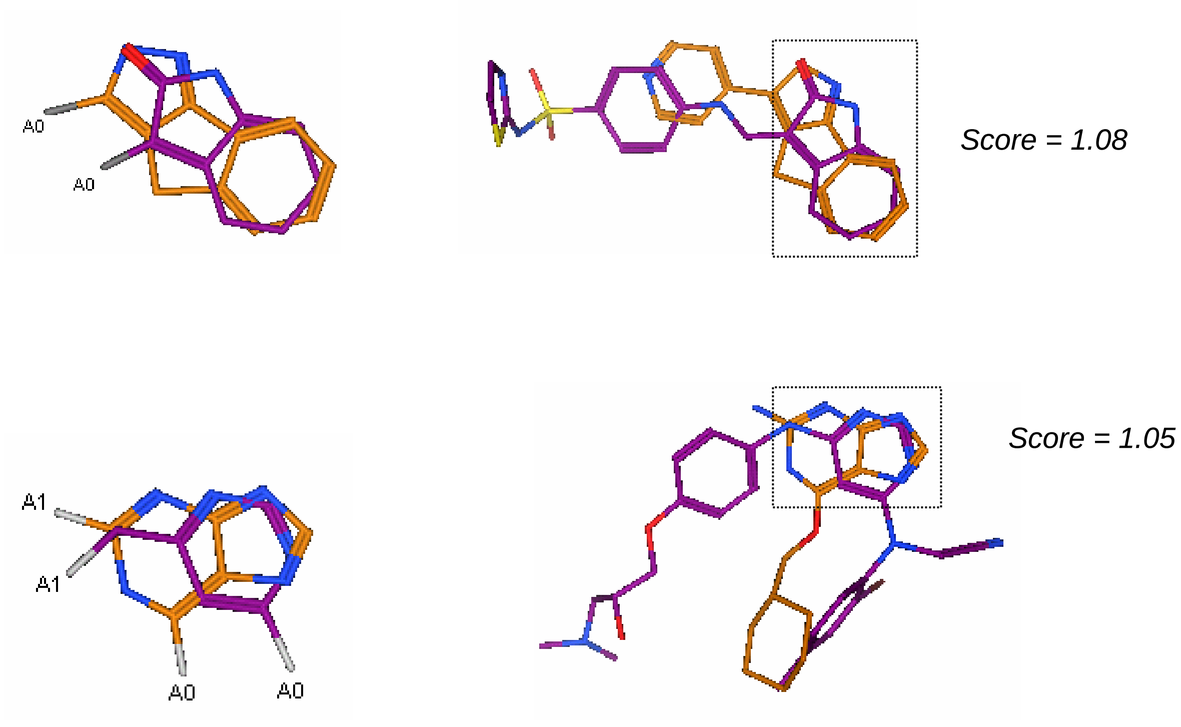
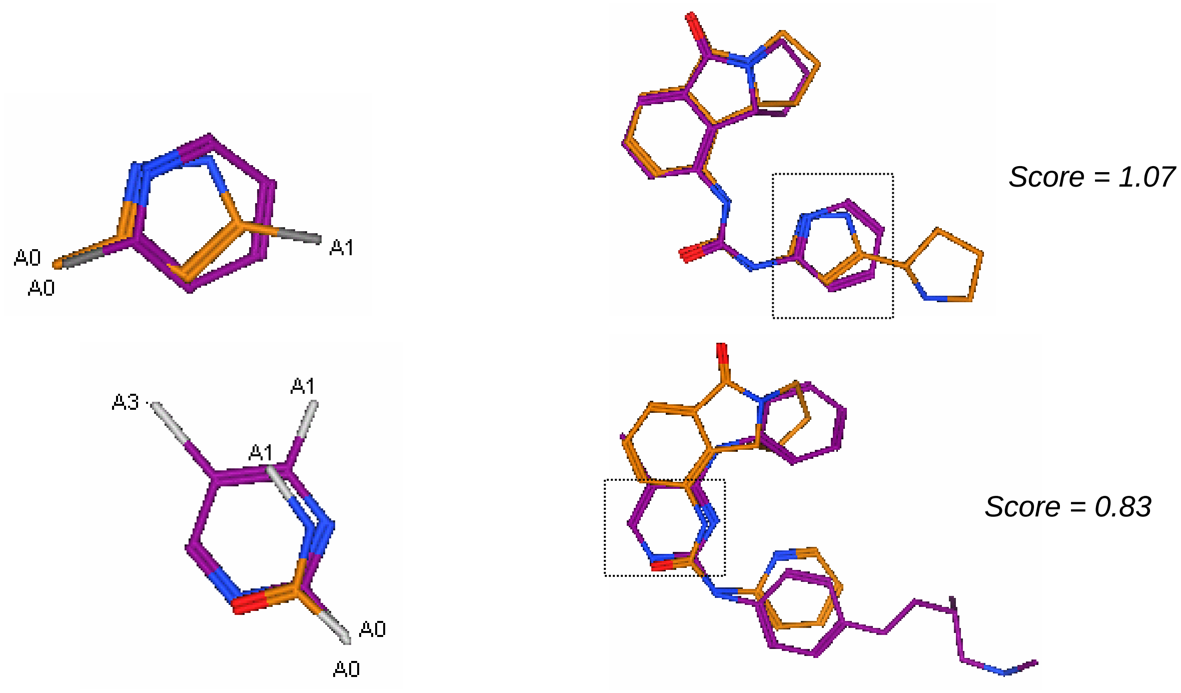


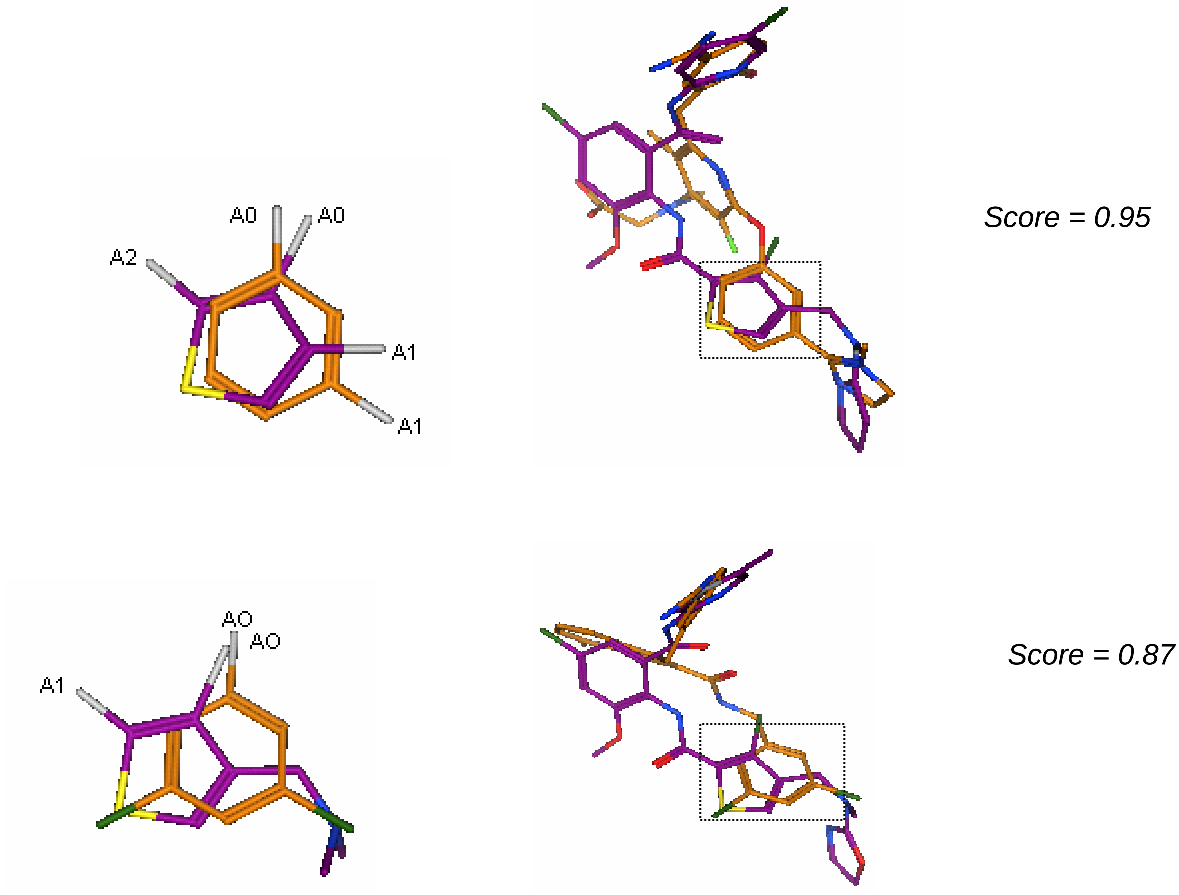
Figure 6. The fragment pair resulting from Figure 5.


A0
A0
Score = 1.08
A1
A1
A0
A0
Score = 1.05

(a)

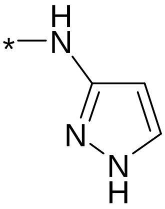
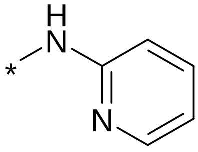
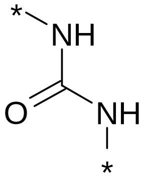
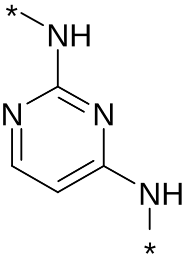

A0
A0
A1
Score = 1.07
A3
A1
A1
A0
A0
Score = 0.83

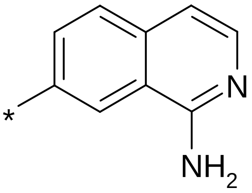
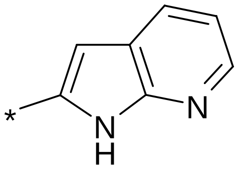
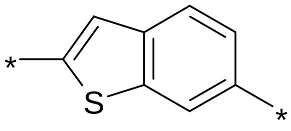
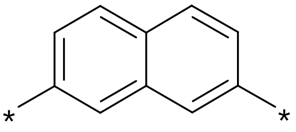
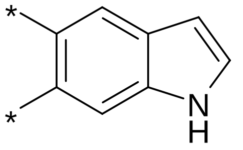
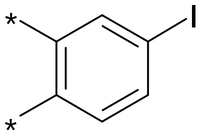
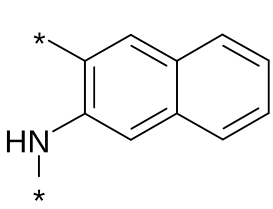
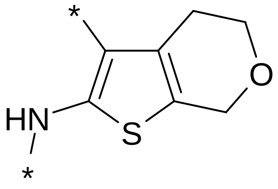
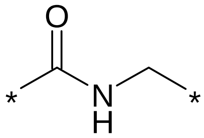
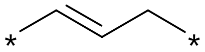
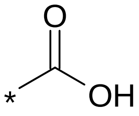
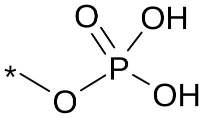
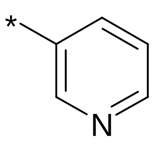
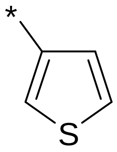
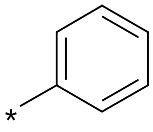
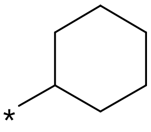
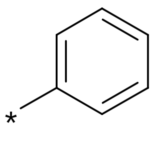
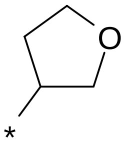
(b)



(c)

Figure 7. Fragment pairs identified from the CDK2 (a), CDK4 (b) and Factor Xa (c) datasets

CDK4 (*)		
CDK4		

Factor Xa (*)		
Factor Xa (*)		
Tyrosine phosphatase		
Tyrosine phosphatase (*)		
Tyrosine phosphatase		
Tyrosine phosphatase (*)		
HIV1 protease		
HIV1 protease		
HIV1 protease		

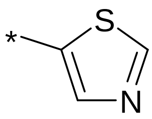
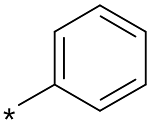
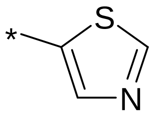
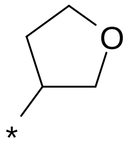
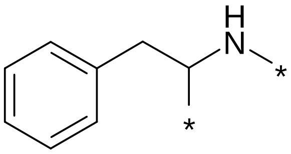
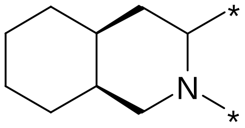
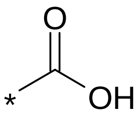
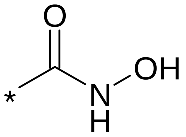
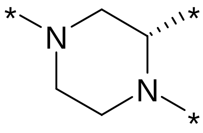
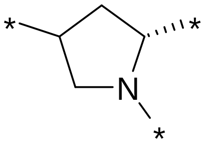
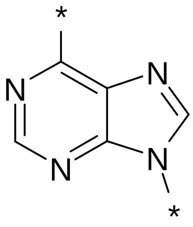
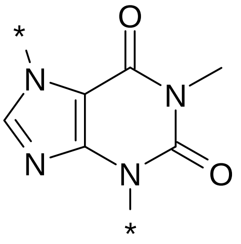
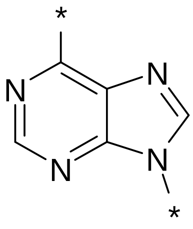
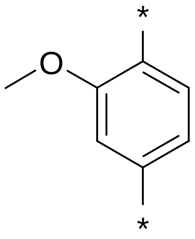
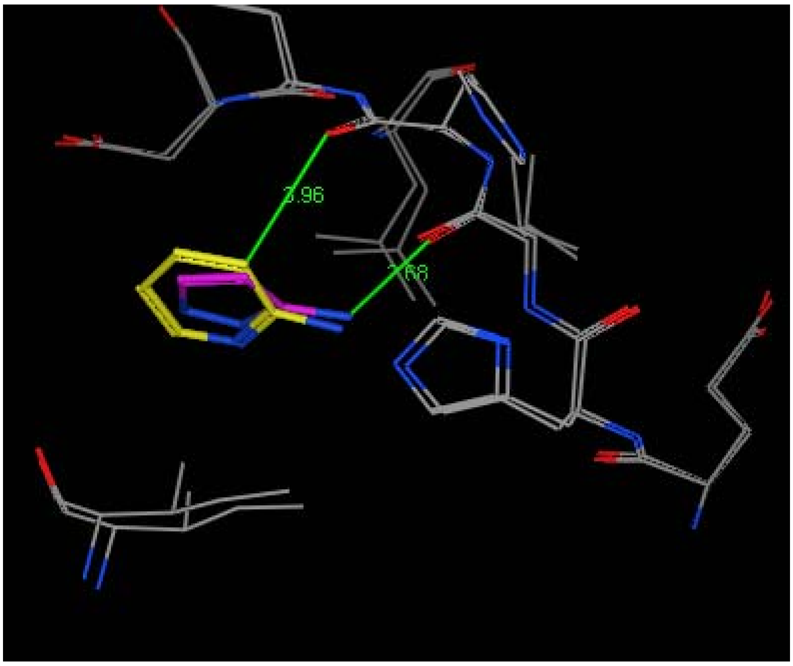
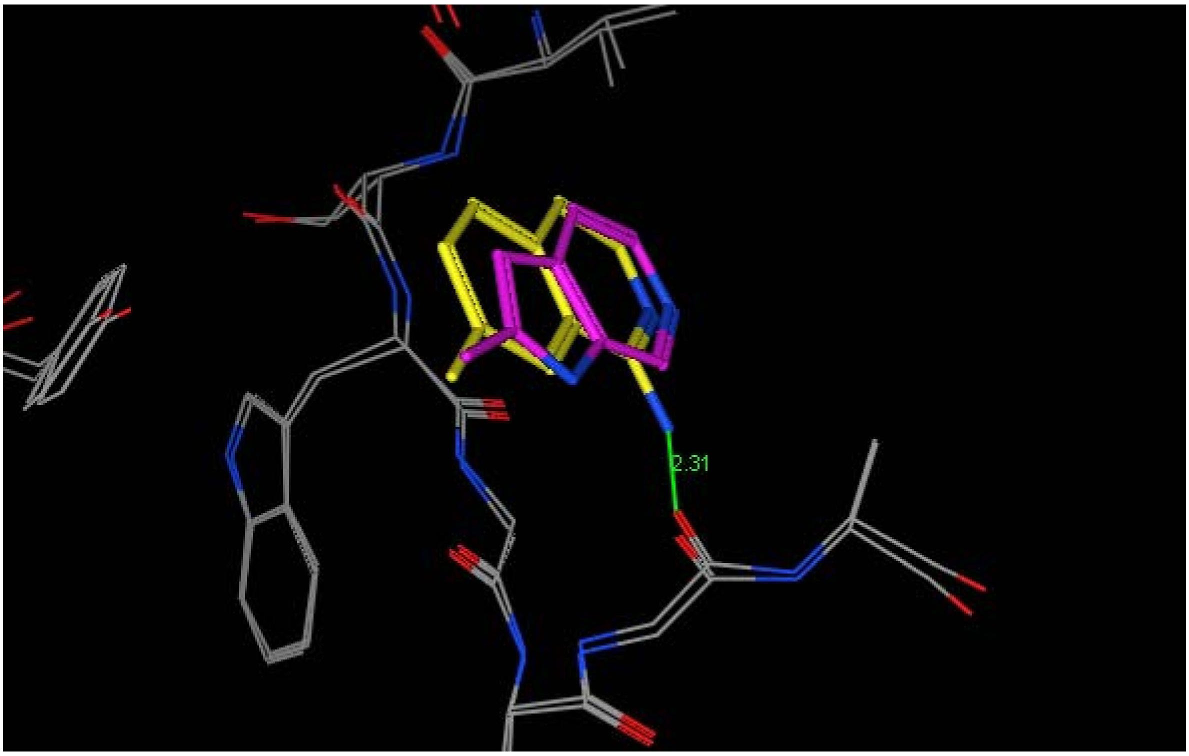
HIV1 protease		
HIV1 protease		
HIV1 protease		
MMP3 (*)		
MMP3		
PDE4		
PDE4		

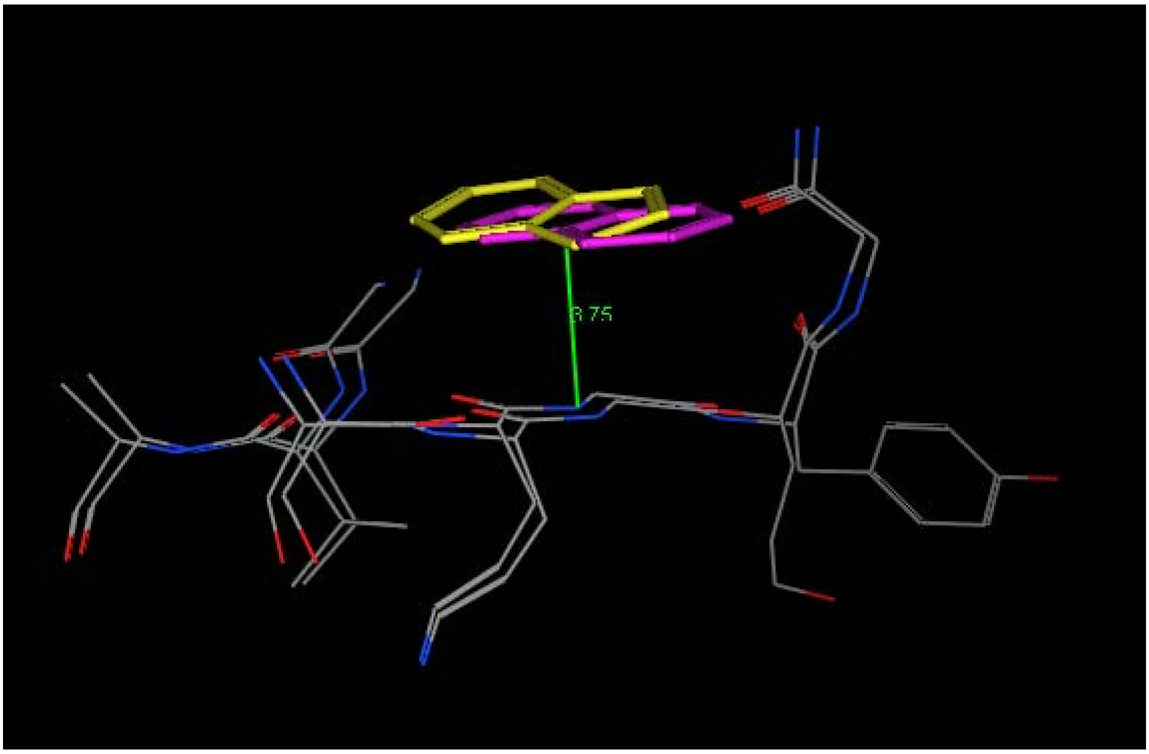
Figure 8. Examples of bioisosteric fragment pairs identified by our procedure. Starred examples are shown in Figure 9.



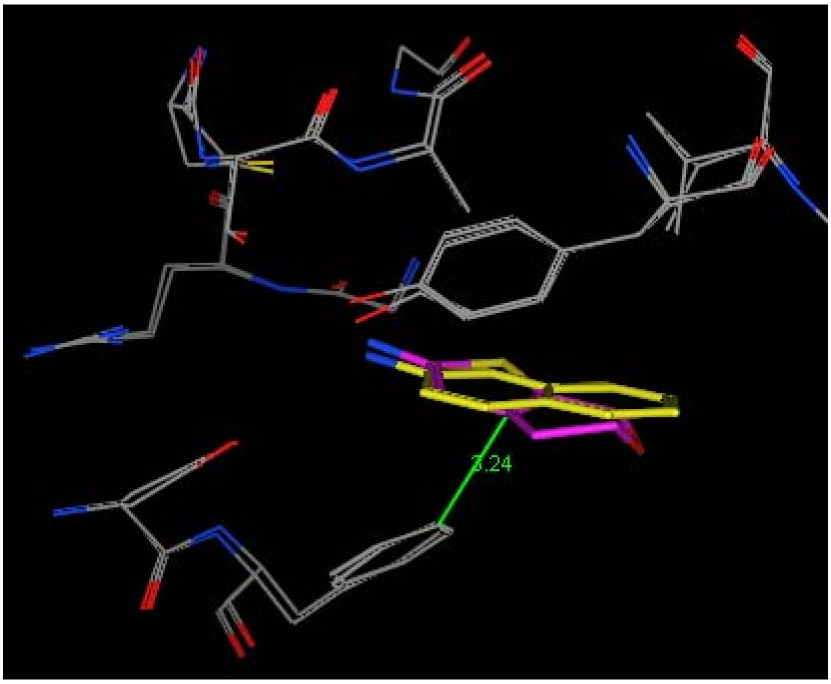
(a) Fragment-pair interactions in CDK4



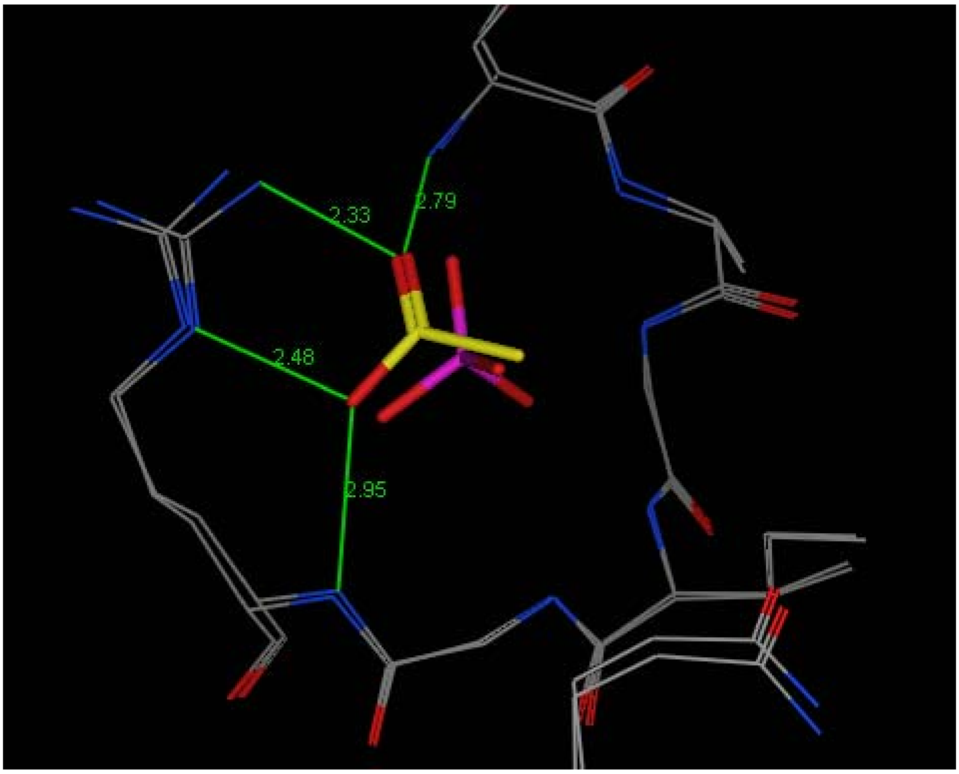
(b) Fragment-pair interactions in Factor Xa



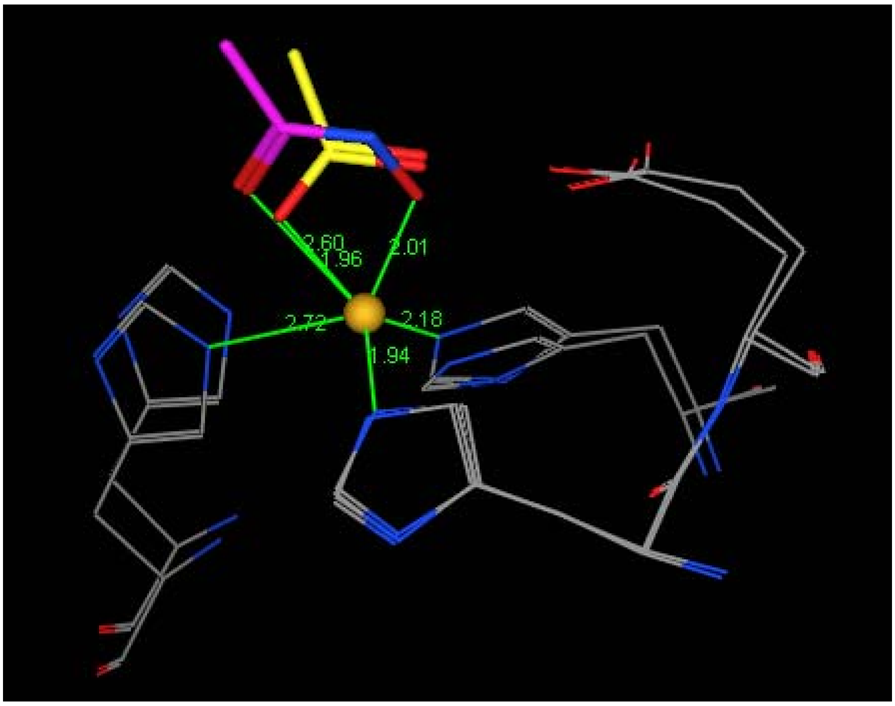
(c) Fragment-pair interactions in Factor Xa



(d) Fragment-pair interactions in tyrosine phosphatase



(e) Fragment-pair interactions in tyrosine phosphatase



(f) Fragment-pair interactions in MMP3

Figure 9. Examples of interactions made by bioisosteric fragment pairs