Nucleic Acids Research Advance Access published May 27, 2011

Nucleic Acids Research, 2011, 1–9
doi:10.1093/nar/gkr287

Swimming into peptidomimetic chemical space using pepMMsMIMIC

Matteo Floris[1], Joel Masciocchi[2], Marco Fanton[3] and Stefano Moro[3,*]

[1]CRS4 – Bioinformatics Laboratory, Parco Sardegna Ricerche, Pula (CA) 09010, [2]CBiB, Université Victor Segalen Bordeaux 2, 33076 Bordeaux Cedex, France and [3]Molecular Modeling Section (MMS), Department of Pharmaceutical Sciences, University of Padova, PD 35131, Italy

Received January 29, 2011; Revised March 22, 2011; Accepted April 12, 2011

ABSTRACT

pepMMsMIMIC is a novel web-oriented peptidomimetic compound virtual screening tool based on a multi-conformers three-dimensional (3D)-similarity search strategy. Key to the development of pepMMsMIMIC has been the creation of a library of 17 million conformers calculated from 3.9 million commercially available chemicals collected in the MMsINC® database. Using as input the 3D structure of a peptide bound to a protein, pepMMsMIMIC suggests which chemical structures are able to mimic the protein–protein recognition of this natural peptide using both pharmacophore and shape similarity techniques. We hope that the accessibility of pepMMsMIMIC (freely available at http://mms.dsfarm.unipd.it/pepMMsMIMIC) will encourage medicinal chemists to de-peptidize protein–protein recognition processes of biological interest, thus increasing the potential of *in silico* peptidomimetic compound screening of known small molecules to expedite drug development.

INTRODUCTION

Protein–protein interactions are central to almost every cellular process from cell motility to DNA replication (1). Alterations in Protein–protein interactions perturb the normal sequence of events in the cell and contribute to conditions such as cancer or neurodegenerative diseases (1). Thus, understanding the normal pattern of Protein–protein interactions can lead to the development of drugs to fight the underlying cause of diseases (1). However, for peptide-based drug design (2), there are several major considerations that limit clinical application such as: (i) rapid degradation by many specific or non-specific peptidases under physiological conditions; (ii) conformational flexibility that allows a peptide to bind to more than one receptor or receptor subtype leading to undesirable side effects; (iii) poor absorption and transportation because of their high molecular mass or lack of specific delivery systems (3).

Peptidomimetics are designed to circumvent some of the above-mentioned problems associated with a natural peptide. Of course, certain other properties, such as receptor selectivity or potency, often can be substantially improved. Hence mimetics have great potential in drug discovery. In an effort to overcome these problems, peptidomimetic drug design has emerged as an important tool for both medicinal chemists and pharmacologists. This approach has evolved into an interdisciplinary scientific endeavour that combines medicinal chemistry, biochemistry, pharmacology and, very recently, chemoinformatics.

In fact, one of the most challenging issues for the future of drug discovery is the capability to incorporate the most crucial protein surface recognition properties, usually displayed in another peptide sequence, into a small organic molecule. With this purpose, we have designed pepMMsMIMIC as a web-oriented peptidomimetics virtual screening tool that, given a peptide three-dimensional (3D) structure, is able to automate a multi-conformers 3D similarity search among 17 million conformers calculated from 3.9 million commercially available chemicals collected in the MMsINC® database (4).

All of pepMMsMIMIC's functions are accessible through a user-friendly web interface as described later in this text.

pepMMsMIMIC PHYLOSOPHY

As anticipated in the 'Introduction' section, pepMMsMIMIC is a public, web-based virtual screening platform with the aim to suggest chemical compounds whose essential elements (pharmacophore) mimic a natural peptide or protein in 3D space, which hopefully retain the ability to interact with the biological target and produce the typical biological effect.

*To whom correspondence should be addressed. Tel: +390498275704; Fax: +390498275366; Email: stefano.moro@unipd.it

Starting from the 3D structure of any Protein–protein/peptide complex, pepMMsMIMIC design process begins by identifying the key residues that are responsible for the Protein–protein recognition process. In this process, the peptide complexity is reduced and the basic pharmacophore model is defined by its critical structural features (peptide annotation points) in 3D space as schematized in Figure 1. All possible peptide pharmacophore feature arrangements can be enumerated to form the basis of a peptide pharmacophore bitstring.

Once generated, the pharmacophore model can be used to screen virtual compound libraries for novel ligands, which present the best similarity to the specific pharmacophore, filtering a database of 3D-conformations based on the positions of the corresponding annotation points derived from each of the ligand conformations. pepMMsMIMIC performs pharmacophore screenings using a multi-conformational version of MMsINC® (multi-confMMsINC) database. Briefly, MMsINC® is a free web-oriented database of commercially available compounds containing around 4 million of non-redundant energetically optimized 3D-chemical structures accessible for virtual screening and chemoinformatic applications (4). In particular, the multi-conf version is characterized by the presence of 17 million of low energy conformers calculated starting from each MMsINC® entry. Also in this case, all possible conformer pharmacophore feature (conformer annotation points) arrangements can be enumerated to form the basis of a conformer pharmacophore bitstring.

To optimize the top peptide mimetic ranking selection, pepMMsMIMIC implements two different scoring functions and one consensus scoring based on the combination of these two. In fact, taking into account that Protein–protein complexes typically exhibit intermolecular interfaces with high electrostatic (chemical) and shape and complementarity, two scoring approaches, such as pharmacophore fingerprints similarity and ultrafast shape recognition (USR) (5), have been implemented to rank and select the best top 200 peptidomimetic candidates.

pepMMsMIMIC WORKFLOW

pepMMsMIMIC's workflow is shown in Figure 2.

The crucial steps of pepMMsMIMIC architecture are detailed as follows.

multi-conf MMsINC: conformers generation

The best five lowest-energy conformers of each MMsINC® entry (including all tautomers and ionic states) are generated by using Rotate ver. 1.0 software

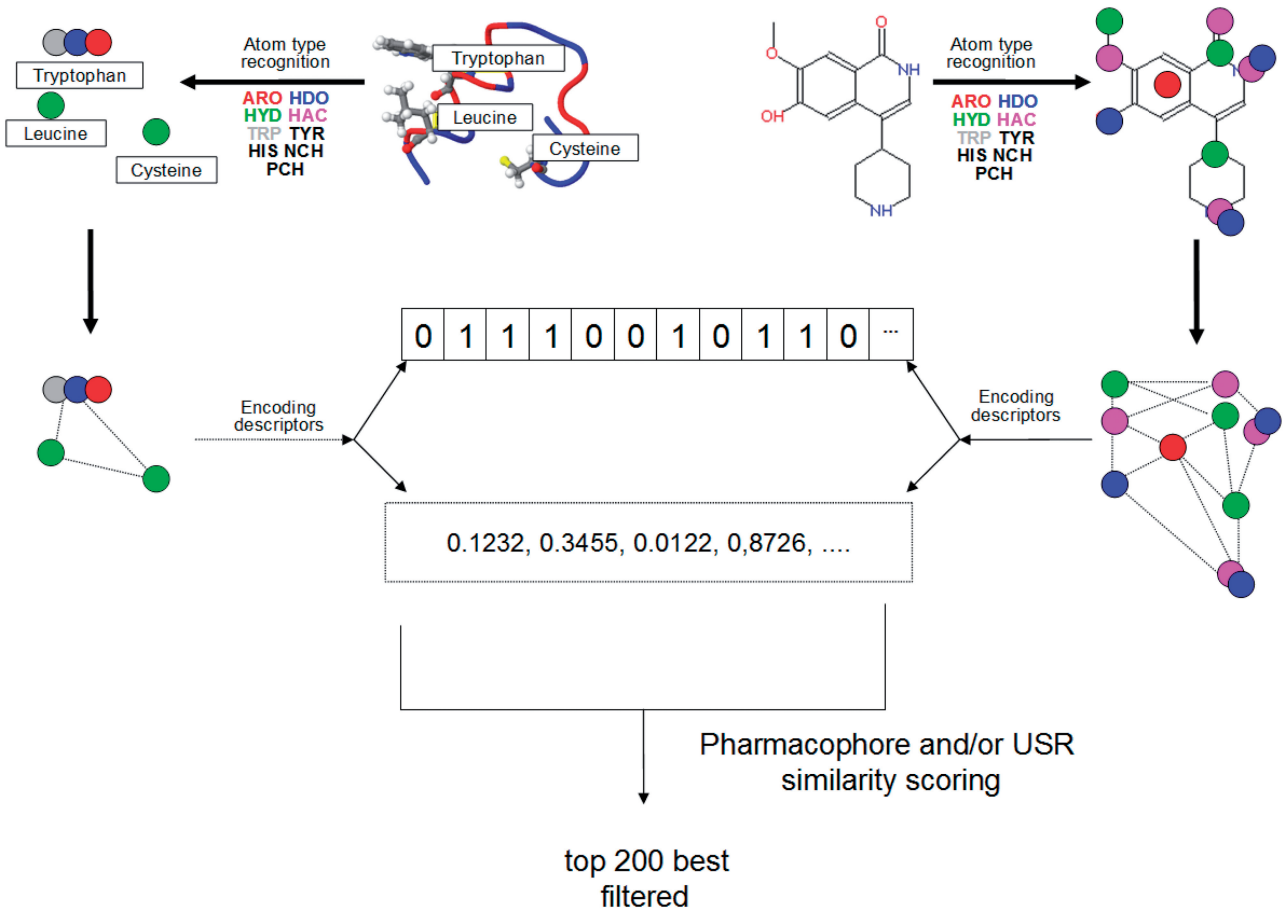


Figure 1. Philosophy behind pepMMsMIMIC's architecture.

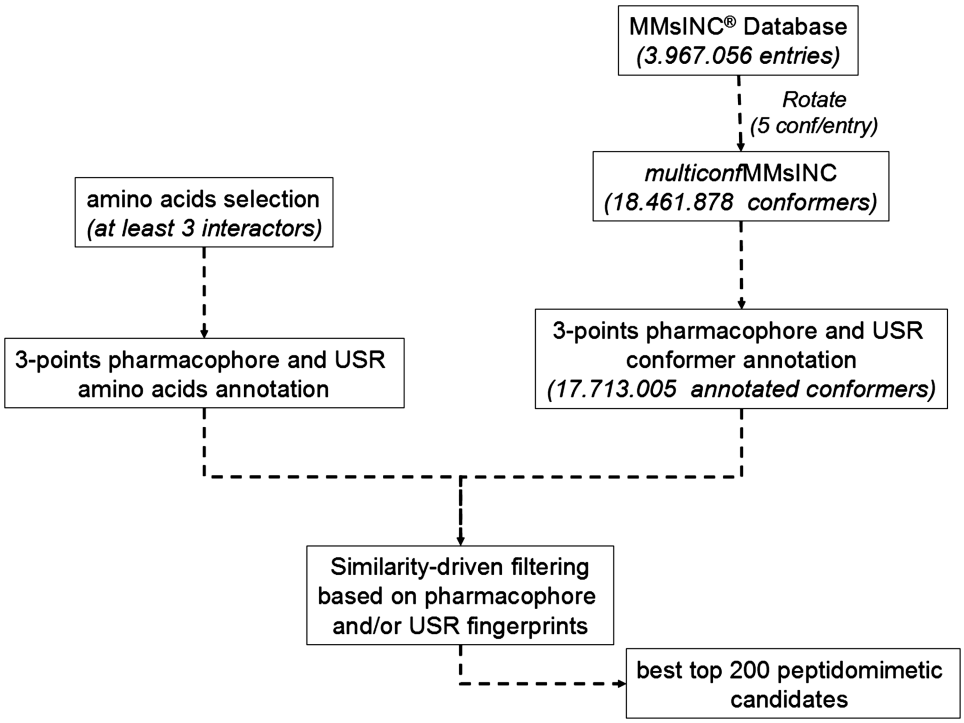


Figure 2. pepMMsMIMIC's workflow.

(http://www.mol-net.de) obtaining an ensemble of 18.461.878 conformers.

Pharmacophore fingerprints generation

Both L- and D-amino acids, as well as phospho-amino acids, are labelled with the following pharmacophoric features: tryptophan side chain, tyrosine side chain, histidine side chain; H-bond acceptor, H-bond donor, positively ionisable, negatively ionisable, aromatic and hydrophobic, as summarized in Table 1.

Both peptide and conformer annotation points ensembles were described in terms of three-point pharmacophores, in which every possible pair of centroids is binned according to the feature distances. There are several works in which pharmacophore fingerprints applied for comparison of protein–ligand recognition in their binding sites are described (6,7). In the present work, we successfully apply this method for the similarity between peptides and ligands pharmacophore fingerprints.

The first criterion used to encode triplets of pharmacophoric points into the pepMMsMIMIC bitstring is based on atom type recognition. All the possible three-point combinations using the above-mentioned nine different centroid types (ARO, aromatic; HYD, hydrophobic; HAC, H-bond acceptor; HDO, H-bond donor; HIS, histidine side chain; NCH, negatively ionizable; PCH, positively ionizable; TRP, tryptophan side chain; and TYR, tyrosine side chain) are encoded in the pepMMsMIMIC bitstring.

Centroids are defined by contiguous atoms with atoms labelled with the same pharmacophoric type (i.e. the six carbon atoms of a benzene ring define an aromatic centroid localized at the centre of the ring). More than one label can be assigned to each atom. Thus, each atom can be part of more than one centroid.

Labels were assigned with an in-house SMARTS mapping tool based on the Chemistry Development Kit (CDK) Java libraries (8,9).

The second criterion used to encode triplets is based on centroid distances. According to the FuzCav method (10), we used a maximum distance cutoff of 14.3 Å and a distance binning defined as follows: [0, 4.8], [4.8, 7.2], [7.2, 9.5], [9.5, 11.9], [11.9, 14.3]. For each class of interaction (i.e. ARO–ARO–PCH, . . . , HDO–HYD–ARO), distance ranges according to the scheme reported below:

ARO–ARO–PCH [0, 4.8][0, 4.8][0, 4.8], . . . , ARO–ARO–PCH [0, 4.8][0, 4.8][11.9, 14.3]
ARO–ARO–PCH [0, 4.8][4.8, 7.2][0, 4.8], . . . , ARO–ARO–PCH [0, 4.8][4.8, 7.2][11.9, 14.3]
ARO–ARO–PCH [0, 4.8][7.2, 9.5][0, 4.8], . . . , ARO–ARO–PCH [0, 4.8][7.2, 9.5][11.9, 14.3]
ARO–ARO–PCH [0, 4.8][9.5, 11.9][0, 4.8], . . . , ARO–ARO–PCH [0, 4.8][9.5, 11.9][11.9, 14.3]
ARO–ARO–PCH [0, 4.8][11.9, 14.3][0, 4.8], . . . , ARO–ARO–PCH [0, 4.8][11.9, 14.3][11.9, 14.3]
ARO–ARO–PCH [4.8, 7.2][0, 4.8][0, 4.8], . . . , ARO–ARO–PCH [4.8, 7.2][0, 4.8][11.9, 14.3]
ARO–ARO–PCH [4.8, 7.2][4.8, 7.2][0, 4.8], . . . , ARO–ARO–PCH [4.8, 7.2][4.8, 7.2][11.9, 14.3]

Using this scheme, each bin is associated with a specific triplet of interaction defined by the type of the vertices comprising the triplet and the relative distances between each centroid pairs. Due to the class definition,

Table 1. List of amino acid side-chain annotations

Aminoacid side chain	Hydrophobic	Positive ionizable	Negative ionizable	H-bond donor	H-bond acceptor	Aromatic	Other
Alanine	V						
Arginine	V	V		V			
Asparagine				V	V		
Aspartic acid			V		V		
Cysteine	V						
Glutamine				V	V		
Glutamic acid			V		V		
Histidine				V	V	V	His ring
D-Histidine				V	V	V	His ring
Isoleucine	V						
Leucine	V						
Lysine		V		V			
Methionine	V						
Phenylalanine	V					V	
Proline	V						
Serine				V	V		
Threonine	V			V	V		
Tryptophan				V		V	Trp ring
Tyrosine				V	V	V	Tyr ring
Valine	V						
O-phosphotyrosine			V			V	
Phosphoserine			V				
Phosphothreonine	V		V				
D-Phosphothreonine	V		V				
D-Alanine	V						
D-Arginine	V	V		V			
D-Asparagine				V	V		
D-Aspartic acid			V		V		
D-Cysteine	V						
D-Glutamine				V	V		
D-Glutamic acid			V		V		
D-Histidine				V	V	V	His ring
D-Isoleucine	V						
D-Leucine	V						
D-Methionine	V						
D-Phenylalanine	V					V	
D-Proline	V						
D-Serine				V	V		
D-Threonine	V			V	V		
D-Tryptophan				V		V	Trp ring
D-Tyrosine				V	V		Tyr ring
D-Valine	V						

ARO–ARO–PCH, ARO–PCH–ARO or PCH–ARO–ARO (triplet) belongs to the same triplets of interaction ARO–ARO–PCH. Every time a triplet is made of two aromatic centroids and one positively ionizable centroid, the first criterion will be used to associate it with the class ARO–ARO–PCH, and the second criterion (based on atom pair distances) will be used to correctly locate the triplet inside the ARO–ARO–PCH class within the pepMMsMIMIC bitstring. The same rule is applied to all three-point pharmacophore (ARO–ARO–ARO, ARO–NCH–PCH, . . . , HYD–HYD–HYD) combinations.

The same classification scheme is adopted for both ligands and peptide residues, with the difference that one or more pharmacophoric points are assigned to the geometric centre of the side chains according to the classification reported in Table 1.

The fingerprint is a vector with possible 19.815 bits; among them, only 12.448 bits were populated in our virtual library.

Finally, only those conformers for which at least three spatially distinct features were retained. We stored both USR descriptors and fingerprints for a total of 17.713.005 conformers.

USR methodology

The USR, introduced by Ballester and Richards (5), is a fast 3D similarity search method based on the assumption that the shape of a molecule is uniquely determined by the relative position of its atoms. The approach is based on moments of distance distributions, and it has been successfully applied to the fast identification of similarly shaped compounds within large molecular databases.

In the USR encoding, the shape of the atomic ensemble (a molecule or any generic set of atoms) is characterized by the distributions of atomic distances to four reference locations: the molecular centroid (*ctd*), the closest atom to ctd (*cst*), the farthest atom to ctd (*fct*) and the farthest

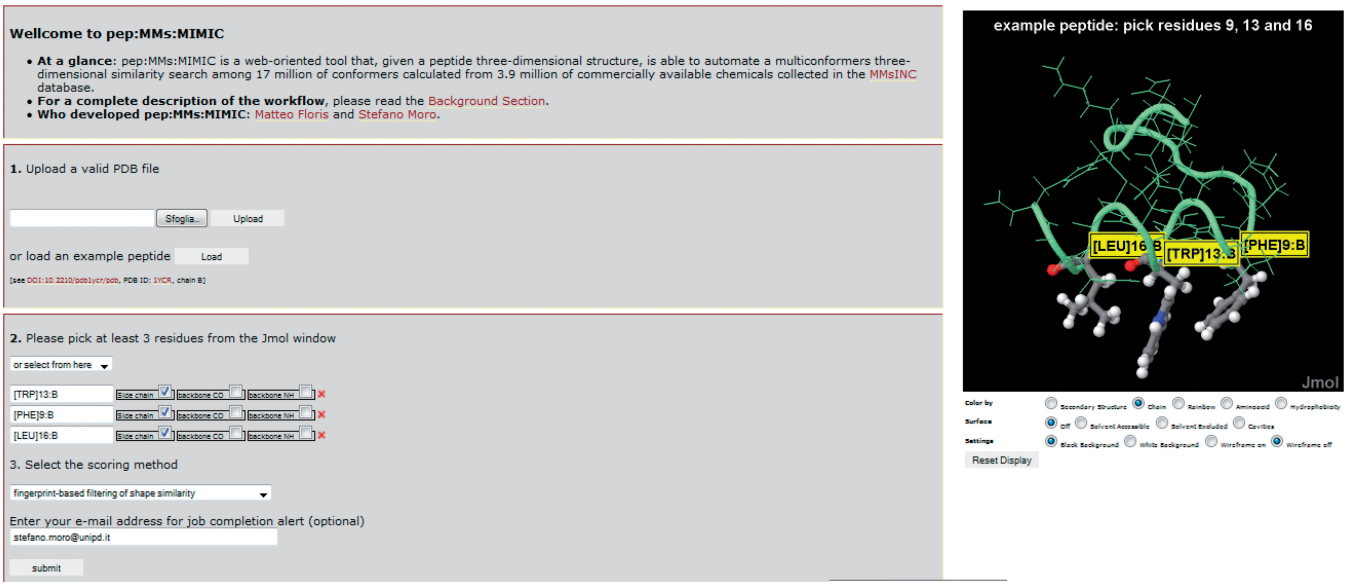


Figure 3. pepMMsMIMIC's main screen.

atom to fct (*ftf*). Overall, each of these distributions is described through its first three vectors. In this way, each molecule (or atomic ensemble) has associated a vector of 12 shape descriptors.

Ballester and Richards (5) proposed in their original paper the use of this method for fast similarity comparisons between macromolecules; here we extend the use of USR method to the comparison between a collection of ligands and a protein atomic ensemble. Our in-house Python implementation of the USR algorithm (available on request to the Authors) has been used to calculate the USR descriptors for the whole data set and it is used for the on-the-fly calculation of protein atoms ensemble defined by the user's selection.

Scoring metrics

Fingerprints focusing methods commonly use similarity coefficients, such as Tanimoto, to retrieve or classify compounds out of a typically large chemical library. However, the results of similarity searching are often systematically affected by bit density differences between reference and database molecules. For example, in a previous study, Tversky similarity calculations showed apparent asymmetric behaviour because optimal search performance was achieved by assigning higher weights on bit settings of reference than database compounds. For Tversky similarity calculations, such biasing effects could be corrected by introducing the weighted Tversky coefficient, which made possible to set relative weights on '1′ and '0′ bits and thereby balance complexity differences between reference and database molecules. In the current version of pepMMsMIMIC, fingerprints similarity measure has been implemented, based on a weighted similarity index (S_w) calculated as below:

$$S_w = c/(c+2.5 \times m)$$

where c is the number of common bits between the peptide query fingerprint and conformer's fingerprint, and m is the count of bits on in the query fingerprint but not conformer's fingerprint.

Four different scoring methods are actually implemented in the current version of pepMMsMIMIC:

(i) shape score (ShS) based on the USR methods;
(ii) pharmacophoric fingerprint similarity (PFS) based on weighted similarity coefficient (S_w);
(iii) combined ShS and PFS filtering: conformers are primarily filtered by using a ShS score threshold (which must be $\geq$0.5) and subsequently sorted by PFS;
(iv) weighted ShS and PFS approach: in order to reduce the number of false positives, we introduced an hybrid scoring function where a weighted combination of the ShS and PFS based on weighted similarity coefficient (S_w) is combined as reported below:

$$\text{ShS and PSF} = (0.4 \times \text{ShS}) + (0.6 \times \text{PFS}).$$

Weight terms were chosen based on a preliminary in-house validation.

QUERYING pepMMsMIMIC

The pepMMsMIMIC is accessible to the public via our web application. It allows users to easily upload a PDB structure and to manage it using Jmol interface (7,8).

The selection of the key residues that are responsible for the Protein–protein recognition process is intuitively carried out either by picking them up from the Jmol interface or by selecting them from the amino acids list available on the left side of the webpage.

For each selected amino acid, users can specifically decide which pharmacophoric feature to include in the annotation process. In the present-version pepMMsMIMIC,

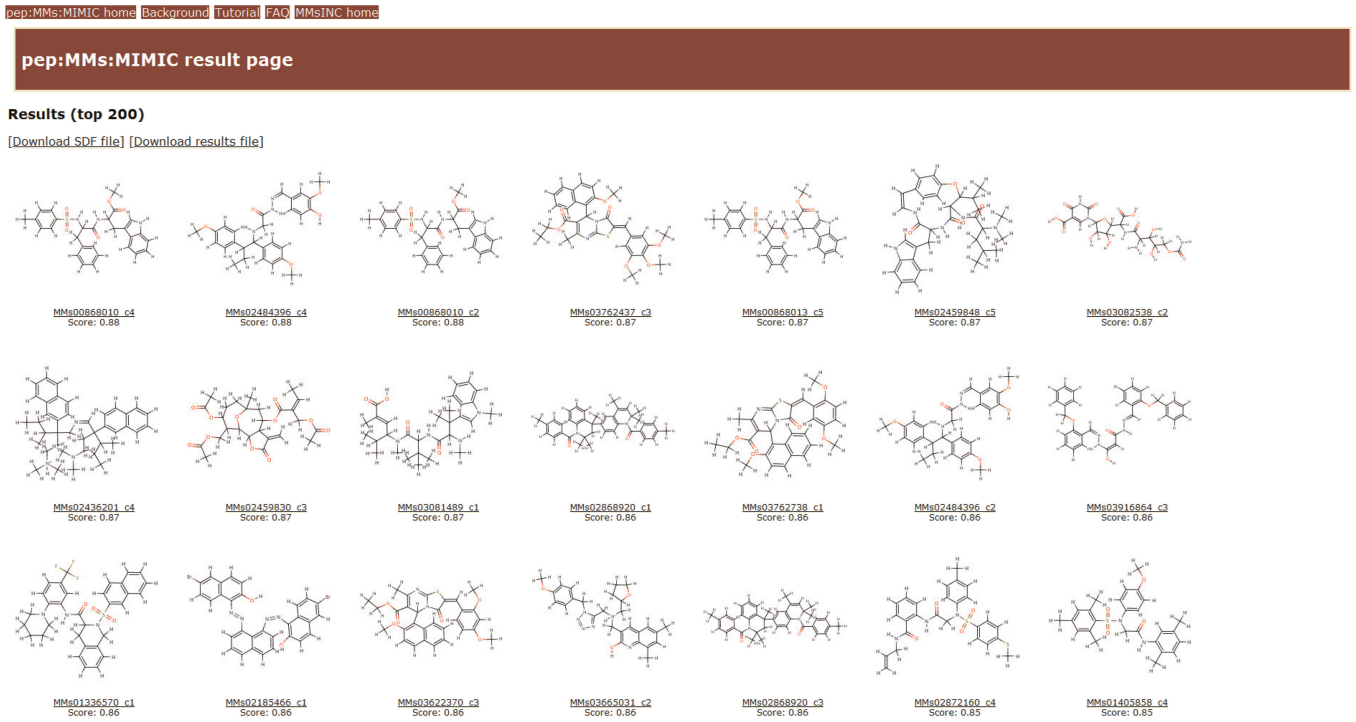


Figure 4. pepMMsMIMIC's peptidometic candidates result screen.

users can select both CO and/or NH interactors deriving from the carbamide bonds of the protein backbone as well the corresponding side chain moiety. At least three interactors are required for starting a search process: this means that in principle one residue alone is enough for a search run if all the three features are selected.

All selected amino acids are clearly labelled in the Jmol interface (Figure 3).

As previously described, before running pepMMsMIMIC search, users can also decide with scoring approach use to rank and select the best top 200 peptidomimetic candidates. If more than one conformers associated with a specific chemical compound are ranked among the top 200, only the best scored conformer will be retrieved. The combined ShS and PFS filtering scoring is selected as default only because it got the best scoring performance in our preliminary validation analysis.

During the search process, the temporary best scored candidate will appear on the main pepMMsMIMIC webpage, and it will be automatically updated as soon as a novel best scored candidate will be found. A complete search run requires from 15 to 20 min, depending on the search method (USR is faster than combined methods).

Displaying peptidomimetic candidates query results

The best top 200 peptidomimetic candidates are finally displayed as soon as the search run is completed. They are sorted in numeric order, from the highest to the lowest score value (Figure 4).

For each result, the pepMMsMIMIC displays the chemical structure of conformer and its MMsCode (as an example: the candidate 'MMs00868010_c4' refers to the conformer number 4 of the MMsINC entry 00868010).

Clicking on the MMsCode of each peptidomimetic candidate the user will get the molecule report of MMsINC database (Figure 5). The report shows basic information about the molecule like the compound type (neutral, tautomer or ionic state), the molecular formula, and its InChI and SMILES representations. The report also contains a 2D image of the molecule, and a 3D movable rendering of molecule shown using Chemis3D (http://chemis.free.fr/mol3d/) Java applet. In addition, the pre-calculated descriptors for the molecule are listed at the bottom of the report. Finally, for neutral molecules the system lists all its tautomers and ions, while for tautomers and ions the neutral state of the molecule is indicated.

Moreover users can easily download, as .sdf file, the structures of all peptidemimetic candidates for further in-house processing and analysis. Single Autodock input files can also be retrieved from each molecule report page, thus providing an easier integration with molecular docking tools.

Implementation

The pepMMsMIMIC system is installed on a server running Linux. The system's web application has been developed in PHP, with some components written in Java and Python. pepMMsMIMIC uses the Chemistry Development Kit (CDK) to perform some of its molecular analyses.

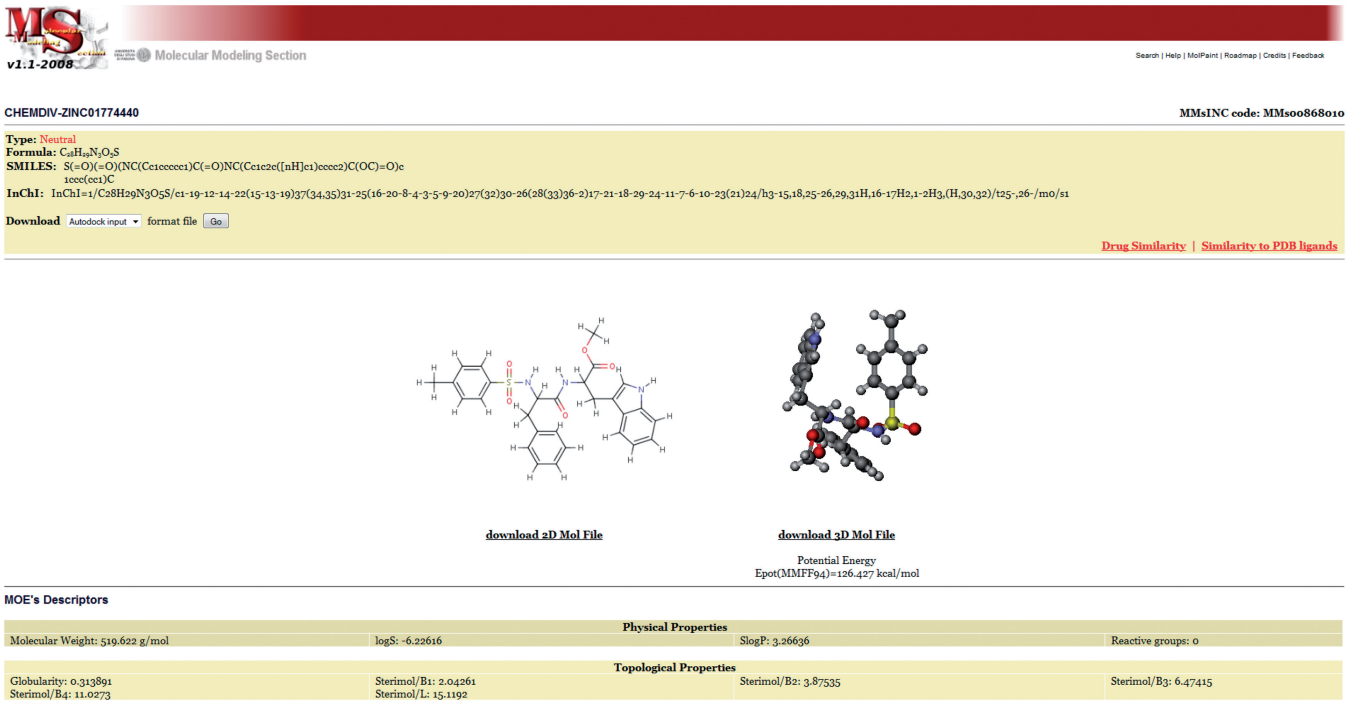


Figure 5. MMsINC's report for molecule MMs00868010.

PRELIMINARY VALIDATION ANALYSIS

With the aim to validate the quality of pepMMsMIMIC candidates selection, five known MDM2/p53 Protein–protein inhibitors, belong to the class of Nutlins (11), have been voluntarily included into MMsINC® database. Their structures are summarized in Table 2. Interestingly, starting from the crystallographic structure of MDM2/p53 complex reported in our demo key-study (PDB code: 1YCR), all five nutlin analogues were ranked among the top 0.6% of the entire multiconfMMsINC® database using the combined ShS and PFS filtering. Although preliminary, these results are surely encouraging. However, further validation analyses are in progress in our laboratories.

CONCLUSION

Gigantic amounts of genomic and proteomic data creates high demand for synthesis and screening of nature-like biopolymers and their more stable modified derivatives. Design and synthesis of peptidomimetics are important because of the dominant position peptide and Protein–protein interactions that play a role in molecular recognition and signalling, especially in living systems.

Examination of the vast literature would suggest that medicinal and organic chemists, who deal with peptide mimics utilize these methods in many different ways. In any case, a variety of methodologies and strategies have been developed and continue to be developed to establish systematic tools for transformation of peptides into peptidomimetics or further into small drug-like molecules. Significant industrial as well as academic resources are invested in this effort and there is still much to learn to optimize these approaches. We envision that peptidomimetic research will continue to be an indispensable tool of structure–activity relationships in drug discovery for the foreseeable future.

We hope that the accessibility of pepMMsMIMIC will encourage chemists to de-peptidize Protein–protein recognition processes of biological interest, thus increasing the potential of *in silico* peptidomimetic candidates screening of known small molecules to expedite drug development. Further integration with consolidated virtual screening tools, such as pharmacophore screening and molecular docking, will be available in the next release of pepMMsMIMIC.

Citing

If you refer to pepMMsMIMIC tool and web interface for your published research, we ask that you please cite this article.

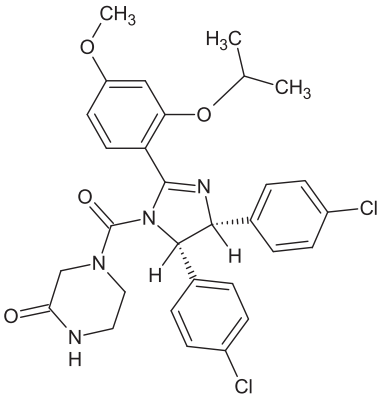
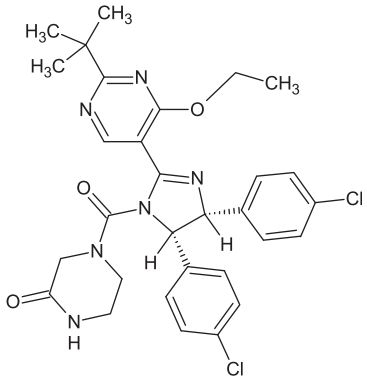
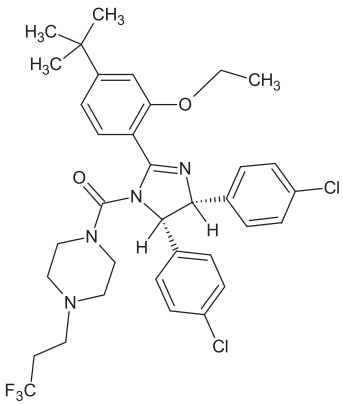
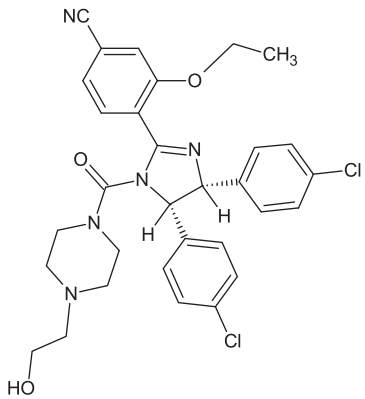
ACKNOWLEDGEMENTS

Stefano Moro is very grateful to Molecular Networks and to Chemical Computing Group for the long and fruitful collaborations.

FUNDING

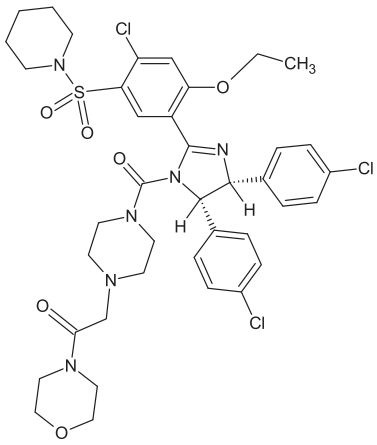
University of Padova, Italy (to S.M.); Italian Ministry for University and Research, Rome, Italy (MIUR,

Table 2. Selected Nutlins as validation test set

Test set	Shape only[a]	Pharmacophore only[b]	Pharmacophore after Shape[c]	Hybrid[d]
	1 475 800 (8.34)	80 365 (0.45)	*42 871 (0.24)*	347 751 (1.96)
	1 316 214 (7.44)	220 946 (1.25)	*110 684 (0.62)*	481 122 (2.72)
	1 469 866 (8.30)	220 945 (1.25)	*110 683 (0.62)*	629 788 (3.56)
	29 643 (0.17)	80 364 (0.45)	*1582 (0.009)*	19 243 (0.11)

(continued)

Table 2. Continued

Test set	Shape only[a]	Pharmacophore only[b]	Pharmacophore after Shape[c]	Hybrid[d]
	29 647 (0.17)	80 363 (0.45)	*1584 (0.009)*	19 244 (0.11)

Values are represented as *n* (%). The absolute position of the best-ranked conformer of each nutlin, and the corresponding 'top X%' position respect the entire *multiconf*MMsINC (with a total of 17.713.005 conformers) have been reported. Values in italic refer to the best performed scoring method.
[a]ShS based on the USR methods.
[b]PFS based on weighted S_w.
[c]Combined ShS and PFS filtering.
[d]Weighted ShS and PFS approach.

PRIN2008: protocol number 200834TC4L_002, to S.M.).
Funding for open access charge: University of Padova.

Conflict of interest statement. None declared.

REFERENCES

1. Zanzoni,A., Soler-López,M. and Aloy,P. (2009) A network medicine approach to human disease. *FEBS Lett*., 2009, 583, 1759–1765.
2. Zhao,L. and Chmielewski,J. (2005) Inhibiting protein–protein interactions using designed molecules. *Curr. Opin. Struct. Biol*., **15**, 31–34.
3. Vagner,J., Qu,H. and Hruby,V.J. (2008) Peptidomimetics, a synthetic tool of drug discovery. *Curr. Opin. Chem. Biol*., **12**, 292–296.
4. Masciocchi,J., Frau,G., Fanton,M., Sturlese,M., Floris,M., Pireddu,L., Palla,P., Cedrati,F., Rodriguez-Tomé,P. and Moro,S. (2009) MMsINC: a large-scale chemoinformatics database. *Nucleic Acids Res*., **37**, D284–D290.
5. Ballester,P.J. and Richards,W.G. (2007) Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem*., **28**, 1711–1723.
6. Mason,J.S., Morize,I., Menard,P.R., Cheney,D.L., Hulme,C. and Labaudiniere,R.F. (1999) New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem*., **42**, 3251–3264.
7. Karnachi,P. and Kulkarni,A. (2006) Application of pharmacophore fingerprints to structure-based design and data mining. In Langer,T. and Hoffmannn,R.D. (eds), *Pharmacophores and Pharmacophore Searches*. Wiley-VCH, Weinheim, Germany, pp. 193–206.
8. Steinbeck,C., Han,Y.Q., Kuhn,S., Horlacher,O., Luttmann,E. and Willighagen,E.L. (2003) The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comp. Sci*., **43**, 493–500.
9. Steinbeck,C., Hoppe,C., Kuhn,S., Floris,M., Guha,R. and Willighagen,E.L. (2006) Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr. Pharma. Design*, **12**, 2111–2120.
10. Weill,N. and Rognan,D. (2010) Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model*., **50**, 123–135.
11. Weber,L. (2010) Patented inhibitors of p53–Mdm2 interaction (2006–2008). *Exper. Opin. Ther. Patents*, **20**, 179–191.