Web Server To Identify Similarity of Amino Acid Motifs to Compounds (SAAMCO)

Fergal P. Casey,[†] Norman E. Davey,[†] Ivan Baran,[‡,||] Radka Svobodova Varekova,[§] and Denis C. Shields*,[†]

UCD Conway Institute of Biomolecular and Biomedical Sciences and Complex and Adaptive Systems Laboratory (CASL), University College Dublin, Dublin, Ireland, Siemens Research Ireland, Dublin, Ireland, and ANF Data, a Siemens Company, Brno, Czech Republic

Received February 19, 2008

Protein−protein interactions are fundamental in mediating biological processes including metabolism, cell growth, and signaling. To be able to selectively inhibit or induce protein activity or complex formation is a key feature in controlling disease. For those situations in which protein−protein interactions derive substantial affinity from short linear peptide sequences, or motifs, we can develop search algorithms for peptidomimetic compounds that resemble the short peptide's structure but are not compromised by poor pharmacological properties. SAAMCO is a Web service (http://bioware.ucd.ie/~saamco) that facilitates the screening of motifs with known structures against bioactive compound databases. It is built on an algorithm that defines compound similarity based on the presence of appropriate amino acid side chain fragments and a favorable Root Mean Squared Deviation (RMSD) between compound and motif structure. The methodology is efficient as the available compound databases are preprocessed and fast regular expression searches filter potential matches before time-intensive 3D superposition is performed. The required input information is minimal, and the compound databases have been selected to maximize the availability of information on biological activity. "Hits" are accompanied with a visualization window and links to source database entries. Motif matching can be defined on partial or full similarity which will increase or reduce respectively the number of potential mimetic compounds. The Web server provides the functionality for rapid screening of known or putative interaction motifs against prepared compound libraries using a novel search algorithm. The tabulated results can be analyzed by linking to appropriate databases and by visualization.

BACKGROUND

We have developed a Web server implementation of a robust and efficient work flow for large-scale screening of small molecule databases against PDB protein structures. The work flow relies on identifying small molecules with substituents that topologically and structurally resemble key amino acid side chains. We demonstrated the utility of this work flow by comparing a large compound database to a large database of peptide regions from the surfaces of protein structures.[1] The Web server implementation provides a simple interface to search preprocessed conformational libraries from three compound databases and return compounds matching the 3D configuration of short peptide sequences belonging to protein structures. The method has 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" that match the input peptide sequences, and (3) estimation of the Root Mean Squared Deviation (RMSD) between the PDB structure of the amino acid sequence and the modeled conformations of the compound. The first two steps correspond to a search for 2D topological similarity (i.e., amino acid side chains and compound fragments with the same atom types and connectivity), while the final step then determines the 3D structural similarity. The Web server work flow is schematically shown in Figure 1.

The server provides the restricted functionality of comparing short regions on protein surfaces or known short linear motifs against small compound databases of biologically active compounds, in order to improve efficiency and analysis of compound matches.

IMPLEMENTATION

Databases and Filters. We have selected three compound databases for inclusion in the SAAMCO Web server: FDA approved drugs,[2] NCI Drug Therapeutic Program (NCI-DTP) database of compounds screened against 60 cancer cell lines,[3] and a database of known protein ligands from CHEMPDB.[4] The first preprocessing step for these databases involves mining for amino acid side chain substituents. The exact procedure is described elsewhere,[1] but in summary we seek terminal fragments of compounds that have full or partial similarity to amino acid side chains (excluding glycine, alanine, and proline). We discovered at least two amino acid-like substituents in 219 drugs out of the 1508 FDA approved drugs, in 5839 compounds out of the 42247 NCI-DTP compounds, and in 1552 ligands out of the 7321 ligands in CHEMPDB.

The next step is to generate conformers for the amino acid-like compounds. This was carried out with Molecular

* Corresponding author e-mail: denis.shields@ucd.ie.
[†] University College Dublin.
[‡] Siemens Research Ireland.
[||] Current address: Havok, Thomas St., Dublin, Ireland.
[§] ANF Data.

10.1021/ci8000474 CCC: $40.75
Published on Web 06/21/2008

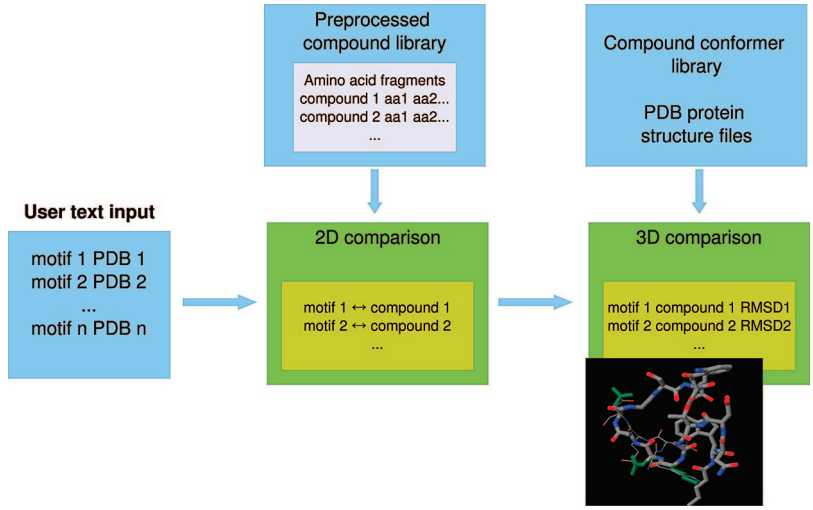


Figure 1. Schematic work flow for the SAAMCO program.

Operating Environment (MOE).[5] Up to 50 of the lowest energy conformers were generated for 218 out of the 219 suitable FDA drugs, since the peptide drug Acthrel was too large for conformation generation. Although the number of conformations chosen is somewhat arbitrary, about half of the FDA drugs resulted in less than 50 conformers to represent their low energy minima, and the other compounds' conformers are the most probable configurations. The candidate set of NCI-DTP compounds was prefiltered by MOE before conformer generation using default filters which generally reduces the number of large compounds with many rotatable bonds and/or rings and increases drug likeness with the intention of enriching the database for more appropriate potential lead compounds and improving conformational coverage. [MOE default filters: molecular weight >600, logP <−4, logP >8, donors+acceptors >12, rotatable bonds >7, single bond chain length >6, chiral centers >4, unconstrained chiral centers >3, rings >8, d-hybrids. Compounds satisfying any of the inequalities are removed from the database. The default metal ion filter was omitted to avoid excluding salts.] The final filtered set of compounds with 3-D coordinates numbered 3248 out of the original 5839 2-D structures. The CHEMPDB database consists of X-ray or NMR 3-D structures in PDB format and so has the advantage of not requiring conformation generation.

The NCI-DTP database was filtered further on the basis of growth inhibition values. We start with a subset of compounds that have pGI_{50} values for almost all of 36 cancer cell lines, prepared previously[6] and available for download. We refined this data set by keeping only those compounds with a significant growth inhibition effect by comparing the estimated pGI_{50} values to the negative logarithm maximum concentration used in the assay (either 4 or 8 where concentrations are molar). If the mean pGI_{50} is greater than the negative logarithm of the maximum concentration and if the standard deviation of pGI_{50} values is greater than 0.1 (ensuring broad weak activity across most of the cell lines or significant activity on a few cell lines), then we define the compound to be active. The combination of these stringent requirements (fairly complete set of pGI_{50} values and activity) leaves 357 NCI-DTP biologically active substances for searching.

The protein structures for comparison are downloaded automatically from the RCSB Protein Data Bank[7] using the supplied PDB identifier, unless the structure is already available on the server. An extra option allows for a protein or peptide PDB structure to be uploaded, in the case that the PDB formatted structure is not available from the RCSB database or has been modified.

INPUT DATA, SETTINGS, AND OUTPUT FORMAT

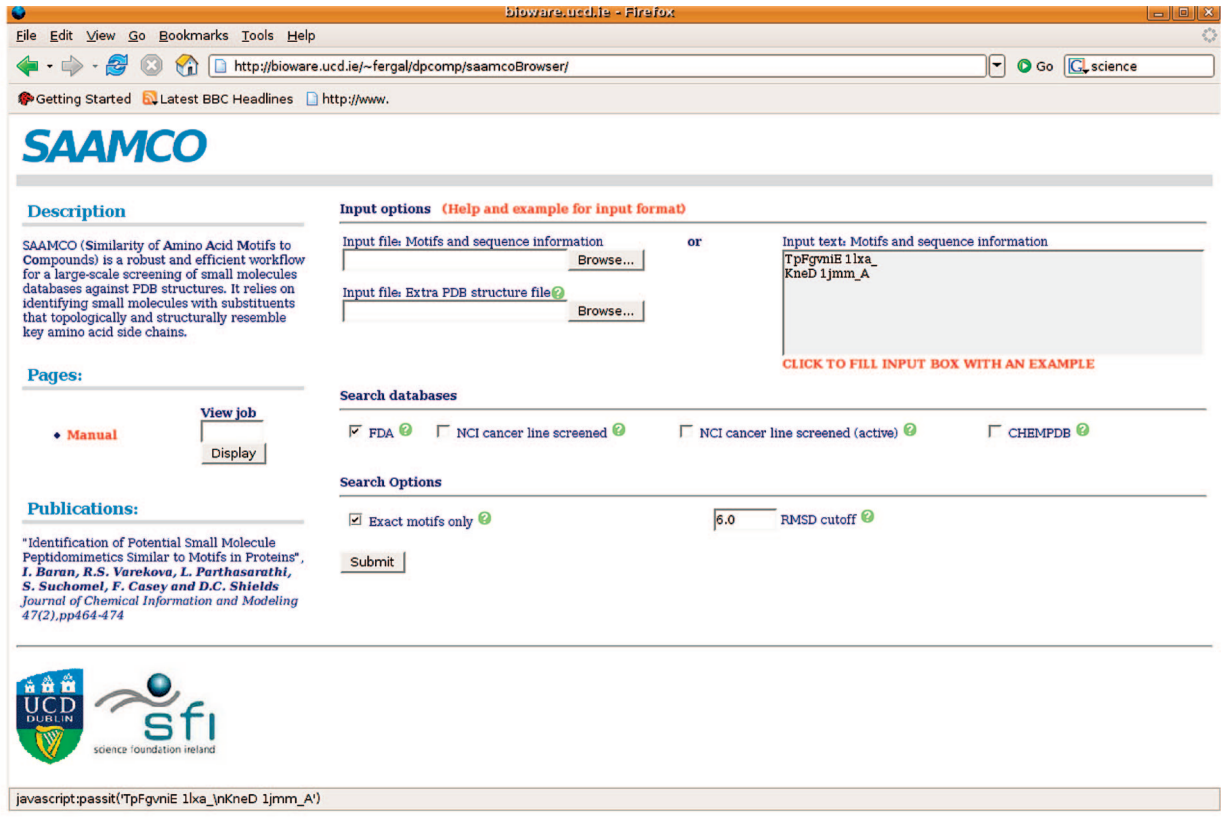
Input Data. The input format consists of lines of text with two whitespace-separated entries on each line: a sequence motif and a PDB structure/chain identifier on which the motif occurs. An example is

```
TpFgvniE 1lxa_
KneD 1jmm_A
```

Note that lowercase one letter amino acid codes denote wildcard positions which we ignore when matching to the compound databases. The full sequence is given to locate the motif in the PDB structure with negligible ambiguity. The PDB 4 alphanumeric character identifier is followed by an underscore and the chain letter for the chain carrying the motif. The chain identifier may be a space as in the first line of this example. The text can either be input as a text file or in a text box on the main page, Figure 2(a). By default, if the text box is nonempty, its contents overwrite the text file contents. For a small example of correctly formatted input text, a link is provided which automatically fills the text box. Another input box is provided to upload a PDB structure file: this is only required if the peptide sequence is not part of a structure in the RCSB Protein Data Bank or if the PDB structure has been specifically modified.

RMSD Comparisons. Comparisons of submitted motifs to one or more of the compound databases differ whether or not the *Exact motifs only* option is selected in the search options of the Web server window. If it is selected, then the submitted sequence, which may contain wildcards (e.g., FpSvS), will be matched completely to compounds that have all the same amino acid-like substituents as the nonwildcard positions (in this example a compound will need to have at least an F and two S like substituents). This is followed by the 3D superposition where all the conformers of the matched

(a)



(b)

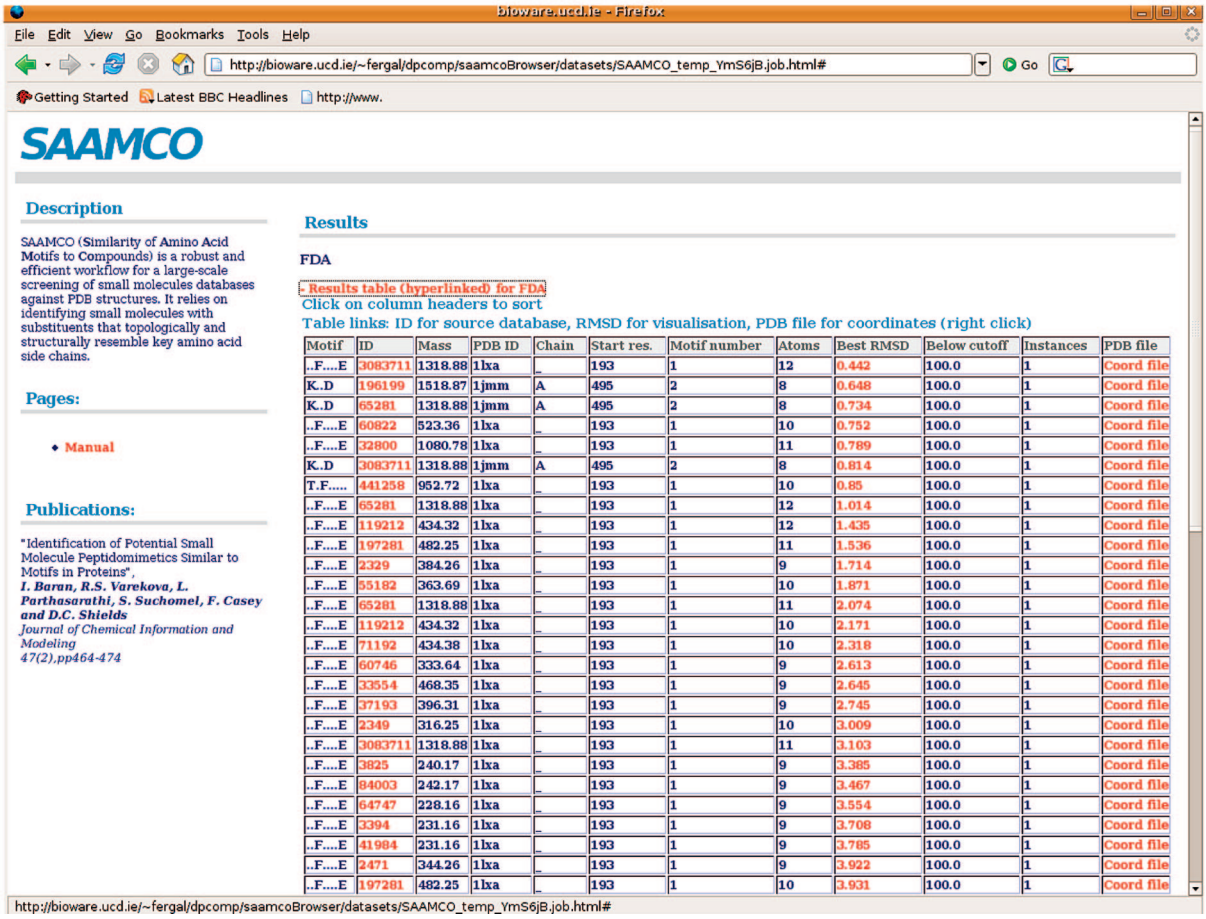


Figure 2. (a) Input boxes and options for SAAMCO server. (b) Output table of results. Each row is a compound-protein match.

compound are used to generate a set of RMSD values. For each conformer, if there are multiple amino acid-like substituents of the same type, all possible superpositions are performed and the lowest RMSD is retained. This type of matching is relatively fast, as for a given set of motifs with two or three nonwildcard positions there will be a reasonably

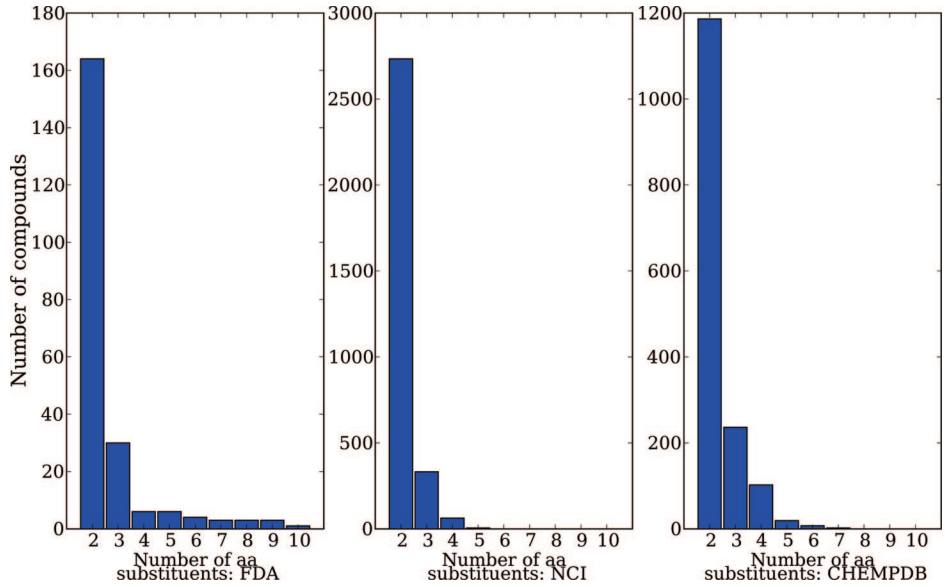


Figure 3. Distribution of number of amino acid substituents per compound in the three input databases.

low number of 2D matches to the drug database. It should be noted that if sequences with more than 4 nonwildcard positions are passed in, it is likely that few 2D matches will be obtained, given that most compounds in the databases have less than 4 amino acid-like substituents, see Figure 3 for the composition of the databases.

When the *Exact motifs only* option is not selected, all subsequences within the passed in protein motif will be searched for a potential 2D match to compounds in the database, which is again followed by an RMSD calculation. As the number of subsequences increases exponentially with the number of nonwildcard positions in the input sequence, care must be taken to limit the number of nonwildcard positions if this option is used. A warning is issued if the total number of subsequences (over all motifs) exceeds 250, and the job is automatically assigned lower priority.

The *RMSD cutoff* input box is used to select an RMSD value above which the matching compound will not be reported in the output table. This reduces the size of the output table but does not reduce computational time significantly, as the same number of superpositions must be performed independently of this cutoff.

Output Format and Examples. The results of the 2-D and 3-D comparisons are compiled into a table where each row is a different compound-protein match. Each row contains the compound and protein IDs, the best RMSD (from all conformers) for each compound-protein pair, the sequence motif (where wildcard positions are replaced with "."), the starting residue for the shared sequence, the number of atoms superimposed, and the percentage of conformers with RMSD less than the cutoff, Figure 2(b).

A separate table is produced for each search database, and the entries are linked by compound ID to the appropriate source database Webpage, i.e. PubChem, NCI-DTP, or CHEMPDB, Figure 2(b). We also link the table's RMSD value to a rotatable representation of the superimposed compound-peptide structure, by means of the Jmol molecular viewer[8] which can be run from the Web browser, if enabled. This allows for a visual assessment of the fit and for an examination of flanking regions of the compound which may render the potential peptidomimetic unacceptable (e.g., bulky or highly charged regions outside of the matched fragments). Alternatively, a downloadable PDB file containing atomic coordinates for the protein sequence matched, the superimposed atomic coordinates for the compound and information that describes which "residues" on the compound are matched to the residues of the protein is accessible for each returned match.

Example output screens showing the return of a true peptidomimetic are shown in Figure 4. The input structure and chain (PDB ID: 1fkn) is for a modified peptide binding to the catalytic site of the human aspartic protease Beta secretase (BACE). The only CHEMPDB compound (CHEMPDB ID: AMK) which matches BACE on three motif residues, with RMSD of 2.48 Angstroms, is in fact a peptidomimetic inhibitor whose design was based on the original peptide sequence.[9] Other examples of true positives returned by the algorithm are potent HIV protease inhibitors (CHEMPDB IDs: OIS, INT, A85) for a query based on the Val-Phe HIV protease recognition sequence in complex (PDB ID: 1ida). They are returned with top-ranking RMSDs of 0.49, 0.53, and 0.6 Å. Another instance of a peptidomimetic that will be correctly identified is the cyclic RGD sequence peptide mimetic with a query for an R.D motif in fibronectin (PDB ID: 1ttg), but the ligand is not in the current version of CHEMPDB.

This relatively small list of 3 compounds is inadequate to fully evaluate objectively the performance of the method. In the future, larger benchmarking data sets for methods such as this may become available as more inhibitors of protein−protein interactions are developed. Such compounds will need to share amino acid substituents in order for this method to be clearly evaluated.

In addition, the underlying method does not identify modified amino acid side chains, amino acid bioisosteres, or nonterminal side chain positions, which can prevent many peptide analogs from being returned. For example, the effective peptidomimetic inhibitors of MDM2 which have halogen atom modifications to crucial aromatic residues[10] would not be recovered in a SAAMCO search. Peptidomimetic inhibitors whose main interaction points consist of backbone hydrogen bonding cannot be recovered, as in the

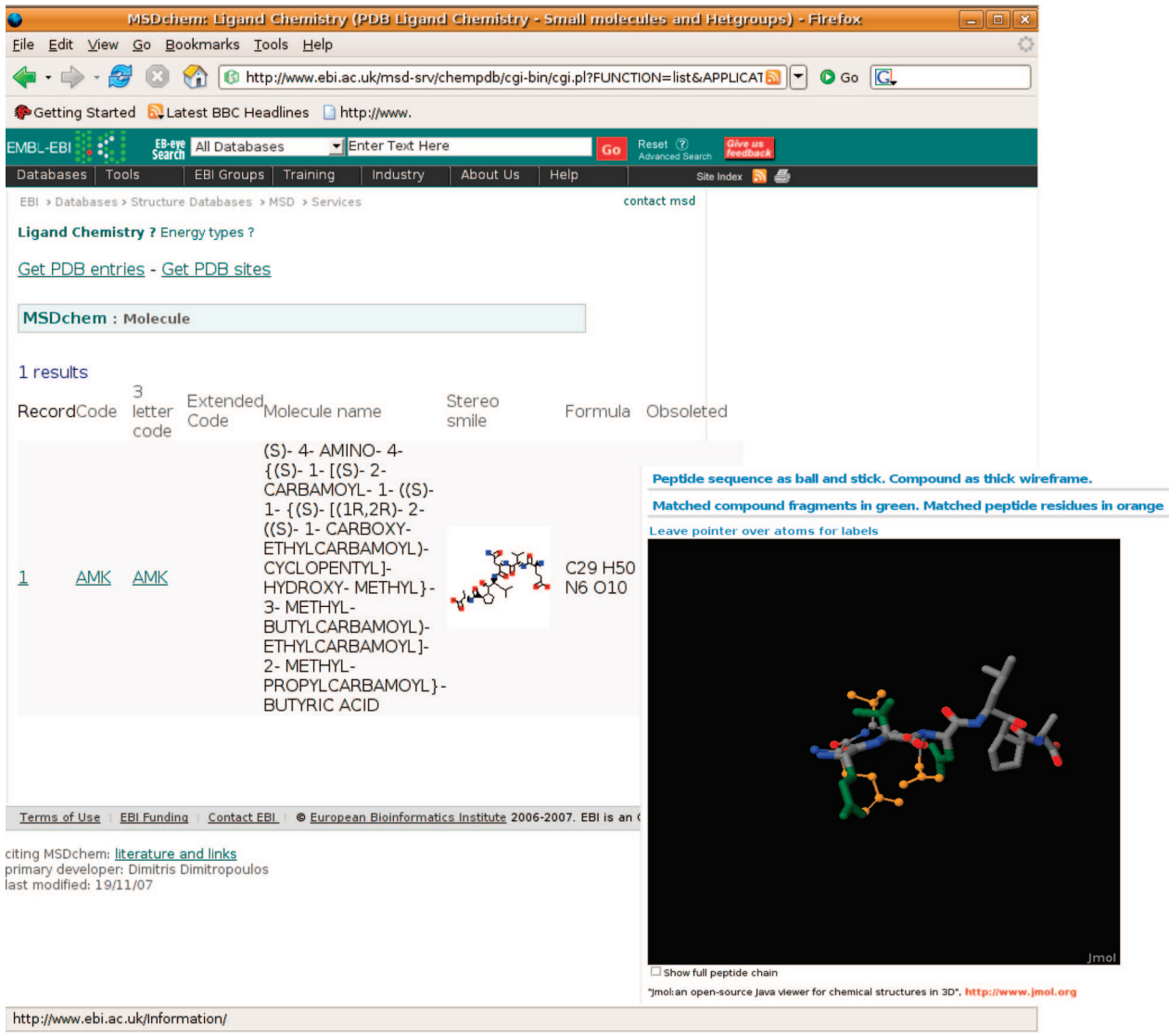


Figure 4. Output screens: CHEMPDB database entry for matched BACE inhibitor (CHEMPDB ID: AMK) and Jmol rotatable molecular viewer of superimposed matched substituents (inset).

case of the matrix metalloprotease stromelysin-1.[11] In other studies, virtual screens have recovered potent inhibitors based only on the shape and interactions within a peptide binding site, and as such the inhibitors may retain very little peptide character.[12,13] Therefore, the speed and small number of returned compounds produced by SAAMCO comes at the expense of being unable to identify more general peptidic features present in many examples of known inhibitors. Nevertheless, returned compounds could be considered potential scaffolds to undergo further modification and optimization of affinity.

CONCLUSION

We have developed a simple and intuitive Web service to search for potential peptidomimetics of user supplied short motifs from three different compound libraries. The Web service engine relies on an efficient algorithm which reduces computation time by first comparing amino acid side chain composition before attempting 3D superposition. The efficiency is further improved by precomputing compound amino acid-like fragments and storing lowest energy conformer coordinates.

As pointed out previously,[1] this method is not intended to supplant other approaches to identifying similarity, since it does not accommodate backbone similarity,[14] nor does it allow for substituents that differ chemically from the natural amino acid substituents. What it does is to search in a very specific search space, thus complementing more general approaches such as pharmacophore based searching. Future versions may incorporate known amino acid side chain bioisosteres into the fragment library which would extend the applicability without significantly compromising performance.

This implementation restricts user input to only motif definitions and PDB structure IDs. This is designed for simplicity and because compound libraries can be preprocessed for efficiency as described. In the future, extra functionality may be incorporated which will allow the user to upload compound libraries for screening. Alternatively, downloadable source code will be made available (along with various dependencies) which allows the user more flexibility in controlling source databases and parameters for fragment mining and searching.

ACKNOWLEDGMENT

The authors wish to thank Laavanya Parthasarathi for database preparation and useful discussion and Richard Edwards and Tibi Simu for testing and suggestions. Funding was provided by the Irish Research Council Embark Initiative and the Science Foundation Ireland.

Availability and Requirements. Web server home page, http://bioware.ucd.ie/~saamco; operating system(s), platform independent; programming language, N/A; other requirements, Java enabled Web browser.

REFERENCES AND NOTES

(1) Baran, I.; Varekova, R. S.; Parthasarathi, L.; Suchomel, S.; Casey, F.; Shields, D. C. Identification of Potential Small Molecule Peptidomimetics Similar to Motifs in Proteins. *J. Chem. Inf. Model.* **2007**, *47*, 464–474.

(2) Orange Book Information Data Files. http://www.fda.gov/cder/orange/obreadme.htm (accessed Aug, 2006).

(3) Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **2006**, *6*, 813–823; http://dtp.nci.nih.gov/branches/btb/ivclsp.html (accessed Feb, 2007).

(4) Dimitropoulos, D.; Ionides, J.; Henrick, K. UNIT 14.3: Using MSDchem to Search the PDB Ligand Dictionary. In *Current Protocols in Bioinformatics*; Baxevanis, A., Page, R., Petsko, G., Stein, L., Stormo, G., Eds.; John Wiley and Sons: Hoboken, NJ, 2006; pp 14.3.1−14.3.3.

(5) Molecular Operating Environment, version 2006.08; Chemical Computing Group Inc.: Montreal, Canada, 2006 http://www.chemcomp.com/ (accessed Aug, 2008).

(6) Schuffenhauer, A.; Brown, N.; Ertl, P.; Jenkins, J.; Selzer, P.; Hamon, J. Clustering and Rule-Based Classifications of Chemical Structures Evaluated in the Biological Activity Space. *J. Chem. Inf. Model.* **2007**, *47*, 325–336.

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

(8) Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org (accessed July, 2007).

(9) Hanessian, S.; Yun, H.; Hou, Y.; Yang, G.; Bayrakdarian, M.; Therrien, E.; Moitessier, N.; Roggo, S.; Veenstra, S.; Tintelnot-Blomley, M.; Rondeau, J.-M.; Ostermeier, C.; Strauss, A.; Ramage, P.; Paganetti, P.; Neumann, U.; Betschart, C. Structure-based design, synthesis, and memapsin 2 (BACE) inhibitory activity of carbocyclic and heterocyclic peptidomimetics. *J. Med. Chem.* **2005**, *48*, 5175–5190.

(10) Sakurai, K.; Schubert, C.; Kahne, D. Crystallographic Analysis of an 8-mer p53 Peptide Analogue Complexed with MDM2. *J. Am. Chem. Soc.* **2006**, *128*, 11000–11001.

(11) Becker, J.; Marcy, A.; Rokosz, L.; Axel, M.; Burbaum, J.; Fitzgerald, P.; Cameron, P.; Esser, C.; Hagmann, W.; Hermes, J.; Springer, J. Stromelysin-1: Three-dimensional structure of the inhibited catalytic domain and of the C-truncated proenzyme. *Protein Sci.* **1995**, *4*, 1966–1976.

(12) Enyedy, I. J.; Ling, Y.; Nacro, K.; Tomita, Y.; Wu, X.; Cao, Y.; Guo, R.; Li, B.; Zhu, X.; Huang, Y.; Long, Y.-Q.; Roller, P. P.; Yang, D.; Wang, S. Discovery of Small-Molecule Inhibitors of Bcl-2 through Structure-Based Computer Screening. *J. Med. Chem.* **2001**, *44*, 4313–4324.

(13) Panchal, R.; Hermone, A. R.; Nguyen, T. L.; Wong, T. Y.; Schwarzenbacher, R.; Schmidt, J.; Lane, D.; McGrath, C.; Turk, B. E.; Burnett, J.; Aman, M. J.; Little, S.; Sausville, E. A.; Zaharevitz, D. W.; Cantley, L. C.; Liddington, R. C.; Gussio, R.; Bavari, S. Identification of small molecule inhibitors of anthrax lethal factor. *Nat. Struct. Mol. Biol.* **2004**, *11*, 67–72.

(14) Goede, A.; Michalsky, E.; Schmidt, U.; Preissner, R. SuperMimic-Fitting peptide mimetics into protein structures. *BMC Bioinformatics [Online]* **2006**, *7*, 11; http://www.biomedcentral.com/1471−2105/7/11 (accessed Mar, 2008).

CI8000474