High-throughput shape similarity screening: Screen3D
Executive Summary

ChemAxon has been providing leading chemical software development platforms and desktop applications for the pharmaceutical, agrochemical, biotechnology industries since 1998. The wide-range of applications provided by ChemAxon includes platform independent & web ready toolkits for chemical database management, drug design and discovery and structural analysis. These functionalities are available through various technologies such as, Web Services, Applets, Partner platforms, Oracle Cartridge, Microsoft products (.Net, SharePoint, etc.) and integration within Workflow management systems including Pipeline Pilot and KNIME. The combination of industry leading performance and user driven and rapid support enables users to perform outstanding work in life science research.

This paper describes a novel approach to ligand-based virtual screening, which includes various current standard approaches and novelties such as flexible 3D alignment of ligands, and a fast similarity search strategy for three dimensional molecular structures. This method identifies the top ranking of molecules based on their 3D structures. Furthermore, we introduce here two methods ('Shape' and 'Match' algorithms), that are available in Screen3D, to compare molecules according to their three dimensional structure.

Screen software suite includes tools for similarity searching with a variety of different fingerprints (ChemAxon’s chemical fingerprint, Pharmacophore fingerprint, ECFP, FCFP) as well as several different dissimilarity metrics and metrics optimization methods. Screen3D opens a way to find bio-equivalent molecules in the presence or absence of crystallographic information, while JKlustor integrates numerous clustering methods that are applicable even for large datasets. A fast lowest energy 3D conformation, conformer generator is also available.

From the case studies, we will demonstrate that Screen3D provides a consistently accurate similarity screening method toward identifying true active compounds. The Match algorithm uses the atomic distance histogram which can be generated from structures regardless of original conformations. This eliminates the preparation of 3D conformation sampling prior to the 3D alignment. From the evaluation result, it will also be shown that Screen3D is also significantly faster than other virtual screening methods. All these features make Screen3D an excellent tool for reliable and rapid virtual screening.

FIGURE 1. Various methods developed at ChemAxon for similarity calculations.
Concept of Screen3D alignment method

Screen3D provides a robust solution for high-throughput 3D similarity screening (Figure 2). The algorithm offers two different methods to calculate similarity measures: The Shape method scores full molecular shape similarity derived from the van der Waals volume overlap of the molecular pairs; The Match algorithm is based on a fully flexible atom-atom or pharmacophore-pharmacophore matching algorithm.

Shape method

Shape algorithm maximizes the intersection of the van der Waals atomic volumes. It pre-processes first the query and the target molecules; 3D coordinates are generated and an atom map is created by using ChemAxon’s extended atom types and predefined pharmacophore features (Figure 3).
Identical atom maps of query and target molecules are superimposed in iterations of successive flexible or rigid alignment and optimization steps. The volume overlap function is maximized – i.e. \((t \cdot q)\) includes atom type weights and summation of volume types of the target and of the query molecules:

\[
\Phi_{tq} = \sum_i \sum_j w(t(i), q(j)) \Phi_{t(i)} \Phi_{q(j)}
\]

The indices \(i, j\) refer to the atom types of the target and query molecules respectively. Weight factors \((w)\) are defined for all type pairs. Default weight factors are 1 for the same and 0 for the dissimilar atom types. \(t(i)\) is the volume assigned to the \(i^{th}\) atom type of the target and \(q(j)\) is the volume assigned to the \(j^{th}\) atom type of the query. Their product is their overlap (Figure 4).

Once volume overlap is maximized 3D Tanimoto similarity is calculated and used for ranking target molecules.

\[
T_c = \frac{\Phi_{qt}}{\Phi_q \Phi_t + \Phi_q \Phi_t - \Phi_{qt}}
\]

Match method

Match algorithm applies a modified quaternion fit to flexibly align molecules. The pre-processing step marks atoms with ChemAxon’s extended atom types and predefined pharmacophore rules, which is followed by the calculations of minimum and maximum possible intramolecular distances between every atom-atom pair (Figure 5).
The conformational sampling is realized via an in-house built method ensuring high-throughput continual conformational scan. Intramolecular distances are collected for the formation of distance range histograms. A single histogram represents the cumulative distance range distribution of the given type of atoms from the selected atom (Figure 6).

Two histograms can be compared using histogram Tanimoto, where each histogram has two identifiers, an atom that is assigned to and an atom type:

$$T_{a,b}^{ta,tb} = \frac{\sum_i \min(h_a^{ta}(i), h_b^{tb}(i))}{\sum_i \max(h_a^{ta}(i), h_b^{tb}(i))}$$

An atomic similarity score is calculated between a single atom of the query \(a\) and that of the target \(b\) using their histograms:

$$S(a, b) = \sum_{ta} \sum_{tb} w(ta, tb) T_{a,b}^{ta, tb}$$

where \(w\) is the weight factor between atomic types (0 for dissimilar and 1 for identical types).

The pre-processing step is followed by a systematic atom-atom matching between query and target atoms. This mapping procedure involves a backtracking algorithm with the following constraints:

- Identity of atom labels are checked
- Distance ranges should remain within a pre-set threshold
- Triangle inequality uses trigonometrical restrictions on three atoms of query and target molecules to assure that their superposition is adequate
- Quaternion Flexible Hybrid Alignment step implicates a modified classical quaternion fit (JMGM, 2007, 25, 595)

For the best alignments, molecular similarity score between the query and the target molecules are calculated as:

$$S = \sum_q S(q, p(q))$$

where \(p(q)\) is the target atom paired with the query atom \(q\).
Screen3D is best used in the following areas

Medicinal chemistry / Application in Virtual screening

• Pair-wise 3D alignment allows superposition in the following ways:
  - Rigid-rigid alignment, having a co-crystal molecule as a query we can superimpose particular conformations onto it. (Target-based virtual screening)
  - Rigid-flexible alignment, having a co-crystal molecule as a query, target molecules can be aligned in a fully flexible mode. (Target-based virtual screening)
  - Flexible-flexible alignment, in the absence of co-crystal ligands, query and target molecules treated in a fully flexible way. 3D alignments provide information for SAR studies. (Ligand-based virtual screening, 3D similarity search)

• Alignments helping pharmacophore perception
  • The alignment of about 5-6 molecules offers an intuitive way for pharmacophore perception and helps SAR analysis and analogue design
  • The alignment of various chemotypes allows the identification of bioisosteric transformations, cores, substructures
  • Free and easy to use interface in Marvin sketch or MarvinSpace
  • Alignments can offer replacements of scaffolds via scaffold hopping

Computational chemistry

• Numerous options are adjustable assuring the effectiveness of the alignment algorithm
  • Available as a command line tool and API
  • Extensive sampling of the conformational space
  • Provides alignments for 3D-QSAR studies
  • Even in the absence of co-crystal structures high-throughput ligand-based virtual screening offers alternatives to rigid-rigid, rigid-flexible and flexible-flexible docking.
Benchmark studies

In-house study

Screen3D was evaluated using Directory of Useful Decoys. This benchmark study involved decoys and active molecules against 40 targets. Enrichment factors were calculated to compare Screen3D to several competitors as seen in the Figure 7. Screen3D was demonstrated to be a high-throughput method that provided similar performance in terms of EFs(1%) to competitors. Screen3D is best used in the following areas.

![Figure 7](image1.png)

**FIGURE 7.** Enrichment factors (EF(1%)) calculated at 1% of the ranked database. Screen3D methods are shown in orange while competitors are coloured green.

Study by Oleg Ursu, University of New Mexico

Virtual screening studies were performed on 12 target molecules. 1% of the available active molecules were randomly selected from ChEMBL, and were used as queries, while the rest of the active molecules were added to a random decoy set of 30,000 molecules. Enrichment factors calculated at 5% of the ranked databases were compared to that of given by the application of ECFP-4. The performance of calculations underlined the effectiveness of virtual screening by shape similarity compared to 2D similarity screening.

![Figure 8](image2.png)

**FIGURE 8.** Enrichment factors (EF(5%)) calculated at 5% of the ranked database. EFs(5%) given by Screen3D is compared to that of ECFP_4 (Extended-Connectivity Fingerprint with a diameter of 4).