A BioSolveIT White Paper

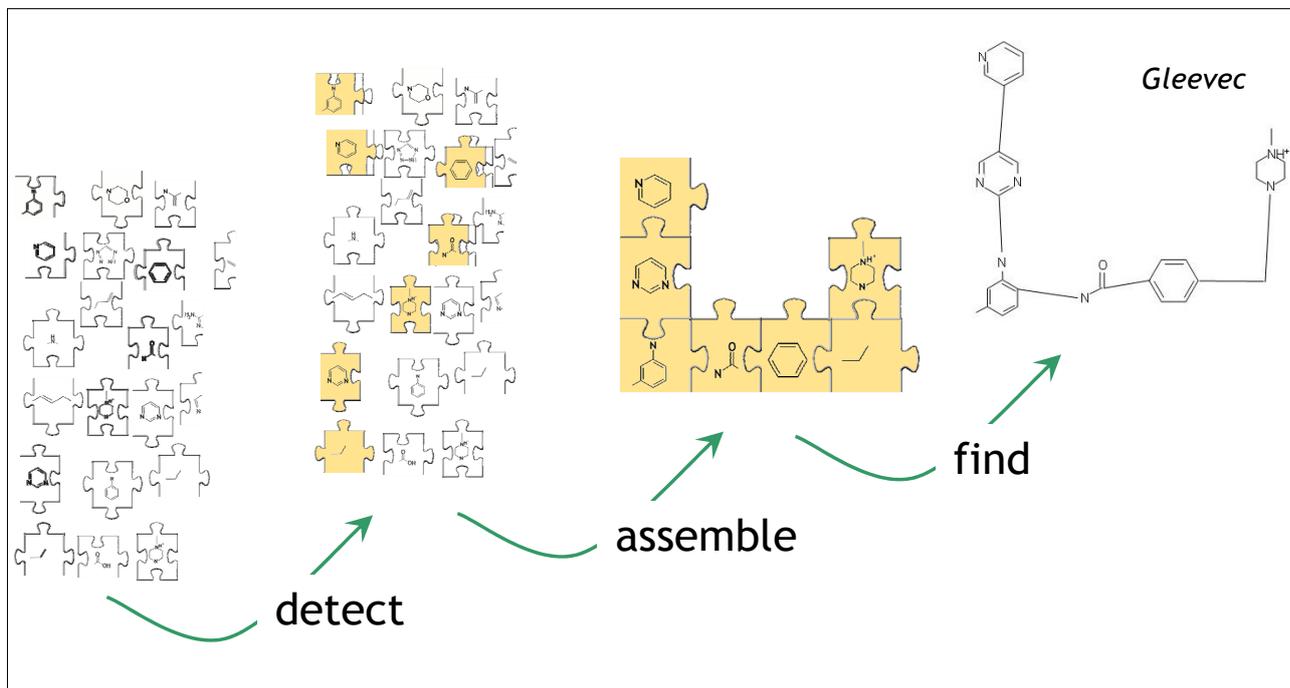# Exploiting MedChem in Virtual Screening

Capitalize on Your Chemists' Know-How

Fragment merging, fragment growing, SAR-by-NMR, the game has many names, which is just one indication of the popularity of fragment-based design methods.

Contents

Authors

Christian Lemmen, Marcus Gastreich, Holger Claußen
BioSolveIT GmbH
An der Ziegelei 79
53757 St. Augustin, Germany
http://www.biosolveit.de

For further information:

contact@biosolveit.de
+49-2241-2525-0

©2009 BioSolveIT

Executive Summary

Virtual products based on your
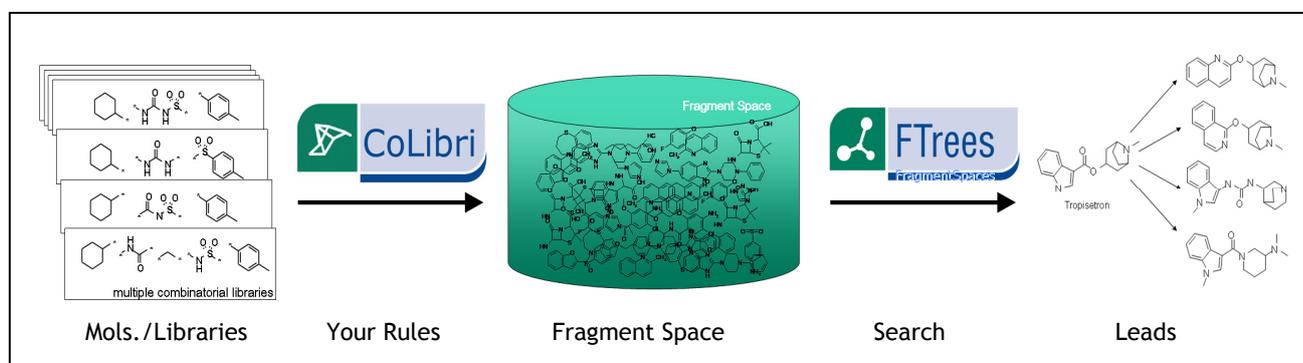in-house chemistry know-how

Imagine you could capitalize on the vast know-how of chemistries developed in your company. There must be hundreds — if not thousands — of protocols for parallel syntheses. This know-how — probably among your company's most valuable assets — easily covers billions of products, of which of course only a small fraction has ever been made. However, the recipes are known, all those compounds are most likely accessible with limited effort, and your chemists will simply know how to make them.

Synthetic access covered

Now further assume there were a method to mine new leads from this space within minutes. "Impossible!" you say? Not at all! In this white paper we show you how easily this can be achieved and provide ample evidence that the procedure does work in practice. It has already led to new active scaffolds in ongoing therapeutic projects in Big Pharma.

Billions of compounds
searched in minutes

Proven to work in Big Pharma

The Big Workflow Picture

1. All your in-house chemistry know-how is captured and stored in a Fragment Space
2. We employ a similarity search method proven to be outstanding at scaffold hopping
3. Applying this method we search the Fragment Space using your unique virtual product assembly
4. Hit lists are reported back to the user — providing synthesis protocols and reagents to each product
5. Visual inspection quickly leads to identifying novel leads from the innumerably large space

5 steps to capitalize on your
company's most valuable asset

| Mols./Libraries | Your Rules | Fragment Space | Search | Leads |

Capturing Chemistries

It all starts with capturing chemistries in a computer-readable fashion. In one way or another it boils down to a description of reaction protocols which range from simple two component reactions to multi-step reactions involving four or more reagents and by-products. In the end the basic principle is the same: a scaffold — potentially with some variations — is formed and a certain number of side-chains off it give rise to a combinatorial explosion in the number of different products.

Assume you have three sets of reagents — for simplicity say 100 each. This would result in a library containing 100 x 100 x 100 = 1 million products. Our software tool CoLibri is capable of taking reagent lists in either SMILES, SD or mol2 format and then applying any sort of modification to these molecules, from simple clipping of protecting groups to more complex ring formations.

The Genesis of a Synthetically Feasible Compound Space

Based on a real-life example from an article in Science (1); we will describe how to represent compound libraries virtually using BioSolveIT technology:

The figure below shows the combinatorial synthesis protocol which Gray et al. (1) used for the discovery of novel CDK2 inhibitors. The authors synthesized 2,6,9-tri-substituted purines from a much simpler 2-, 6-, or 9-substituted purine scaffold by solid-phase amination or alkylation reactions and a subsequent acidic cleavage:
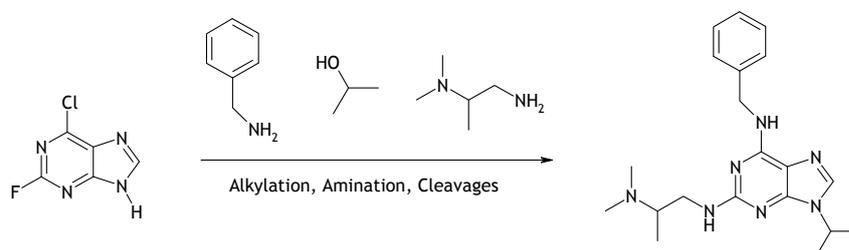
Representing this protocol in the computer is a simple procedure. What is essential are educts, products, and a formalized description of the newly formed bonds.

We will term the purine scaffold a "core", whereas the reagents for substitution will supply the "R-groups" (i.e. the residues in the resulting products — not to be confused with the R-notation in chemical formulas). Where exactly a bond is created is defined on the basis of dummy or "linker atoms". The core and educts therefore need to be equipped with these.

In Step 1, CoLibri is used to replace the amine-H at position 9 and halogens at positions 2 and 6 at the core for linker atoms. The linker atoms are denoted R1 to R3 below. The corresponding SMIRKS-like CoLibri rules for the core are a simple three-liner in which a dot denotes the cleavage of a bond and the [n*] notation introduces the linker atoms R*n*.

```
[Cl][*]      >>   [Cl].[*][1*]
[H:1][N:2]   >>   [H:1].[N:2][2*]
[F][*]       >>   [F].[*][3*]
```

Applying these transformations to the purine core looks like this:
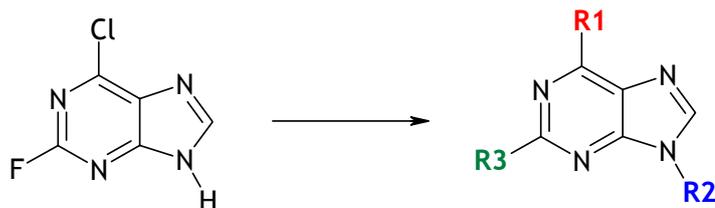
With 3 more rules, CoLibri prepares the educts for the substituents: it clips the amine and alcohol H-atoms to form the "naked" R-groups as follows:

Three more lines of code define the R-group clipping

```
A [H:1][N:2]([H:3])[*:4]  >>   [H:1].[1*][N:2]([H:3])[*:4]
B O[C][*]                 >>   O.[C]([*])[3*]
C [H:1][N:2]([H:3])[*:4]  >>   [H:1].[1*][N:2]([H:3])[*:4]
```

Finally, we need to define how clipped fragments may be recombined ("linker compatibility"). This is a simple ASCII file.
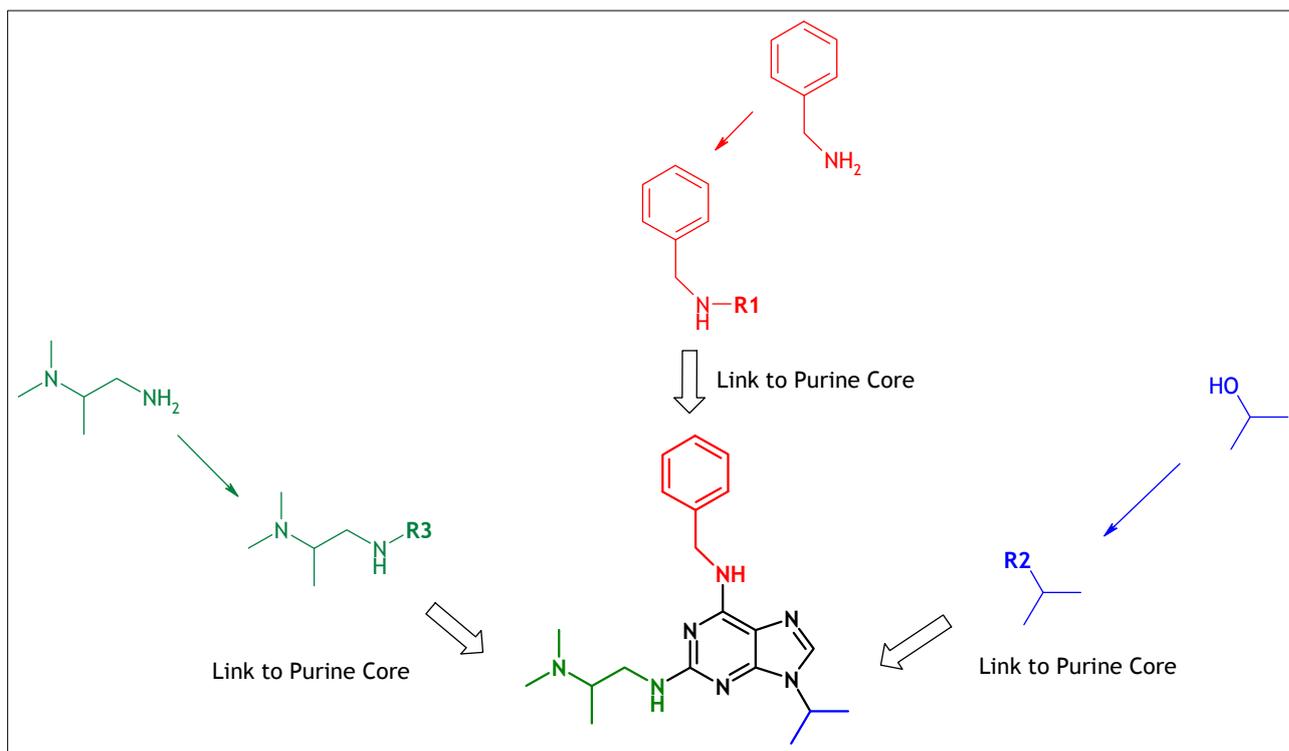
Reagent list expansion quickly leads from 1,330 to 70 million products

The original Science publication reports 19 reagents for R1, 7 for R2, and 10 educts for R3 resulting in 19 x 7 x 10 = 1,330 products. We combined the reagents lists for R1 and R3 and added 309 additional primary amines. Furthermore we extended R2 by 274 alcohols from the same source. All this was done with fine medicinal chemistry expertise. Our resulting search space consists of 70 million compounds (499 x 281 x 499 = 69,969,281) for this single reaction protocol.

Now, how about more reactions? ➜ the KnowledgeSpace[TM]

So far we processed 82 different pub-
lished synthetic protocols from the
Journal of Combinatorial Chemistry.
These cover a variety of targets
(GPCR, Protease, Kinase) but also
purely Chemistry-Driven Libraries.

Distribution of Targets in KnowledgeSpace[TM]

- Chem. Driven
- GPCR
- Protease
- Kinase
- Other

**KnowledgeSpace[TM] -
82 libraries with 10^{10} virtual
compounds — free of charge
www.biosolveit.de/download/
?product=knowledgespace**

The KnowledgeSpace[TM] consists of 10,879 unique fragments that were
approved by a medicinal chemist by eye and deemed to be of signifi-
cance and drug like nature. The properties of the fragments are mainly
respect the 'Fragment Rule of 3' proposed originally by Astex, with
filters applied for the properties of logP, donors and acceptors and
molecular weight. Thus 90% of fragments have a molecular weight
between 100-250 Daltons. Half of the molecules in the space contain
at least 1 ring structure with approximately the remaining two quar-
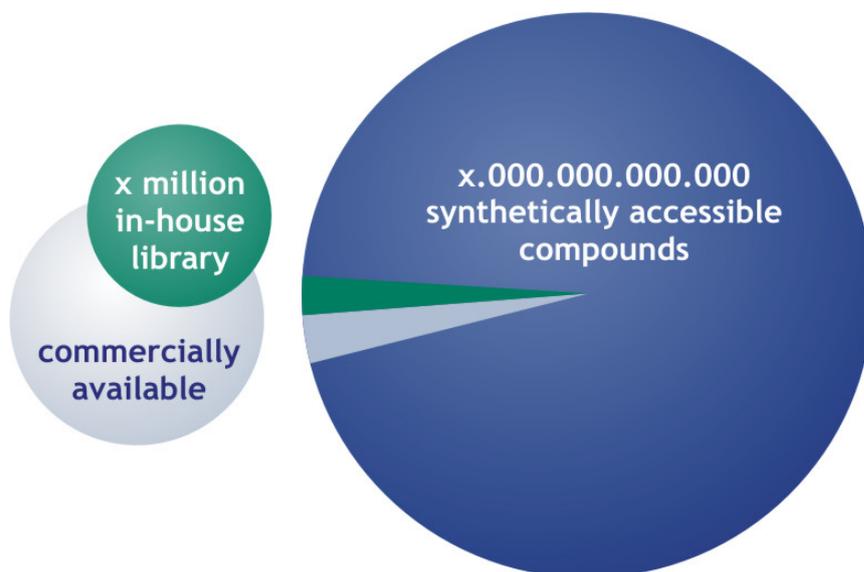ters equally split between having no ring structure or 2 rings.

~11k unique MedChem-
approved fragments

These ~11k fragments in the KnowledgeSpace[TM] can be combined fol-
lowing the synthetic feasibility rules to produce 11,725,417,388 virtual
products that can actually be made! The KnowledgeSpace[TM] has been
validated to retrieve known drugs and drug like molecules (2). The
huge advantage is not only that this Fragment Space comes free of
charge with our software, but also it is open access. So you can use
this space to add your own libraries. You can modify the space accord-
ing to your needs. Nothing is written in stone, you are in charge.

An extensible chemistry space
for enhanced IP-value

Now you are ready to process all your in-house reactions.

Summary & Obvious Advantages

x million in-house library

commercially available

x.000.000.000.000 synthetically accessible compounds

No matter how big your in-house
library and no matter how many com-
pounds you acquire to add to it, it will
only be a tiny fraction of what your
chemists are capable of synthesizing

Well the biggest advantage is entirely obvious. The compound spaces
that become accessible (both virtually and synthetically) are huge. If
you compare to any vendor catalog or any in-house repository, no mat-
ter how big, these are tiny compared to what you now have available
at your fingertips.

Scriptable, visually supported

One space, no redundancies

Everything is based on the easy-to-read chemistry description standards SMARTS/SMIRKS which allow all kinds of substructure detection and replacement. Using the CoLibri procedure, we transform the raw input data into a Fragment Space. Also we are already working on the support of reaction-protocol input with a graphical user interface.

Output: the original protocol and reagents

The entire CoLibri procedure is scriptable and supported by 2D visualization of substructure matches. The real power of the mechanism obviously comes from its ability to process hundreds of protocols as above and make them accessible as a single enclosed Fragment Space. Here it is important to note that CoLibri is able to remove the redundancy from a dataset by representing duplicate fragments in the input using only one representative instance and maintaining a lightning fast, hashkey-based lookup table to map any results data back onto the original input. This way CoLibri reports not only virtual products to the user but is also able to annotate these results with the chemistry library protocol and the particular reagents that form a product.

RECAP-shredding as another means to arrive at giant Fragment Spaces

On a Side Note, Fragments by Shredding

Another very well known way to obtain Fragment Spaces, the so-called RECAP-approach, has been described way back in 1998 by Lewell et al. (JCICS v38:511ff) and later improved by Schneider and co-workers (JCAMD v14:487ff). The basic idea here is to shred molecules in pieces following a set of rules from retosynthetic analysis. In essence you split those bonds that with some likelihood can be formed applying certain chemistries.

As a thought-experiment, if you split 2 molecules A and B at the same type of bond (say A equals a1-a2, B equally b1-b2 and the bond in between is of the same chemical type) then you can recombine the 4 fragments a1, a2, b1, and b2 already in 4 different ways, namely a1-a2, a1-b2, b1-a2, and b1-b2. Imagine the combinatorial explosion if you split molecules repeatedly at different bonds and apply this procedure to many molecules.

	Input origin of		likelihood of synthetic access	IP-value
	fragments	connection rules		
shredding	public compounds	published rules	good	low
	corporate compounds	published rules	better	medium
	corporate compounds	corporate rules	better	high
reactions	public reagents	published reactions	best	low
	public reagents	corporate reactions	best	medium
	corporate reagents	corporate reactions	best	high

You may now be confused. How is this different from the approach described above? How can I best preserve my IP? Are there some guidelines as to what I should use and when? All of this is addressed in the following table.

Free download

In essence best likelihood of synthetic access and highest IP-content can be expected if you encode you corporate MedChem knowledge. However, various other – less cumbersome – options are available.

CoLibri supports shredding as well and comes with an improved set of rules along the lines of the old RECAP-idea. It has been tested on large vendor-libraries as well as small focused sets. Also, from our academic partner the Center for Bioinformatics (ZBH) of the University of Hamburg you can download free of charge a shredding-based Fragment Space called BRICS (3) that works just as well with the searching technology which we will describe next.
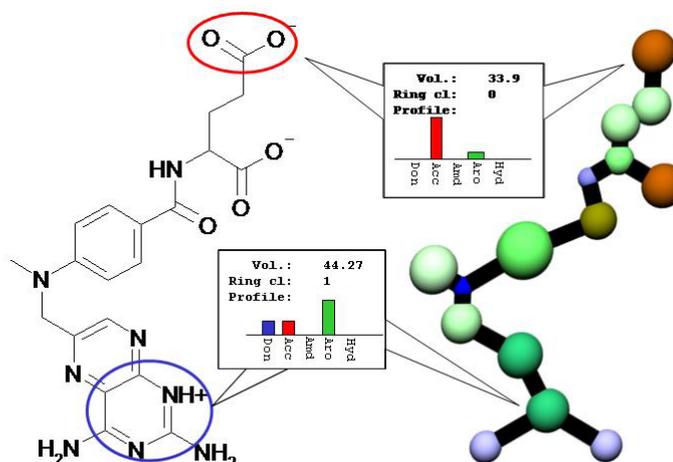
Scaffold Hopping: FTrees

Now that the in-house Fragment Space is generated, the Feature Tree software ("FTrees") can be used to perform similarity searches in this space (4).

A Feature Tree represents the molecule as a so-called reduced graph — essentially a tree:

FTrees advantages:
o   excellent enrichment
o   orthogonal to others
o   proven to hop scaffolds

Similarity by alignment — your chemists will understand HOW they are similar!

It reduces functional groups as well as rings to single nodes. The physico-chemical properties of the substructure represented by a node are stored in a chemistry property profile for that node. The overall topology is preserved in the Feature Tree in that nodes representing fragments that are connected in the molecule are also connected in the Feature Tree. Now, if two molecules are both represented by their respective Feature Tree, FTrees is able to calculate from among all the topology preserving mappings of the two Feature Trees the one that gives the highest possible similarity value. How is the similarity value calculated? Put simply, if two nodes are mapped then the Tanimoto distance of their respective property profiles gives a local similarity, and the overall similarity is just the normalized sum of these local similarities.

An FTrees mapping of methotrexate (above) onto dihydrofolate (FTrees similarity overall = 0.96). Substructures of the same color are mapped onto each other. Only the highlighted mappings are less than the maximum local similarity 1.0

8
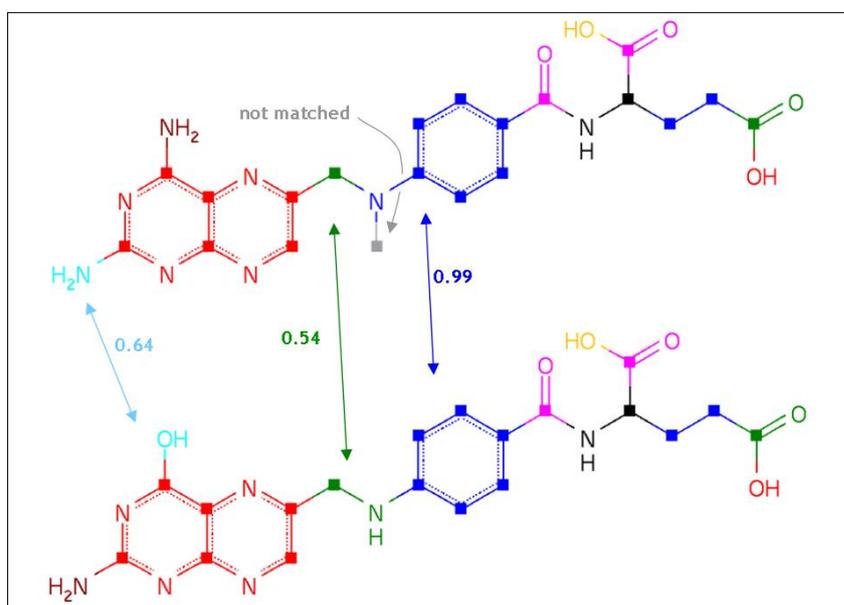
FTrees screening is fast, and time is money

FTrees calculates this optimum similarity value within less than a millisecond, so 1,000 molecules can be compared within a second, a vendor catalog with some 100,000 compounds during a coffee break and any size of in-house collection with millions of compounds overnight — on one single CPU. Needless to say, multiple processors speed up the process almost linearly.

Fuzzy description enables the detection of remote similarities

Yet, most importantly, the virtue of the method is not merely its speed: due to the fuzzy description of the molecule, ignoring 3D, and focusing only on the presence or absence of certain functionalities in roughly the right place, FTrees is able to detect remote similarities and is prized for its ability to jump chemical classes (7). FTrees has also been shown to be among the top performers when it comes to enrichment rates and is found to be orthogonal to other descriptors, providing a different view on the problem and thereby retrieving different types of active molecules. That is why FTrees is often found to be a very valuable complement to other similarity screening software.

FTrees complements other search methods

Searching the Impossible: FTrees Fragment Spaces

FTrees in Fragment Spaces: unrivaled potential

It gets even better: FTrees not only has all these fantastic attributes that make it the perfect choice for similarity searching, but it can also be used to search within Fragment Spaces. In the same way as for a whole molecule, we can represent the fragments in a Fragment Space as Feature Tree fragments.

If we provide a query molecule as input, an extension module called FTrees-FS is capable of searching the Fragment Space; the result will be a list of the most similar product molecules generated on the fly. This is done by recursively detecting a sizeable set of most similar fragments and assembling multiple fragments to virtually grow a set of molecules from the Fragment Space (5).

Two fragments are represented by two Feature Trees. Forming a "bond" between the link atoms leads to a bigger Feature Tree, which translates back to a new molecule. The history behind its formation is preserved.
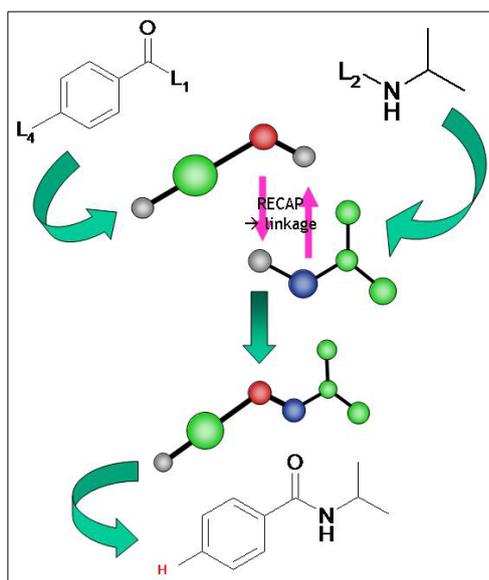
The method thus detects similar molecules in a deterministic fashion (i.e. no random element is involved). The resulting molecules are guaranteed to be the ones most similar to the query molecule.

Summary & Advantages

Innumerably large spaces can be searched

FTrees-FS is suitable for searching innumerably large combinatorial chemistry spaces for retrieving virtual products that are similar to a query molecule.

9

Recent Application Successes at Big Pharma

In the paper *'Similarity Searching and Scaffold Hopping in Syntheti- cally Accessible Combinatorial Chemistry Spaces'* by Markus Böhm et al. (6), the authors describe the generation, validation and application of the procedure outlined above to a large proportion of Pfizer's com- binatorial chemistry protocols.

A total of 358 combinatorial libraries were converted into a single con- cise Feature Tree Fragment Space comprising a total of 3,000,000,000,000 (3 trillion) virtual products. This Fragment Space was then validated in a variety of ways. In summary, with a sample set of 1,790 queries (5 randomly chosen for each protocol), it was possible to retrieve three or more queries on the top 100 ranks for 99% of the protocols. Considering the vast number of products in the space, this is literally akin to finding a needle in a gigantic haystack.
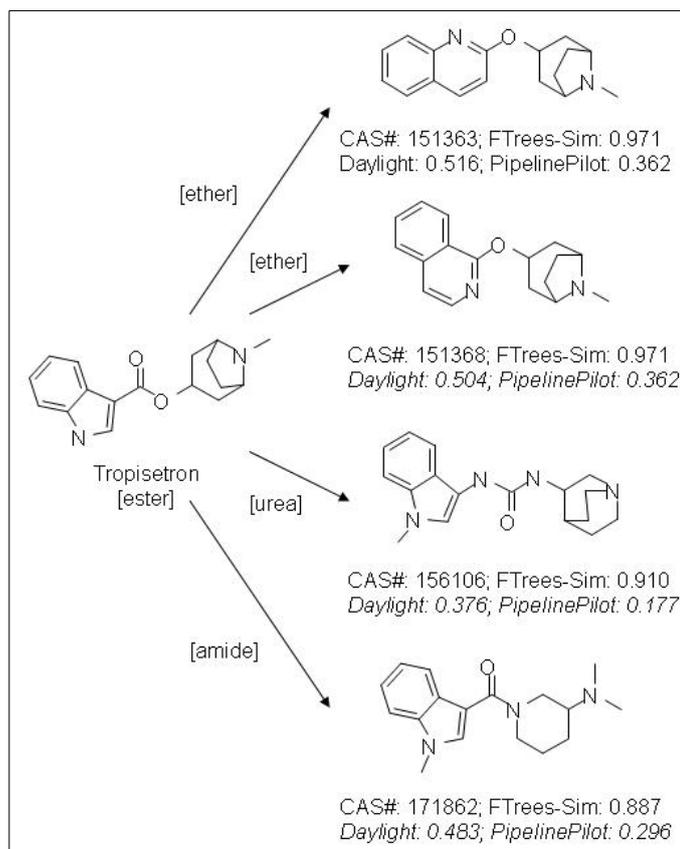
When applied to searching a sample set of 1,661 compounds from the WDI, 91% retrieved a compound with similarity of 0.9 or higher, dem- onstrating that the Fragment Space covers a broad range of drug-like molecules. 90% of these searches had a search time of less than 20 minutes on a single CPU.

Also the results covered a broad range of different chemistries in that 50% of all protocols were employed at least once to form the top- ranking product for a search. Below is an example case from an ester query:

Most interesting is of course the ability to scaffold-hop from one active hit to another attractive series. The most interesting hits were found in the range of similarities between 0.9 and 0.95. In other searches at Pfizer, a central pyrrolo-indole scaffold was replaced by an indanyl

Pfizer's success: mined from trillions of compounds within a few minutes

Drug-like molecules covered

A variety of chemistries covered

FTrees finds reasonably high similarities where others fail [graphic based on (6)]

FTrees hops scaffolds

piperazine ring, a central ketone group was substituted by an amide bond linker, or a phenothiazine heterocycle was replaced by a phenyl-indole scaffold — to mention just a few scaffold hops.

For the serotonin 5-HT3 receptor two marketed drugs were used as queries which produced active hits originating from a variety of chemistries such as ether and amide bond formation or a urea reaction. Unfortunately, some of the more exciting hits could not be disclosed by the authors. Interestingly for a sizeable number of cases the hits produced by FTrees-FS had quite low Daylight or Pipeline Pilot (FCFP4) fingerprint similarities, which further underlines the uniqueness of these results. Not only are these methods a no-go for virtual libraries of this size (10^{12}) due to sheer size, but also these other methods would only have retrieved these solutions ranked worse than a few billion others because of the low similarity scores.

In another recent study, Uta Lessel of Boehringer Ingelheim presented at the 8th International Conference on Chemical Structures (8) two successful applications of Feature Tree Fragment Space searches based on combinatorial library protocols. Based on a sizeable number of combichem protocols, the so-called BI-CLAIM Fragment Space was generated on the basis of roughly 1,600 scaffolds and about 30,000 unique reagents. Thousands of compounds were actually synthesized for each of these protocols, which however amounts to only a tiny fraction of the 500,000,000,000 (500 billion) virtual products covered by BI-CLAIM. The typical workflow described in this presentation has two parts. First a literature active is taken to search the space and produce in the order of a few thousand hits. Then a shape filter is applied in order to provide a first pass validation of the hits, and finally the results are grouped by scaffold and visually inspected. Part two of the workflow is to manually select the most promising scaffolds and design focused libraries around them or purchase prototypes of those compounds if commercially available. If activity is found and confirmed in these series, then one or more rounds of refinement based on traditional medicinal chemistry are applied. The researchers from Boehringer Ingelheim reported on a GPCR project and a proteinase project where these procedures quickly led to nanomolar inhibitors in novel series.

The third success story is from a Boston-based biotech company called Arqule, demonstrating that the proposed way of capitalizing on your chemists know-how is by no means affordable only to Big Pharma. Arqule also took a different approach to encode their corporate Fragment Space. Instead of using CoLibri they used the Daylight™ reaction toolkit to do the transformations as outlined above. So you see that our software can be feed with data from other sources. There are no 'hidden secrets' in the CoLibri files.

Arqule presented at the Spring ACS-meeting 2009 in Salt Lake City about 'Reagent-Based Fragment Space for Hit Generation' (9). Their space comprises roughly 1200 reaction schemes covering more then 15 years of MedChem experience. As part of Arqules Kinase Inhibitor Platform (AKIP™), Feature Tree Fragment Space searches are routinely performed. Results are provided through a web-interface with the product, the reaction protocol, and the reagents even with availability and price tag. The ease of use, rapid response time and superiority of results make this a very desirable tool, not only for modelers but also for the medicinal chemists themselves.

More interesting hits were not disclosed

Other methods incapable of finding FTrees-FS hits

Boehringer Ingelheim rapidly found nanomolar GPCR inhibitors with novel scaffolds

ArQules 15+ years of experience covered in a single concise virtual space

Summary & Advantages

Virtual combinatorial library spaces can be built and effectively searched by the CoLibri and FTrees-FS software packages. The vast know-how of chemistries within a company can successfully be exploited, thus capitalizing on your IP. Unlike other de novo methods, the virtual products can actually be made as their synthetic protocol is known, and the reagents are available. Scaffold hopping is one of the strengths of the Feature Tree software — and this can be brought to full effect here as the vastness of the search space covers an innumerably large number of products based on quite sizeable numbers of scaffolds. The searches can be done within minutes and have been shown to produce active hits from novel chemical series in Big Pharma.

Exploit the vast IP resources within your company !

**Don't settle for just searching what you already have
        but in addition access what you can make !**

References

1. Gray, N.S. et al., Science **281**, 533, 1998
2. Lemmen, C. at MIPTEC'08, Basel, Switzerland, 2008
   http://www.biosolveit.de/conferences/talks/2008-10-16_MipTec_CL.pdf
3. Degen, J. et al.; ChemMedChem, **3**:1503-1507, 2008
4. Rarey M, and Dixon, S.; *JCAMD* **12**:471-490, 1998
5. Rarey M. and Stahl, M; *JCAMD,* **15**:479–520, 2001
6. Böhm, M. et al.; *J. Med. Chem* **51**:2468–2480, 2008
7. Briem, H. and Lessel, U.; PD3 **20**:231, 2000
8. Lessel, U.; *J. Chem. Inf. Model.*, **49**, 270-279, 2009
9. Rojnuckarin, A. at ACS'09, Salt Lake City, USA, 2009