Beilstein-Institut Molecular Informatics: Confronting Complexity, May 13th - 16th 2002, Bozen, Italy

KNOWLEDGE-BASED LEAD FINDING BY MATCHING CHEMICAL AND BIOLOGICAL SPACE

KARL-HEINZ BARINGHAUS*, THOMAS KLABUNDE, HANS MATTER, THORSTEN NAUMANN AND BERNARD PIRARD

Aventis Pharma Deutschland GmbH, LG Chemistry-Computational Chemistry, Industriepark Hoechst, Building G 878, D-65926 Frankfurt am Main, Germany
E-Mail: karl-heinz.baringhaus@aventis.com

Received: 3rd July 2002 / Published: 15th May 2003

ABSTRACT

This paper describes a target family-related lead finding approach, which consists of capturing public and proprietary information to build a biological and a chemical space. Computational tools to assemble these spaces as well as appropriate techniques to match them are covered. Three recent applications in the field of kinases, ion channels and GPCRs exhibited already improved lead finding capabilities compared to traditional approaches.

INTRODUCTION

Molecular Informatics is usually involved in several parts of the value chain of drug discovery, mainly of course in lead finding and lead optimization through appropriate techniques, such as for instance HTS data analysis, database mining and pharmacophore modeling (1). However, the competitive pressure in the pharmaceutical industry requires a reduction of project cycle-times and therefore an increase in productivity and efficiency.

In addition, pharmaceutical companies are faced with the challenge of translating genomic information into new, innovative medicines. More than a thousand potential new drug targets have emerged from the sequencing of the human genome, but currently available drugs only target approximately 500 different proteins (2). In order to effectively cope with genomic research, an improvement in lead finding is needed.

A good strategy would be a molecular informatics driven knowledge based approach, whereby chemical information and expertise within target families are acquired and applied. Thereby, a

match of chemical and biological information within target families is achieved (3). This article highlights computational methods towards such a target family approach and contains three recent examples in this field.

MATCHING OF CHEMICAL AND BIOLOGICAL SPACE

Our target family related approach consists of biased libraries, which were built by matching chemical and biological space. Therefore, the intersection of biological structures and functionalities with chemical structures and properties is derived to perform a knowledge driven biased library design. This allows an extraction of common structural features for target families out of a more or less infinite chemical space. Thereby, chemical libraries could be built, enriched by preferred features of biologically active compounds (Fig. 1).

Target family related structural features are identified by a combined 2D and 3D analysis.

The 2D approach is based on a collection of biologically active compounds and consists mainly of similarity and substructure searching and of the analysis of common frameworks and fragments.

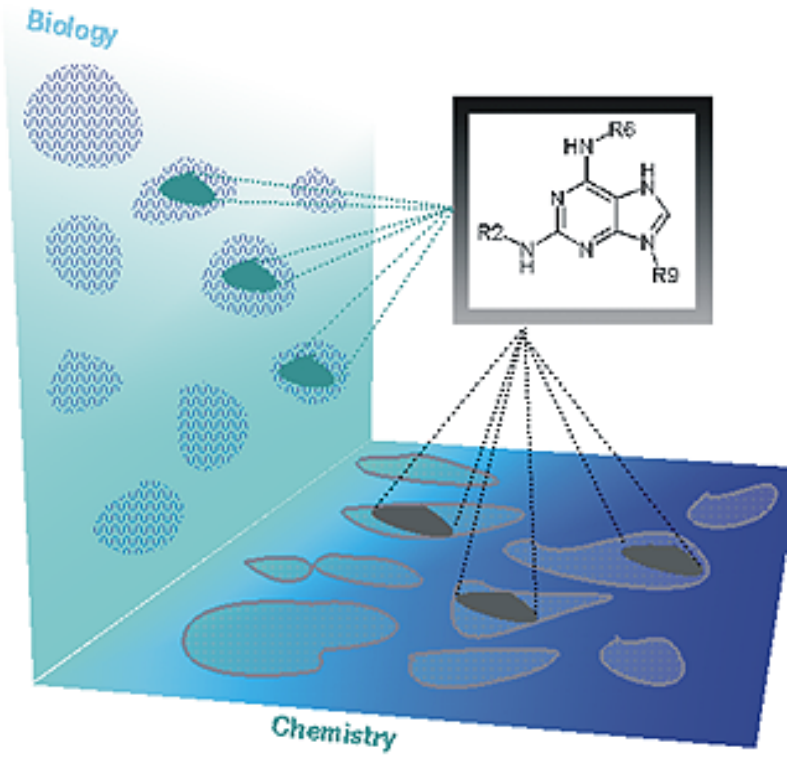

Figure 1. Matching biological structures and functionalities with chemical structures and properties.

The 3D approach relies on homology models, pharmacophores and binding sites, which are

used for virtual screening and for our target family landscape approach (Fig. 2).

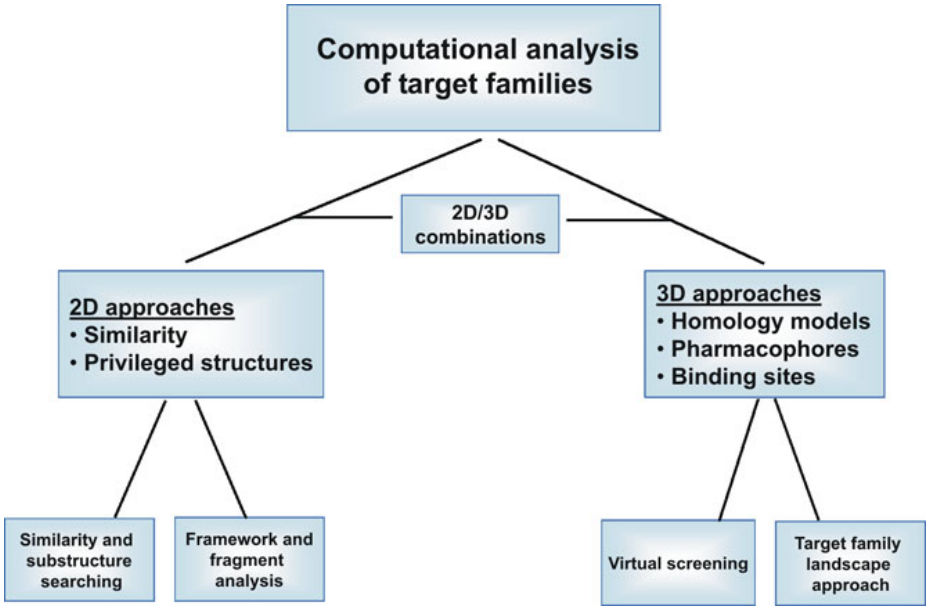

Figure 2. Computational tools to analyze target families for a knowledge driven design approach.

Three of these techniques are outlined here in more detail.

Bemis and Murcko (4) published the topological framework analysis in 1996.

They analyzed shapes of existing drugs from a commercial database to extract drug-related molecular frameworks. A graph theoretical approach was used to decompose molecules into rings and non-cyclic side chains. Linkers and rings together form the framework of a molecule (Fig. 3).

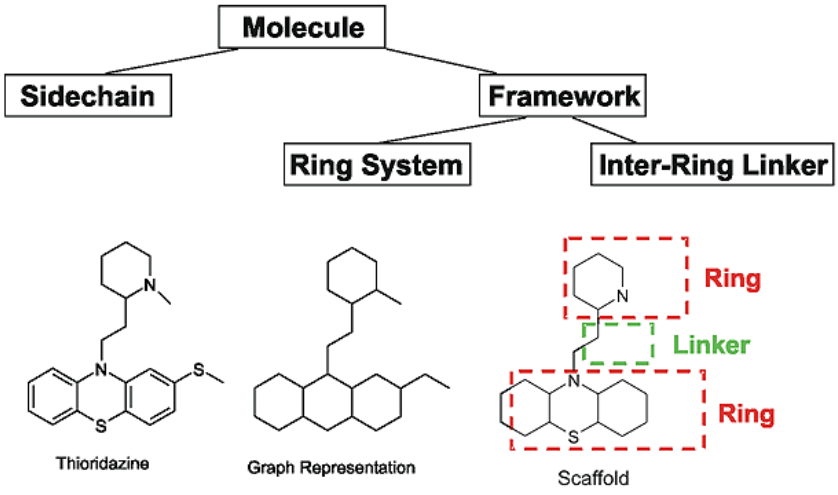

Figure 3. Topological framework analysis.

In this example, acyclic side chains of thioridazine are removed leaving a framework composed of two rings and one interring linker.

Application of such a topological framework analysis on biologically active compounds within a target family reveals access to privileged substructures for activity. By conversion of these frameworks into appropriate scaffolds for synthesis, target family related libraries can be built.

The fragment analysis (5) is based on the RECAP algorithm, published in 1998. This retrosynthetic combinatorial analysis procedure begins with a collection of active molecules and then fragments these molecules using any of the 11 retrosynthetic reactions. For example, Cisapride is cleaved into four fragments based on three different bond cleavage types (Fig. 4).

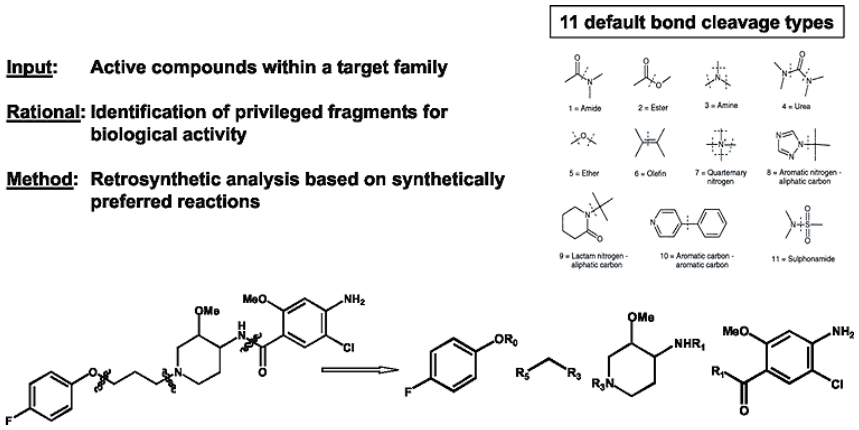

Figure 4. RECAP: Retrosynthetic Combinatorial Analysis Procedure

Resulting fragments are usually clustered and transformed into sets of monomers for subsequent library design. Because the monomers come from biologically active compounds there is a high likelihood that new designed molecules from them will contain biologically interesting motifs.

Our target family landscape approach to classify target family related proteins is a four step procedure (6) (Fig. 5), starting with the selection of representative structures from the Brookhaven and our internal database, followed by an alignment of all structures from 2D sequence and 3D structure similarity. This alignment and all subsequent analysis are focused on the 3D binding site only.

Protein-ligand interactions are then calculated using the GRID force field. They somehow reflect the biological similarity and dissimilarity of proteins from a ligand perspective. A subsequent statistical analysis of the attractive interactions by chemometrical tools, like a principal component analysis, highlights those areas that reveal most differences.

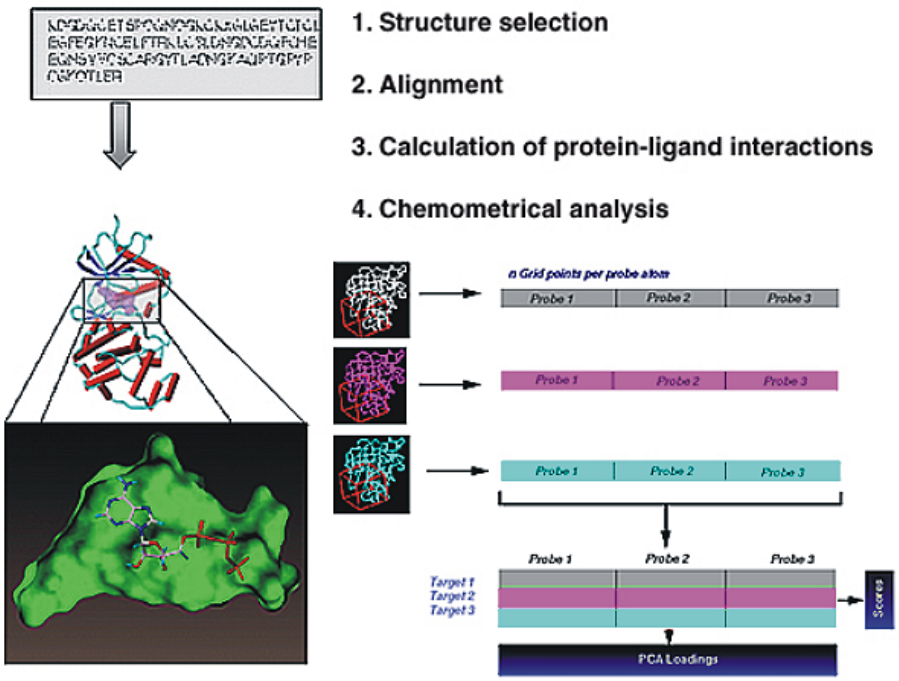

Figure 5. Target family landscape analysis.

This quantitative analysis of GRID-derived molecular interaction fields leads then to a classification of proteins and to a detailed understanding of which structural features are responsible and most characteristic for a particular protein. Hence, this is an excellent description of biological space.

The chemometrical analysis (PCA) of this landscape approach is schematically shown in Figure 5. Starting from the GRID interaction field for one single GRID probe and one protein a vector is constructed. Subsequent interaction fields for other probes are concatenated to this vector to result in a longer vector containing x GRID times n probe points. Similar vectors are derived for all proteins resulting in an X matrix with one single row per protein.

After block scaling to normalize the probe interactions the X matrix is analyzed using PCA or CPCA. Thereby, the data matrix X is approximated by the product of two smaller matrices, scores and loadings. Scores reflect the similarity of proteins, whereas loadings highlight differences in 3D space in terms of molecular interactions (7).

We performed this target family landscape approach to classify kinases based on their ATP binding site (8). We used 26 kinase structures and aligned them to 1ATP, a cyclo AMP dependent kinase as template. GRID interaction fields were calculated through the amide, carbonyl and dry probe. Only attractive interactions were considered for the subsequent chemometrical analysis. The PCA score plot with the first relevant component on the x-axis and

the second on the y-axis is shown in Figure 6. We call this plot the target family landscape of kinases.

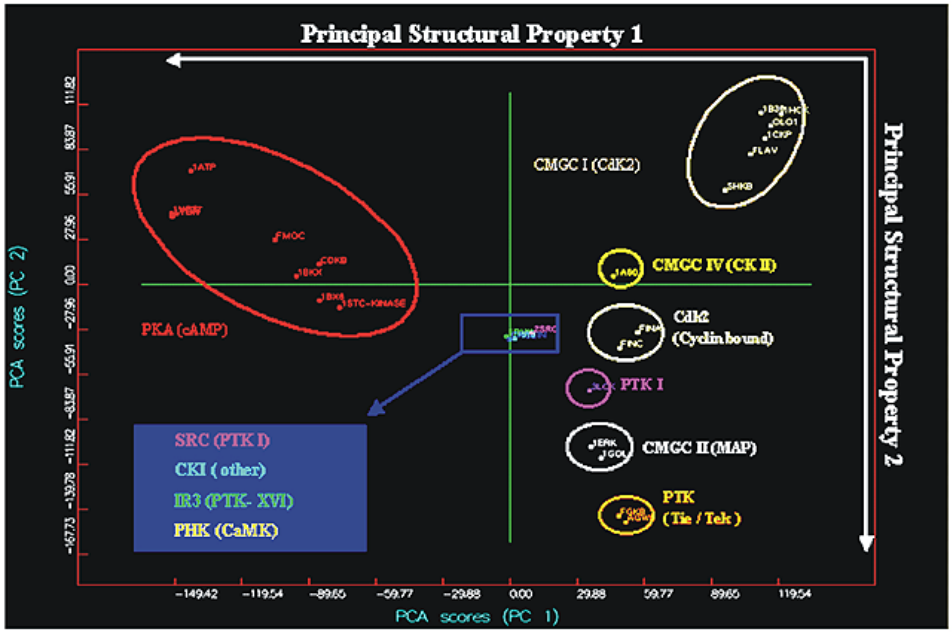

Figure 6. PCA score plot of the kinase landscape.

The first principle component separates CDK and MAP receptor kinases on the right with positive PC1 score values from the family of PKA kinases on the left.

The second principal component allows to separate between MAP and other receptor kinases with negative PC2 scores and the CDK family showing positive PC2 scores.

Kinase subfamily selectivity differences are explained by the corresponding PCA loadings plot (Fig. 7). The loadings of the first principal component of the dry probe clearly describe favorable and selective interactions of the family of PKA and CdK2 kinases. Blue contours indicate hydrophobic regions in space with preferences for PKA, while red contours highlight selective interactions for CdK2 (Fig. 7A). This principle component 1 is dominated by selectivity subsites in the kinase purine and hinge-binding region.

The loadings of the second principal component of the carbonyl probe outline favorable and selective interactions to the CdK2 and PTK1 kinase subfamilies. Blue contours favor hydrogen bond acceptors for CdK2 ligands, whereas red contours highlight selectivity interactions for the PTK1 subfamily (Fig. 7B). This principle component 2 is mainly driven by structural differences in the phosphate binding area.

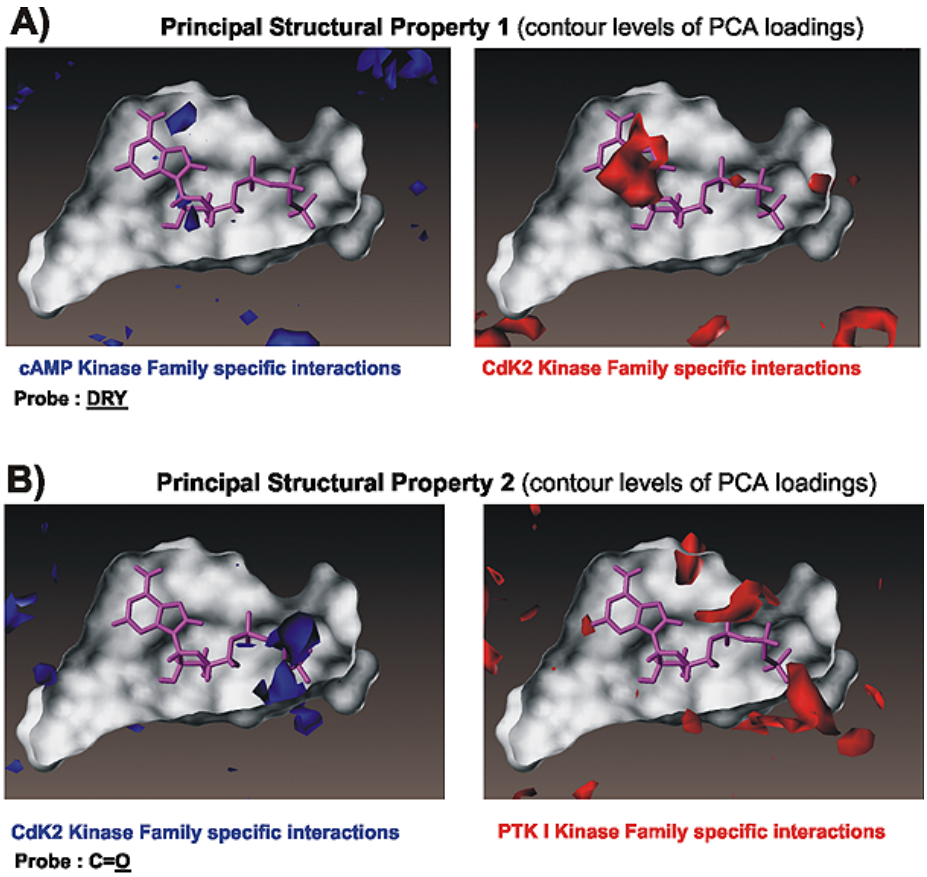

Figure 7. PCA loadings plot of kinases. A) First principle component of the GRID DRY probe highlights selective interactions for PKA and CDK2 kinases. B) Second principle component of the Carbonyl probe differentiates between CdK2 and PTK1 kinases.

EXAMPLES: KINASES, ION CHANNELS AND GPCRS

The design of our kinase library is based on the derived kinase landscape, on pharmacophore models and on privileged frameworks and fragments, which were derived from proprietary and public kinase projects. This kinase specific information was then applied to build our first kinase directed library by cherry picking in our compound collection and by purchasing additional samples (Fig. 8). We are currently and continuously improving this biased library by designing new proprietary kinase scaffolds, from which small compound libraries are built.

Screening of this focused library against new kinase targets like NIK, yielded almost 10-times higher hit rates compared to whole library screening. Such focused screening enables derivation of initial SAR models suitable for subsequent compound optimization. Virtual screening then gives access to the entire compound collection.

Such derived kinase knowledge is applicable in lead optimization of compounds as well. For instance, our initial IκB lead compound was lacking kinase specificity.

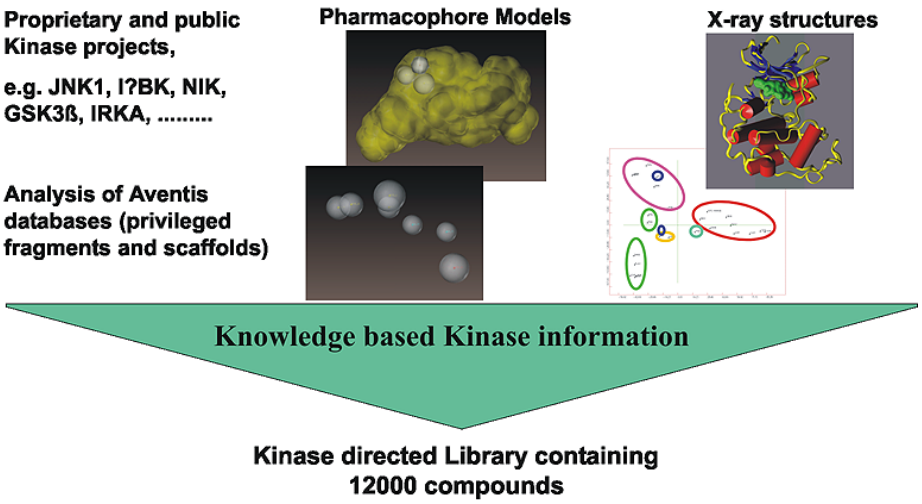

Figure 8. Design of a kinase focused library.

We were able to improve this selectivity by applying our kinase landscape. In this way, we identified a hydrophobic pocket, which is most important to gain selectivity.

Ion channels are of potential interest not only for cardiovascular diseases. However, appropriate high throughput assays to test several hundred thousand compounds against a particular ion channel are still lacking sufficient signal to noise ratios (9). Therefore, a biased ion channel library is of high interest for lead finding in that field.

Our ion channel library, containing 15412 compounds, was composed on public and proprietary ion channel lead compounds. Similarity searches and virtual screening based on pharmacophores and homology models yielded most representatives of this library. Addition of several small combinatorial libraries of ion channel privileged scaffolds then leads to our biased library (Fig. 9). Such QSAR driven refined models or related pharmacophores are well suited for a knowledge-based optimization of certain potassium channel inhibitors. We are constantly improving these models and adding new models as well.

G-protein coupled receptors comprise a large protein family sharing a conserved trans-membrane structure composed of seven trans-membrane helices (11). GPCRs are located at the surface of the cell and are responsible for the transduction of an endogenous signal into an intracellular response. The natural ligands of this receptor family are extremely diverse, for instance biogenic amines, amino acids, lipids, peptides and proteins or nucleosides and nucleotides.

As GPCRs are quite important biological targets, we decided to improve our lead finding capabilities in this field by building a GPCR biased library.

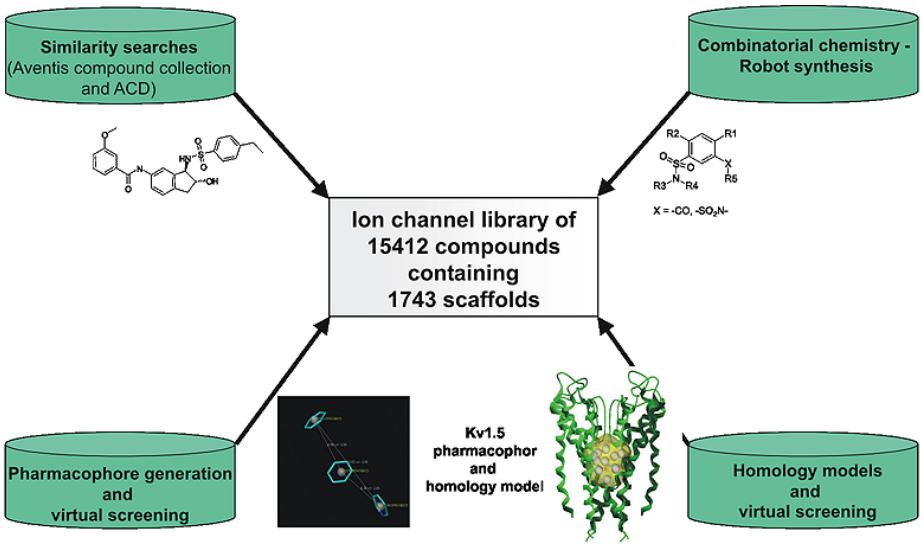

Figure 9. Building of the ion channel library.

Our knowledge based Aventis GPCR library was built upon a GPCR database containing 20,000 public and proprietary GPCR ligands. Fragment and framework analysis revealed privileged GPCR motifs from which some were turned into small combinatorial libaries. Further compound selection by virtual screening yielded our GPCR biased library (Fig. 10). This library still needs improvement, because of the high diversity of potential GPCR ligands.

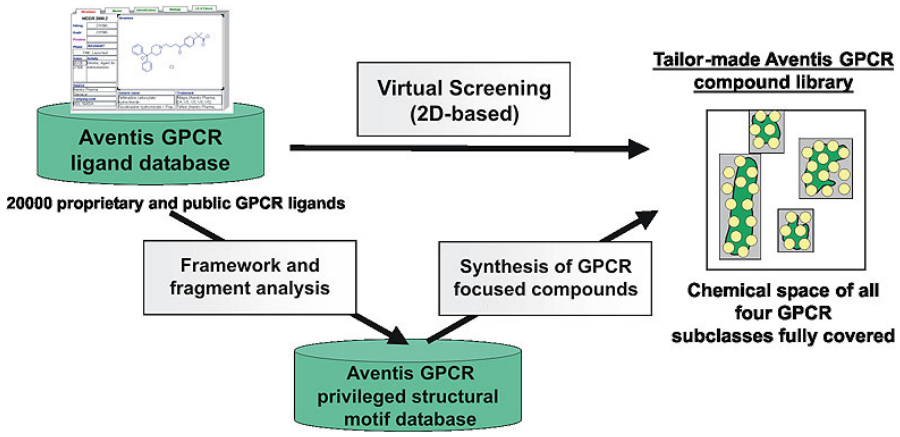

Figure 10. 2D-Based Approach to a Targeted GPCR Library.

The overall outcome of our lead finding strategy of Urotensin II receptor antagonists is outlined in Table 1. We observed an almost 20-times higher success rate by our rational approach in comparison to HTS. The hit rate of our GPCR directed library, however, was only slightly higher than by random screening. This is not disappointing, because that initial GPCR biased library was lacking ligands against peptidic GPCRs.

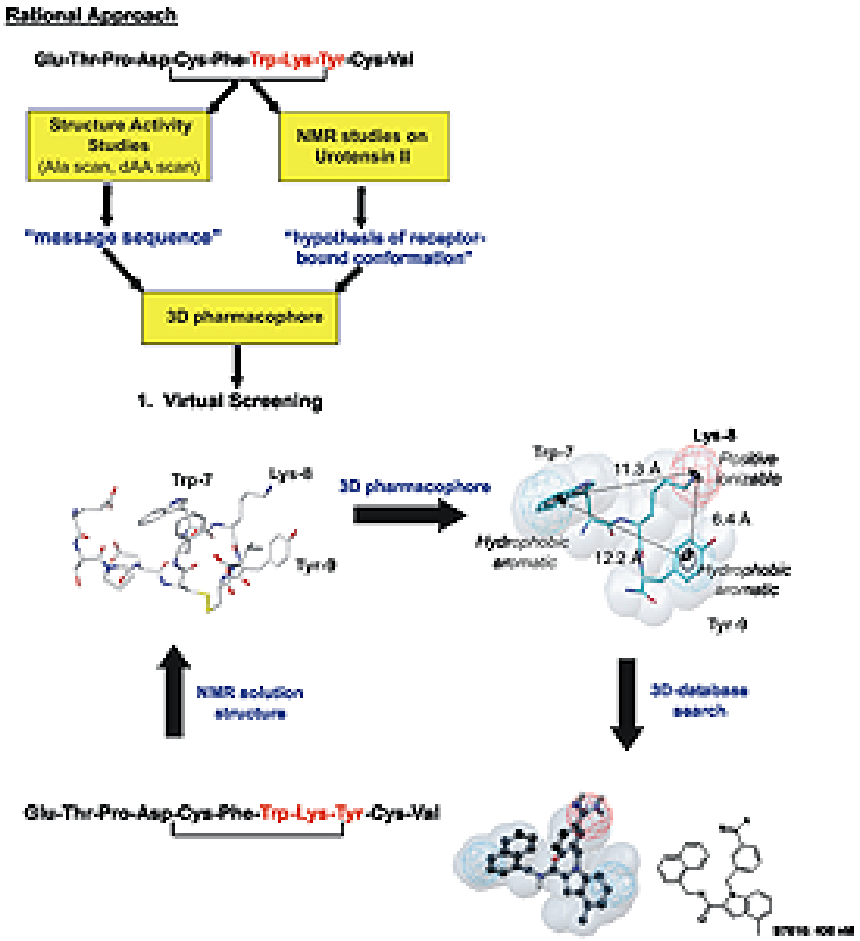

Figure 11. Rational Approach to Identification of Non-Peptidic Urotensin II Receptor Antagonists

We applied recently a knowledge-based lead finding strategy in the GPCR field. We were particularly interested to identify non-peptidic Urotensin II receptor antagonists. Urotensin II, a peptide comprised of 11 amino acids and a disulfide bridge, is the most potent vasoconstrictor known and therefore is of therapeutic interest (12). A combined rational and screening lead finding strategy was applied, whereby a part of our whole library and of course our GPCR biased library was screened. In addition, a rational approach by gathering structure-activity information through an alanine scan and by NMR studies on Urotensin II was performed (Fig. 11). Thereby, the message sequence tryptophan, lysine and tyrosine was successfully identified and a hypothetical receptor bound conformation of Urotensin II was elucidated, from which a 3D pharmacophore was constructed for virtual screening (13).

The spatial 3-dimensional arrangement of the message sequence tryptophan, lysine and tyrosine was elucidated from the NMR solution structure of Urotensin II.

By assuming this as the bioactive conformation a 3-point pharmacophore was built, containing two hydrophobic features from the aromatic moieties of tryptophan and tyrosine and a positive

ionizable center from the terminal amino group of lysine. In addition to these three features, the shape of the message sequence was included in the pharmacophore. Subsequent virtual screening returned S7616 as most active Urotensin II receptor antagonist with an IC_{50} of 400nM (Fig. 11).

CONCLUSIONS

Knowledge based lead finding is achieved by matching chemical and biological space through collecting target family related ligands and identification of privileged structural motifs, building distinct 3D pharmacophores, and 3D classification of binding sites to build target family biased libraries for focused screening to find better hits faster.

ACKNOWLEDGEMENTS

The authors thank Clemens Giegerich, Anna Gorokhov, Sven Grüneberg, Gerhard Heßler, Robert Jäger, Andreas Kugelstadt and Stefania Pfeiffer-Marek for fruitful discussions. We also like to thank our collaborators, Gabriele Cruciani from Perugia, Gerhard Klebe and the Cambridge Crystallographic data center, the GMD, now Fraunhofer institute, and BioSolveIT.

REFERENCES AND NOTES

[1] Bajorath, J. (2001). Rational drug discovery revisited: interfacing experimental programs with bio and chemo-informatics. *Drug Discov. Today* **6**:989-995.

[2] Drews, J. & Ryser, S. (1997). The role of innovation in drug development. *Nat. Biotechnol.* **15**:1318-1319.

[3] Wess, G., Urmann, M., Sickenberger, B. (2001). Medizinische Chemie: Herausforderungen und Chancen. *Angew. Chem.* **113**:3443-3453.

[4] Murcko, M. A. & Bemis, G. A. (1996). The properties of known drugs: 1. Molecular frameworks. *J. Med. Chem*. **39**:2887-2893.

[5] Lewell, X. Q., Judd, D. B., Watson, S. P., Hann, M .M. (1998). RECAP-Retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with usefule applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **38**:511-522.

[6] Matter, H. & Schwab, W. (2000). A view on affinity and selectivity of nonpeptidic matrix metalloproteinase inhibitors from the perspective of ligands and target. In *Molecular Modeling and Prediction of Bioactivity*; Gundertofte, K., Jorgensen, F. S., Eds.; Kluwer: New York, pp 123-128.

[7] Westerhuis, J. A., Kourti, T., Macgregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemom.* **12**:301-321.

[8] Kastenholz, M. A., Pastor, M., Cruciani, G., Haaksma, E. E. J., Fox, T. (2000). GRID/ CPCA: A new computational tool to design slective ligands. *J. Med. Chem.* **43**:3033-3044.

[9] Naumann, T. & Matter, H. (2002). Structural classification of protein kinases using 3D molecular interaction field analysis of their ligand binding sites: Target family landscapes. *J. Med. Chem.* **45**:2366-2378.

[10] Xu, J., Wang, X., Ensign, B., Li, M., Wu, L., Guia, A., Xu, J. (2001). Ion-channel assay technologies: quo vadis? *Drug Discov. Today* **6**:1278-1287.

[11] Doyle, D. A., Cabral, J. M., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., Chait, B. T., MacKinnon, R. (1998). The structure of the potassium channel: Molecular basis of K^+ conduction and selectivity. *Science* **280**:69-77.

[12] Flower, D. R. (1999). Modelling G-protein-coupled receptors for drug design. *Biochimica et Biophysica Acta* **1422**:207-234.

[13] Wess, J. (1998). Molecular Basis of Receptor/G-Protein-Coupling Selectivity. *J. Pharmacol. Ther.* **80**:231-264.

[14] Ames, R. S., Sarau, H. M., Chambers, J. K., Willette, R. N., Aiyar, N. V., Romanic, A. M., Louden, C. S., Foley, J. J., Sauermelch, C. F., Coatney, R. W., Ao, Z., Glover, G. J., Wilson, S., McNulty, D. E., Ellis, C. E., Elshourbagy, N. A., Shabon, U., Trill, J. J., Hay, D. W. P., Ohlstein, E. H., Bergsma, D. J., Douglas, S. A. (1999). Human Urotensin-II is a potent vasoconstrictor and agonist for the orphan receptor GPR14. *Nature* **401**:282-286.

[15] Flohr, S., Kurz, M., Kostenis, E., Brkovich, A., Fournier, A., Klabunde, T. (2002). Identification of nonpeptide Urotensin II receptor antagonists by virtual screening based on a pharmacophore model derived from structure-activity relationships and nuclear magnetic resonance studies on Urotensin II. *J. Med. Chem.* **45**:1799-1805.