Project Number JPD - 0001

Prediction of New Pharmacophore Scaffolds

A Major Qualifying Project Report

Submitted to the Faculty and Staff of

WORCESTER POLYTECHNIC INSTITUTE

for requirements to achieve the

Degree of Bachelor of Science

in Chemistry

by

_____

Nicole Lounsbury

_____

Raha Moussavi-Aghdam

Date: 4/27/2011

Approved:

_____

Professor Dittami

Abstract

A computational method known as the Similarity Ensemble Approach (SEA) was used to predict drug-target interactions for molecular scaffolds previously made in the Dittami lab. Through three different analyses of this SEA data, the most promising scaffolds were determined. These scaffolds will then be synthesized and the predictions will be validated through a biological assay.

Table of Contents

List of Figures

List of Tables

Acknowledgments

We would like to thank Professor Dittami for all of his help and guidance with this project. We would also like to thank Alicia Morgan for all of her help in the lab and with this paper.

Introduction

The overall goal of this work is the discovery and synthesis of original molecular scaffolds as new pharmacological leads that can be used to develop therapeutic agents to treat a wide range of biological targets.

The Dittami group[1] previously had uncovered a fast and efficient method for assembling complex multicyclic scaffolds using a photoinitiated intramolecular ylide-olefin cycloaddition method. The group has prepared, characterized, and published reports on a number of interesting scaffolds, some of which bear resemblance to the skeletal frameworks of biologically important natural products. To date, none of these scaffolds have been subjected to the scrutiny of biological testing. The group assembled a diverse array of novel compounds resulting from a tandem photoinitiated intramolecular ylide-olefin cycloaddition reaction. The syntheses of these compounds are short and efficient, often proceeding through common intermediates. More importantly, they are readily adaptable to the preparation of new scaffolds incorporating new ring systems and substituents. The majority of these products arise from photolysis of the corresponding aryl vinyl ether, aryl vinyl amine or aryl vinyl sulfide photo-precursors which incorporate either a pendant alkene or heterodipolarophile side chain. Most of this work has been done on either of two structural motifs: Type I or Type II photo-precursors, which differ in placement of the pendant dipolarophile side chain.

A goal of this group now is to identify the potential of these scaffolds as pharmacological agents. A secondary goal is to improve and expand upon the procedures we have developed for synthesis of these scaffolds.

Using a computational approach, biological targets have been predicted for these cyclic compounds. These targets were identified using the Similarity Ensemble Approach (SEA). SEA

is a statistics-based chemoinformatics approach that can be used to predict new targets for FDA-approved small molecule drugs and pharmaceutical compounds. The basic approach, pioneered by the Shoichet group at the University of California, San Francisco[2], relies on the premise that structurally similar molecules should exhibit similar biological activity. Thus, by comparing compounds to ligands which are known to bind to proteins, it should be possible to predict new drug-target associations. Their approach utilizes two types of fingerprinting (ECFP4 and Daylight) to assess structural similarity. They compared 3,365 US Food and Drug Administration (FDA) drugs to 65,241 ligands annotated in sets representing 246 protein targets from the MDL Drug Data Report (MDDR) database. Twenty-three previously unknown associations were found via this analysis and validated by bioassay. Therefore, SEA is promising for determining side-effects of existing drugs and discovery of new drugs. [2]

The prediction of targets for and the evaluation of the biological activities of our novel scaffolds are part of a collaboration with the Shoichet group who used the SEA approach to compare our scaffolds to the ligands from the MDDR database. The results of this computational analysis were expressed using expectation values (E-values). Validation of these biological target predictions will then be performed experimentally. The predicted associations will be tested via biological assay under The Psychoactive Drug Screening Program (PDSP) at the University of North Carolina, Chapel Hill, directed by Dr. Bryan Roth.

There are four specific aims for this project:

Specific Aim I Submit our structures to the Shoichet group for chemoinformatics assessment using the similarity ensemble approach (SEA) to predict drug-target associations.

Specific Aim II Synthesize the most promising molecular scaffolds and submit to the

Roth PDSP screens to test the predictions. In addition, in conjunction with the Roth group, we will do random screening of compounds against a broad cross section of G-protein coupled receptors to see if we uncover other targets not predicted by the SEA analysis.

Specific Aim III Use the bioassay data obtained to design a new round of targets for synthesis. These will incorporate bioisosteric replacements for suspected pharmacophores in the molecule to enhance therapeutic activity. New scaffolds will be submitted for additional screening against more specific targets as appropriate.

Specific Aim IV Evaluate new procedures to enhance the depth and breadth of the Photoinitiated Intramolecular Ylide-Olefin Addition Reaction for synthesis of new and novel pharmacophore scaffolds.

Project Background

Scaffold Syntheses

Our group has assembled a diverse array of novel compounds resulting from a tandem photoinitiated intramolecular ylide-olefin cycloaddition reaction. The syntheses of these compounds are short and efficient often proceeding through common intermediates. More importantly, they are readily adaptable to the preparation of new scaffolds incorporating new ring systems and substituents. Our goal now is to select and resynthesize a representative cross section of these compounds for biological testing.

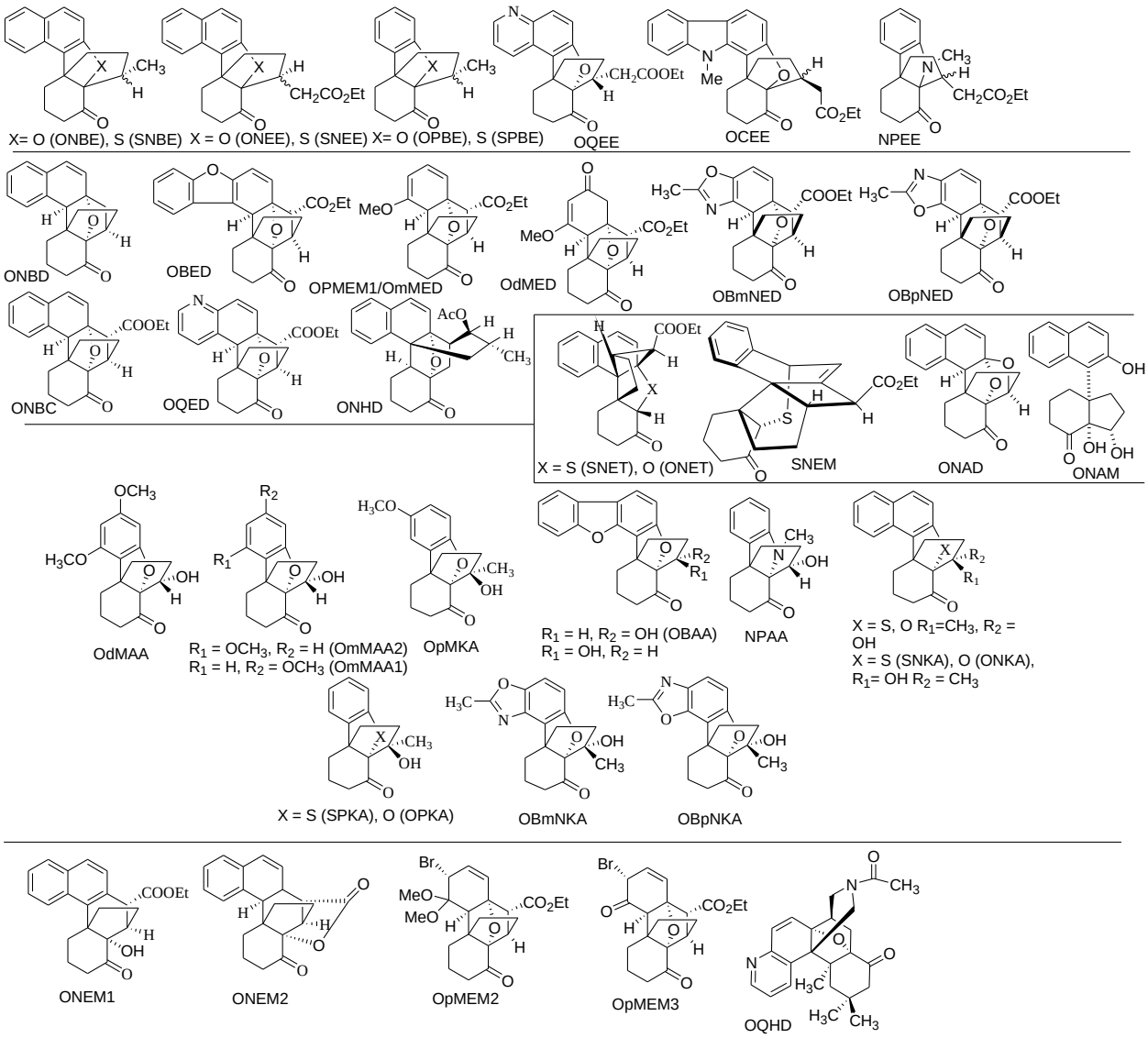


Figure 1: Representation of the molecular scaffolds created using photoinitiated intramolecular ylide-olefin cycloaddition with assigned code names (e.g. ONBE)

Since we have previously made and characterized every structure shown, we are assured success in gaining access to samples for testing. Figure 1 shows a representative set of the kinds of products we have made.

As noted earlier, the majority of these products arise from photolysis of the corresponding aryl vinyl ether, aryl vinyl amine or aryl vinyl sulfide photoprecursors which

incorporate either a pendant alkene or heterodipolarophile side chain. Most of our work has been done on either of two structural motifs: Type I or Type II photoprecursors, which differ in placement of the pendant dipolarophile side chain. The synthesis of each type is shown in Figure 2.

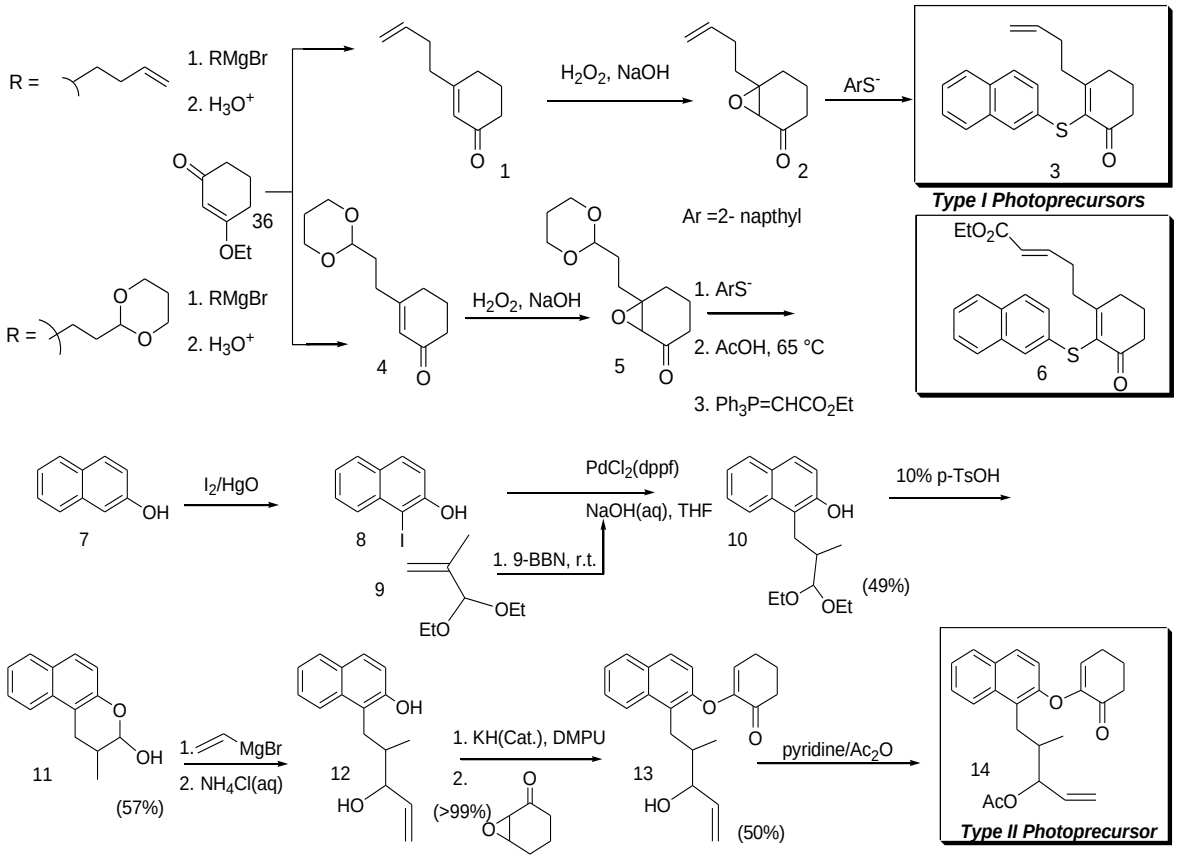


Figure 2: Synthesis of Type I and Type II photoprecursors from which the scaffolds were derived

Via the aforementioned routes we can assemble photoprecursors with a multitude of aromatic groups and pendant dipolarophiles, as shown in Figure 3.

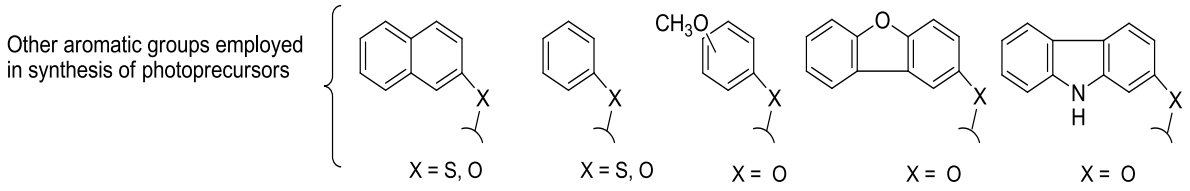


Figure 3: Aromatic groups that were utilized in the synthesis of the scaffolds

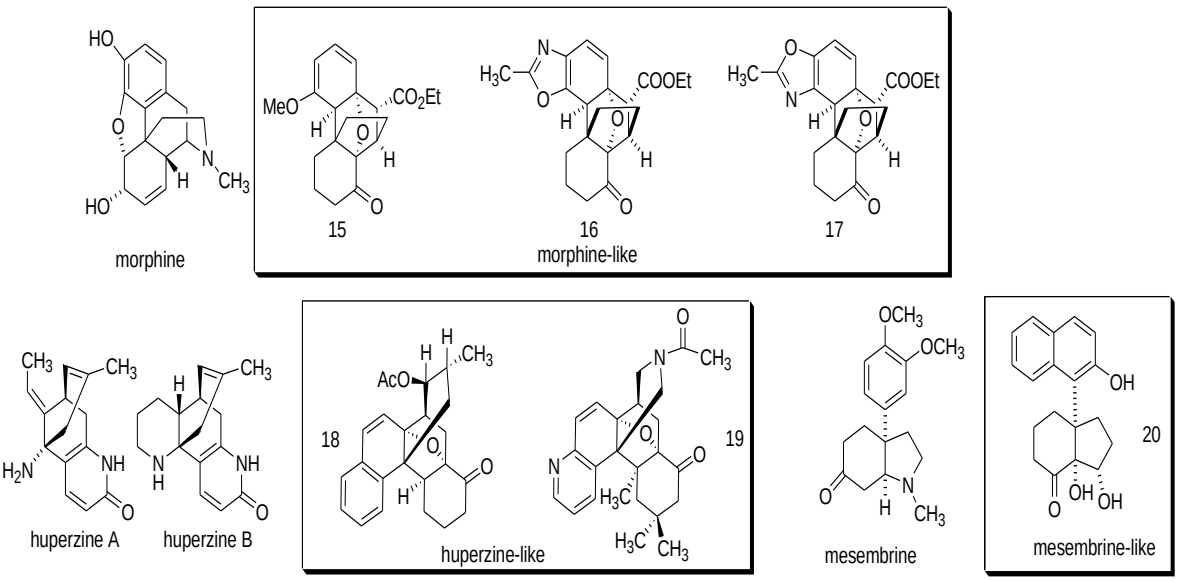


Figure 4: Examples of synthesized scaffolds which share similarity with naural products

Some examples of scaffolds resulting from the photolysis of these systems which incorporate features of natural products are shown in Figure 4. Notably morphine is an important analgesic and the huperzine alkaloids have been heralded as important acetylcholinesterase inhibitors with potential for treatment of Alzheimer's and Myasthenia Gravis.

In addition to changes in the aromatic component, there are several functional group transformations that are readily employed to convert our photoproducts into more diversely substituted products. These include reactions at the carbonyl center, carboxylic acid ester conversions and reactions at the alkene center. Examples can be seen in Figure 5: products like **21** are readily converted to either the alcohol or lactone while products like **23** are easily modified via electrophilic addition, and products such as **22** and **24** should be readily converted to the angularly substituted 6,5 ring system by reductive cleavage.

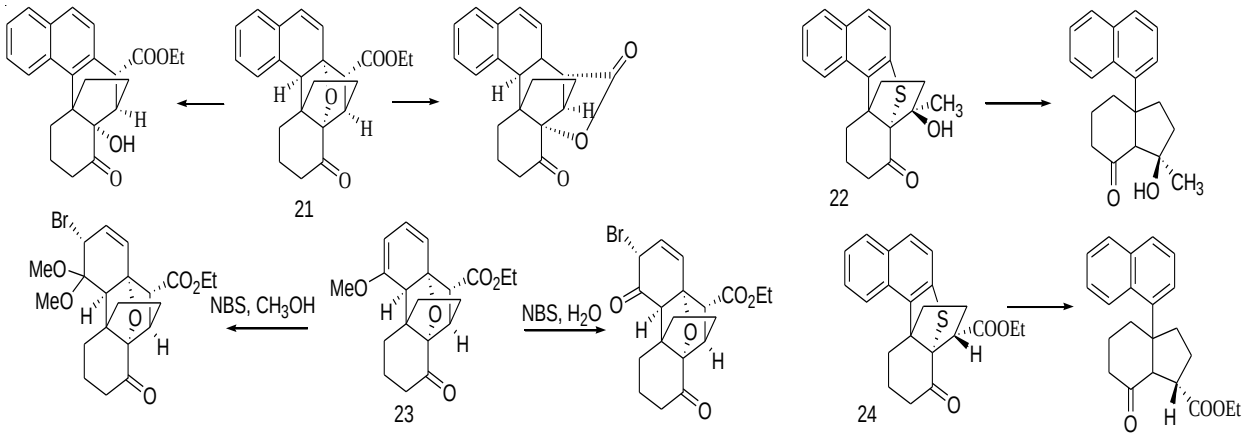


Figure 5: Examples of various functional group transformations that were performed

Using these photoprecursors, 40 scaffolds were synthesized and the structure verified by NMR, and in some cases, X-ray crystallography. The structures of these scaffolds were sent as "SMILES" files with assigned code names (e.g. ONBE, see Figure 1) to the Shoichet group in San Francisco for analysis using the SEA program. SMILES, or Simplified Molecular Input Line Entry System, is a form of line notation used to define a molecule. Instead of using numbers, however, it uses a predefined code of strings in order to easily and quickly describe complex molecules.[3]

SEA Data Analysis

Molecular Fingerprints

As stated earlier, the Similarity Ensemble Approach (SEA) was utilized to assess the structural similarity of our scaffolds to ligands known to bind to a diverse array of proteins. To do so, these compounds were compared using what are known as molecular fingerprints. Molecular fingerprints are representations of chemical structures that are useful for similarity searching, clustering, and classification. There are numerous fingerprinting models which concentrate on different aspects of chemical structures such as connectivity or topology. This concept is best described in comparison to a person's fingerprint. Every person has a characteristic fingerprint, yet fingerprints by themselves have no other meaning.

The most popular connectivity fingerprint algorithm is available through Daylight Chemical Information Systems[4] and is referred to as Daylight and is the next evolution from structural keys. Structural keys were the first type of screen used for high-speed screening of chemical databases. These structural keys are represented by Boolean arrays, arrays in which the presence or absence of specific structural patterns are identified in a true/false pattern.[4] The problem with the structural keys was that they were extremely specific. This specificity would screen out many chemicals that may have proved useful and generated many "false hits." The Daylight fingerprint algorithm addresses the lack of generality within the structural keys by eliminating the idea of pre-defined patterns. Daylight examines chemical structures and is able to identify patterns of the chemical, however the meaning of the patterns are not well defined. This allows for the patterns of each chemical to be compared to the patterns of other molecules to check for structural similarity without prioritization.

While the Daylight algorithm has been used in both high throughput screening and similarity searching methodology for decades, another method which examines the chemical topology fingerprints has become popular in the last few years. Specifically, the extended connectivity fingerprints (ECFP) are a novel class of topological fingerprints developed for structure-activity modeling.[5] These fingerprints use atom information obtained from the Daylight atomic invariants rule to identify atoms. The atomic invariants rule are six innate properties which are not dependent on the fashion in which the atoms of the chemical are numbered. These properties are:

- The number of immediate neighbors that are "heavy" (non-hydrogen) atoms
- The valence minus the number of hydrogens
- The atomic number
- The atomic mass
- The atomic charge
- The number of attached hydrogens.[5]

Often the additional property of whether the atom is connected to at least one ring is included. However, a novel feature of this ECFP fingerprint is that is does not follow a linear pathway like Daylight but uses a series of circular bond lengths to capture the environment of an atom in a pattern. This is reflected in the naming of these ECFP fingerprints. Each name is followed by a number which is the effective diameter of the largest structural feature and is equal to twice the number of iterations performed. For instance, ECFP4 indicates that two iterations were performed and the largest possible fragment will have a width of four bonds while ECFP6 indicates that three iterations were performed producing the largest possible fragment with six

bond widths. A second novel feature is that a pre-defined number of iterations are performed as opposed to continuously defining unique patterns. This allows the ECFP patterns to contain both positive and negative structural information crucial for analyzing molecular activity.

Each chemical structure receives a unique score for both the Daylight and ECFP fingerprint based on the atomic invariants rule. For Daylight these properties are linear and for ECFP they are circular. In order to suggest similarity between two compounds, the fingerprint scores, which range from zero to one, must be similar.

E-Values

Similarity between fingerprint scores is made statistically significant by expected values (E-values). This parameter describes the number of hits one can expect to see by chance when searching a database of a particular size; essentially describing the background noise. For example, an E-value score of one assigned to a hit can be interpreted as meaning that in a database it is expected to see a single match with a similar score by chance. The closer to zero the E-value is, the more significant the match between fingerprint scores. Any match with an E-value less than $1x10^{-10}$ is considered significant and suggests that the match is not likely to be occurring by chance.[6]

Drugability

In addition to the structural similarity, we also decided to analyze the chemophysical properties of the molecules to establish if there was a connection between our compounds and ligand sets in terms of these physical properties. These properties (such as ClogP, ClogD, and Topological Polar Surface Area) can be used as predictors for biological activity of compounds.

This composite data can provide information about a compound's "drugability" or tendency to interact with biological targets.

Among the values that are useful for calculating the drugability of a compound are ClogP, ClogD7.4, and the Topological Polar Surface Area. Log P, also known as the partition coefficient, is defined as the log of the ratio of a compound in n-octanol to its concentration in water. This value is a measure of how hydrophilic a compound is and also how well a compound would absorb in the body. Higher Log P values correlate to lower absorption, which is important when designing medicinal compounds.[7] ClogP is a method of quickly calculating Log P, which breaks the compound into fragments and calculates the fragment values based on the bonding environment.[8]

While Log P is calculated for compounds when they are neutral, Log D, or the distribution coefficient, uses the same calculation, except it includes both the neutral and charged forms of the compound. If the compound is neutral, then the Log P equals the Log D; however, when the molecule is charged, the degree to which it partitions itself in octanol may differ. Log D varies with pH, so for drugability, the Log D that is often used is at physiological pH, which is 7.4.[9] ClogD uses the same algorithm as ClogP.

Topological Polar Surface Area (TPSA) is calculated by adding the contributions of the polar fragments on the surface of a compound.[10] This differs from Polar Surface Area, which involves optimizing the geometry and creating a 3D molecule, and therefore takes longer amounts of time to calculate.[11] TPSA relates to how well a molecule transports across membranes, and is important when considering oral bioavailability.[12] This value is also especially useful regarding intestinal absorption and crossing of the blood-brain barrier. It has

been found that drugs with TPSAs less than 60 $Å^2$ are absorbed more than 90% in the intestines while drugs with TPSAs greater than 140 $Å^2$ are absorbed less than 10%.[13]

Lipinski's "Rule of 5" was developed to determine if compounds with medicinal properties would be good orally active drugs. The rule states that poor absorption of drugs is more likely when there are more than 5 hydrogen bond donors (OHs or NHs), there are more than 10 hydrogen bond acceptors (nitrogen or oxygen atoms), the molecular weight is greater than 500 daltons, and the Log P is greater than 5.[14] This means that drugs with good oral activity are smaller and more hydrophilic than non-orally active drugs.[15] These rules were determined by calculating values for drugs in clinical Phase II with good permeability and absorption. While most medicinal compounds obey the rule, there are a few classes of orally active drugs which do not fit within these guidelines, which include antibiotics, antifungals, vitamins and cardiac glycosides.[14] Recently, there have been extensions to the rule, with one paper by Ghose et al. stating that good absorption is more likely when the Log P is between -0.4 and 5.6, the molecular weight is between 160 and 480 daltons, the molar refractivity is between 40 and 130 and the number of atoms is between 20 and 70.[16] A separate paper by Oprea et al. says that there is good absorption when the Log P is between -3.5 and 4.5, the Log D at pH 7.4 is between -4 and 4, there are no more than 10 non-terminal single bonds, there are no more than 8 hydrogen bond acceptors and there are no more than 4 rings.[15]

Data Analysis Methodology

The compounds synthesized by the Dittami groups were submitted to Michael Keiser of the Shoichet group as "SMILES" files with corresponding codenames (e.g. ONBE, see Figure 1). These compounds were then analyzed by both the Daylight and ECFP4 fingerprinting methods as a part of the Similarity Ensemble Approach. This data was posted to online databases.

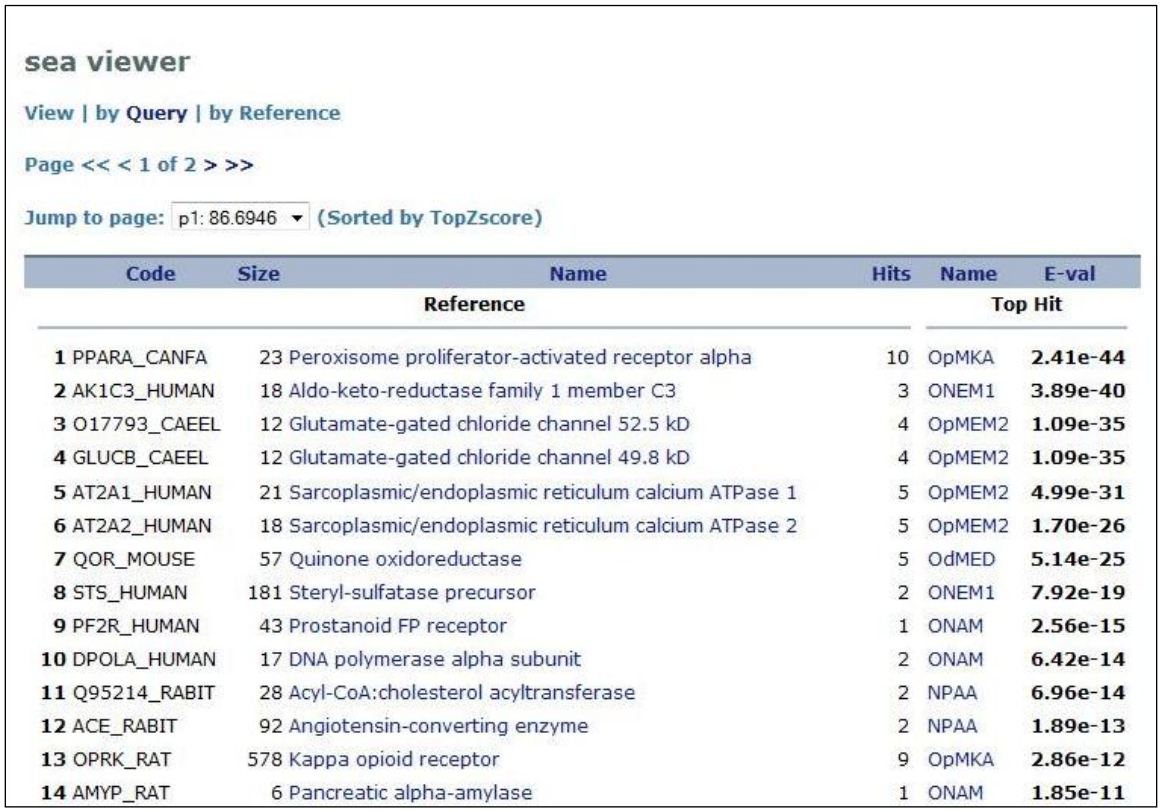
Two separate yet simultaneous approaches were taken to analyze the data. The first approach used the E-values to obtain an initial group of compounds for which drugability data was independently calculated and these values used to create a separate "hit" list. The second approach focused on the promiscuity of the proteins and compounds and the E-Values of the synthesized compounds. These lists were then cross-referenced to provide a master "hit" list of the compounds which will undergo biological testing.

Typically, there is not a single strict threshold that is used for deciding which E-values are significant. In their Nature paper, Keiser used a $1x10^{-10}$ threshold as a rough cutoff. However, according to Keiser, typically E-values smaller than $1x10^{-5}$ are used in practice. For the purposes of this data analysis, E-values of $1x10^{-7}$ or smaller were considered.

As stated previously, each of our scaffolds was compared to ligand sets for a diverse array of proteins. Within the online database, for each predicted protein-scaffold interaction, the ligands were ranked by either an ECFP4 or Daylight value, depending on which database it occurred in. This value, the Tanimoto coefficient (Tc), was between 0 and 1, with 0 being dissimilar compounds based on fingerprinting results and 1 being identical compounds.[17]

The data comparing the scaffolds to known biological targets were split into two databases based on which fingerprinting method was used for the comparison – Daylight or ECFP4. The data from this analysis was transferred into a spreadsheet and sorted by E-values.

Any compound with a significant E-value, less than $1x10^{-7}$, was selected for further analysis (Figure 6).



sea viewer

View | by **Query** | by Reference

Page << < 1 of 2 > >>

Jump to page: p1: 86.6946 (Sorted by TopZscore)

Code	Size	Name	Hits	Name	E-val
		Reference		**Top Hit**	
1 PPARA_CANFA	23	Peroxisome proliferator-activated receptor alpha	10	OpMKA	**2.41e-44**
2 AK1C3_HUMAN	18	Aldo-keto-reductase family 1 member C3	3	ONEM1	**3.89e-40**
3 O17793_CAEEL	12	Glutamate-gated chloride channel 52.5 kD	4	OpMEM2	**1.09e-35**
4 GLUCB_CAEEL	12	Glutamate-gated chloride channel 49.8 kD	4	OpMEM2	**1.09e-35**
5 AT2A1_HUMAN	21	Sarcoplasmic/endoplasmic reticulum calcium ATPase 1	5	OpMEM2	**4.99e-31**
6 AT2A2_HUMAN	18	Sarcoplasmic/endoplasmic reticulum calcium ATPase 2	5	OpMEM2	**1.70e-26**
7 QOR_MOUSE	57	Quinone oxidoreductase	5	OdMED	**5.14e-25**
8 STS_HUMAN	181	Steryl-sulfatase precursor	2	ONEM1	**7.92e-19**
9 PF2R_HUMAN	43	Prostanoid FP receptor	1	ONAM	**2.56e-15**
10 DPOLA_HUMAN	17	DNA polymerase alpha subunit	2	ONAM	**6.42e-14**
11 Q95214_RABIT	28	Acyl-CoA:cholesterol acyltransferase	2	NPAA	**6.96e-14**
12 ACE_RABIT	92	Angiotensin-converting enzyme	2	NPAA	**1.89e-13**
13 OPRK_RAT	578	Kappa opioid receptor	9	OpMKA	**2.86e-12**
14 AMYP_RAT	6	Pancreatic alpha-amylase	1	ONAM	**1.85e-11**

Figure 6: Screenshot of Daylight fingerprint results sorted by decreasing E-Value in context of the biological target

SEA and Drugability Analysis

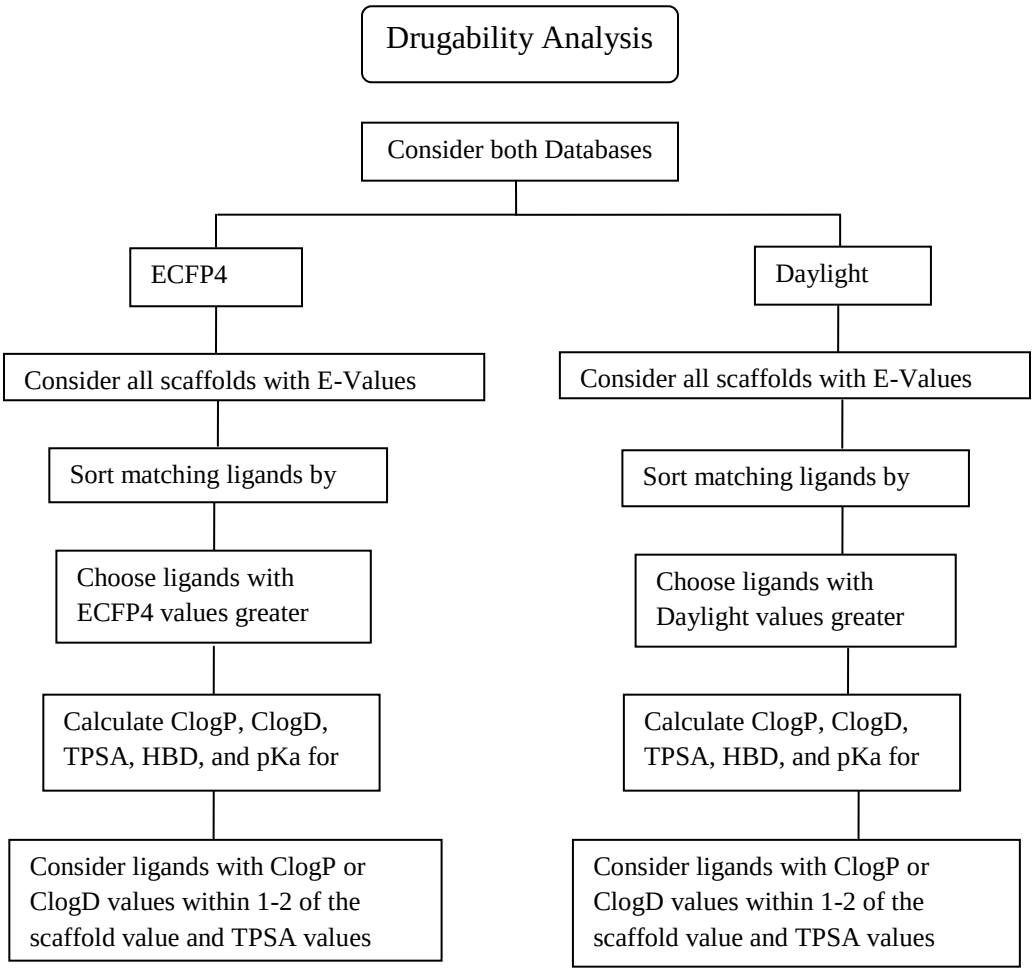


Figure 7: Flowchart of Drugability Analysis

Both the Daylight and ECFP4 databases were separately examined for the drugability analysis. When looking at the data in a database, we sorted the biological targets by E-value (Figure 7). E-values with an exponent of $1x10^{-7}$ or lower were considered to be good matches and were the only matches considered from that point forward.

For every predicted protein-scaffold interaction, the set of known ligands associated with the target were examined and analyzed (Figure 8). All of the ligands shown were examined and were subsequently sorted (targets in the ECFP4 database were sorted according to ECFP4 value

and target in the Daylight database were sorted according to Daylight value). Once these ligands were sorted by their respective comparison value, a cut-off value was established to narrow down the list of ligands to no more than the top twenty ligands (sorted by ECFP4 or Daylight value). The cut-off values were between 0.4-0.5 for Daylight and 0.1-0.2 for ECFP4.

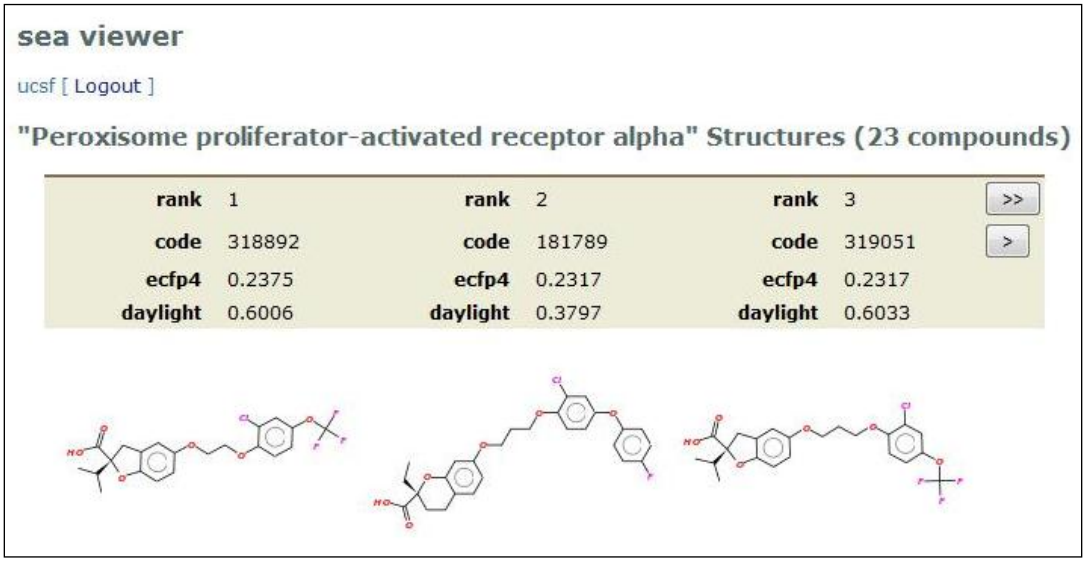


Figure 8: Screenshot of the top ranked known ligands for the Peroxisome proliferator-activated receptor alpha target protein used for comparison to scaffolds

A number of calculations were then performed on the selected ligands. These calculations included the ClogP, ClogD at pH 7.4, TPSA (Topological Polar Surface Area), TPSA at pH 7.4, Molecular Weight, HBD (Hydrogen Bond Donors), and HBD at 7.4, and pKa.

The calculations for every ligand in a set were compared to the calculations of its associated scaffold. Ligands whose ClogP and ClogD values were within 1-2 and whose TPSA value was within 10 of the scaffolds were chosen as the best matches for biological testing.

Promiscuity and E-Value Analysis

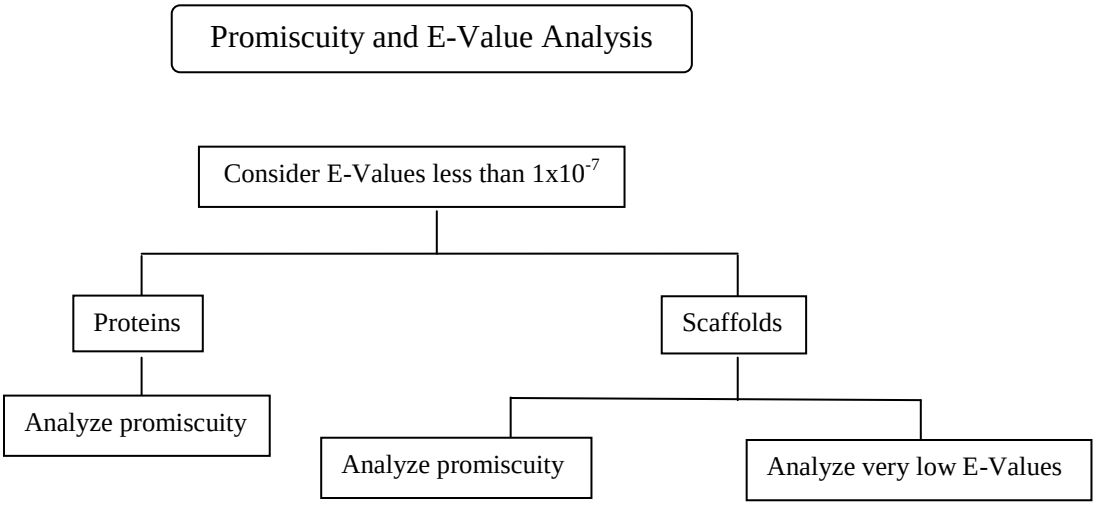


Figure 9: Flowchart of Promiscuity and E-Value Analysis

To compare the most promiscuous proteins and synthesized compounds (Figure 9), a new spreadsheet was created with all the predicted protein-scaffold interactions with an E-value with an exponent of $1x10^{-7}$ or less. Data from the Daylight database and ECFP4 database were considered separately. For each predicted interaction, the protein code, the protein name, the synthesized compound code, and its E-value were recorded.

Proteins that were predicted to interact with multiple scaffolds with very low E-values were recorded. Similarly, scaffolds that were predicted to interact with multiple proteins (especially those that were predicted to interact with multiple similar proteins) with very low E-values were recorded. Scaffolds that were only predicted to interact with one protein but had extremely low E-values were also considered and recorded.

Results

SEA and Drugability Analysis

Based on the drugability data, the following scaffolds were determined to be the best matches to the ligand sets of certain proteins:

Table 1: Best Scaffolds Based on Drugability Analysis

Compound Name	Protein Name	E-value	Database
ONBE	NAD-dependent histone deacetylase SIR2	$3.81x10^{-19}$	ECFP4
ONKA	NAD-dependent histone deacetylase SIR2	$2.57x10^{-15}$	ECFP4
OpMKA	Delta opioid receptor	$6.08x10^{-8}$	Daylight
ONEM1	Estradiol 17-beta-dehydrogenase 1	$6.05x10^{-8}$	Daylight
OdMED	Sarcoplasmic/endoplasmic reticulum calcium ATPase 1	$1.53x10^{-16}$	Daylight
ONAM	DNA polymerase alpha subunit	$6.42x10^{-14}$	Daylight
ONEM1	DNA polymerase alpha subunit	$9.27x10^{-12}$	Daylight
OpMKA	Kappa opioid receptor	$2.86x10^{-12}$	Daylight
OpMKA	NADH-ubiquinone oxidoreductase chain 4	$1.48x10^{-9}$	Daylight
OmMAA1	NADH-ubiquinone oxidoreductase chain 4	$1.68x10^{-9}$	Daylight
OdMAA	NADH-ubiquinone oxidoreductase chain 4	$7.40x10^{-10}$	Daylight
OpMKA	Mu opioid receptor (OPRM_CAVPO)	$1.74x10^{-7}$	Daylight
OpMKA	Mu opioid receptor (OPRM_HUMAN)	$2.33x10^{-7}$	Daylight

Promiscuity and E-Value Analysis

In the Daylight database, 28 proteins had predicted interactions with scaffolds with E-values with an exponent of $1x10^{-7}$ or less. A total of 60 predicted protein-scaffold interactions had an E-

value with an exponent of $1x10^{-7}$ or less. Of all the listed proteins, seven proteins were chosen because they had the most predicted protein-scaffold interactions with the lowest E-values. These proteins were:

- Peroxisome proliferator-activated receptor alpha
- Glutamate-gated chloride channel 52.5 kD
- Glutamate-gated chloride channel 49.8 kD
- Sarcoplasmic/endoplasmic reticulum calcium ATPase 1
- Sarcoplasmic/endoplasmic reticulum calcium ATPase 2
- Quinone oxidoreductase
- NADH-ubiquinone oxidoreductase chain 4

Table 2: Proteins with high promiscuity and low E-Values in the Daylight database

Protein Name	**Number of Interactions**	**Lowest E-Value**
Peroxisome proliferator-activated receptor alpha	9	$2.41x10^{-44}$
Glutamate-gated chloride channel 52.5 kD	4	$1.09x10^{-35}$
Glutamate-gated chloride channel 49.8 kD	4	$1.09x10^{-35}$
Sarcoplasmic/endoplasmic reticulum calcium ATPase 1	3	$4.99x10^{-31}$
Sarcoplasmic/endoplasmic reticulum calcium ATPase 2	3	$1.7x10^{-26}$
Quinone oxidoreductase	3	$5.14x10^{-25}$
NADH-ubiquinone oxidoreductase chain 4	4	$7.04x10^{-10}$

A number of synthesized compounds were selected because of their promiscuity or because they had a predicted protein-scaffold interaction with a low E-value. Compounds with high promiscuity are shown in the table below.

Table 3: Compounds with high promiscuity and low E-Values in the Daylight database

Compound Name	Number of Interactions	Lowest E-Value
OmMED	2	$2.50x10^{-21}$
OdMED	6	$5.14x10^{-25}$
ONAM	8	$1.06x10^{-15}$
ONEM1	7	$3.89x10^{-40}$
OpMEM2	5	$1.09x10^{-35}$
OpMKA	8	$2.41x10^{-44}$

Compounds with a predicted protein interaction with a low E-value are shown below.

Table 4: Compounds with low E-Values in the Daylight database

Compound Name	E-Value
OdMAA	$5.74x10^{-39}$
OmMAA1	$5.73x10^{-40}$
OmMAA2	$4.79x10^{-39}$
ONBE	$7.71x10^{-39}$
ONEE	$1.01x10^{-36}$
ONKA	$8.70x10^{-40}$
OPBE	$1.10x10^{-39}$

In the ECFP4 database, three proteins had predicted protein-scaffold interactions with E-values with an exponent of $1x10^{-7}$ or less. A total of seven predicted interactions had an E-value with an exponent of $1x10^{-7}$ or less. Of all the listed proteins, one protein was chosen because it had the most predicted protein-scaffold interactions with the lowest E-values. This protein was:

- NAD-dependent histone deacetylase SIR2

Table 5: Proteins with high promiscuity and low E-Values in the ECFP4 database

Protein Name	Number of Interactions	Lowest E-Value
NAD-dependent histone deacetylase SIR2	4	$3.81x10^{-19}$

No compounds in this list were part of more than one predicted protein-scaffold interaction.

Only two drug-target interactions appeared as predicted protein-scaffold interactions in both databases. These are ONEM2 for Aldo-keto-reductase family 1 member C3 and ONAM for NAD-dependent histone deacetylase SIR2. Both of these predicted interactions had relatively high E-values though, with exponents > $1x10^{-9}$. This is noteworthy; that with so many predicted protein-scaffold interactions, only two appear in both databases and neither of them are particularly good matches.

Conclusions

After analyzing the predicted protein-scaffold interactions based on drugability, E-value and promiscuity, the following four scaffolds were chosen as the most promising.

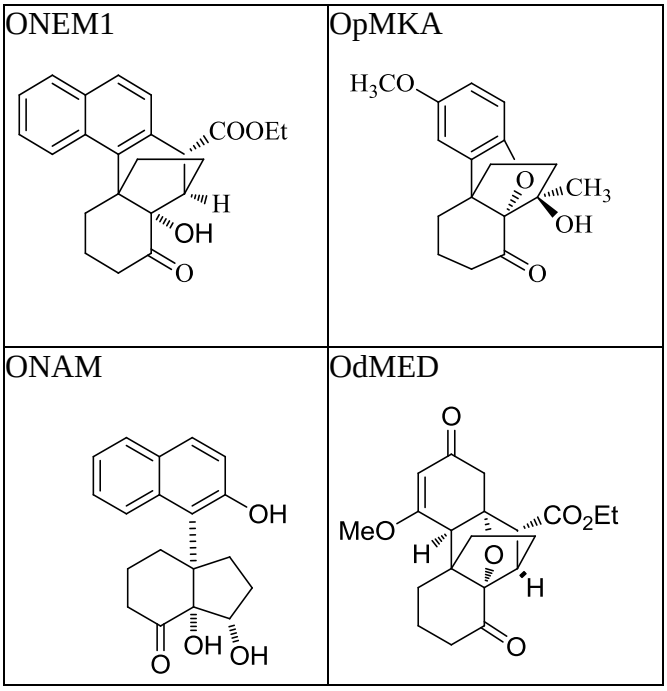


Figure 10: Structures of the best scaffolds

In addition, the following proteins were chosen as the best targets based on E-value, promiscuity and protein function analysis: Aldo-keto-reductase family 1 member C3, due to its possible role in asthma and cell growth,[18] and Sarcoplasmic/endoplasmic reticulum calcium ATPases 1 and 2, due to their potentials as therapeutics for heart failure. [19, 20]

It is recommended that the previous four scaffolds be synthesized and the protein-scaffold predictions be validated via biological assay, with the most interesting protein targets. Once the scaffolds have been confirmed to bind to the targets, new derivatives should be synthesized and investigated as potential drugs.

References

[1] Dittami, J.P.; Nie, X-Y; Nie, Hong; Ramanathan, H.; Buntel, C.; Rigatti, S.; Bordner, J.; Decosta, D.; Williard, P. *J. Org. Chem.* 1992, *57* , 1151

[2] Keiser, M.J.; Setola, V.; Irwin, J.J.; Laggner, C.; Abbas, A.I.; Hufeisen, S.J.; Jensen, N.H.; Kuijer, M.B.; Matos, R.C.; Tran, T.B.; Whaley, R.; Glennon, R.A.; Hert, J.; Thomas, K.L.H.; Edwards, D.D.; Shoichet, B.K.; Roth, B.L. *Nature.* 2009, *462*, 175-181

[3] Weininger, D. *J. Chem. Inf. Comput. Sci.* 1988, *28*, 31-36

[4] Daylight Toolkit, version 1, Daylight Chemical Information Systems: Mission Viejo, CA, 1987.

[5] J. Chem. Inf. Model., **2010**, 50, 742-754

[6] Pro. Natl. Acad. Sci USA, **1990**, 87, 2264-2268

[7] *ClogP Calculation*; http://www.organic-chemistry.org/prog/peo/cLogP.html

[8] *ClogP User Guide*; http://www.biobyte.com/bb/prod/40manual.pdf

[9] *ADME Tutorial: CLogP versus CLogD*; http://www.cmbi.ru.nl/edu/bioinf4/adme-Prac/clogd.shtml

[10] Fernandes, J.; Gattass, C.R. *J. Med. Chem.* 2009, *52* (4), 1214-1218.

[11] Remko, M. *Chem. Pap.* 2006, *61* (2), 133-141.

[12] Ertl, P.; Rohde, B.; Selzer P. *J. Med. Chem.* 2000, *43* (20), 3714-3717.

[13] Ertl, P. Polar Surface Area in *Molecular Drug Properties,* R. Mannhold (ed), Wiley-VCH, 2007, 111-126.

[14] Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. *Adv. Drug Delivery Rev.* 2001, *46* (1-3), 3-26.

[15] Oprea, T.I.; Davis, A.M.; Teague, S.J.; Leeson, P.D. *J. Chem. Inf. Comput. Sci.* 2001, *41* (5), 1308-1315.

[16] Ghose, A.K.; Viswanadhan, V.N.; Wendoloski, J.J. *J. Comb. Chem.* 1999, *1* (1), 55-68.

[17] Keiser, M.J.; Roth, B.L.; Armbruster, B.N.; Ernsberger, P.; Irwin, .J.; Shoichet, B.K. *Nat. Biotechnol.* 2007, *25* (2), 197-206.

[18] AKR1C3 aldo-keto reductase family 1, member C3 (3-alpha hydroxysteroid dehydrogenase, type II) [*Homo sapiens*]; http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd=ShowDetailView&list_uids=8644.

[19] Teucher, N.; Prestle, J.; Seidler, T.; Currie, S.; Elliot, E.B.; Reynolds, D.F.; Schott P.; Wagner, S.; Koghler, H.; Inesi, G.; Bers, D.M.; Hasenfuss, G.; Smith, G.L. *Circulation*. 2004, *110 (*23), 3553-3559.

[20] Kawase, Y.; Hajjar, R.J. *Nat. Rev. Cardiol.* 2008, *5*, 554-565.