Fractal Box Counting Applied to Structure-Property Relationships

L. Weber, M. Almstetter, M. Cappi, T. Fuchs, S. Hess, K. Illgen, A. Treml, P. Zegar

Morphochem AG, Gmunder Str. 37-37A, 81379 München, e-mail: lutz.weber@morphochem.de, phone: +4989780050, fax: +498978005555

1 Introduction

The concept of fractal dimensions has been developed to describe various scaling properties of natural objects by mathematical equations, based on the observation that the structure-property relationship of an object can be independent on the scaling that is used to measure its properties [1-3]. Thus, fractals are defined as "self-similar" at every scale, measured by using appropriate fractal dimensions.

Fractal theory was successfully applied to describe and yield decision-making tools in various industries. In drug discovery, fractal dimensions have been used in clustering data sets [4], protein surfaces [5], binding kinetics [6], DNA structure [7], analysis of biopsies [8] and compound library design [9]. Here, we will describe the utility of fractal dimensions to evaluate the structure-property relationship of a complete combinatorial library of small molecules together with their biological activities.

The described, publicly available data set of 15'840 compounds *vs* 5 enzymes has been used by several groups as a test set for the predictive power of their computational methods.

The set of 5*44*16*24 activity data, together with the smiles of the starting materials can be found at [10].

The box counting method as a means to determine fractal dimension towards evaluating structure-property relationships of combinatorial libraries was implemented. Therefore, a C++ program, FractalBoxes, was written and is available upon request. The program can read bit-maps or tab-delimited data sets. Bit-maps were used to verify the correct calculation of fractal dimensions for n-dimensional pictures of typical fractal test case objects (e.g. the Koch snowflake).

The input structure-property data file consists of a header line describing the data, such as a compound identifier, the coordinates of the respective compound in the n-dimensional space of the combinatorial library such as $A(1\ldots N_A)B(1\ldots N_B)C(1\ldots N_C)D(1\ldots N_D)E(1\ldots N_E)$ as well as the property values found for the individual property types and starting material combinations. The program than allows the display, as

well as selecting and evaluating subsets of the data. Thus, it is possible to select and deselect coordinates, enabling to study their effect of certain parts of the combinatorial library such as for example individual starting materials or combinations of starting materials.

The calculation of the fractal dimension by the box counting method is straightforward: an n-dimensional box is spanned by the n-types of starting materials used. For example, the smallest box with a size of one covers one compound or combination of starting materials, a box size of two (2*2*2) covers 8 compounds for an 3-dimensional space and so forth. The box is than counted as filled if at least one active compound below a given threshold is found within this box. If no active compound is found within the box of the defined size, the box is considered as empty.

Thus, in our example, a compound is considered as active and hence a box, within which this combination is found, is considered as filled, if its actual IC50 is below or equal the chosen value step. Selecting a value step of 10 will result in considering compounds active, if their respective IC50 value is below or equal 10µM. To avoid distortions from SAR patterns that might emerge from the random initial arrangement of the compounds along the dimensions $X(1\ldots N_X)$ the calculation of the fractal dimension was repeated several times by randomizing the coordinates, such that the respective property values are carried over to the new coordinates. For our current example of A(1…44)B(1…16)C(1…24) reaction products, the minimum of the box size **r** is 1*1*1, the maximum size is 16*16*16. Randomizing the coordinates ensures that all combinations are considered. For each value step, the number of filled boxes **N(r)** are counted and log(**N(r)**) is displayed versus the box size using log(1/**r**). The slope of the resulting plot gives the box counting $\mathbf{d_b}$ fractal dimension.

$$\mathbf{d_b} = \lim_{r->0} \mathrm{Log}(\mathbf{N(r)})/\mathrm{Log}(1/\mathbf{r})$$

Calculating the fractal dimension using various upper concentration limits gives the dependency of the fractal dimension on the IC50's of the reaction products.

2 Results

2.1 Box counting

Fractal dimensions were calculated for all enzymes in relation to the observed IC50's of the respective reaction products (Fig. 4). Each such $\mathbf{d_b}$ value is the average of 100 runs on the same data set, randomizing the AxByCz coordinates.

The concentration dependency of fractal dimensions is quite different for the different enzymes. A fractal dimension of $\mathbf{d_b}$ = 1 is observed for tryptase at approximately 5µM, for Factor Xa at 15µM, for u-PA at 33 µM, for trypsin at 50 µM. It was not observed for chymotrypsin in the measured concentration range. Below a certain concentration, a marked drop of $\mathbf{d_b}$ is observed: starting approximately below 40 µM for tryptase, 40µM for Factor Xa, and 60µM for u-PA, while for trypsin and chymotrypsin a rather flat curve is seen.

We also studied the influence of individual starting materials Ax, By or Cz on $\mathbf{d_b}$ by removing all combinations that contain the respective individual starting material from the data set and calculating the percent decrease $\Delta\mathbf{d_b}$. For example, this was calculated for the Factor Xa data set both at 15µM ($\mathbf{d_b}$ = 1) and 80µM concentration, see below.

In the same way, we calculated the effect of removing pairs of starting materials from the data set such as AxBy, ByCz and AxCz. An additive inhibitory effect of the respective two individual starting materials should result in finding them also in the list of high scoring starting materials. If high scoring pairs are found that are composed of starting materials that do not score high as individuals, a synergistic effect must be assumed. Thus, the cooperative activity of pairs of starting materials can be observed in a straightforward way.

2.2 Factor Xa inhibitors

Evaluating the impact of individual starting materials A, B or C on the fractal dimension confirmed this first assessment. Removing B13 at the 80µM concentration level had the highest impact and lowers the fractal dimension by 17.5%, next are B11 (5.8%), B1 (3.3%), C19 (1.7%), followed by C2, C24, and C15. At 15µM where $\mathbf{d_b}$ = 1, the impact of B13 grows to 30.8%, followed by C2 (19.5%), B1 (9.6%), A39 (5.3%), C23 (4.0%), and A9, C4, A12, A42, A43, C22. Thus, the increasing percentage of active reaction products that contain A39, B13 or C2 at lower concentrations ensures that these starting materials participate in forming an active reaction product and that the activity is not due to the activity of the respective starting materials themselves. Also it is noteworthy that A39, B13 and C23 score high, which is reflected in their participation in the most active starting materials. Thus, $\mathbf{d_b}$ reflects well the SAR for these Ugi-3CR products. However, C23 is only second after C2, raising the question of why the linear combination A39B13C2 is not the most active reaction product. This becomes even more evident when evaluating

the impact of removing pairs of starting materials on $\mathbf{d_b}$. Here, the pair B1C2 had the highest score with 9.8%, followed by B13C2 (6.7%), A39B13 (4.7%), and than B13C23, B13C4, A12B13, B13C22, A9B13, A42B13, A43B13, A6B13, B13C24, A37B13, B13C5, B13C14, B13C17, B13C19. The combination A39B13 scoring third as well as the following combinations that show the variability of both the aldehyde and the isonitrile component around B13 make sense in respect to the Ugi-3CR SAR. In contrast, the pairs B1C2 and B13C2 follow a different SAR around C2 that is the result of the formation of a different backbone as already published by us earlier [12], giving raise to piperazinones A43B1C2 (IC50 1.0µM) and A23B1C2 (IC50 1.9µM) and A6B13C2 (IC50 2.5µM). In this SAR, the isonitrile component is fixed and only B and C are variable.

It is interesting to note that the summary contribution to the decrease of $\mathbf{d_b}$ of the three classes of starting materials A, B and C is almost equal – the sum of all decreases for A is 39.1%, for B is 42.4% and for C is 42.1%. However, when normalized with the number of starting materials in the respective class, B/16 is 2.65%, C/24 is 1.75% and A/44 is 0.89%. This reflects the importance of the B substituent for the S1-binding protein pocket giving a large part of the affinity for a substrate or inhibitory ligand. C is assumed to bind into factor Xa's lipophilic S4 pocket and A into the very shallow S2 pocket, giving a lower increase in affinity that both B or C.

2.2 Trypsin inhibitors

In the trypsin data set at 50µM ($\mathbf{d_b} = 1$) B13 is, like in the factor Xa data, the starting material that contributes most to active reaction products with a 33.6% decrease of $\mathbf{d_b}$. This is followed by C2 (20.5%), B16 (9.9%), C7 (5.1%), B11 (5.0%), A39 (3.3%), C17 (3.2%) and A25, A1, A37, A34, C19, A11, C1, A7, A42, A33, A44, A9, A36, C5, C3, C4, A26, A12. Evaluating the impact of removing a pair of starting materials on $\mathbf{d_b}$, B13C2 (10.6%) is followed by B16C2 (5.1%), B13C7 (4.1%), A39B13 (3.3%), and with lower influence the pairs A37B13, B13, C17, B11C2, A9B13, A33B13, A34B13, B13C4, A11B13, A1C2, A44B13, B13C3, A12C2. This is again the result of two separate, the Ugi-3CR and the piperazinone, SAR relationships like for factor Xa data.

2.3 u-PA inhibitors

In the u-PA data set at 33µM ($\mathbf{d_b}$ = 1) B11 has a dominating influence (49.2%), this is followed by A44 (8.9%), C15 (7.0%), C21 (6.1%), A41 (4.7%), B13 (4.7%), A37 (4.6%), followed by C23, A34, A23, B16, C24, C12, C1, C16, A19, C20, A38, C22.

B11C15 had the highest impact among pairs of starting materials with 2.8%, followed by B11C21, A34B11, A41B11, A41B11, B11C24, A23B11, B11C12, A37B11, B16C1, B11C10, B11C16, B11C17, A1B11. This order of pair does not at all reflect the impact of formaldehyde A44 as a single component. Most likely, with formaldehyde and B11 an unrelated side product is formed that exhibits a low, but frequently occurring activity. In follow-up experiments, an alternative product was isolated and confirmed to be N-(4-amindinophenyl)glycine methylester.

2.4 Tryptase inhibitors

In the tryptase data set we can only observe a weak dependency on the nature of the amine in the following order at 80µM: B13 (3.5%), B11 (3.4%), B1 (3.3%), B3 (3.3%), B16 (3.2%) and B9 (2.1%), respectively. At 5µM ($\mathbf{d_b}$ = 1) C2 has the highest impact with 25.0%, B1 (18.0%), B3 (10.8%), B13 (7.6%), A44 (6.7%), B16 (4.9%), A37 (4.3%), C23 (3.3%), C24 (2.9%), C19 (2.5%), followed by C1, A28, A34, A14, B11, A41, A23, A36.

Evaluating the impact of removing pairs of starting materials: B1C2 (11.8%), B3C2 (7.3%), B13C2 (4.0%), B16C2 (2.3%), A44B16 (2.0%), followed by A28C2, A44C23, A37B13, A44B1, A14C2, A23C2, A37B1, A37B3, A44B3, A44B13, A36C2, A44C19, B1C24, A44C1, A34C2, A37C23, A37C24.

Clearly, these pairs of starting materials suggest a predominant effect of the piperazinone scaffold, substituted by B1, B3, B13, B16 at N-1 as potential tryptase inhibitors. This finding was also confirmed by the resynthesis of such amidinophenyl substituted piperazinones.

2.5 Chymotrypsin inhibitors

Different to the above four proteases, chymotrypsin does select for substrates that have small lipophilic P1 side chains such as alanine. This is also reflected in the IC50 data set where, contrary to the above proteases, no amidines are part of the active reaction products. Also, one of the two most active product reaction mixtures

(A19B14C1 <0.1µM; A2B15C7 <0.1µM) could give rise to at least two possible compounds where beside the expected Ugi-3CR an imidazopyrrole A19B14C1 can be formed [13].

3. Discussion

It has been demonstrated previously that the Ugi-3CR reaction provides a useful chemical scaffold for the design of serine protease inhibitors. N-substituted 2-substituted-glycine N-aryl/alkyl-amides have been identified as factor Xa, factor VIIa or thrombin inhibitors [13]. The three variable substituents of this scaffold, provided by the amine, aldehyde and isonitrile starting materials, are spanning a favorable pyramidal pharmacophoric scaffold that can fill the S1, S2 and S3 pockets of the respective protease. Thus, the 15'840 member complete combinatorial library of Ugi-3CR reaction products that been assembled allows exploring in more detail the influence of the three variable substituents of this scaffold on the structure-activity relationship versus various serine proteases.

The synthesized compound library consists of crude reaction products, thus the amount of the expected Ugi-3CR in the respective well can vary between 0 and 100%. In addition, in certain combinations of starting materials the formation of alternative scaffolds can occur, as discussed above. Only reaction products with sufficiently interesting activities were resynthezised and purified on larger scale to verify the found activities, which will be reported elsewhere. It should be noted that the use of large combinatorial compound libraries of unpurified reaction products is not anymore a preferred way of generating compound for general high-throughput screening in pharmaceutical companies. However, for certain applications this method may still be of interest, especially for target focused libraries and using biological assays that are robust enough.

A useful method to evaluate the assay results for such crude reaction products has to to deal with various errors. Thus, the uncertainty that an expected product is present in a significant amount as well as the error of biological assay itself has to be taken into account. In addition, unexpected side products may yield active compounds. However, due to the fact that we are using a systematic combinatorial library, these errors can be addressed by a statistical, systematic analysis, that can usually not be applied for small compound libraries or collections of "random" screening compounds. In this work we have explored, if the analysis of the fractal dimension by a box counting method can aid in decision making and the analysis of the obtained results.

Therefore, we calculated various fractal dimensions $\mathbf{d_b}$ for the data set of starting material combinations versus each protease at various concentrations. In general, a filled box has the fractal dimension 3, a square has a dimension of 2 and a line the dimension 1. At high concentrations more compounds are active and hence the dimension grows. If all compounds are active ($\mathbf{d_b}$ = 3), no conclusion can be made on the SAR. Thus, for high-throughput screening (HTS) it appears meaningful to select a concentration where a certain compound family shows activity and others do not show activity. While for most HTS a compound concentration of 10µM is selected, this concentration is certainly dependent on the nature of the compound library, target and assay. With the results obtained in this study we propose to set a useful HTS compound concentration where $\mathbf{d_b}$ = 1 is achieved (thus for example for tryptase, factor Xa, u-PA, and trypsin at 5, 15, 33 and 50µM, respectively).

Fractal dimensions can aid to evaluate the structure-activity relationships in combinatorial libraries of crude reaction products in various ways:

- at $\mathbf{d_b}$ = 1 it is possible to extract those starting materials / substituents that contribute most to the activity of the respective reaction products or compounds.
- $\Delta\mathbf{d_b}$ gives a measure to rank-order starting materials or pairs of starting materials and make appropriate proposals for the resynthesis of individual, pure compounds or follow-up compound libraries.
- In a well-behaved SAR, the $\Delta\mathbf{d_b}$ of individual starting materials / compounds has to be matched by finding an appropriate $\Delta\mathbf{d_b}$ for a pair of these starting materials / compounds. Otherwise it is likely that a side product with an alternative SAR or other reasons are responsible for such activities.

In conclusion, box counting allows evaluating biological data obtained from large combinatorial libraries of unpurified reaction products. The obtained results can be used to limit expensive purification and resynthesis efforts to a very limited number of statistically significant compounds. Therefore, assessing fractal dimensions by box counting can contribute to an efficient hit finding process for known or novel targets.

4. References

[1] B.B. Mandelbrot, *Fractals and Chaos: The Mandelbrot Set and Beyond*, Springer-Verlag, New York, 2004.

[2] B.B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, New York, 1977, 1982, 1983.

[3] H. Jürgens, H.-O. Peitgen, D. Saupe, *Chaos and Fractals: New Frontiers of Science*, Springer, New York, 1992.

[4] First Workshop on Fractals and Self-similarity in Data Mining: Issues and Approaches in conjunction with 8th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, July 23 - 26, 2002, Edmonton, Alberta, Canada, http://www.isi.edu/~adibi/FractalKDD02/.

[5] L.A. Kuhn, M.A. Siani, M.E. Pique, C.L. Fisher, E.D. Getzoff, J.A. Tainer, *J. Mol. Biol.* **1992**, *228*, 13-22.

[6] H. Butala, A. Sadana, *Fractal Analysis of Binding Kinetics on Biosensor Surfaces*, in *Dekker Encyclopedia of Nanoscience and Nanotechnology*, **2004**, 1191 – 1202.

[7] B. Audit, C. Vaillant, A. Arneodo, Y. D'Aubenton-Carafa, C. Thermes, *Journal of Biological Physics*, **2004**, *30*, 33-81.

[8] N. Dioguardi, F. Grizzi, P. Bossi, M. Roncalli, *Anal. Quant. Cytol. Histol.*, **1999**, *21*, 262-266.

[9] D.K. Agrafiotis, D.N. Rassokhin, *J. Chem. Info. Comput. Sci.*, **2002**, *42*, 117-122.

[10] www.modlab.de

[12] K. Illgen, S. Nerdinger, T. Fuchs, C. Friedrich, L. Weber, E. Herdtweck, *Synlett,* **2004**, *1*, 53–56.

[13] K. Groebke, L. Weber, F. Mehlin, *Synlett*, **1998**, *6*, 661-663.

[14] L. Weber, *Current Medicinal Chemistry*, **2002**, *9*, 1241-1253.

5. Figures and Legends

Figure 1. Starting materials, aldehydes, A series

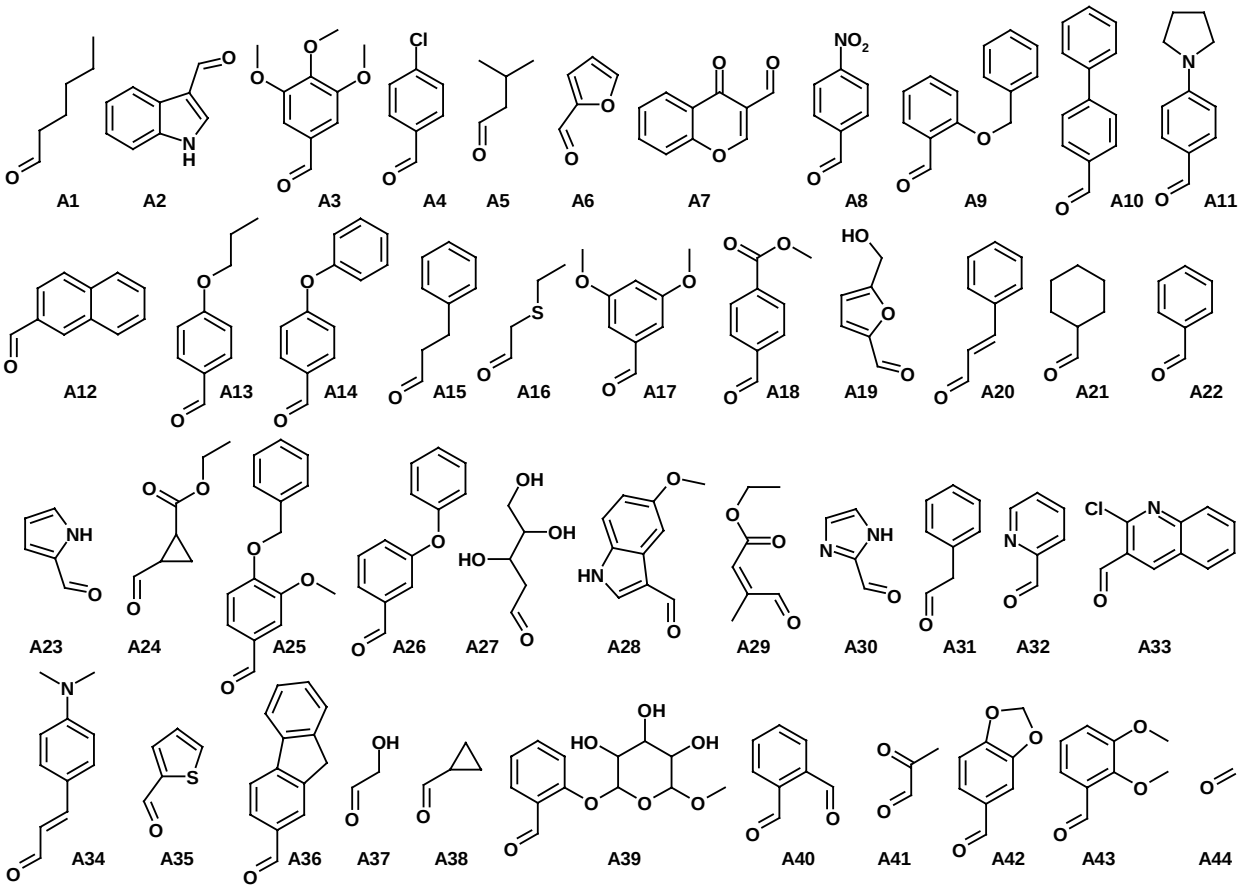

Figure 2. Starting materials, amines, B series

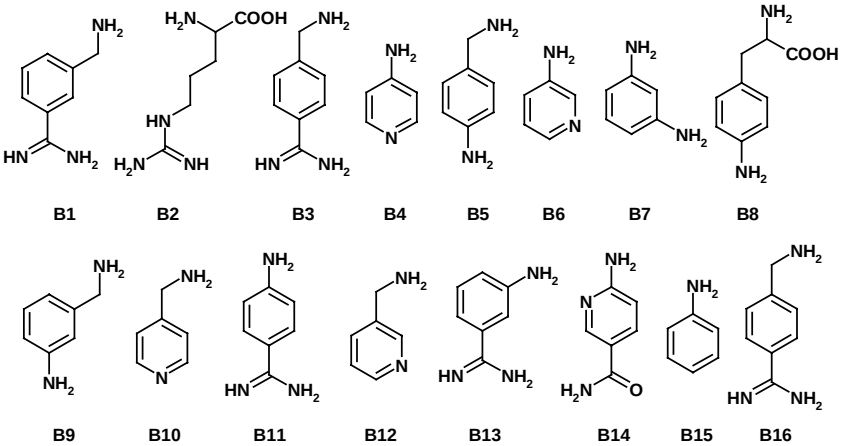

Figure 3. Starting materials, isonitriles, C series

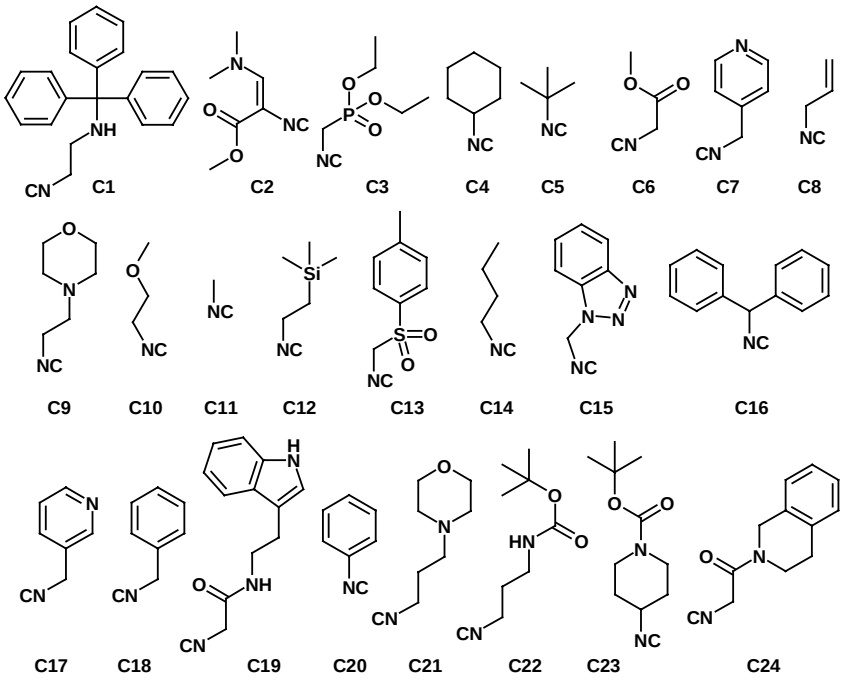

Figure 4. Fractal dimensions in relation to the IC50's of reaction products