Design of Virtual Libraries of Umami-Tasting Molecules

Martin G. Grigorov,* Hedwig Schlichtherle-Cerny, Michael Affolter, and Sunil Kochhar

Nestlé Research Center, Vers-chez-les-Blanc, P.O. Box 44, CH-1000 Lausanne 26, Switzerland

Received December 11, 2002

The introduction of molecular tools in food research offers the possibility to the food industry to benefit from the experience gained in the field by pharmaceutical companies. In this work we are showing how in silico virtual screening techniques based on molecular similarity were applied for identifying novel umami-tasting compounds. The results obtained suggest that 5′-ribonucleotides and monosodium glutamate might elicit the fifth basic taste via the same molecular mechanism. New algorithms were developed and used in this work, such as the dimension reduction of data sets by singular value decomposition and the introduction of the correlation dimension as a natural dimension of a chemical space. It is shown that the representations of molecular data sets in chemical spaces possess self-similar properties, characteristic of fractal objects.

INTRODUCTION

During this decade, drug discovery research was subject to radical changes due to the integration of novel and powerful tools of automated gene sequencing and expression, parallel and combinatorial chemical synthesis, and bioinformatics.[1] Today's drug discovery projects are aimed at efficient implementation of high-throughput screening of large combinatorial libraries of chemicals on multiple protein targets.

Currently, the majority of protein targets are G-protein-coupled receptors (GPCRs) and ion-channels, responsible for the flow of information between the cells of living organisms. It was established that GPCRs and ion channels are also involved in the processing of external photo-, mechanical-, or chemostimulations. Among the external chemical stimuli the ones leading to taste perception and olfaction are essential for living organisms. It is generally accepted that humans recognize five basic taste sensations—salty, sour, bitter, sweet, and umami.[2] Umami taste is of particular interest for the food industry. It is described by a Japanese word, which roughly translates into savory or delicious, and is defined by the taste of its prototypical stimulus, monosodium-L-glutamate (MSG). Another specific aspect of umami taste is its enhancement by the 5′-ribonucleotides inosine 5′-monophosphate (IMP) and guanosine 5′-monophosphate (GMP). In far-eastern countries natural protein hydrolysates (soy or fish sauces), with high content of free MSG, were known for centuries to enhance food palatability and flavor. MSG is currently widely used as a taste enhancer both in processed foods and in home cooking.

Recently, several G-protein-coupled receptors have been proposed to transduce the taste of glutamate. Nelson et al.[3] describe a mammalian amino acid taste receptor which functions as a broadly tuned L-amino acid sensor, with ligand affinity increasing synergistically in the presence of IMP. In a similar work Li et al.[4] indicate that in humans this same receptor has higher affinities for glutamate. According to other authors there is evidence that umami taste is mediated, at least in rodents, either by a taste-specific form of the metabotropic glutamate receptor — type 4 (taste mGluR4)[5] or by an ionotropic glutamate receptor (iGluR) variant.[6]

To perform a computer-assisted identification of new umami-tasting compounds, possibly ligands of these receptors, in silico structural biology techniques could not be used, due to the lack of target protein structures. Therefore, the implementation of in silico screening should be based on molecular similarity techniques.

MATERIALS AND METHODS

Molecular Data Set. Amino acids and small peptides hold a central place in biology. A long evolution made the 20 naturally occurring amino acids become the building blocks for many small ligands of a wide variety of receptors. The molecular diversity offered by a library of peptides with limited molecular weights is thus very large in terms of possible biological activities. Any molecular representation of such a library in which some reference compounds are found to form well-defined clusters is therefore of highest interest. Virtual screening can be implemented in such a space with satisfactory resolution among classes of different activities. Along this line of thought we decided to build a reference virtual library of 8420 members containing all the amino acids and all the di- and tripeptides and to investigate the clustering of some selected molecules, depending on the descriptors used. Such a reference library is particularly relevant in the context of umami taste as few dozens of di- and tripeptides were described in the literature to elicit nearly the same taste as the prototypical monosodium salt of glutamic acid (MSG).[7−12] We have considered in our investigation all of the peptides reported at least once in the literature to evoke this taste. We have included in a training set monosodium glutamate and the umami-tasting tripeptides listed in Table 1. In difference, 5′ ribonucleotides IMP and GMP as well as the umami-tasting dipeptides in Table 1 were kept in a test set as query molecules for validating the approach.

* Corresponding author phone: +41 21 785 89 39; fax: +41 21 785 85 49; e-mail: martin.grigorov@rdls.nestle.com.

10.1021/ci025652g CCC: $25.00
Published on Web 05/16/2003

Table 1. Dipeptides and Tripeptides Reported in the Literature To Evoke the Umami Taste

no.	peptide	reference
1	Asp-Asp	Tamura et al.
2	Asp-Glu	Tamura et al.
3	Glu-Asp	Arai et al.
4	Glu-Glu	Arai et al.
5	Glu-Leu	Arai et al.
6	Glu-Lys	Tamura et al.
7	Glu-Ser	Arai et al.
8	Glu-Thr	Arai et al.
9	Lys-Gly	Tamura et al.
10	Thr-Glu	Nogushi et al.
11	Ala-Asp-Ala	Ohyama et al.
12	Ala-Glu-Ala	Ohyama et al.
13	Asp-Glu-Leu	Frérot et al.
14	Asp-Glu-Ser	Noguchi et al.
15	Glu-Asp-Glu	Noguchi et al.
16	Glu-Asp-Phe	Firmenich
17	Glu-Asp-Val	Noguchi et al.
18	Glu-Glu-Glu	Noguchi et al.
19	Glu-Glu-Ile	Frérot et al.
20	Glu-Glu-Leu	Frérot et al.
21	Glu-Gly-Ala	Noguchi et al.
22	Glu-Gly-Ser	Noguchi et al.
23	Glu-Leu-Glu	Frérot et al.
24	Gly-Asp-Gly	Ohyama et al.
25	Gly-Glu-Gly	Ohyama et al.
26	Ile-Glu-Glu	Noguchi et al.
27	Leu-Asp-Leu	Ohyama et al.
28	Leu-Glu-Glu	Frérot et al.
29	Ser-Glu-Glu	Noguchi et al.
30	Val-Asp-Val	Ohyama et al.
31	Val-Glu-Val	Ohyama et al.

Although it was clearly established that only L-stereoisomers of all of the peptides are evoking umami taste, in our study all molecular structures were considered as flat graphs. This intrinsic limitation of molecular similarity techniques using 2D descriptors was expected to lead to a certain amount of false positive hits in the virtual screening. Nevertheless, in silico screening is aimed at enriching subsets of molecular libraries in positive hits in comparison to random selections in the same libraries, and therefore a relatively high rate of false positives could be accepted. This is compensated by a low-cost and fast way to evaluate huge numbers of candidates. Such a compromise can be tolerated especially in discovery projects where only limited structure–activity data is available.

Another limitation in our work was to consider only the nonionized forms of MSG and all of the peptides. In reality, it was found out that only one dissociation form of MSG elicits umami taste.[11] This form dominates the proportion of MSG species existing in aqueous solution at pH 6 and results from the negative ionization of the two carboxylic groups and the positive ionization of the α amino group. We relied on the very nature of the molecular descriptors being used to take implicitly into account the ability of the structures to accept or to donate protons, especially in the case when 2D triangular fingerprints were used.

Representations of Molecules. The experience of researchers in the area of medicinal chemistry led to the formulation of the similarity "hypothesis" by Maggiora and Johnson.[13] Indeed, there is evidence that "similar" molecules generally induce similar biological effects. The notion of similarity in biological activities is well defined as it is related to the experimentally measurable effect of a molecule on an artificial bioassay or on a living organism. In difference, the concept of molecular similarity is much more controversial and emerged in recent years in relation to the advent of combinatorial chemical synthesis techniques. The central problem resides in the definition of a chemical space in which molecular similarity measures will be highly correlated with bioactivity. Certainly, the concept of similarity of a set of molecules is meaningful only with respect to a specified representation of the molecules, that is why the controversy about molecular similarity should find its solution in the theory of molecular representations.[14] Currently, there is a lack of a universal molecular representation correlated with biological activity. Rather, it seems that the particular molecular representation to be used is strongly dependent on the problem to be solved.

A variety of computable molecular properties have been described in the literature,[15] which can be divided in three main groups. First, the 1D ("one-dimensional") descriptors contain information about the entire molecule, such as the octanol–water partition coefficient logP, or the molecular weight (MW). Second, the topological (2D—"two-dimensional") descriptors reflect only the neighborhood relationships of atoms in a molecule. Third, 3D—"three-dimensional" descriptors take into account the spatial arrangement of atoms. Although 3D descriptors are based on a physically more realistic molecular representation, several validation studies indicated that similarity selection and design of virtual libraries is efficient enough when using only 1D and 2D descriptors.[16−20] The 1D and 2D descriptors have moreover the advantage to be computable faster than 3D descriptors.

For the purposes of our work we have developed a software able to compute four different sets of descriptors starting with the molecular connection table, coded by a SMILES string.[21] These descriptors are Ghose-Cripen atom types,[22] several Kier-Hall topology descriptors,[23] HOSE-codes descriptors of chemical environment,[24] and 2D triangular fingerprints.[25] In our work we have used the original 120 Ghose-Cripen atom types. Additionally, we calculated Kier-Hall descriptors of path, cluster, path/cluster, and chain subgraph type up to the order 10 as well as Kier's kappa descriptors. The HOSE strings coded for the chemical environment of a central atom, surrounded by up to four successive atomic shells, by inspecting the topologically closest neighbors first. According to predefined construction rules it is possible to build an exhaustive list of all possible HOSE strings. However, to be able to represent an arbitrary set of molecules by a reasonably restricted list of relevant codes, we generated only the HOSE codes for the reference library composed out of all naturally occurring amino acids and all di- and tripeptides. We ended up with a list of 1497 unique codes, which we further ranked lexicographically in a way that we were able to assign an integer index k to each entry in this list. Consequently, to represent an arbitrary molecular structure, we used the counts of every unique HOSE code from the reference list in a given molecule. When using 2D triangular fingerprints we proceeded exactly as we did for identifying the unique HOSE codes—only 1848 unique fingerprints were found in the set of 8420 amino acids and di- and tripeptides. Consequently, we systematically ordered these unique fingerprints in a table in a way that an index k could unequivocally be associated to each of them.

Thus, we were able to represent any molecule of interest by the counts of every unique fingerprint from the reference list in this same molecule.

Data Reduction Algorithm. *Generalities.* In the previous paragraph we have shown that a set of m molecules can be represented in terms of molecular descriptors by a molecule-by-descriptor matrix $\boldsymbol{M}$ composed of m row vectors

$$\boldsymbol{M} \equiv \{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_m\}^T \qquad (3.1)$$

where each row vector $\boldsymbol{X}_i \equiv \{x_{i1}, x_{i2}, ..., x_{in}\}$ consists of n descriptors $\{x_{ij}\}$ associated to that particular molecule. From a geometrical point of view, each vector $\boldsymbol{X}_i$ defines a point in an n-dimensional vector space, so that a point in this space can represent every molecule. In many applications it is often highly informative to assess molecular neighborhoods visually. Therefore computational techniques allowing one to map the original n-dimensional vector space to a low-dimensional one, preferably two- or three-dimensional, are of practical interest. Even when it is impossible to obtain the mapping in two or three dimensions, any dimension reduction would lead to a significant decrease of computational costs for molecular similarity assessment. In the recent literature, several such methods have been discussed, and the most widely used is principal component analysis (PCA). Recently, singular value decomposition (SVD), a technique commonly used in linear algebra for matrix factorization,[26] attracted the interest of researchers in chemoinformatics. In the first publication,[27] SVD coupled with the minimization package TNPACK has been used as a projection protocol in the analysis of chemical databases. It was found that in several cases SVD projection preserved the original pairwise distances sufficiently well to notice only a marginal improvement by further applying the minimization procedure. Another paper appeared recently, describing a method for computing chemical similarity[28] inspired largely by a patented document retrieval method termed as latent semantic indexing (LSI).[29] Latent semantic indexing relies essentially on singular value decomposition to overcome the synonymy problem or the equivalent problem of correlated descriptors in chemoinformatics. Latent semantic indexing is even more useful as it generates through SVD a low-dimensional projection of the original data. This projection can be formulated as the transformation of the original rank r molecule-by-descriptor matrix $\boldsymbol{M}^{mxn}$, $m \geq n$ to the canonical form

$$\boldsymbol{M} = \boldsymbol{U}.\boldsymbol{D}.\boldsymbol{V}^T \qquad (3.2)$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, that is $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}^m$ and $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}^n$, and $\boldsymbol{D} = diag\ (\sigma_1, ..., \sigma_n)$, $\sigma_i > 0$, for $1 \leq i \leq r$ and $\sigma_j = 0$ for $j > r + 1$. The columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, are referred to as the *left- and right-singular vectors*, respectively, and the singular values of $\boldsymbol{M}$ are defined as the diagonal elements of $\boldsymbol{D}$ which are the nonnegative square roots of the n eigenvalues of $\boldsymbol{M}\boldsymbol{M}^T$. The important dyadic decomposition of $\boldsymbol{M}$ is given by

$$\boldsymbol{M} = \sum_{r}^{l=1} \sigma_{\boldsymbol{l}} \cdot u_l \cdot v_l^{\,T} \qquad (3.3)$$

Singular value decomposition is sometimes called *Eckart-Young decomposition* after the two authors who published in 1936 a paper entitled: *"The approximation of one matrix by another of lower rank"*.[30] The title of this work explains probably at best the true meaning of SVD. By using the dyadic decomposition (3.3), Eckart and Young have shown that among all rank-k matrices $\boldsymbol{N}$ the matrix $\boldsymbol{M}_k$, constructed from the first k singular triplets of $\boldsymbol{M}$, is the closest one to $\boldsymbol{M}$ in *the least-squares sense*:

$$\min_{rank(N)=k} ||\boldsymbol{M} - \boldsymbol{N}||_F^{\,2} = ||\boldsymbol{M} - \boldsymbol{M}_k||_F^{\,2} = \sigma_{k+1}^{\ \ 2} + ... + \sigma_n^{\ 2} \qquad (3.4)$$

Somewhat similar results were obtained as early as 1907 by Schmidt[31] and later revisited by Mirsky;[32] therefore, SVD is also called the Schmidt-Mirsky transform. The property (*3.4*) is the one which led to a wide use of SVD as a general procedure for data compression in diverse areas ranging from image processing[33] to attractor reconstruction in the theory of dynamical systems.[34]

An arbitrary molecule, coded by a set of descriptors, can be represented as a molecule-by-descriptor matrix $\boldsymbol{\mu}^{1xn}$

$$\boldsymbol{\mu} \equiv \{x_{11}, x_{12}, ... , x_{1n}\} \qquad (3.5)$$

and can be projected in the space spanned by the first k left singular vectors, taken in rank order

$$\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}^T U_k D_k^{\ -1} \quad k = 1, ... ,m \qquad (3.6)$$

where the singular values appearing on the diagonal of the matrix $\boldsymbol{D} = diag\ (\sigma_1, ...\sigma_n)$, $\sigma_i > 0$ are playing the role of scaling factors. Equation 3.6 demonstrates how molecular queries can be formulated and projected in the transformed space. After such a processing is carried out, similarity assessment between the query molecule and the structures in the database can be conducted by a variety of methods. The most frequently used similarity measures are the Euclidean distance, which we used in our work, or the normalized dot product. The latter turns out to be the well-known cosine similarity in the case when the singular values are omitted as scaling factors in eq 3.6.[28] In our work similarities deduced from Euclidean distances clearly outperformed the ones obtained by the dot product method.

Sparse Matrix Factorization with SVD. The classical SVD decomposition technique, that we have just described, is not very demanding in computational resources. However, a problem could arise when molecules are represented by counts of a large number of features. In this case SVD becomes prohibitively expensive in memory storage. A way to circumvent this difficulty is to notice that most of the associated molecule-by-descriptor matrices are with a very sparse structure, i.e., most of their elements are zeros. For instance the molecule-by-descriptor matrix based on the HOSE codes description of the library of 8420 amino acids and di- and tripeptides has a total of 12′574′800 (8420 × 1497) entries, 93% of which are zeros. Similarly the molecule-by-descriptor matrix based on the 2D triangular fingerprints has a total number of 15′561′160 (8420 × 1848) elements of which 82% are zeros. In such situations it is much more efficient to apply iterative Krylov-subspace-based SVD algorithms,[35] such as the implicitly restarted Arnoldi algorithm,[36] and to compute only a limited part of the singular spectrum, restricted to the few largest eigenvalues.

Separating Signal from Noise by SVD. It is generally recognized that with varying the number of singular triplets regarded as significant one can control the separation of signal from noise when applying SVD. In previous applications of the method to textual documents processing[29] or chemoinformatics[28] it was established that the choice of the number of singular triplets *k* influences the level of fuzziness of searching. Larger values of *k* produce better approximations of the original data with the drawback of taking into account irrelevant features (noise), whereas smaller values of *k* lead not only to improved generalization but also to loss of resolution. How much of the spectrum should be computed is therefore a controversial issue. From a purely theoretical point of view, based on the decomposition given by eq 3.4, it seems clear that the best rank-*k* approximation of the original matrix should be based on the first *k* nonzero singular values. However, in many practical applications the criterion for separating useful information from noise by considering only the nonzero singular values is not readily applicable. In fact, real-world singular spectra exhibit few large singular values followed by an extended flat region of much smaller and slowly decreasing nonzero ones. In this situation it was proposed to take into account only the singular values not smaller than a certain ratio of the largest one. The more realistic Monte Carlo SVD method consists of the generation of singular spectra for a large number of synthetic random matrices of the same size as the original one and to compare these spectra to the spectrum of the real data.[38] Only the part of the spectrum arising above the bundle of synthetic ones should be considered as relevant and noise-free. Although this last methodology is statistically sound it could not be applied to very large matrices. The reason is that SVD factorization of the real data alone stands as a computational challenge. In the next paragraph we propose a fast automatic method able to select the relevant singular values when applying SVD to an arbitrary data set. Such an algorithm is of particular interest for the proper balance between fuzziness and resolution when applying the technique to real data sets.

Numerical Subroutines for SVD. We have used the SVD routine included in the Silicon Graphics mathematical library for transforming the relatively small descriptor-by-molecule matrices based on Ghose-Crippen atom types and Kier-Hall topological indices. In the case of the larger matrices based on HOSE codes and 2D triangular fingerprints we have calculated only the first 15 singular triplets by applying the sparse SVD routines included in the library ARPACK, based on the iterative implicitly restarted Arnoldi algorithm. For improving performance when screening a huge number of candidates, we have used the parallel version of the library, ported to our Linux cluster computing facility.[37]

Fractal Objects and Their Characterization. There is an emerging evidence that real data sets are statistically self-similar and, thus, fractal.[39,40] This is an important insight since it allows a compact statistical description of these data sets. It is our aim to demonstrate in this paragraph that a parallel can be drawn between the theory of fractals and the design and characterization of chemical libraries, by considering some statistical measures of fractal data distributions.

Let us turn now to our problem where we have a projection of a compound collection of *m* molecules in the space spanned by the left singular vectors, associated to an approximation of the original descriptor matrix *M* at rank *k*. Further let us assume that a grid is used to partition this *k*-dimensional space in cells of size *r*, so that we can compute the frequency p_i with which data points fall into the *i*th cell. Using these frequencies we can further compute D_q, the generalized fractal dimension of the *q*th order[41]

$$D_q(r) = \lim_{r \to 0} \frac{I_q}{log1/r} = \frac{\partial I_q}{\partial\ log\ r} \tag{4.1}$$

where I_q is the Renyi information of the *q*th order

$$I_q(r) = \frac{1}{1-q} \cdot log \sum_{i=1}^{N(r)} p_i^q \tag{4.2}$$

where $N(r)$ is the total number of cells occupied by points of the data set. Clearly, eqs 4.1 and 4.2 express the important property of self-similar fractal sets, namely that the change of their *q*th order statistical moments across different scales remains invariant. Among the infinity of fractal dimensions defined by eq 4.1, the *Hausdorff dimension* ($q = 0$), the *information dimension* ($\lim_{q \to 1} D_q$), and the *correlation dimension* ($q = 2$) are widely used for characterizing the information content of data sets.[42] It is now well established that the existence of a finite, possibly noninteger Hausdorff information or correlation dimensions is indicative of the self-similar, scale-invariant, fractal nature of a given data set.

In practice, the most important fractal dimension, the correlation dimension, is computed by a method due to Grassberger and Procaccia.[43] This method relies on “sphere counting” in difference to the slower classical “box-counting” algorithm derived from the definitions (4.1) and (4.2). In the scheme of Grassberger and Procaccia, the correlation integral $C_k(r)$ of the data set is computed as a function of the *correlation length r*

$$C_k(r) = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j=i+1}^{k} \theta(\mathrm{r}\ \text{-}||\boldsymbol{u}_i - \boldsymbol{u}_j||) \tag{4.3}$$

where $\theta(\boldsymbol{x})$ is the Heavyside function, $\theta(\boldsymbol{x}) = 0$ if $\boldsymbol{x} < 0$ and $\theta(\boldsymbol{x}) = 1$ if $\boldsymbol{x} \geq 0$. Here the correlation length *r* has a meaning of a radius of the neighborhood around the phase space point $\boldsymbol{u}_i$, while $||\boldsymbol{u}_i - \boldsymbol{u}_j||$ denotes the distance between any pair of points *i*, *j* in the space spanned by the first *k* left singular vectors. It is possible to use different norms to evaluate this distance. We have used the Euclidean norm. From the definition it can be seen that the correlation integral is the cumulative probability distribution of pairs of molecules in the SVD-transformed descriptor space. Geometrically it reflects the intrinsic compactness or expansiveness of the representation of a given chemical library. Therefore it can be used as an intrinsic measure of diversity in a molecular data set.

Grassberger and Procaccia proved that at sufficiently small correlation lengths, in the so-called *scaling region*[43]

$$C_k(r) \sim r^{\,d} \tag{4.4}$$

where *d* is termed the *correlation exponent*. The correlation exponent is calculated as the slope of the plot of ***log*** $C_k(r)$ versus ***log*** *r*. It has been found out that the correlation exponent *d* of self-similar data sets reaches a saturation value (levels off) and stays approximately constant when the dimension of the phase space is increasing.[43] In contrast, when the data set is purely random, the correlation exponent increases in a monotone way every time a dimension is added. The saturated correlation exponent is an estimation of the *correlation dimension* of the data set as defined in eq 4.1. It is assumed that it equals the number of independent variables *df* (degrees of freedom), needed to characterize a given data set. In most cases the correlation dimension reveals to be an noninteger number, and therefore different alternatives exist about approximating at best the number of independent variables *df*. In the case of dynamical systems, Takens[44] demonstrated that $df = 2d + 1$, but it is generally accepted that there is no need to take *df* so large, and it is sufficient to consider the closest integer greater than the correlation dimension *d*.[45]

RESULTS AND DISCUSSIONS

Searching for an Optimal Chemical Space. The molecule-by-descriptor matrices were computed for the reference virtual library composed of amino acids and peptides by using the four molecular representations discussed previously, i.e., the ones based on counts of Ghose-Crippen atom types, on Kier-Hall topological descriptors, on counts of HOSE codes for atomic chemical environments, and finally on counts of 2D triangular fingerprints. The resulting matrices were then submitted to singular value decomposition. In Figure 1a we have depicted the evolution of the singular spectra when increasing the rank *k* of the SVD-based approximations associated with the four molecular representations under scrutiny. It can be deduced that singular spectra allowed reliable separation of signal and noise only in the case when Kier-Hall indices were used. Indeed, the first singular value in this case is three orders of magnitude greater than the singular values falling in the flat part of the spectrum. In what concerns the molecular coding with Ghose-Crippen atom types it is clear that the original matrix could not be compressed accurately to low dimensions as the corresponding singular spectrum is decreasing in a monotone way beyond rank 10. In the cases of the molecular representations based on HOSE codes and 2D triangular fingerprints the singular spectra are relatively flat so that the signal-to-noise separation is not sufficiently sharp. To gain further insights on the last two difficult cases, we investigated whether saturation occurs in the correlation exponents when increasing the dimensionality in the respective column spaces, spanned by the left singular vectors. This is illustrated in Figure 1b, depicting the evolution of the correlation exponents by increasing the number of singular triplets, taken in rank order. The data shown supports our conclusion that it is difficult to compress the matrix using Ghose-Crippen atom types descriptors. The reason for this is currently investigated and will be communicated elsewhere. In the three remaining cases of coding, the correlation exponent reaches a clear saturation and thus gives evidence to the self-similar, fractal nature of these representations of the reference library. Interestingly, the estimated correlation dimensions are all nearly equal to 2.6, independently of the molecular descriptors being used. Applying the Takens rule ($df = 2 \times d + 1 = 2 \times 2.6 + 1 \sim 6$) we concluded that an upper bound of six independent variables is needed to describe this library. On the other hand, according to the method of Fraedrich,[45] only three variables would suffice for this. We checked numerically that adding more than three degrees of freedom is not altering significantly interpoint separations, and we decided to adopt the three-dimensional (3D) chemical space as the one in which the data set unfolds optimally.

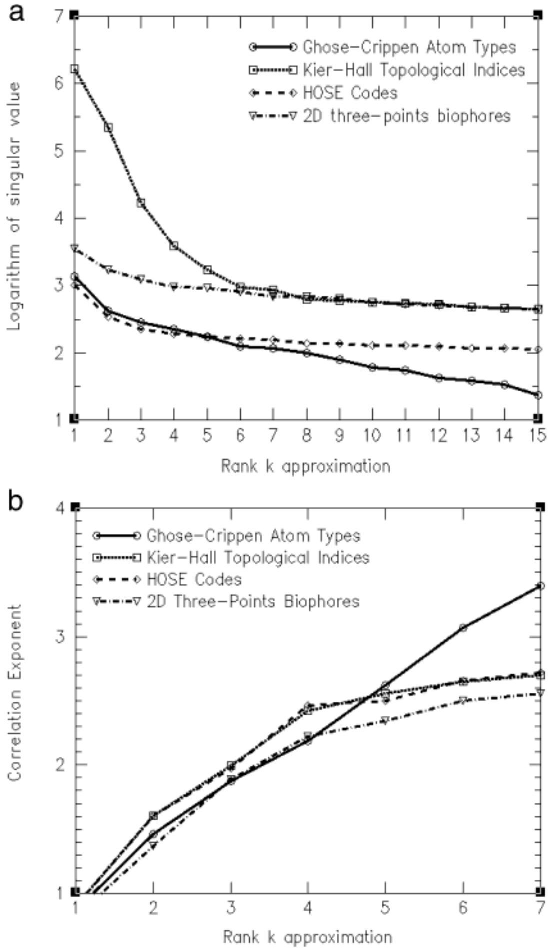

Figure 1. Evolution of singular values (a) and correlation exponents (b) versus the rank *k* of SVD-transformed chemical spaces. The figure was generated by using Xmgr data plotting software.[51]

We further investigated in which one of the four types of molecular representations the region occupied by the training-set compounds had the most compact shape in comparison to the overall size of the reference library composed of amino acids and peptides. We performed this to find out that in one of the cases the umami-tasting peptides were clustered in a relatively restricted volume in comparison to the space occupied by the whole reference library. In this way we were able to set up a simple virtual screen consisting in picking as a highly ranked candidate every query molecule projected in the volume occupied by the training-set compounds. In Figure 2 we have represented the projections of the reference virtual library in the 3D space spanned by the first three left singular vectors. These projections result from coding based on Ghose-Crippen atom types (Figure 2a), Kier-Hall topological descriptors (Figure 2b), HOSE codes (Figure 2c), and 2D triangular fingerprints (Figure 2d). Clearly, the training-set compounds, denoted by crosses, are significantly clustered when using the 2D triangular fingerprints, while the grouping is poor when applying the remaining three molecular representations.

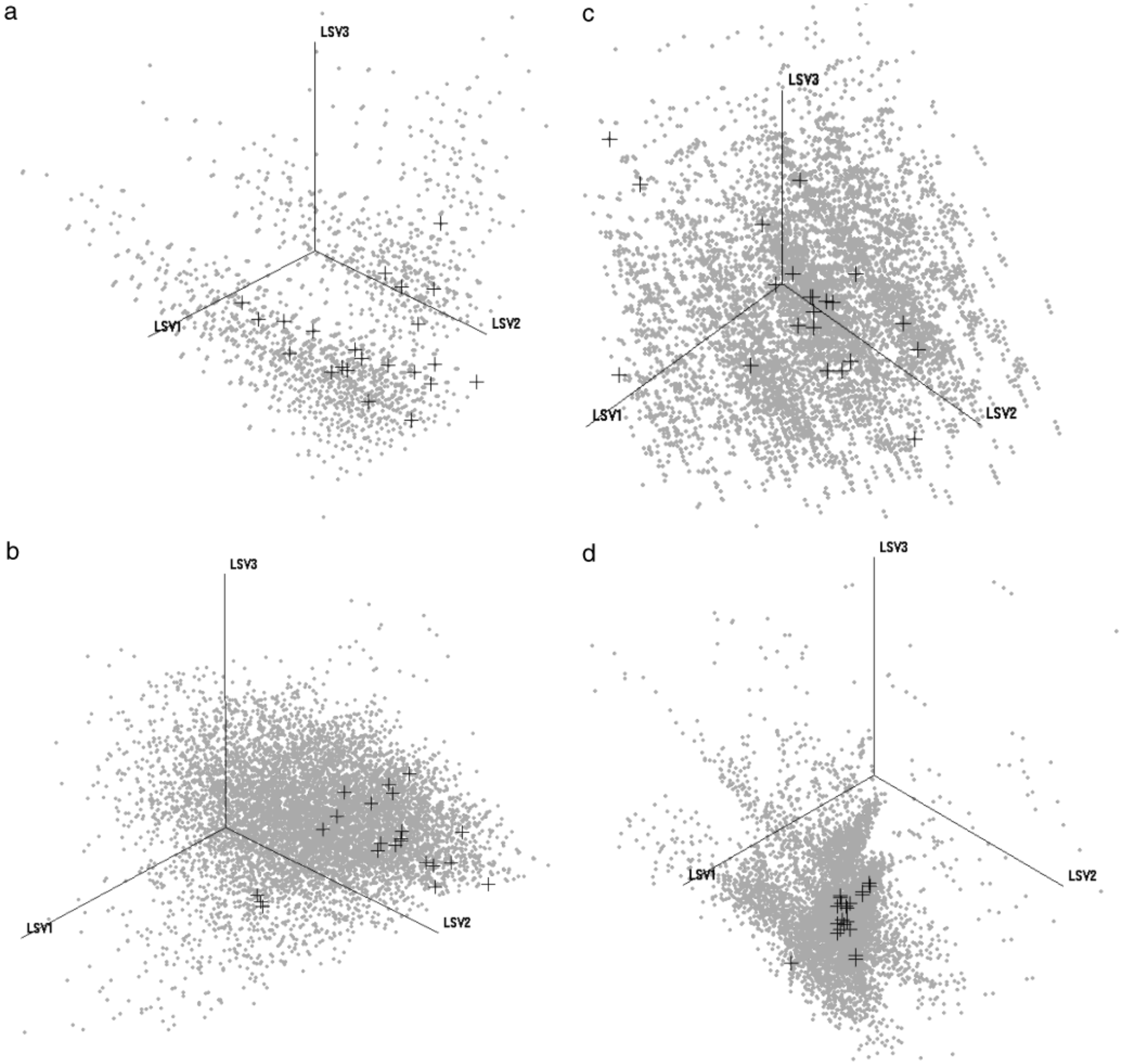

Figure 2. Projections of the reference peptide virtual library in a 3D space spanned by the first three left singular vectors (LSV) based on counts of Ghose-Crippen atom types (a), Kier-Hall topological descriptors (b), counts of HOSE codes (c), and counts of 2D triangular fingerprints (d). Points represent all of the amino acids and all di- and tripeptides, whereas crosses denote MSG and the training-set umami-tasting tripeptides (cf. Table 1). The figure was generated with Xgobi data visualization software.[52]

Outlier Identification. The grouping of the training-set compounds is even more marked if some of the umami-tasting peptides are labeled as outliers. Indeed one can easily identify in Figure 2d three distant clusters each containing two members as well as the single tripeptide Glu-Asp-Phe (16). The tripeptides Glu-Asp-Glu (15) and Glu-Glu-Glu (18) form the cluster the most distant from the main group of umami-tasting training-set compounds. A more closely situated cluster contains the tripeptides Asp-Glu-Ser (14) and Ser-Glu-Glu (29), while an equally distant doublet contains the tripeptides Val-Asp-Val (30) and Val-Glu-Val (31). One of the outliers was confirmed inactive by a dedicated panel of sensory specialists.

To quantitatively characterize the clusters after outliers were removed, we computed first the volume delimited by the convex hull spanned by the representative points of the training-set compounds. Consequently, we calculated the volume of the convex hull of the whole reference library composed of amino acids and di- and tripeptides. The convex hull of a set of points is the smallest convex set that includes these points, where in two dimensions the convex hull reduces to a convex polygon. We used the software qhull[46] to compute the related convex hulls in 3D spaces. We further estimated the ratios of the convex hull volumes, computed by the software, to find out that in the 3D chemical space based on 2D triangular fingerprints the training-set compounds occupy only 0.4% of the volume of the whole reference library (when excluding the outliers from the analysis). In difference, the respective ratios rose above 17% when using the remaining three molecular codings.

In Figure 3 we have provided a magnified view of the region occupied by the training-set compounds as it appears

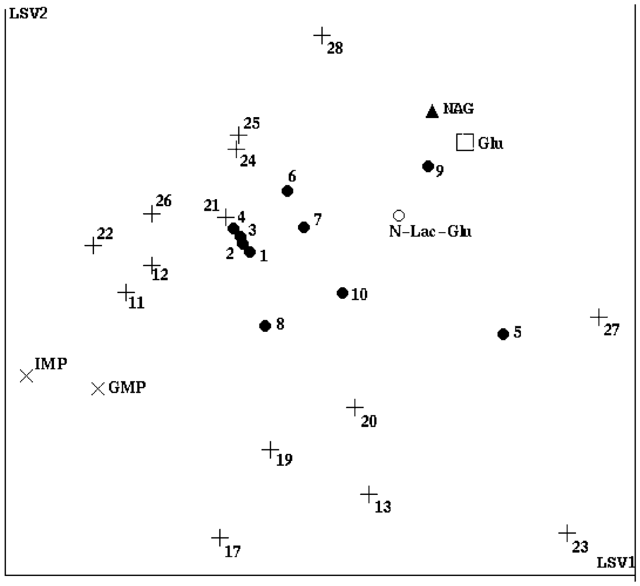

Figure 3. Magnified overview of the region occupied by the training-set molecules in the optimal 3D chemical space based on counts of 2D triangular fingerprints. Dark filled circles (●) represent umami-tasting dipeptides (cf. Table 1), and crosses (+) represent umami-tasting tripeptides (cf. Table 1), whereas IMP and GMP are denoted by (×) signs. MSG is symbolized by a square (□), while *N*-lactoylglutamate is represented by a small circle (○). *N*-Acetylglycine is depicted with a filled triangle (▲). The figure was generated with Xgobi data visualization software.[52]

in Figure 2c. We already stated the fact that the optimal unfolding of the whole data set (amino acids and di- and tripeptides) occurs in three dimensions. In difference with Figure 2c however, in Figure 3 the representation is in the two-dimensional space spanned by the first two left singular vectors. The third dimension was omitted, as it does not add any significant amount of information on neighborhood relationships. To attain a sufficient resolution in Figure 3, we have displayed glutamic acid and only 14 (out of the 21) tripeptides appearing in Table 1, thus excluding the seven tripeptides that we have previously labeled as outliers. Tripeptides containing at their C-terminus a highly hydrophobic residue (13, 17, 19, 20, and 27) tend to cluster in the lower right corner of Figure 3, whereas the ones with a more polar residue at this position are grouped in the upper left one.

Validation. Once we found a molecular representation in which the reference compounds were clustered, we proceeded with a three-stage validation study of the approach. For this, three molecular queries were projected in the target 3D chemical space.

The first query was composed out of all dipeptides listed in Table 1. They were found to map among the umami-tasting tripeptides but somewhat closer to glutamic acid. Highly similar dipeptides such as Asp-Asp (1), Asp-Glu (2), Glu-Asp (3), and Glu-Glu (4) form a characteristic cluster appearing almost in the middle of Figure 3. In their vicinity maps the tripeptide Glu-Gly-Ala (21). The dipeptide that maps closest to MSG is Lys-Gly (9).

As a second query we choose the amide of glutamic acid and lactic acid (N-Lac-Glu). This amide was a relevant test as it was recently described in a patent[47] and a publication[12] to evoke umami taste. The amide projects also in the vicinity of MSG, thus confirming model's validity.

Finally, a third query contained the 5′ ribonucleotides IMP and GMP, which are structurally quite different from MSG, though eliciting the same taste. Interestingly, these molecules map also in the vicinity of the glutamate-like tasting peptides and MSG indicating that 5′ ribonucleotides could elicit umami taste through interaction with the same binding site to which umami-tasting compounds do attach. To our knowledge this is the first time that an indication is provided that ribonucleotides, peptides, and MSG may evoke the fifth basic taste via a common molecular mechanism.

Identification of an Umami-Taste Sapophore. In analogy to the concept of pharmacophore, the term sapophore was used to designate the minimal number of chemical features needed to be present on a molecule, in order for this one to elicit a particular taste, which in our case is umami taste.[48] To gain further insights in the arrangement of the relevant chemical features necessary for activity, we analyzed the information provided by the right singular vectors. Right singular vectors can be thought of as generalized descriptors. Indeed, each right singular vector is represented by the weighted sum of the original descriptors, i.e., the frequency of occurrence of each of the 1848 2D triangular fingerprints in every molecule of the analyzed reference library. The linear coefficients (weights) contained in the most important first right singular vector can be viewed as the relative frequencies with which the original descriptors are present on average in the data set. To give evidence to this fact, we have first calculated the cumulated frequencies of occurrence of the 1848 2D triangular fingerprints among the 8420 members of the peptide reference library. This was achieved by summing column by column all elements in the molecule-by-descriptor matrix based on counts of 2D triangular fingerprints. The cumulated frequencies were then normalized by the number of rows in the same matrix. We have depicted these frequencies in Figure 4a, whereas in Figure 4b we have shown the coefficients of the first right singular vector. It is interesting to notice that the coefficients of the first right singular vector appear like the filtered and noise-cleaned cumulated-frequencies spectrum shown in Figure 4a. The dominant peaks in the two figures lie at exactly the same positions, corresponding to 2D triangular fingerprints largely over-represented in the molecular data set. However, a small difference exists between parts (a) and (b) of Figure 4 in the region of 2D triangular fingerprints of indices higher than 1600. This can be explained by the fact that the spectrum of frequencies in Figure 4a reflects a cumulative statistics on the data set, whereas the first right singular vector in Figure 4b presents an SVD-averaged statistics on this same data set. The right singular vectors associated to the first few statistically significant singular triplets obtained by SVD are therefore a valuable data analysis and data cleaning tool, complementing commonly used statistical techniques.

At this point it was certainly of interest to find the precise 2D triangular fingerprints characterizing the umami-tasting compounds only and differentiating them from the members of the reference peptide library. We established that the dominant peak appearing in Figure 4a,b represents a 2D triangular fingerprint in which an acceptor site is distant by two to three bonds from an acceptor/donor site itself distant by four to five bonds from a hydrophobic site. The hydrophobic and acceptor sites are themselves separated by four to five bonds. The precise mapping of this 2D triangular

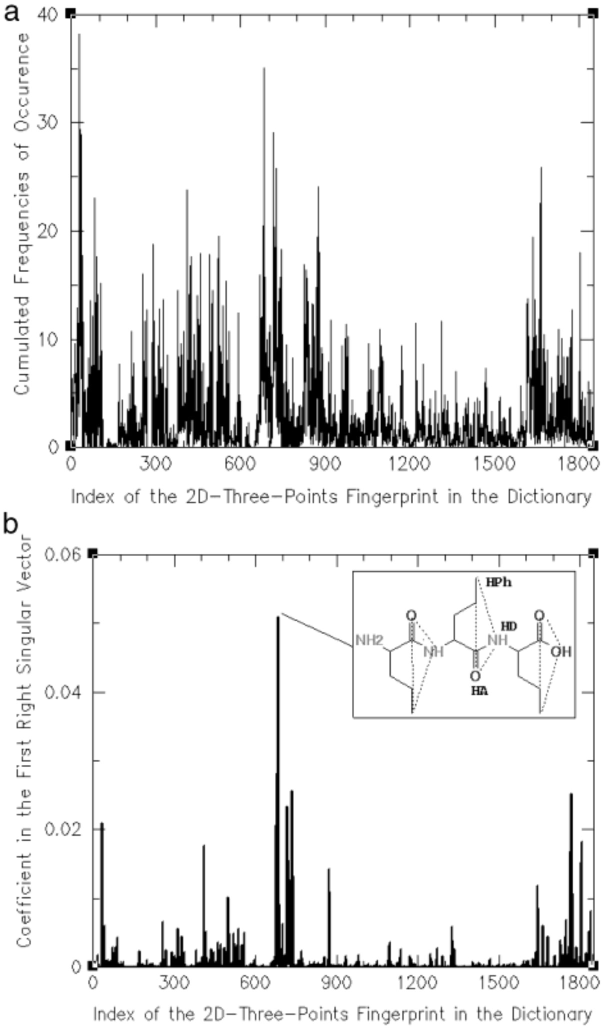

Figure 4. Statistics on the occurrence of 2D triangular fingerprints in the reference virtual peptide library. The cumulated frequencies of the 1848 2D triangular fingerprints are depicted in (a), whereas the coefficients of the first right singular vector are shown in (b). The figure was generated with Xmgr data plotting software.[51]

fingerprint is shown overlaid in Figure 4b on a typical tripeptide atomic framework, which is by far the dominant one in the reference library we have used. On this diagram we have depicted peptide side-chains as *n*-propyl residues because the mean topological length of the side chain among the 20 natural amino acids is nearly 3.5 bonds. This fact provides an explanation of the high incidence of precisely this fingerprint in the analyzed data set.

We consequently produced a smaller data set containing only the umami-tasting molecules, i.e., the 10 dipeptides, the 14 tripeptides (seven outliers were excluded), IMP and GMP, and of course MSG. We have submitted the associated molecule-by-descriptor matrix to singular value decomposition and in Figure 5 we have depicted the resulting first right singular vector. Comparing Figure 5 with Figure 4a,b shows that the peptide-specific peak is missing in Figure 5, but several new peaks appeared instead. In Figure 5 we have labeled them 5.I, 5.II, and 5.III, respectively. The first peak 5.I, appearing at index 858, is associated with a 2D triangular fingerprint coding for an acceptor atom distant at two to three bonds from a bulky atom, which is itself distant at four to five bonds from a second acceptor atom. The first and second acceptor sites are separated themselves by six to seven bonds. To find out how this fingerprint maps on typical umami-tasting molecules we took as examples IMP and MSG, lying at the two extremes in Figure 3. The respective projections are shown in Figure 6a where in bold we have delineated the glutamate-like framework defined by the mapping of this fingerprint on IMP. The same framework is emerging once more when we turn to investigate how another two characteristic fingerprints map on the same two compounds. These two similar fingerprints are represented by two closely situated peaks, 5.II and 5.III, appearing at indices 1657 and 1659 in Figure 5. These triplets contain an acceptor atom distant at six to seven bonds from another acceptor atom, which is itself situated at a variable distance from an atom acting both as acceptor or donor. This third ambivalent atom is situated at six to seven bonds from the first one in the triplet. The two triplets differ slightly in their structure by the fact that the variable distance separating the second atom from the third is longer by two bonds in the first triplet.

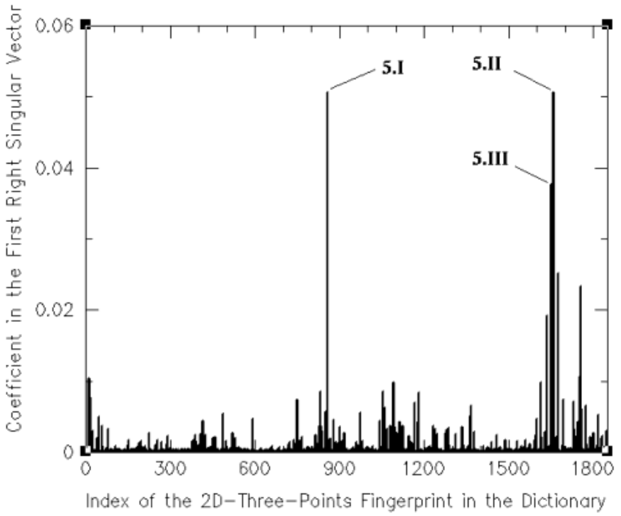

Figure 5. Coefficients of the first right singular vector of the molecule-by-descriptor matrix representing the 27 umami-tasting molecules coded with 2D triangular fingerprints. The figure was generated by using Xmgr data plotting software.[51]

It is important to mention that the two identified sapophores, depicted in Figure 6a,b, could be used as very fast virtual screening tools, complementing the selection by similarity in the SVD-transformed chemical space, based on the left singular vectors. In fact, they can be used as filters when hundred of thousands of chemicals should be processed, prior to the mapping in the target chemical space. To apply these filters it is necessary to verify that a query molecule possess the required chemical sites, positioned at the correct topological distances (bond counts).

Applications. We applied the methodology exposed in this work to the identification of new glutamate-like tasting molecules. To this end we constructed an exploratory virtual library by coupling nonpolar natural amino acids with naturally occurring carboxylic acids. The couplings were constructed by manipulating the SMILES codes representing the amino acids and the carboxylic acids in order to obtain a virtual collection of all possible amides. The respective molecule-by-descriptor matrices were calculated by using counts of 2D triangular fingerprints as molecular descriptors. We further applied eq 3.6 to map the representative points

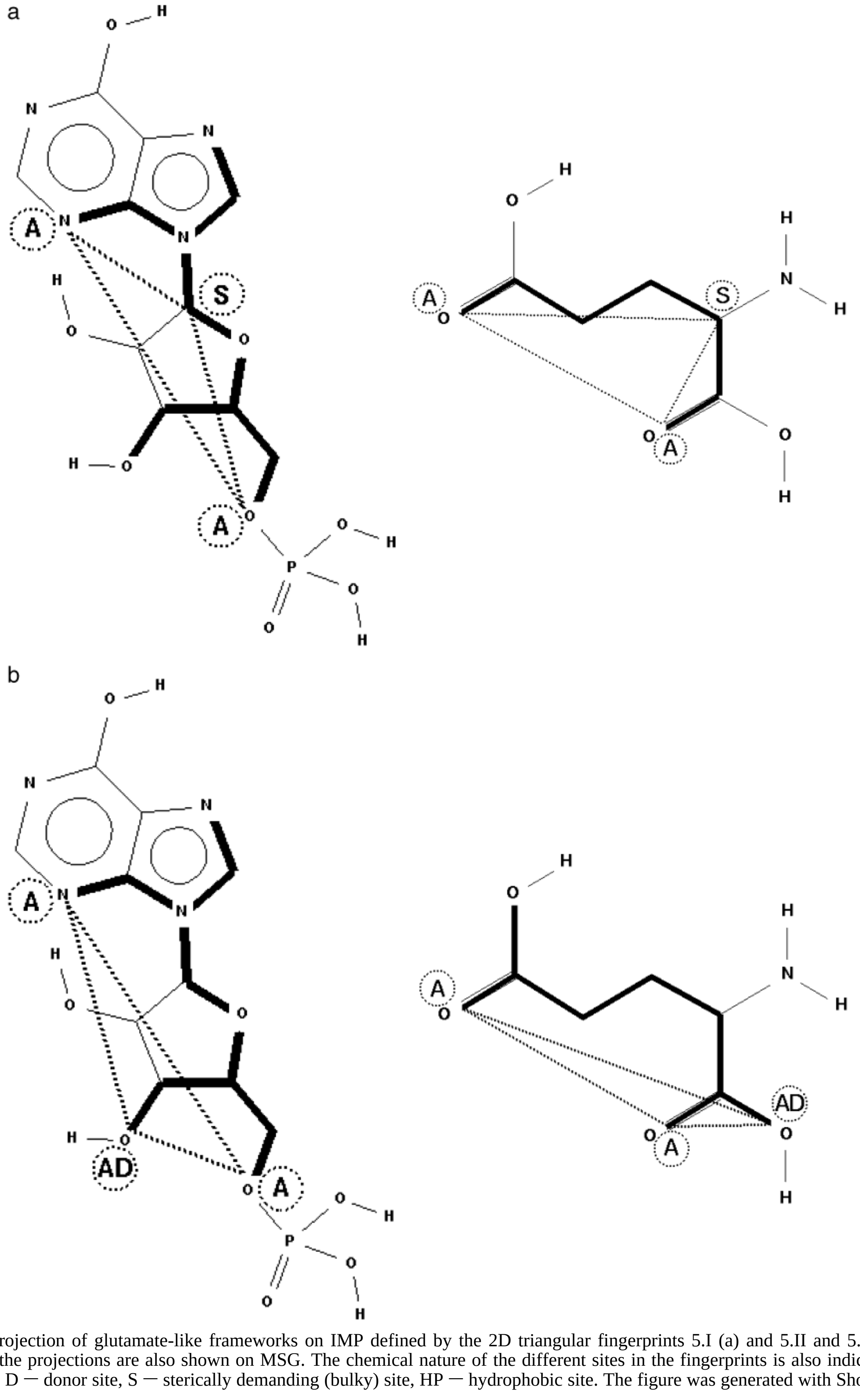

Figure 6. Projection of glutamate-like frameworks on IMP defined by the 2D triangular fingerprints 5.I (a) and 5.II and 5.III (b). For comparison, the projections are also shown on MSG. The chemical nature of the different sites in the fingerprints is also indicated: A — acceptor site, D — donor site, S — sterically demanding (bulky) site, HP — hydrophobic site. The figure was generated with Showcase SGI software.

of the exploratory library in the 3D space spanned by the left singular vectors, resulting from the SVD of the molecule-by-descriptors matrix of the reference peptide library. Molecular similarity was evaluated in the SVD-transformed space by using the Euclidean distance between every two points. In this way we were able to retrieve 10 closest neighbors of MSG to finally isolate the lead compound *N*-acetylglycine (cf. Figure 3). The molecule elicited umami taste at a 5-fold higher taste threshold than MSG.[49] However, it was found to be food-grade[50] and inexpensive and was shown to elicit pure notes of umami without off-flavors, according to its evaluation by a dedicated, trained panel of sensory specialists. Investigations on the chemical modifications of the lead structure that can help to attain lower taste thresholds are currently underway.

CONCLUSIONS

We have presented in this work an application of molecular similarity techniques to the selection of active leads among the members of virtual libraries of potential umami-tasting molecules. Molecules were represented by standard descriptors—counts of Ghose-Cripen atom types, Kier-Hall topological descriptors, counts of HOSE codes, and counts of 2D triangular fingerprints. Dimensionality reduction of the data sets was carried out by singular value decomposition (SVD). Chemical spaces were estimated to have an optimal dimension of three using a method based on the correlation exponent of a statistical data set. Correlation exponent was also found to be a useful measure of the degree of compactness of a data set. Thus, we are proposing it as a new measure of the diversity in a chemical library. Selection of active leads was performed by defining a target volume in the representation of a reference virtual library composed of all amino acids and all possible di- and tripeptides. The target volume was defined as the one where training-set umami-tasting molecules were mapped by SVD. Validation studies established that test-set active molecules were also projected in the same location. 2D triangular fingerprint descriptors were found to be the most successful descriptors in our study. When using them, the target volume condensed at most, compared to the volume of the whole reference library. We estimated volume ratios by calculating the volumes delimited by the convex hulls of the target volume and the whole reference library, respectively. Finally, high-potential candidate molecules were selected in an exploratory virtual library of amino acid amides. This was achieved by retrieving the closest Euclidean neighbors of MSG in the 3D space spanned by the left singular vectors of the molecule-by-descriptor matrix defined by counts of 2D triangular fingerprints. Among these, the compound *N*-acetylglycine was chosen as the most promising lead, due to its good taste activity, cost, and previous applications in food industry. A tentative sapophore for umami taste was identified through the analysis of the first right singular vector of the molecule-by-descriptor matrix defined by the same descriptors. We propose such an analysis of right singular vectors as an important source of structural information in the applications of SVD to chemoinformatics. Such an analysis led us to the finding that a glutamate-like sapophore is present on the ribonucleotide inosine 5′-monophosphate (IMP), indicating that this type of compound might evoke umami-taste through the same molecular pathway as MSG.

ACKNOWLEDGMENT

We are thankful for the useful discussions we had on the subject with Drs. Marcel-Alexandre Juillerat and Johannes Le Coutre. We are grateful to Dr. Marcel-Alexandre Juillerat who suggested several improvements to the manuscript.

REFERENCES AND NOTES

(1) Beeley, L. J.; Duckworth, M. D.; Southan, C. The impact of genomics on drug discovery. *Prog. Med. Chem.* **2000**, *37*, 1−43.

(2) Bellisle, F. Glutamate and the Umami taste: sensory, metabolic, nutritional and behavioral considerations. A review of the literature published in the last 10 years. *Neurosc. Behav. Rev.* **1999**, *23*, 423−438.

(3) Nelson, G.; Chandrashekar, J.; Hoon, M. A.; Luxin, F.; Zhao, G.; Ryba, N. J. P.; Zuker, C. S. An Amino acid taste receptor. *Nature* **2002**, *416*(6877), 199−202.

(4) Li, X.; Staszewski, L.; Xu, H.; Durick, K.; Zoller, M.; Adler, E. Human receptors for sweet and umami taste. *Proc. Natl. Acad. Sci.* **2002**, *99*(7), 4692−4696.

(5) Chaudhari, N.; Landin, A. M.; Roper, S. D. A metabotropic glutamate receptor variant functions as a taste receptor. *Nat. Neurosci.* **2000**, *3*(2), 113−119.

(6) Brand, J. G., Receptor and transduction processes for umami taste. *J. Nutr.* **2000**, *130*(4S), 942S.

(7) Arai, S.; Yamashita, M.; Noguchi M.; Fukimaki, M. Tastes of L-glutamyl oligopeptides in relation to their chromatographic properties. *Agric. Biol. Chem.* **1973**, *37*, 151−156.

(8) Noguchi, M.; Arai, S.; Yamashita, M.; Kato, H.; Fujimaki, M. Isolation and identification of acidic oligopeptides occurring in a flavour potentiating fraction from a fish protein hydrolysate. *J. Agric. Food Chem.* **1975**, *23*(1), 49−53.

(9) Ohyama, S.; Ishibashi, N.; Tamura, M.; Nishizaki, H.; Okai, H. Synthesis of bitter peptides composed of aspartic acid and glutamic acid. *Agric. Biol. Chem.* **1988**, *52*, 871−872.

(10) Tamura, M.; Nakatsuka, T.; Tada, M.; Kawasaki, Y.; Kikuchiokai, H. The relationship between taste and primary structure of “delicious peptide” (Lys-Gly-Asp-Glu-Glu-Ser-Leu-Ala) from beef soup. *Agric. Biol. Chem.* **1989**, *53*(2), 319−325.

(11) Van Den Oord, A. H. A.; Van Wassenaar, P. D. Umami peptides. Assessment of their alleged taste properties. *Z. Lebensm.-Unters. Forsch. A* **1997**, *205*(2), 125−130.

(12) Monastyrskaia, K.; Lundstrom, K.; Plahl, D.; Acuna, G.; Schweitzer, C.; Malherbe, P. *Br. J. Pharmacol.* **1999**, *128*(5), 1027−1034.

(13) Johnson, M. A.; Maggiora, G. M. in *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; J. Wiley and Sons: New York, 1990; pp 1−13.

(14) *Fundamentals of Molecular Similarity*; Carbó-Dorca, R., Gironés, X., Mezey, P. G., Eds.; Kluwer Academic/Plenum Publishers: New York, NY, 2001.

(15) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors − Methods and Principles in Medicinal Chemistry Series*; Mannold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley-VCH: Weinheim and New York, 2000; Vol. 11.

(16) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of “Molecular Diversity” Descriptors. *J. Med. Chem.* **1996**, *39*(16), 3049−3059.

(17) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*(3), 572−584.

(18) Potter, T.; Matter, H. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *J. Med. Chem.* **1998**, *41*(4), 478−488.

(19) Matter, H.; Pötter, T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*(6), 1211−1225.

(20) Matter, H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **1997**, *40*(8), 1219−1229.

(21) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36. Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(22) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure−activity relationships. *J. Comput. Chem.* **1986**, *7*, 565−568.

(23) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; John Wiley and Sons: Chichester, 1986.

(24) Bremser, W. HOSE − a novel substructure code. *Anal. Chim. Acta* **1978**, *103*(4), 355−365.

(25) Fisanick, W.; Cross, K. P.; Rusinko, A., Jr. Similarity searching on CAS Registry substances. 1. Global molecular property and generic atom triangle geometric searching. *J. Chem. Inf. Comput. Sci.* **1992**,

32(6), 664−74. Gerber, P. In *Fundamentals of Molecular Similarity*; Carbó-Dorca, R., Gironés, X., Mezey, P. G., Eds.; Kluwer Academic/Plenum Publishers: New York, NY, 2001.

(26) Golub, G. H.; Van Loan, C. F. *Matrix Computations*, 3rd ed.; John Hopkins University Press: Baltimore, MD, 1996.

(27) Xie, D.; Tropsha, A.; Schlick, T. An efficient projection protocol for chemical databases: singular value decomposition combined with truncated-Newton minimization. *J. Chem. Inf. Comput. Sci.* **2000**, *40*(1), 167−177.

(28) Hull, R. D.; Singh, S. B.; Nachbar, R. B.; Sheridan, R. P.; Kearsley, S. K.; Fluder, E. M. Latent semantic structure indexing (LaSSI) for defining chemical similarity. *J. Med. Chem.* **2001**, *44*(8), 1177−84.

(29) Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Harshman R.; Landauer, T. K.; Lochbaum, K. E.; Streeter, L. A. Computer Information Retrieval using Latent Semantic Structure. U.S. Patent 4,839,853 issued June 13, 1989; Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; Harshman R. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391−407.

(30) Eckart, C.; Young, G. The approximation of one matrix by another of lower rank. *Psychometrika* **1936**, *1*, 211−218.

(31) Schmidt, E. Zur theorie der linearen und nichtlinearen integralgleichungen. I Teil. Entwicklung willkürlichen funktionen nack system vorgeschriebener. *Math. Ann.* **1907**, *63*, 433−476.

(32) Mirsky, L. Symmetric gauge functions and unitarily invariant norms. *Quart. J. Math*. **1960**, *11*, 50−59.

(33) Andrews, H. C.; Patterson, C. L. Singular Value Decompositions and Digital Image Processing. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 1976; Vol. ASSP-24, pp 26−54.

(34) Broomhead, D. S.; King, G. P. Extracting qualitative dynamics from experimental data. *Physica D* **1986**, *20*, 217−236.

(35) Lanczos, C. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Stand.* **1950**, *45*, 255−282.

(36) Arnoldi, W. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.* **1951**, *9*, 17−29.

(37) Maschhoff, K. J.; Sorensen D. C. P_ARPACK: An efficient portable large scale eigenvalue package for distributed memory parallel architectures. In *Applied Parallel Computing in Industrial Problems and Optimization*, volume 1184 of *Lecture Notes in Computer Science*; Wasniewski, J., Dongarra, J., Madsen, K., Olesen, D., Eds.; Springer-Verlag: Berlin, 1996.

(38) Allen, M.; Smith, L. A. Monte Carlo SSA: Detecting irregular oscillations in the presence of coloured noise. *J. Clim.* **1996**, *9*, 3373−3404.

(39) Mandelbrot, B. *The fractal geometry of nature*; W. H. Freeman: New York, 1977.

(40) Schröder, M. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. W. H. Freeman and Company: New York, 1991.

(41) Renyi, A. On measures of entropy and information. *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics, and Probability*; 1960; pp 547−561.

(42) Grassberger, P.; Procaccia, I. Measuring the Strangeness of Strange Attractors. *Physica D* **1983**, *9*, 189−208.

(43) Grassberger, P.; Procaccia I. Characterization of strange attractors. *Phys. Rev. Lett.* **1983**, *50*, 346−349.

(44) Takens, F. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence*; Rand, D. A., Young L. S.; Eds.; Springer-Verlag: Warwick, 1981; pp 366−381.

(45) Fraedrich, K. Estimating the dimensions of weather and climate attractors. *J. Atmos. Sci.* **1986**, *43*(5), 419−432.

(46) Barber, C. B.; Dobkin, D. P.; Huhdanpaa, H. *The Quickhull algorithm for convex hull*; Technical Report GCG53; The Geometry Center, University of Minnesota: July 1993.

(47) Frerot, E.; Escher; S. D. Flavored products and a process for their preparation; Patent US5780090.

(48) Shamil, S.; Birch, G. G.; Mathlouthi, M.; Clifford, M. N. Apparent molar volumes and tastes of molecules with more than one sapophore. *Chem. Senses* **1987**, *12*(2), 397−409.

(49) Grigorov M.; Schlichtherle-Cerny H.; Affolter M.; Kochhar S.; Juillerat M.-A. European Patent Application 02076575.6.

(50) Juichiro M. Prevention of off-flavoring of frozen milk. Patent JP56010012.

(51) Turner P. J.; Stambulchik, E. Xmgr software; http://plasma-gate.weizmann.ac.il/Xmgr

(52) Swayne, D. F.; Cook, D.; Buja, A. XGobi: Interactive Dynamic Data Visualization in the X Window System. *J. Comput. Graph. Stat.* **1998**, *7*(1), 113−130.

CI025652G