ELSEVIER

Visualizing substructural fingerprints

Robert D. Clark,* David E. Patterson,* Farhad Soltanshahi,*
James F. Blake,† and James B. Matthew†

**Tripos, Inc., St. Louis, Missouri, USA*
†Pfizer Research Laboratories, Groton, Connecticut, USA

Substructural fingerprints have proven very useful for chemical library and diversity analysis, but their high dimensionality makes them poorly suited to principal components analysis and to standard nonlinear mapping methods. By using a combination of optimizable K-dissimilarity selection (OptiSim™) and a modified stress function that suppresses effects of distances that fall beyond a characteristic horizon, it is possible to relax principal components analysis coordinates into more consistently meaningful projections from fingerprint space into two dimensions. The nonlinear maps so obtained are useful for characterizing combinatorial libraries, for comparing sublibraries, and for exploring the distribution of biological properties across structural space.

Keywords: molecular diversity, fingerprints, visualization, nonlinear mapping, principal components analysis, combinatorial libraries, library design, optimizable K-dissimilarity, OptiSim

INTRODUCTION

Substructural fingerprints are binary vectors (bit sets) in which each element is set to 1 or 0 to indicate the presence or absence, respectively, of some substructural element in the corresponding molecular structure. The mapping is one to one for the substructure keys distributed by MDL,[1] whereas Daylight[2] fingerprints are hashed such that particular bits can be set by any of several different, unrelated substructures. UNITY®[3] fingerprints are qualitatively intermediate, in that only related substructures, e.g., alkyl fragments, get hashed together.

Fingerprints originally were developed to speed up 2D searches of chemical databases,[4] but recent work has made it clear that such fingerprints also work remarkably well for assessing similarities and differences between molecules in a biochemically meaningful way.[5-9] Because the bit-string operations underlying their manipulation are very fast, fingerprints are particularly appealing as tools for dealing with the large amounts of data produced by the high-throughput screening (HTS) and combinatorial chemistry programs currently under way at many pharmaceutical companies. In particular, one would like to present the relationship between sets of fingerprints in such a way that the full power of human pattern recognition can be brought to bear for elucidating structure-activity relationships.

Unfortunately, fingerprints do not lend themselves naturally to visualization, in part because of their high dimensionality. Indeed, it seems likely that their high dimensionality is directly related to their good neighborhood behavior—the fact that molecules with very similar fingerprints are very likely to exhibit similar biochemical properties.[7] There are simply too many ways for large numbers of compounds to be mutually distinct to be conveyed with complete accuracy in any low dimensional display space.

A second complication lies in the fact that the Euclidean distances to which people are accustomed are not the best way to measure distances in fingerprint space. This is because any particular substructure (e.g., a pyrazole ring) is much more relevant in terms of medicinal chemistry when it is found in one or both of two molecules than when it is absent from both. Hence, distances (dissimilarities) between two fingerprints are more meaningfully assessed[10,11] using the Soergel[12] distance d given by:

$$d(a, b) = 1 - T(a, b) = \frac{\|a \cup b\| - \|a \cap b\|}{\|a \cup b\|} \qquad (1)$$

where a and b are the fingerprints of interest, the double bars indicate cardinality, and $T(a,b)$ is the Tanimoto similarity coefficient. Note that this distance measure runs from 0 to 1, and that bits that are set to 0 in both fingerprints do not contribute. Taken together, these considerations serve to reduce the effective dimensionality around each fingerprint, which helps to counteract the "curse of high dimensionality" referred to earlier.

Our own experience[7] and that of others[5] indicates that two molecules separated by a Soergel distance of 0.15 or less (corresponding to a Tanimoto similarity coefficient of 0.85 or more) are likely to exhibit biological activities within two orders of magnitude of each other, which makes them substantially redundant in terms of HTS. Hence, 0.15 generally is used as an exclusion radius when selecting subsets from a combinatorial library.

Color Plates for this article are on pages 527–532.

Corresponding author: Robert D. Clark, Tripos Inc., 1699 South Hanley Avenue, St. Louis, MO 63144, USA. Tel. 314-647-1099; fax: 314-647-9241. *E-mail address*: bclark@tripos.com (R.D. Clark).

Journal of Molecular Graphics and Modelling 18, 404–411, 2000

655 Avenue of the Americas, New York, NY 10010

1093-3263/00/$–see front matter
PII S1093-3263(00)00065-6

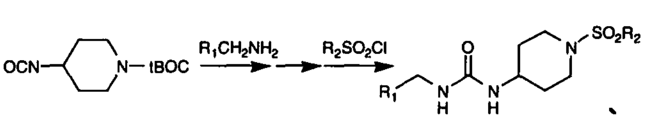

Figure 1. Virtual reaction defining the sulfonylpiperidine urea combinatorial library.

METHODOLOGY

The Sulfonylpiperidine Urea Library

Consider, for example, the virtual library defined by the reaction shown in Figure 1, which could be used as a platform from which to design generic screening sublibraries. The 4-aminopiperidine scaffold upon which the full library is built is not commercially available, but it is a known compound. A UNITY substructure search of commercially available[13] reagents was run and the candidate reagents obtained were screened in ChemEnlighten™[14] for desirable physical properties.

UNITY 2D searches were restricted to molecules containing no more than 10 rotatable bonds, and reagents containing the substructural fragments listed in Table 1 were excluded by using the *-notlist* option in *dbsearch.* Note that a moderate level of potentially interfering functionality (e.g., single free hydroxyl groups) was permitted, the assumption being that a modest investment in protection and deprotection chemistry could be accommodated. The primary amine and sulfonyl chloride hitlists obtained were then loaded into ChemEnlighten databases and filtered for the physical property limits listed in Table 2. A total of 308 distinct primary amines passed the filters, as did 154 sulfonyl chlorides, so the full library encompassed 47,432 products.

The filters applied were chosen with an eye toward generating products with generally drug-like properties[15] and succeeded reasonably well: 91% of the products in the resulting library had a molecular weight ≤550 (68% ≤ 500), and 95% returned a CLogP[16] of 5.0 or less. Most contained one or two aromatic rings (38% and 46%, respectively).

Additional filters are, of course, involved in creating "real" libraries, but those used here are stringent enough to ensure that the distribution of substructural features in the resulting library is realistic. In addition, they produce a range of products that illustrate the behavior of visualization methods at hand. The product library also is realistic in that it is flexible enough to explore an interesting range of binding site geometries, but not so flexible that tight binding is likely to be precluded by the entropic cost of "freezing out" rotatable bonds.

OptiSim Subsets

It is not necessary to project data points for all 47,432 products from fingerprint space simultaneously to get a good idea of the various structural relationships that exist between the compounds that make up the library. Indeed, it is impossible to fully resolve that many points even in three dimensions, let alone in the two dimensions to which one is restricted in print. Instead, one can view a subset selected in such a way that it is representative of those compounds not shown, and which provides a useful mechanism for "drilling down" to any required level of resolution.

This can be accomplished by examining a random sample, which is quite efficient if the structures are uniformly distributed or if one is looking at more than 10 or 20% of all the compounds in a given data set. Unfortunately, combinatorial libraries often are rather unevenly distributed across the region of fingerprint space spanned by each, in that distances between clusters of related products vary depending on the relative structural complexity of the substituents (alkyl *vs* phenyl *vs* azoles) and the nature of their linkage to the combinatorial core, as does the "density" of each cluster. Hence, a random sample large enough to cover the space adequately tends to produce at least one area where the point density is too high to be useful for evaluating the co-localization and segregation of, for example, activity classes.

Subsets obtained by applying OptiSim methodology[17–19] to a large library are more informative, however, in that they are representative enough to give an good sense of the distribution of structures within a library, yet diverse enough to accurately convey its coverage of the available structural space. Such selection sets are built up by pulling the best representative from a series of candidate subsamples and adding it to the set of compounds already selected. Subsample sizes k of three to five generally work well, so creating selection sets is very fast. Using OptiSim selection also is convenient in that the library need not be fully enumerated: selection can instead be made directly from a combinatorial definition, e.g., from a combinatorial SLN[20] (CSLN).

An initial subset of 300 compounds was drawn from the sulfonylpiperidine urea library by running OptiSim with an exclusion radius (distance below which compounds are considered redundant) of 0.15 and a subsample size $k = 3$.

Working from a subset has the side benefit of reducing the effective dimensionality of the problem to a considerable degree, because the underlying level of dimensional complexity is always less than the number of compounds being examined. Here, that translates to a potential reduction from 988 dimensions (the number of bits in a standard UNITY fingerprint) to 300 or less.

Combinatorial Sublibraries

Combinatorial sublibraries were generated by applying the OptiSim[17] extension illustrated in Color Plate 1. The process is seeded by choosing one product at random, which specifies the first reagent pair A_1B_1. At each step, new reagents are chosen at random from the list of those available, and the products produced from each by reaction with the complementary reagents that have already been specified are examined. That reagent whose products compare most favorably to the sublibrary that has been built up so far are added to the selection list for the appropriate reagent. What exactly "most favorable" means is very flexible; it may simply mean most diverse, but also can involve considerations of cost or synthetic compatibility.

In Color Plate 1, the subsample size k is set to 3 for illustrative purposes, and a 3 × 4 pattern has been specified. Compound A_1B_1 is selected at random to seed the process. Reagent candidates a_{21}, a_{22}, and a_{23} then are considered by comparing $a_{21}B_1$, $a_{22}B_1$, and $a_{23}B_1$ to A_1B_1. The candidate that produces the best set of products (most diverse, cheapest, best average expected activity, etc.) specifies A_2. In the next step, three candidate reagents B are selected: b_{21}, b_{22}, and b_{23}. Each

Table 1. Substructure exclusions included in the files specified by the *–notlist* option in 2D UNITY searches

UNITY query	SLN for excluded substructures	Targets
CH2N[f]H2	CHN[not=NHC(=O)].C[not=C:Any]NH	Polyamines
CS(=O)(=O)Cl	CHN[not=NHC(=O)]	Free amines
	S(=O)(=O)Hal. S(=O)(=O)Hal	Polysulfonyl halides
Both	C(=O)OH	Free acids
	C(=O)[f]	Carboxylate salts
	C(=Het)Hal	Reactive halides
	OH.OH	Polyols
	C(=Het)NH.C(=Het)NH	
	N[not=NHC(=O)]HN[not=NHC(=O)]H	Hydrazines
	C(=Het)N.C(=Het)N.C(=Het)N	Peptides
	C[is=C~Any=:Any]HZ{Z:Cl,Br,I}	Activated halides
	N(~O[f])~O[f]	Nitro compounds
	F.F.F.F.F.F.F	Perfluoroalkyls > C2
	CCCCCCCCH3	Long alkyls
	H[I=2]	Heavy isotopes
	H[I=3]	Heavy isotopes
	C[I=13]	Heavy isotopes
	C[I=14]	Heavy isotopes
	N[I=15]	Heavy isotopes
	S[I=35]	Heavy isotopes
	P[I=32]	Heavy isotopes

Table 2. Statistics and secondary filters applied to primary reagent lists

	Primary amines		Sulfonyl chlorides	
Property	Cut-off	Passed	Cut-off	Passed
Single structure	—	436	—	178
Molecular weight	200	361	350	163
Molecular volume ($Å^3$)	190	363	255	165
ClogP	2.6	370	5.0	168
Aromatic ring count	1	394	2	171
Combined filters	—	308	—	154

candidate will now give rise to two products—A_1b_{2i} and A_2b_{2i}—which are evaluated against A_1B_1 and A_2B_1.

Selections from the reagent lists alternate until one of the specified block dimensions is reached; the corresponding reagent then is skipped over until the full block is filled out. Once a block is completed, a new seed is chosen by picking *k* candidate *compounds* at random and comparing them to the products in the blocks that have already been specified. The process continues as for the first block until the required number of products have been specified or no valid selections remain.

Note that no products from reactants selected for earlier blocks are considered in selecting the seed product (e.g., A_4B_5 in Color Plate 1) that starts a new block, and that all products in preceding blocks are considered when evaluating candidates for subsequent blocks. In Color Plate 1, for example, similarity of $a_{42}B_5$ to A_2B_3 may militate against the selection of a_{42} as A_4.

Three 200 member sublibraries were created using a combination of customized code in SYBYL®[21] programming language (SPL) and commercially available functions from the Legion™ combinatorial builder module of SYBYL. The value of *k* was set to 5 and block dimensions were set to 1 × 1 ("cherry picking," which is identical to ordinary OptiSim selection), 10 × 5 ("four blocks"), or 20 × 10 ("single block") for primary amines and sulfonyl chlorides, respectively.

Reagent subsamples were chosen at random with uniform probability from among those for which no anticipated product fell within an exclusion radius of 0.10 of any product already specified. Candidate reagents were selected with replacement and so could be selected for inclusion in several different blocks. In fact, only 32 primary amines are called for in the "four blocks" design, because four contributed to two different blocks and three appeared in three blocks. No sulfonyl chlorides were used more than once, so the design would require a total of 52 reagents *vs* the 30 used in the single block design.

Roulette wheel selection[22] weighted by price, supplier, etc., can easily be incorporated into the subsample selection process, as can categorical exclusion criteria such as physical property cut-offs ("druggability").[15]

For the libraries described here, candidate reagents were rated simply on the basis of diversity. In particular, the MiniMax criterion was used to select the best candidate at each stage: that reagent was selected for which the maximum Tanimoto similarity to *any* already-specified product was smallest. Other metrics (e.g., smallest average cosine coefficient) can be used in place of MiniMax Tanimoto, and nonstructural criteria can be incorporated into the fitness function if desired.

A thorough characterization of the library designs obtained using OptiSim in this way is beyond the scope of this article, but several salient points bear mentioning.

- Replacement of "bad" reagents that slip past the filters simply entails rerunning the corresponding step in the analysis while including products specified at *subsequent* steps when evaluating replacement candidates; replacing B_4, for example, would involve comparison of its products with A_5B_8, $A_{10}B_{4}$., etc., as well as with A_1B_1 and A_3B_2.
- Extension to reactions involving more than two reagents is straightforward.

Perhaps most interesting is the use of roulette wheel selection in place of uniform random sampling for choosing subsample candidates. Introducing a particular bias (e.g., toward cheaper reagents) when deciding which subsample of reagents to consider next can produce quite different results from those produced by adding analogous terms to the fitness function used to select the "best" candidate from each subsample.

Note that sublibraries obtained in this way are both representative *and* diverse, in the same sense that OptiSim selection sets are.[18,19] For any given block layout, the balance between the two characteristics is set by the value chosen for *k*: smaller subsample sizes give more representative sublibraries and larger subsample sizes give more diverse ones.

RESULTS

Principal Components Analysis and Nonlinear Mapping Projections

Principal components analysis (PCA) has seen extensive use in diversity analysis.[23,24] Color Plate 2A shows the projection obtained by extracting the first two principal components from the fingerprint space for the 300-compound OptiSim selection set described earlier. This subset includes 11 compounds that have no neighbors within a Soergel radius of 0.3, beyond which biochemical similarity falls off rapidly; their positions in the plot are highlighted as open circles. It is not at all obvious by inspection of the principal components projection that these 11 compounds are structurally isolated. In fact, they all tend to fall into the central areas of the map.

Color Plate 3 includes the corresponding structures, which are numbered in the order in which they were brought into the OptiSim selection set; "**X**" denotes the shared piperidine core.

The PCA map can be modified to better reflect the real pairwise distances within the data set by applying a nonlinear mapping technique (NLM) developed originally by Sammon[25] and subsequently extended by Kowalski and Bender[26] and by others.[27-29] In this approach, the PCA coordinates are perturbed so as to minimize some stress function. Color Plate 2B shows the result of doing this for the sulfonylpiperidine ureas using Sammon's original stress function *S*:

$$S = \sum_i \sum_{j>i} \frac{(d^*_{ij} - d_{ij})^2}{d_{ij}} \qquad (2)$$

where $d_{ij}{}^*$ is the distance between points *i* and *j* in the projection, and d_{ij} is the distance between *i* and *j* in the original space. Here, we are interested in the Soergel distance.

The isolated points have been displaced toward the edge of the map, which is clearly desirable. This improvement comes, however, at the cost of reducing the anisotropy of the map—the distinctive shape of a PCA projection is characteristically reduced or lost altogether in generating a nonlinear map from a high-dimensional space, particularly for data sets as inherently symmetrical as combinatorial libraries.

Many near neighbors in the fingerprint space also are near neighbors in both projections (not shown), but many have been pulled apart in the PCA or the NLM projection, or in both. Examples include the other 10 compounds highlighted in Color Plate 2A and B, which have been paired up by similarity; their structures are also shown in Color Plate 3. The Soergel distances separating **12** from **20**, **10** from **14**, **19** from **21**, **4** from **8**, and **16** from **18** are 0.243, 0.249, 0.271, 0.304, and 0.339, respectively. These separations are small enough to imply a substantial potential for similarity in biological activity but large enough that differences in potency can be expected to exceed 100-fold. Such pairs form the bridges that link structural islands of biological activity, so getting an accurate presentation of their relationship to each other is critically important.

A Modified NLM

Unfortunately, the relatively large separations which dominate the NLM in Color Plate 2B are precisely those that carry the least amount of useful information; it is the *local* similarity that matters most. Once the Soergel distance between two fingerprints gets much beyond 0.4, one can conclude that the corresponding structures are different, but not really how different they are.[30]

This consideration can be incorporated into the NLM by modifying the stress function[31] so that each compound only "sees" compounds that lie within a neighborhood of radius *h* around it. This can be done by replacing each of the distance terms in the numerator of Equation 2 with the distance *h* to the horizon whenever two compounds are far apart (Equation 3):

$$S = \sum_i \sum_{j>i} \frac{(\min(h, d^*_{ij}) - \min(h, d_{ij}))^2}{d_{ij}} \qquad (3)$$

Sacrificing long-range interactions in this way allows the NLM to relieve stress by unfolding. This is illustrated in Color Plate 4, which shows NLM plots created by minimizing the modified stress function defined in Equation 3 as h is reduced from 0.65 to 0.3. Compounds that do not fall within the horizon of *any* other compound in the subset being examined cannot be placed meaningfully into the projection and so are set off to the edge of the plot (shaded circles in Color Plate 4C and D). Two compounds—**2** and **13**—are excluded at $h = 0.4$ (Color Plate 4C), but **12** and **20** remain well separated, as do, to a lesser extent, **4** and **8**. Upon contracting the horizon still further to $h = 0.3$, the remaining nine isolated compounds are pushed off the map, whereas all five problem pairs cluster appropriately.

The acid test for any visualization method is its ability to order structures in a way that makes sense to a medicinal chemist. Color Plate 5 again shows the projection for the 300 compound OptiSim selection set at $h = 0.3$, but with different compounds highlighted to illustrate the rather "natural" layout of substructures produced by the introduction of an horizon.

As one might expect from the chemistry involved in production of the respective reagents, benzenesulfonyl chlorides and benzylamines dominate the pools of available reagents. Their mutual prevalence is reflected in the dense clump of diaryls (e.g., **22** and **23**) in the upper left quadrant. Those rare products such as **33** and **34**, which lack aryl groups altogether, cosegregate in the sparsely populated area to the right of center in the map, whereas alkylamino arylsulfonamides **26**, **32**, **38**, and **39** occupy the center and center left. Arylamino alkanesulfonamides **35–37** fall into the upper right quadrant, with the more aliphatic **35** positioned toward the bottom of the cluster, near the nonaryl **33** and **34**. Thiophenes and azoles (e.g., **27–31**) appear in the lower left quadrant. Compound **28** is a particularly distinctive compound and so shows up at the periphery of the plot, near the less unusual 5-isoxazolylthiophene-2-sulfonamide **27**. The "reasonableness" of such distributions is intuitively appealing to medicinal chemists but difficult to quantify.

Comparing Combinatorial Sublibraries

Relationships between two or more libraries are best visualized by projecting them into a common NLM, but using fingerprints from all 600 compounds in the individually selected, four block, and single block sublibraries described earlier produces an unnecessarily overcrowded map. Instead, 100 compounds were drawn at random from each sublibrary. The three samples obtained then were pooled and projected together using $h = 0.3$ to create the map shown in Color Plate 6.

This plot clearly supports the expected conclusion[32] that the sublibrary of individually selected compounds (cherry-picking design) is the most diverse, whereas the single block design is the least diverse and, concomitantly, the most redundant. One indication of this is the eight representatives from the cherry-picking library that appear along the edge of the plot, indicating that they fall beyond the horizon of any other compound in the sublibraries. By contrast, only two such outliers (**41** and **54**) were produced by the four block design, and only one (**53**) by the single block sublibrary. In addition, the individually selected compounds are clearly more evenly spread in general. Finally, note the redundancy indicated by the large clumps of single block compounds that surround **42**, **46**, and **48**.

These points probably could be gleaned from summary statistics calculated "blind" using pairwise distances or other numerical data. It is hard to imagine, however, that any such analysis would detect the significant undersampling of compounds evident in the upper right quadrant circumscribed by **51**, **52**, and **55**, particularly in the single block design (large green symbols). The need to identify such diversity "holes" by direct inspection has been a major impetus behind developing tools for the effective visualization of fingerprint space.

Visual comparisons of such projections also provide a way to assess trade-offs in optimality among factors such as coverage, diversity, synthetic efficiency, cost, and redundancy across variations in sublibrary design parameters (e.g., subample size k in the.OptiSim design strategy described here).

Projecting Biological Activity into Fingerprint Space

Analyses carried out on small literature data sets have clearly shown that 2D fingerprints exhibit good neighborhood behavior,[7] but it would be useful to have a less abstract demonstration of this point. To accomplish this, we examined the results of assaying a generalized screening library of proprietary kinase inhibitors against a specific target enzyme, then applying our combination of PCA and modified NLM projection to fingerprints for 300 compounds drawn at random from the pool of inactives together with 100 randomly selected actives. The plots obtained are shown in Color Plate 7A–C, with actives indicated by red symbols and inactives by blue symbols. Color Plates 7A and B show the PCA and direct (no horizon) NLM projections for this data set, whereas the plot in Color Plate 7C was obtained with $h = 0.3$.

There is much more structural diversity among compounds in the kinase data set than is found in the sulfonylpiperidine library, with 80% of the pairwise distances between the fingerprints from the kinase library in excess of the maximum pairwise separation (0.714) seen in the combinatorial one. The large number of long-range interactions involved reduces the extent of "rounding up" possible in this case when going from the principal components projection (Color Plate 7A) to unmodified NLM (Color Plate 7B).

A handful of inactive compounds fall into the cluster of actives that includes compounds **56–60**, and **70** and **71** are juxtaposed in both Color Plates 7A and 7B despite the large pairwise separation between them (0.861) in fingerprint space. Applying our modified NLM procedure with $h = 0.3$ (Color Plate 7C) removes **61**, **62**, **71**, and other outliers—i.e., compounds with no neighbors within a Soergel distance of 0.3—into the frame of the plot and purges the inactives from the large cluster of actives to the left of the plots. Moreover, the stress drops from 9,034 to 36 in going from Color Plate 7B to 7C. Other compounds have been highlighted as light blue squares to illustrate how imposing the horizon affects their distribution relative to one another.

A greater proportion of inactives (56%) show up as outliers in Color Plate 7C than is the case for the actives (30%), indicating that the distibution of "hits" is gratifyingly nonrandom. Of greater interest, however, are the several islands of activity set off from one another by intervening stretches of inactives: good neighborhood behavior implies that such islands will be relatively free of inactives, although it does not preclude the existence of multiple islands. Nor does it imply that the scale of coupling between activity and structure will be the same everywhere. Indeed, some of the "shorelines" in Color Plate 7C are much more sharply defined than others. Cases in which structural changes as simple as adding a methyl group produce a dramatic drop or increase in biological activity represent extreme instances of this, but they in no way disprove the existence of the islands themselves or the continuity of the activity—*and lack thereof*—on either side of such boundaries.

Direct examination of the underlying structures shows that each island represents a more-or-less different patent estate from the large island at the left of the plot, particularly for those farther afield (proprietary data not shown). Some of the compounds that make up the smaller islands are quite active and so may represent new lead areas of chemistry ripe for more thorough exploration.

The inactive compounds make a key contribution to this plot by defining the "shores" of the islands of activity. Note, in fact, that the activity islands are not completely surrounded by inactives. The unbounded edges of the islands may suggest synthetic directions to take that could extend the scope of the chemistries involved. The exact nature of such direction is very context dependent and is best for identifying the structures near the unbounded edge. Finding activity for compound **26** in Color Plate 5 with respect to some (hypothetical) target receptor would suggest synthesis of methoxymethyl or hydroxyethylcyclohexyl homologs, or of hydroxymethylcyclopentyl or hydroxymethyltetrahydrofuranyl amine analogs, for example. Finding activity for **28**, on the other hand, would suggest synthesis of pyridone or furanyl analogs. A quick similarity search carried out against known inactives would then show whether such compounds do represent a real boundary in structural space.

Again, it is difficult to imagine any summary statistic that could accomplish this as effectively as does direct visual inspection of Color Plate 7C.

Projecting Pharmacophore Models into Fingerprint Space

A four-point pharmacophore model for the target enzyme was formulated in connection with the kinase research project. When this pharmacophore hypothesis was employed as a query in a UNITY 3D flex search, it "hit" 67% of the actives and 26% of the inactives, but only 1% of the more generalized database of drug-like molecules represented by Chapman and Hall's Directory of Pharmacological Agents. Color Plate 7D shows the plot obtained by applying our modified NLM procedure (h = 0.3) to an initial PCA for all actives that matched the proposed pharmacophore together with "hits" from the same number of inactives selected at random.

The actives in Color Plate 7D are distributed in a very similar pattern to those in Color Plate 7C, indicating that the query captures something quite real about available binding sites on the target enzyme. The similarity between the two maps testifies to the value of using PCA to get consistent starting coordinates and to how robust the unfolding by the modified NLM is. Moreover, the general disorganization of the inactives away from the islands of activity indicate that such "hits" are probably nonspecific, in that the structural classes to which they belong characteristically present the pharmacophore of interest.

Two compounds (**61** and **71**) that are outliers in Color Plate 7C show up in doubleton "islands" in Color Plate 7D. This is because all compounds "hit" by the query were used to generate the latter map, whereas only one of each pair happened to get selected for the random sample used to generate the former. The two pairs fall well off to the right in Color Plate 7D, reflecting their isolation from other "hits" in structural (fingerprint) space.

DISCUSSION

The roots of the inadequacy of both PCA and standard NLM for projecting combinatorial libraries from fingerprint space down into two dimensions become clearer when one considers some details of how compounds in such libraries are typically distributed in structural space and illuminates the reason that introducing an horizon is so effective.

To begin with, the useful dynamic range of the Soergel distances within a combinatorial library is limited if there is any scaffold. The smallest distance between any two of the 300 compounds shown in Color Plates 2–5, for example, is 0.163, whereas the largest distance is only 0.714. This is less than a four-fold range, yet it spans the spectrum from near redundancy in an HTS context to essentially no expected relationship in biochemical activity.[33]

In addition, the high dimensionality of fingerprints makes it easy to generate nearly symmetrical relationships that cannot be displayed accurately in two dimensions. All 21 pairwise Soergel distances between compounds **1**, **2**, **3**, **5**, **6**, **9**, and **11** (Color Plate 3), for example, fall between 0.424 and 0.527. In other words, they form a slightly irregular six-dimensional simplex. Even a tetrahedron, which is only a three-dimensional simplex, cannot be projected into two dimensions without severe distortion. Absent interactions with other points, a perfectly regular six-dimensional simplex will be projected as a regular heptagon—hence the tendency toward round, isotropic maps when "ordinary" NLM is applied in this situation.

That long-range, high-dimensional relationships do exist within these data sets is clear from the PCAs used to derive starting points for the NLM. The first and second principal components obtained for the sulfonylpiperidine library (Color Plate 2A) capture only 5.8% and 4.9%, respectively, of the total variance in the corresponding fingerprints, for example. Extending the projection up to 10 components (dimensions) only captures 28.7% more, for a total of 43.6%. It would take a reduced descriptor space of 62 dimensions to capture 85% of the variance for this data set. PCA statistics from the more diverse kinase data set (Color Plate 7A) are even more daunting: the first two components capture 14.5% of the total variance, the first 10 components capture 34%, and 108 components are required to account for 85% of the original fingerprint variance.

Our modified NLM procedure could, of course, be initiated using random starting coordinates, which in many cases would produce projections with comparably low stress. The key reason to use principal components is not their explanatory power but the continuity they bring to projections obtained from overlapping subsets: random initialization would obliterate the commonalities of pattern between Color Plates 7C and 7D, for example.

Cutting off long-range effects in these projections by introducing an horizon allows the maps to relax, essentially by letting them unfold. For the modified NLM maps for the 300-compound subset, for example, the total stress S falls sharply as the horizon shrinks—from 5,151 for h = 1.0 to 4747, 2403, 1151, and 253 for h = 0.65, 0.50, 0.40, and 0.30, respectively.

This reduction comes in part from defining away long-range stress, but it also can be interpreted as eliminating distracting sources of long-range noise that are irreconcilable anyway. Less information actually is discarded than one might expect: the 9,292 pairwise distances that fall within the horizon of 0.4 used to create Color Plate 4C imply that, on average, each compound "sees" about 136 neighbors; 233 neighbors, on average, fall within 0.5 of each compound, and 57 fall within an horizon of 0.3.

Fifty-seven compounds can still support some relatively high-dimensional relationships, however. It is evident from the data presented here that fingerprint spaces defined by chemical libraries in general and by combinatorial libraries, in particular, are locally "flat" networks embedded at all angles in a mostly empty space, somewhat like the snowflakes making up a snowdrift. That they can be unfolded while preserving local detail and connectivity seems reasonable, given the constraints that chemical connectivity and feasibility of synthesis put on incremental structural changes and the vast diversity that is synthetically accessible. The result is that the local dimensionality around any single compound usually is much lower than is that of the library as a whole.

Setting an NLM horizon at or near a Soergel distance of 0.3 defines neighborhoods within which the *effective* dimensionality is low enough that meaningful projection into two dimensions is possible. It is fortunate that this natural scale of unfolding conserves relationships between individual structures and between structural classes, while also making possible informative projections of biological activity into the unfolded structural space that results. This certainly will not be the case for all high-dimensional descriptor spaces; where it does hold true, however, the method described herein may prove more generally useful.

ACKNOWLEDGMENTS

This research was supported in part by SBIR grant 1R43GM58919 from the National Institutes of Health. Drs. Richard Cramer, Trevor Heritage, and Denise Beusen of Tripos, Inc., provided encouragement and helpful discussion, as did Peter Fox of Tripos and Dr. Robin Spencer of Pfizer Research Laboratories. We also would like to thank our anonymous reviewers for their helpful suggestions.

REFERENCES

1 MDL Information Systems, Inc., 146000 Catalina Street, San Leandro, CA 94577

2 Daylight Chemical Information Systems, Inc., 27401 Los Altos, Mission Viejo, CA 92691

3 UNITY is distributed by Tripos, Inc., 1699 S. Hanley Rd., St. Louis, MO 63144

4 Willett, P. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* 1998, **38**, 983–996

5 Brown, R.D., and 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**, 572–584

6 Matter, H., and Lassen, D. Compound libraries for lead discovery. *Chem. Oggi* 1996, **6**, 9–15

7 Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D., and Weinberger, L.E. Neighborhood behavior: A useful concept for validation of molecular diversity descriptors, *J. Med. Chem.*1996, **39**, 3049–3059

8 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**, 1219–1229

9 Wild, D.J., and Blankley, C.J. Comparison of 2D fingerprint types and hierarchy level selecection metrhods for structural grouping using Ward's clustering. *J. Chem. Inf. Comput. Sci.* 2000, **40,** 155–162

10 Willett, P., and Winterman, V. A comparison of some measures for the determination of inter-molecular structural similarity. *Quant. Struct.-Act. Relat.* 1986, **5**, 18–25

11 Barnard, J.M., and Downs, G.M. Clustering of chemical structures on the basis of two-dimensional similarity measures. *J. Chem. Inf. Comput. Sci.* 1992, **32**, 644–649

12 Gower, J.C. Measures of similarity, dissimilarity and distance. In: *Encyclopedia of statistical sciences, Volume 5,* Kotz, S., and Johnson, N.L., Eds., John Wiley & Sons, New York, 1985, pp. 397–405

13 Available Chemicals Directory is distributed by MDL Information Systems, Inc., 146000 Catalina Street, San Leandro, CA 94577

14 ChemEnlighten is a registered trademark of Tripos, Inc., St. Louis, MO 63144

15 Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 1997, **23**, 3–25

16 CLogP is a product of BioByte, Inc., Pomona Corporation

17 Patent pending. OptiSim is a registered trademark of Tripos, Inc., 1699 S. Hanley Rd., St. Louis, MO 63144

18 Clark, R.D. OptiSim: An extended dissimilarity selection method for finding diverse representative subsets. *J. Chem. Inf. Comput. Sci.* 1997, **37**, 1181–1188

19 Clark, R.D., and Langton, W.J. Balancing representativeness against diversity using optimizable K-dissimilarity and hierarchical clustering. *J. Chem. Inf. Comput. Sci.* 1998, **38**, 1079–1086

20 Ash, S., Cline, M.A., Homer, R.W., Hurst, T., and Smith, G.B. SYBYL line notation (SLN): A versatile language for chemivcal structure representation. *J. Chem. Inf. Comput. Sci.* 1997, **37,** 71–79

21 SYBYL is distributed by Tripos, Inc., St. Louis, MO 63144

22 Judson, R. Genetic algorithms and their use in chemistry. In: *Reviews in computational chemistry, Volume 10* Lipkowitz, K.B., and Boyd, D.B., Eds., VCH Publishers, New York, 1997, pp. 1–73

23 Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K., and Moos, W.H. Measuring diversity: Experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* 1995, **38**, 1431–1436

24 Shemetulskis, N.E., Dunbar, J.B. Jr., Dunbar, B.W., Moreland, D.W., and Humblet, C. Enhancing the diversity of a corporate database using chemical database

clustering and analysis. *J. Comput.-Aided Mol. Design* 1995, **9**, 407–416

25 Sammon, J.W. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* 1969, **C-18**, 401–409

26 Kowalski, B.R., and Bender, C.F. Pattern recognition-II. Linear and nonlinear methods for displaying chemical data. *J. Am. Chem. Soc.* 1973, **95**, 686–692

27 Domine, D., Devillers, J., Chastrette, M., and Karcher, W. Non-linear mapping for structure-activity and structure-property modelling. *J. Chemometrics* 1993, **7**, 227–242

28 Hudson, B., Livingstone, D.J., and Rahr, E. Pattern recognition display methods for the analysis of computed molecular properties. *J. Comput.-Aided Mol. Design* 1989, **3**, 55–65

29 Agrafiotis, D.K. Stochastic algorithms for maximizing molecular diversity. *J. Chem. Inf. Comput. Sci.* 1997, **37**, 841–851

30 Flower, D.R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* 1998, **38**, 379–386

31 Patent pending

32 Gillet, V.J., Willett, P., and Bradshaw, J. The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 1997, **37**, 731–741

33 Delaney, J.S. Assessing the ability of chemical similarity measures to discriminate between active and inactive compounds. *Mol. Diversity* 1995, **1**, 217–222