Multivariate Data Analysis Using D-Optimal Designs, Partial Least Squares, and Response Surface Modeling: A Directional Approach for the Analysis of Farnesyltransferase Inhibitors

Elie Giraud,† Claude Luttmann, François Lavelle, Jean-François Riou, Patrick Mailliet, and Abdelazize Laoui*

Rhône-Poulenc Rorer S.A., CRVA, 13, quai Jules Guesde, F-94403 Vitry-sur-Seine Cedex, France

Received September 24, 1999

We have investigated the combined use of partial least squares (PLS) and statistical design principles in principal property space (PP-space), derived from principal component analysis (PCA), to analyze farnesyltransferase inhibitors in order to identify "activity trends" (an approach we call a "directional" approach) and quantitative structure−activity relationships (QSAR) for a congeneric series of inhibitors: the benzo[*f*]perhydroisoindole (BPHI) series. Trends observed in the PCA showed that the descriptors used were relevant to describe our structural data set by clearly identifying two well-defined structural subclasses of inhibitors. D-Optimal design techniques allowed us to define a training set for PLS study in PP-space. Models were derived for each biological assay under evaluation: the in vitro Ki-Ras and cellular HCT116 tests. Each of these assay-based sets was subdivided once more into two subsets according to two structural classes in this BPHI series as revealed by the PCA model. The response surface modeling (RSM) methodology was used for each subset, and the corresponding RSM plots helped us identify "activity trends" exploited to guide further analogue design. For more precise activity predictions more refined PLS models on constrained PP-spaces were developed for each subset. This approach was validated with predicted sets and demonstrates that useful information can be extracted from just a few very informative and representative compounds. Finally, we also showed the potential use of such a strategy at an early stage of an optimization process to extract the first "activity trends" that might support decision making and guide medicinal chemists in the initial design of new analogues and/or lead followup libraries.

Introduction

Ras proteins play an important role in signal transduction process involved in cell proliferation. Mutated and constitutively activated Ras proteins have been found in 30% of human cancers.[1,2] p21-Ras proteins (Ha, N, Ki4A, and Ki4B) are small guanine nucleotide-binding proteins that undergo a series of posttranslational modifications which include the prenylation of Ras by the protein farnesyltransferase (Ftase).[3] These posttranslational modifications promote cell membrane association of Ras which is mandatory for its biological activity, e.g. cell growth and tumorigenesis. Since farnesylation of mutated Ras proteins is required for cell transformation,[4] inhibition of Ftase emerged as an attractive target for therapeutic intervention in the cancer field and prompted significant effort to discover potent inhibitors of Ftase as novel anticancer agents.[5]

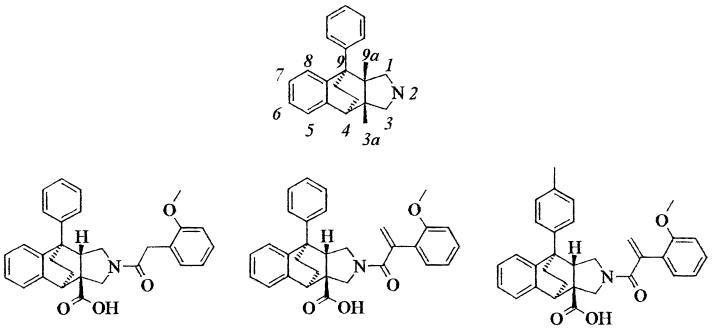
Ftase is a bisubstrate, FPP and p21-Ras protein, enzyme. The inhibition of its activity may be obtained with compounds able to mimic and/or to compete with either FPP or p21-Ras proteins. The discovery of some p21-Ras CAAX tetrapeptides[6] inhibiting Ftase stimulated intensive rational drug design of CAAX mimics leading to peptidic,[7] pseudopeptidic, and peptidomimetic inhibitors.[8−10] Besides drug design, intensive screening efforts of natural products[11] and chemical libraries[12,13] led to the discovery of potent FPP-competitive nonpeptidic inhibitors of Ftase. Several lead series have been published so far.[14] RPR104213 (Figure 1), a submicromolar FPP-competitive inhibitor of Ftase, is a representative of one of these discovered lead series. Since RPR104213 was our initial most potent Ftase inhibitor exhibiting significant inhibition of Ha-Ras processing in whole cells (THAC hamster fibroblasts[9]), we focused on this particular series of benzo[*f*]perhydroisoindole (termed BPHI) derivatives for further optimization. Starting from RPR104213, this optimization process led to an analogue, RPR115135 (Figure 1), with improved enzymatic and cellular Ftase inhibition and then to RPR130401 (Figure 1) exhibiting evidence of in vivo cytostatic activity in animal models.[15]

Concomitant with this effort a statistical analysis was performed to obtain predictive models on both enzymatic Ki-Ras/SPA and HCT116 colony formation biological tests by means of a novel combined method using principal properties derived from principal component analysis,[16] D-optimal designs,[17] and partial least squares[18] methods. The objective of such an approach is to unveil or evaluate untested compounds in this series leading to new interesting compounds with improved biological and pharmacological profile or potential backups. The generated models may also help in gaining a new insight in the molecular features determining the biological profile and to point out new untested combinations of substituents to lead to another structural optimum by a directional approach.[19]

* Corresponding author. Tel: 33 1 55 71 3831. Fax: 33 1 55 71 8063. E-mail: Abdelazize.Laoui@Aventis.com.

† Present address: Synt:em, Parc scientifique G. Besse, 30000 Nîmes, France.

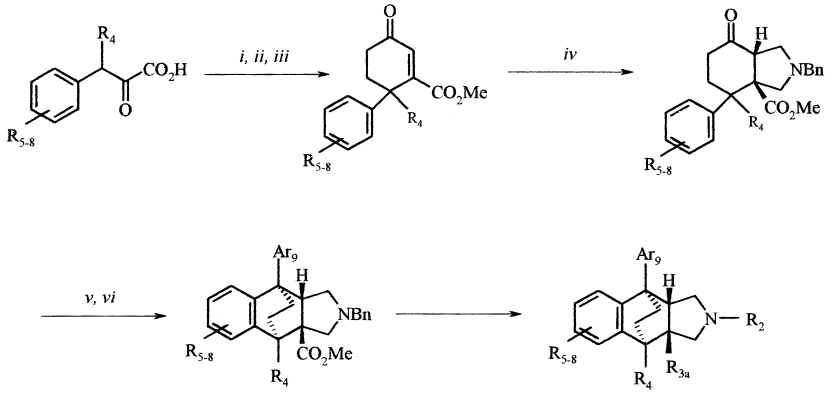
10.1021/jm991166h CCC: $19.00
Published on Web 04/14/2000



$IC_{50}(\mu M)$	(+/-) RPR104213	(+) RPR115135	(+) RPR130401
Ftase/KiRas/SPA	0.2	0.05	0.008
HCT116 cloning	> 10	0.72	0.045
HCT116 *vivo*	n.d.	almost inactive	cytostatic activity

Figure 1. Numbering of BPHI and chemical structures and biological activities of (±)-RPR104213, (+)-RPR115135, and (+)-RPR130401.

Scheme 1. General Synthesis of BPHI[a]



[a] (i) Methyl vinyl ketone, KOH, MeOH (reflux); (ii) *p*-TsOH, toluene (reflux); (iii) MeI, DBU, acetone (reflux); (iv) Me_3Si-CH_2-N(Bn)-CH_2O-*n*-Bu, TFA (cat.), CH_2Cl_2 (rt); (v) ArMgX ($CeCl_3$), Et_2O, toluene (0 °C to rt); (vi) excess CF_3SO_3H, CH_2Cl_2 (0 °C to rt).

This paper describes the validation of this new approach in statistical molecular design as a decision making tool in lead optimization for project chemists. We also investigate the applicability of this modeling approach at an early stage of lead discovery process.

Chemistry

Various chemical routes have been developed to synthesize BPHIs.[20] A full paper, which will detail the different synthetic pathways, the in vitro structure−activity relationships, and the in vivo biological activity of BPHIs, will be soon submitted for publication.[21]

As outlined in Scheme 1, the more versatile approach successively involved as key steps: (1) annelation reaction of substituted phenylpyruvic acids with methyl vinyl ketone leading to 6-arylcyclohexen-4-onecarboxylates; (2) 1,3-dipolar cycloaddition of *N*-*n*-butoxy-*N*-trimethylsilyloxybenzylamine leading to 4-aryl-7-oxoperhydroisoindole-3a-carboxylates; (3) introduction of the second aromatic ring (usually by Grignard reaction), followed by intramolecular Friedel−Crafts reaction to 9-aryl-2-benzyl-4,9-ethano-2,3,3a,4,9,9a-hexahydro-1*H*-benz[*f*]isoindole-3a-carboxylates, whose side chain at position 2 and functional group at position 3a of the BPHI skeleton were further modified by conventional methods.

Computational Methods

We used chemometrics and molecular modeling to better understand the structural features affecting the biological responses. Usually, to gain insight into structure−activity relationships (SARs) one changes a substituent one at a time. This leads to set of molecules which may not contain information enough for ranking the importance of individual features in affecting biological activity. The use of design strategies to select molecules of interest is particularly valuable as it permits to select only a few molecules for further consideration in such a way that they contain the widest information useful to derive the most reliable QSAR models. Usually such a method uses fractional design (FD) or fractional factorial design (FFD)[22,23] to select molecules of interest. Within FD one has to consider as many molecules as there are substitution sites and substituents possible. D-Optimal[24,17] designs can be used as an alternative to FD in constrained situations, e.g. when some regions of the variable space are excluded or when the data set is discrete as with molecular structures. The D-optimality criterion consists of determining the *n* experiments (or structures) in the accessible domain which contain together the maximum amount of information. In other words the D-optimal algorithm selects a few points optimally spanning the domain of interest.

Clementi et al.[17] and Wold et al.[18] suggest an overall chemometric strategy for drug design studies which relies on design in latent variables and partial least square (PLS) modeling. The PLS algorithm appears to be the most appropriate tool for establishing the quantitative relationship between a biological activity profile and a matrix of structural descriptors, especially in detecting the structural features affecting the biological activity and in providing reliable predictions. The combination of methods such as principal component analysis (PCA) and PLS seems to be particularly appropriate in QSAR both for the exploratory analysis of the structural data and for establishing the quantitative relationship under the same descriptor space. The selection of a designed set of informative molecules to explore at best the structural variation in the series ensures the reliability of derived models. In addition to predictions we exploited these models using the response surface modeling (RSM) technique[25−26] to point out trends, favorable directions, and activity regions, i.e. optimums. The principle of RSM is to assume that the variation of the PLS predicted responses is functionally related to the factors, i.e. the principal properties used. A polynomial model is postulated to link the responses to the factors using linear, quadratic, and interaction terms. The polynomial coefficients are estimated by a PLS method. The RSM is a powerful tool which allows an easy interpretation of the models by visual inspection of 2D or 3D plots, especially for finding optimums or pointing out the interplay between different factors.

All these calculations were performed with SIMCA 6.01[27] and MODDE 4.0.[28]

Data Sets. 1. Structural Data Set. A set of 484 BPHIs was selected based on the activities in the Ki-Ras and HCT116 assays (IC_{50}s < 100 μM). Out of this set there was one subset of 229 compounds having Ki-Ras activity data and a second one of 127 compounds which had HCT116 data. Two major structural subclasses coexist in the data set: a first subclass of 363 compounds exhibiting an acid or ester group at position 3a (Figure 1) and a second subclass of 121 structures exhibiting an amide or analogue group at the same position. All 2D structures were converted to 3D using the CONCORD program.[29] SMILES[30] files or the 3D SD files[31] were used for descriptor calculation.

2. Calculated Descriptors. 2D topological and physicochemical descriptors were calculated using the Cerius[2] software[32] on the whole set of selected structures. The nature of the descriptors used and the corresponding references are available as Supporting Information. Most of the descriptors describe steric, electronic, and lipophilic properties as well as topological and information indices. In addition to these descriptors calculated by the Cerius[2] program we added the octanol/water partition coefficient (ClogP) for the whole set of molecules using the Daylight Software[33] and the polar surface area (PSA) based on Palm's work[34−35] and slightly modified in-house.[36]

In total a set of 45 descriptors were used.

Results

The following strategy was applied to obtain models that predict directions to guide lead optimization based on trends in SARs as well as good activity prediction models. We first analyzed the initial data set by PCA to check if our descriptor space is representative of the data set under study, then we used D-optimal design to point out informative compounds on which PLS models were calculated in principle property space (PP-space). This was performed by dividing the original data set into subsets according to the assay of interest (in vitro Ki-Ras or cellular HCT116) and to structural subclasses. Second, to obtain not only trends in SARs but also more accurate predictions, we developed more refined models on more constrained PP-space. Finally, we also checked that using such an approach some interesting trends for initial optimization stages could already be derived at an early stage of the optimization process. The results of each step are presented below.

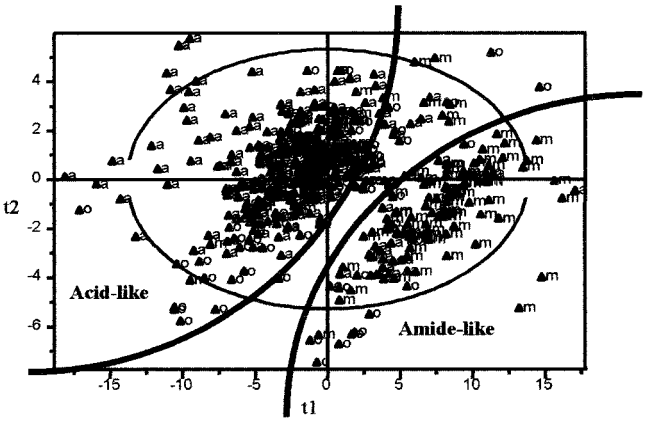


Figure 2. Score plots of PLS-DA model for Ki-Ras data. Acid-like cluster members are represented by **a** (acid function at position 3a of BPHI skeleton) and **o** (ester function at position 3a of BPHI skeleton). Amide-like cluster members are represented by **m** (amide, hydroxamate, and hydrazide at position 3a of BPHI skeleton).

Data Set Analysis. Using these 45 descriptors we performed a PCA with autoscaling to unit variance on the whole set of 484 compounds that extracted three significant principal components explaining 88% of the variance in this descriptor matrix. The PCA score plot (not shown) shows two major clusters. By analyzing these in more detail one can notice that each cluster is representative of a chemical BPHI subfamily. One subfamily contains an acid or ester group at position 3a, and the second one contains an amide or related chemical group at this same position. A PLS model, using log transformed responses, on the entire set of compounds described by the original variable set could not be achieved (R^2 and Q^2 were too low). This may be due to the observed groupings in the PCA score plot. We thus decided to use PLS discriminant analysis[28] (PLS-DA) to check if it could discriminate between the two clusters for each of the biological tests Ki-Ras and HCT116. The compounds were divided into two groups according to their observed biological data: class 1 for IC_{50}s less than 1 μM and class 2 for IC_{50}s greater than 1 μM. A three-component model could be obtained for the set of 229 molecules having Ki-Ras data and a two-component model for the 127 molecules with HCT116 data. The explained variations in the descriptive matrix are respectively 82% and 80% (cumulative R^2X), and the explained variations in the responses are respectively 71% and 69% (cumulative R^2Y). The internal predictive power is 68% (cumulative Q^2) for the former model and 67% for the second one. Figure 2 shows the PLS-DA t_1,t_2 score plot for Ki-Ras data set. One can notice two well-separated groupings with few "misclassified" compounds. Moreover the third t score adds significant information (an additional 8% of explained variance) and allows a better separation leading to even less misclassified compounds. Similar groupings can be observed as well with the HCT116 data set (plot not shown). This clearly shows that the compound's descriptive variables for each assay-based compound set discriminate the two chemical subfamilies.

Table 1. Summary of D-Optimal Statistical Parameters

					D-optimal design			
assay	subsets	explained variance R^2	predicted variance Q^2	model	no. of selected structures	G_{eff} (%)	condition number	predicted/ observed
Ki-Ras	acids	0.96	0.84	polynom 2°	16	70	5.5	81/106
	amides	0.94	0.80	polynom 2°	12	79	4.6	60/80
HCT	acids	0.81	0.70	polynom 2°	11	68	4.1	45/49
	amides	0.97	0.84	polynom 2°	13	80	7.0	36/44
Ki-Ras	initial set	0.97	0.69	polynom 2°	6	69	7.6	26/31

According to these results the set of BPHI structures, represented in PP-space, were divided into two distinct subfamilies for the studies described hereafter and each biological test was considered separately.

Combined PCA−D-Optimal−PLS Approach. In this section we will describe the PLS models based on a set of informative structures extracted from a D-optimal design using principal components (PPs) as molecular descriptors. As noticed in the PCA models mentioned above the three principal components, also called principal properties (PPs), still retain a high degree of the descriptive power of the original data. Because these PPs are limited in number and are mathematically orthogonal, they are particularly well-suited as design variables. We thus decided to adopt a strategy combining PCA and D-optimal design to extract the most informative molecules to be feed in the PLS study as a training set.

1. Reference PCA Models. Two reference PCA models were performed: one for the acid-like compound set (363 compounds) and one for the amide-like compound set (121 compounds). The analysis on the first data matrix 363 × 45 (compounds × descriptors) and the one on the second data matrix 121 × 45 (compounds × descriptors) resulted in respectively four-and three-component models with 90% of explained variation for both models. The first three components of the acid-like model, here denoted t_1, t_2, and t_3, are dominant and account for 85% of the variance. An extra fourth component for this model resulted in a slightly higher cumulative percentage of variance explained (approximately 5% increase). However, we consider models containing more than three components less convenient for interpretation. Thus, the resulting three components for both models were used for the all design studies presented hereafter.

2. D-Optimal Design of Ki-Ras and HCT116 Responses. For these studies we considered only compounds with in vitro Ki-Ras values ranging from 0.001 to 100 μM and cellular HCT116 data from 0.08 to 100 μM, the limit of 100 μM being considered as totally inactive. This led to respectively two assay-based sets of 216 and 124 structures out of the initial 484 BPHI compounds.

As mentioned above, an initial PLS model performed on the entire set of molecules did not lead to a satisfactory predictive model for each biological data set. We decided to subdivide each of the above-mentioned assay-based sets into subsets with respect to the presence of an acidic group (or related) or an amide group (or related) at the 3a position of the BPHI bicyclic system leading to four subsets. "Acid-like" compound sets include acids and esters. "Amide-like" compound sets include pure amides, hydroxamates, and hydrazides. On each of these subsets a D-optimal design was performed.

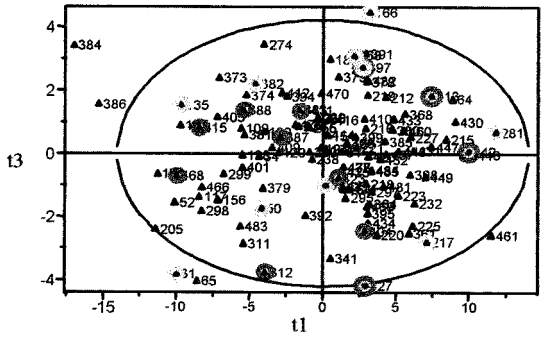


Figure 3. Molecules selected by D-optimal design for amide-like family in PP-space (t_1 and t_3 scores) for both Ki-Ras and HCT116 biological responses, respectively shown in black and gray.

All four D-optimal designs resulted in second-degree polynomial models with three linear terms (i.e. the PCA t scores t_1, t_2, t_3), three quadratic terms (i.e. t_1^2, t_2^2, t_3^2), and three cross-terms (i.e. $t_1 \times t_2$, $t_1 \times t_3$, $t_2 \times t_3$). The summary statistics of these D-optimal designs are listed in Table 1. The distribution of selected compounds for the amide-like set is shown in Figure 3. D-Optimal yielded four subsets of 16 acids and 12 amides for the Ki-Ras data set and 13 acids and 11 amides for HCT116 that were used as training sets for deriving corresponding PLS models.

The high explained variances as well as the predictive power for all these models (R^2 ranging from 81% to 97% and Q^2 ranging from 69% to 84%) show that they are representative of the corresponding data set, therefore validating the four subsets defined as described above. Two quality criteria for such design models are accessible within MODDE and were examined: the $G_{efficiency}$[37,38] (G_{eff}) and the condition number. G_{eff} compares the efficiency of a D-optimal design to a fractional factorial (=100%). A G_{eff} above 60% is recommended. The condition number is a measure of the orthogonality of the design: the lower this number the better the design. Orthogonality is indicated by the value 1. A value below 8 is recommended. The data shown in Table 1 (G_{eff} respectively of 70% and 79% for the Ki-Ras models and 68% and 80% for the HCT116 models as well as condition numbers in the range of 4−7.0) denote that the D-optimal design efficiencies are satisfactory for all these models.

These models allowed the prediction of biological corresponding activities for the remaining compounds of each subset with respectively 76%, 75%, 92%, and 82% of compounds correctly predicted within 1 log unit of the observed value (see Table 1).

3. RSM: Mapping Regions of Interest. The RSM methodology was used to interpret the models by scrutiny of 2D or 3D plots displaying the biological response variation in the local PP-space. The postulated

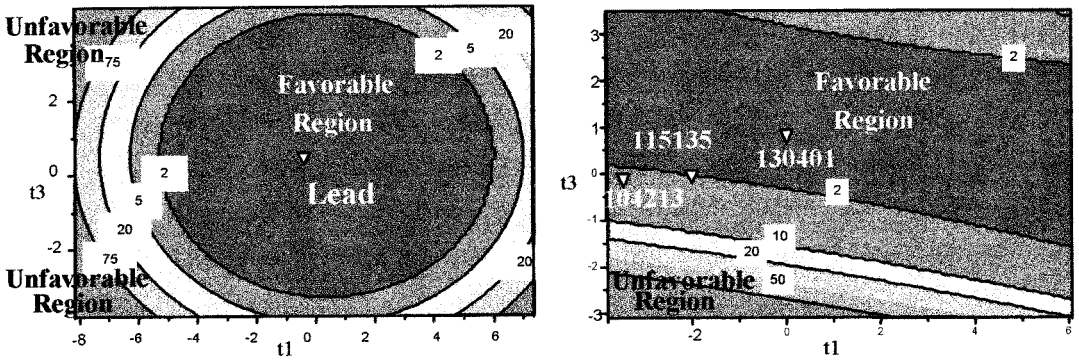


Figure 4. RSM graphs pointing out favorable and unfavorable regions in the PPs (t_1 and t_3 scores) for the acid-like family for both Ki-Ras (left) and HCT116 (right) biological responses.

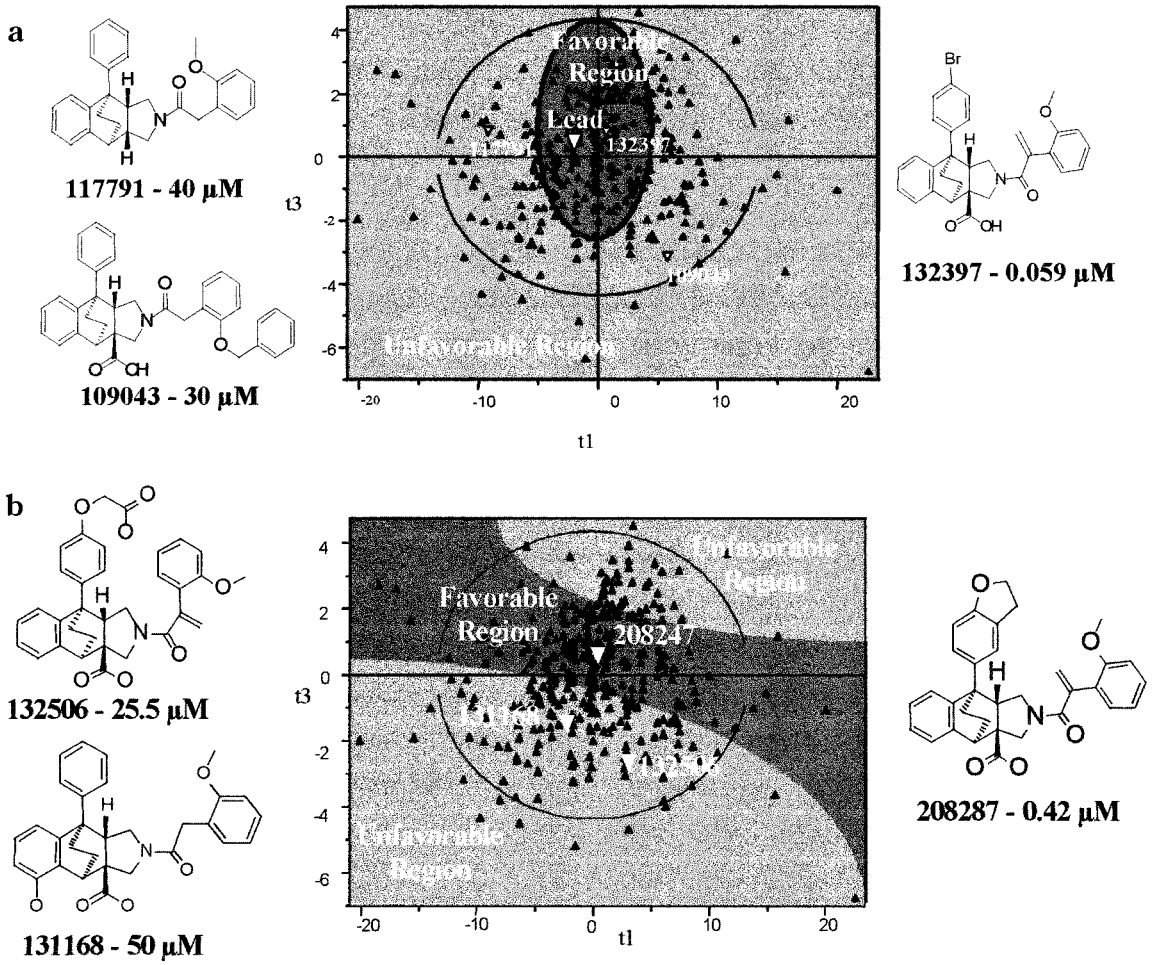


Figure 5. RSM regions of Figure 4 back-projected in reference PP-space for the acid-like family: (a) Ki-Ras and (b) HCT116.

nonlinear model including the quadratic and interaction terms allowed some flexibility to check the impact on the predictive power of these terms. We observed that by omitting such terms statistical parameters of the resulting models were significantly affected (data not shown). According to this it seems that the biological responses for our compounds are nonlinear. We used this tool in order to identify the subspace of interest for Ki-Ras and HCT116 responses and the promising "directions" for the design of new BPHI analogues. In Figure 4 are shown the acid-like family contour plots for t_1 and t_3 scores for the biological Ki-Ras and HCT116 response models (contour values in μM). We can notice for Ki-Ras that the favorable region for active molecules is concentrated in the center of the plot and very unfavorable regions are found at the corners. In the HCT116 response contour plot the favorable region is distributed along a narrow channel and a very unfavorable region is identified for highly negative t_1 and t_3 values. Interestingly when we plot the three reference compounds RPR104213, RPR115135, and RPR130401, the direction in the RSM plot is in agreement with their relative potency (see Figure 1). In addition we also observe that the favorable region for the HCT116 response is embedded in the Ki-Ras one. As a consequence, as we optimize the in vitro Ki-Ras activity, the model predicts that we may also optimize the cellular HCT116 activity.

The reliability of these models developed with a few informative compounds was checked by projecting com-

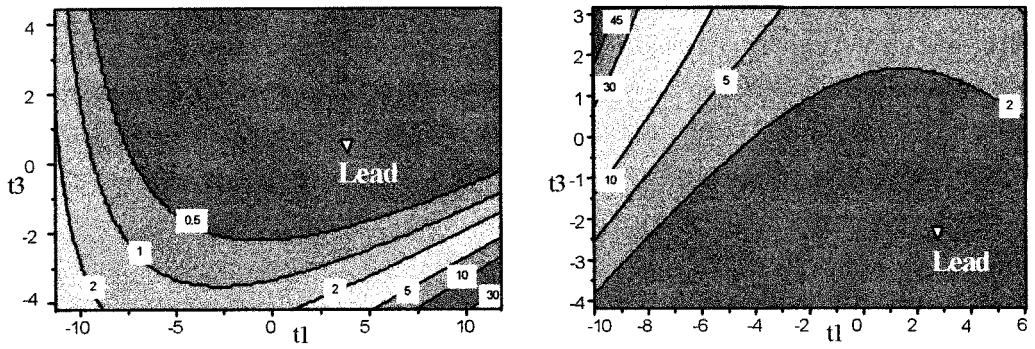


Figure 6. RSM graphs pointing out favorable and unfavorable regions in the PPs (*t* scores) for the amide-like family for both Ki-Ras (left) and HCT116 (right) biological responses.

pounds back onto the corresponding reference PP-space. Figure 5a shows the acid-like compounds (Figure 4) back-projected onto the reference PP-space for the Ki-Ras model. We observed that compounds having *t* scores corresponding to favorable RSM regions indeed were all active. No inactive structure was located in this region. By scrutiny of some structures spread over different regions we can identify the structural characteristics representative of the favorable and unfavorable regions. Favorable structures bear an acid group at position 3a and small ortho substitution (methoxy group) at the phenyl ring of the side chain, whereas for unfavorable ones, the absence of the carboxylic group at position 3a and the steric hindrance of the ortho substitution are the main disfavorable structural characteristics. Similarly, Figure 5b shows the corresponding plot for HCT116 response. An aromatic ring at position 2 and the substitution nature of the aromatic ring at position 9 discriminate between active and inactive compounds. A more subtle modification at position 5 (see compound RPR131168) led to very unfavorable compounds.

A similar analysis is performed for the amide-like family. Figure 6 shows the RSM plots for Ki-Ras and HCT116 in the same t_1 and t_3 PCA scores (PPs). These figures reveal an important feature: an opposite trend in the RSM plots for favorable activity "directions". This "anticorrelation" between Ki-Ras and HCT116 responses is a major difference with respect to the acid-like subset. The t_3 component is the main component that reveals this phenomenon. Indeed, while the Ki-Ras activity is optimal for positive t_3 values, it is greatly unfavorable for the HCT116 one. Conversely, a good HCT116 activity is obtained for negative t_3 values and positive t_1 values, while poor Ki-Ras responses are observed for these values. To get a deeper insight into the molecular characteristics involved, we investigated the influence of the various original descriptors on the PCA t_3 component. Figure 7 shows the corresponding loadings plot. It appears that seven descriptors are highly influencing t_3 and thus are involved in the explanation of this biological profile. As expected it is not just a single factor which is responsible for this specific profile but the interaction between all these factors. However, just by taking each variable separately into account we may be able to draw some conclusions about the main factors involved. As HCT116 activity increases with negative t_3 values, more potent compounds may be obtained by controlling Hbdonor, polar surface area, and Cic descriptors while trying to design compounds with larger negative values for the

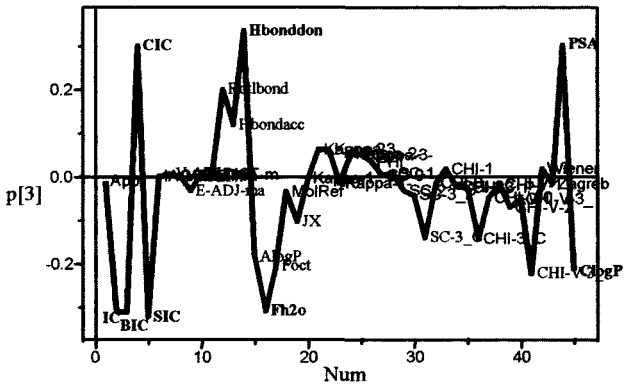


Figure 7. Third PCA loading *p*[3] plotted versus descriptors.

connectivity (IC, BIC, SIC) and desolvation free energy (Fh2o) descriptors.

4. PLS on Regions of Interest. Initially, due to the unavailability of Ki-Ras protein, the first compounds were evaluated using Ha-Ras as protein substrate. We then evaluated the ability of inhibitors to prevent the farnesylation of a Ki-Ras related peptide, which is more relevant according to clinical status of Ras. Consequently, not all BPHI analogues were tested in the Ki-Ras and HCT116 tests. Additionally, in the course of the lead optimization process, we turned our focus only on "pure" acid and "pure" amide compounds. So the examination of more "localized" activity regions for the corresponding subsets, i.e. excluding the related chemical group at position 3a of the BPHI scaffold, is of interest to virtually screen proposals with the aim to identify only those supposed to lead to very potent compounds as potential prototypes for in vivo evaluation. We derived such refined models by excluding those compounds lying in unfavorable areas for each corresponding structural subset. The following PLS models were thus developed in a standard approach by using the original variable space.

Table 2 shows the statistics for the corresponding PLS models. The high explained variances as well as the predictive power for all these models (R^2X ranging from 66% to 80% and R^2Y from 59% to 78%) show that they are representative of the corresponding data set. The values for the predictive power (Q^2) are around 0.6, thus clearly showing that prediction of Ki-Ras and HCT116 biological responses is possible. The error for such predictions is within 0.5 log unit.

The HCT116 models led us to select 12 molecules (predicted activity of less than 1 μM): 5 belonging to the PLS domain, 3 at the frontier of this domain, and 4

Table 2. Summary of the Statistics for Refined PLS Models for Pure Acid and Pure Amide Sets

	Ki-Ras		HCT116	
	acid	amide	acid	amide
R^2X	0.79	0.80	0.77	0.66
R^2Y	0.64	0.59	0.78	0.66
Q^2	0.59	0.56	0.69	0.60
R^2V	(0; 0.06)	(0; 0.05)	(0; 0.07)	(0; 0.07)
Q^2V	(0; −0.03)	(0; −0.05)	(0; −0.09)	(0; −0.06)
Nb of molecules	86	58	31	41

Table 3. Validation Set

	Ki-Ras (μM)		HCT (μM)	
compd	pred	obsd	pred	obsd
1	0.05	0.17	0.96	0.88
2	0.07	0.13	1.20	0.99
3	0.10	0.10	1.30	5.60
4	0.17	0.09	1.80	5.10
5	0.23	0.13	1.20	5.10

out of the PLS expertise domain. These latter have predicted biological activities that are extrapolated, and thus predictions may be rather speculative. All but one were well-predicted. The failure to predict well one molecule can easily be interpreted by the fact that it is totally out of the "expertise" domain of the QSAR model. Activity predictions by model extrapolation are obviously less precise than interpolation, thus more subject to failures.

Such QSAR models were also used to rank structures designed with the aid of the in-house crystal structure of Ftase and FT protein−inhibitor complexes. Five BPHI 3a-amide analogues were designed based on crystal structure and were found located in the most favorable regions of RSM plots for both Ki-Ras and HCT116. The refined models described above predicted these compounds at around 100 nM. Thus they were proposed for synthesis. The observed biological data confirmed these predictions (see results in Table 3).

Initial Set and "Crude" QSAR. We also investigate the applicability of this modeling approach to a real case project at an early stage of lead optimization. This implies the rapid identification of the most interesting as well as informative compounds and more importantly the unfavorable activity regions with a rather small starting data set. If successful, such an approach would be an interesting decision making tool to guide chemistry efforts.

To check if there is a potential of deriving "initial" QSAR models out of a small starting data set and to point out the trends regarding favorable and unfavorable activity regions, we performed an a posteriori analysis on a very small set of 31 compounds (23 actives and 8 inactives) identified at the early stage of this project. The same directional approach as described above was performed. Six molecules were selected by D-optimal (2 inactives and 4 with moderate activity) as the training set for a PLS model. This model's statistics (R^2X_{cum}, R^2Y_{cum}, Q^2_{cum}) are excellent (see Table 1). Using this model a RSM plot (Figure 8) helped us identify one favorable region with molecules having Ki-Ras activities under 5 μM and two unfavorable regions with Ki-Ras activities over 20 μM. The essential SAR trends are rather well-predicted. Indeed, we retrieved the initial lead in the most favorable region (Ki-Ras response under 5 μM), and all the molecules belonging to this region are well-estimated. Conversely, the molecules lying in the unfavorable regions are all inactive. Although based solely on six training compounds, the model is reliable and the corresponding RSM pertinent.

Usually at an early stage of the lead optimization one concern is to identify those features of the lead structure that are mandatory for the activity of interest. It is obvious that one exploitation of such a model is to help identify the minimum requirements for the lead scaffold. As one can see in Table 4, the side chain at position 2 and an aromatic group at position 9 are absolutely critical to obtain active compounds. As we have identified the scaffold of interest, we sought the main SAR trends. For example, the substitution of the 3a position by an aminocarbonyl group instead of a carboxylic or carboxamide decreases dramatically the activity. These results prove that such an approach can be successful even at an early stage of the lead optimization process.

Discussion

In the course of improving the potency of our FPP-competitive Ftase inhibitors (BPHI series) we have investigated the use of a particular QSAR approach, a combined D-optimal−PLS method in PP-space. In any QSAR study the critical point is to obtain a reliable model useful for any analogue design purpose. Such a model relies on an appropriate training set as well as on relevant molecular descriptors.

During the initial data analysis phase based on a PCA (PCA score plot not shown) two clusters consistent with the structural variation observed at position 3a of the BPHI scaffold were identified. The explanation for the few misclassified compounds is straightforward. All

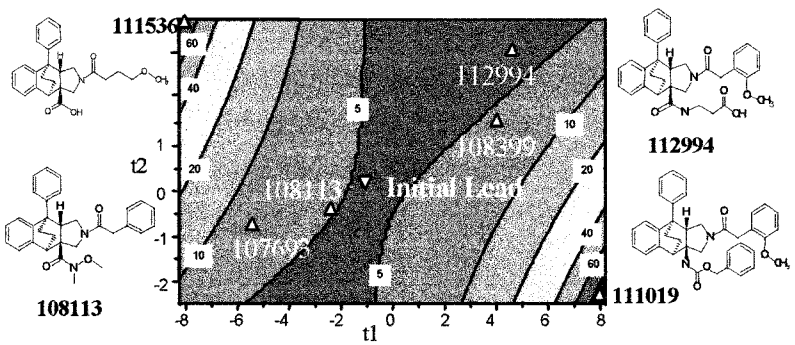
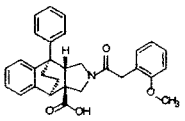
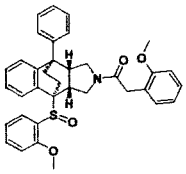
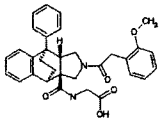
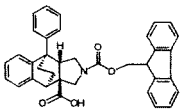
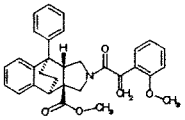
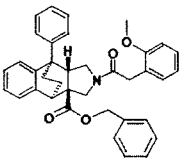
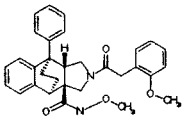
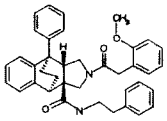
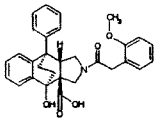
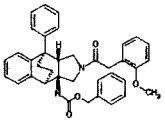
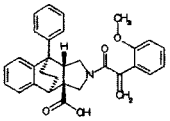
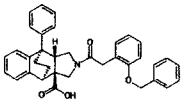
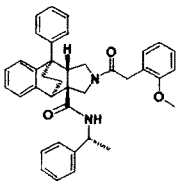
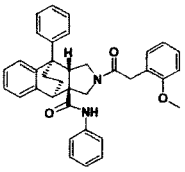
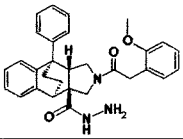
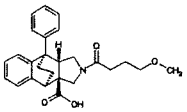


Figure 8. RSM graphs pointing out interesting and unfavorable regions in the PPs (*t* scores) of the initial set for the Ki-Ras response.

Table 4. SAR Data for the Directional Approach on the Initial Set

Favorable Region			Unfavorable Region		
Cmpd	Structure	Ki-Ras SPA (μM)	Cmpd	Structure	Ki-Ras SPA (μM)
110234		0.14	107065		n.d.
113258		0.738	112125		n.d.
113342		n.d.	114912		n.d.
113401		0.42	114081		n.d.
114235		0.25	111019		100
115135		0.025	109043		32
110990		2.08	113579		n.d.
114306		2.45	111536		100

these compounds possessed both chemical groups, i.e. the acidic and amide groups, and thus PCA was unable to unambiguously attribute these structures to a specific structural cluster. PLS-DA is particularly suited for a data set for which some groupings may be embedded as it will reinforce them. PLS-DA plot for t_1 and t_2 scores (Figure 2) confirms the trend observed for the PCA. These results are consistent with the intuitive classification done by project chemists when analyzing both structures and biological responses and thus reinforce the confidence in our descriptor set as representative of our structural data set. According to these results and to the fact that data sets with groupings may lead to difficulties in developing a QSAR model, the data set was divided into subgroups with respect to the structural variation at position 3a of the BPHI scaffold and to the biological responses for further analysis. Indeed the attempt to develop a good PLS model on the whole set of compounds was unsuccessful.

We thus investigated the development of PLS models on the structural subsets for each of the enzymatic and cellular biological data sets using a training set of

information-rich structures identified by D-optimal design experiments, structural descriptors being principal properties (i.e. latent variables) that are well-suited for such techniques. The counterpart of representing compounds in PP-space is that one loses easy interpretability; i.e. it will be more difficult to interpret the factors influencing potency by the original descriptors. But here this is not a major drawback as the ultimate objective of the present study is evaluating a whole set of virtual structures by a sound QSAR model rather than detecting the actual factors influencing potency and consequently fine-tuning the analogue design based on this knowledge.

The strategy consisting of subsetting the whole data set according to the biological response and to the observed structural classes followed by an experimental design to extract the most informative compounds proved to be very efficient in producing sound models (Table 1). According to these results the models can be used with confidence for activity predictions and for use as tools to virtually screen new proposals provided that they are all within the expertise domain of the models.

We also investigated the analysis of activity trends by the RSM methodology. This structure−activity visualization tool revealed a very interesting feature for the amide-like subset. Whereas the Ki-Ras and HCT116 responses are correlated for the acid-like compounds (Figure 4), they are anticorrelated for the amide-like compounds (Figure 6). Although we have identified some major factors influencing HCT116 activity (Figure 7), it is not so straightforward to design compounds with the appropriate characteristics. But the analogue design can be performed in an indirect way by plotting new proposals onto these RSM plots, and in this way one would point out those structures that are potentially of interest for synthesis according to the Ki-Ras/HCT116 profile identified.

Another objective of this work was to identify very potent HCT116 compounds as potential prototypes for an in vivo evaluation and to rank compounds designed with the aid of the in-house Ftase crystal structure. With this objective in mind, we developed refined PLS models to lead to more accurate predictions, within 0.5 log unit. To do so we constrained the PP-space to a limited portion by excluding compounds lying in previously identified unfavorable "activity regions". In this way, we excluded observations responsible for some noise, preventing the setting of a model on the whole initial set and thus leading to the combined method described in this paper. This procedure led to high-quality PLS models for "pure" acids and "pure" amides, using this time original descriptors and thus allowing an easier interpretation in term of structure−activity. Prediction sets were generated and some compounds identified for further development, highlighting the real predictive power of our derived models (see Table 3).

Furthermore, we have shown that such a combined D-optimal−PLS−RSM approach is also applicable at an early stage of an optimization process using a limited set of compounds as a start.

We demonstrate in this paper that the use of multiple designs and multiple models does not lead to any loss of information or understanding. It is just a consequence of our knowledge that a model to have any validity must be based on, and derived from, a set of homogeneous data. The same multivariate and interpretation tools that are used to find the clusters can be used to classify new compounds. Hence predictions and optimizations are achieved as easily with multiple models as with a single one. Moreover, since the separate cluster models are much simpler than any "global" single model (we coud not derive in our case...), the clustering also simplifies the interpretation and the understanding of the results. The idea to divide structure space into clusters is a natural thing to do, considering that this has been done in organic chemistry for a very long time. Finally, one can foresee an application of such an approach combining experimental design, RSM, and multivariate analysis for designing focused, information-rich libraries for lead exploration and/or optimization.

Acknowledgment. We thank Dr. Lennart Erickson for kindly providing a preprint of his publication[19] and D. Clark for his modified version of the polar surface area calculation.

Supporting Information Available: List of the 45 2D descriptors used in this study with their literature references. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Barbacid, M. ras Genes. *Annu. Rev. Biochem.* **1987**, *56*, 779−827.
(2) Bos, J. L. ras Oncogenes in Human Cancer: A Review. *Cancer Res.* **1989**, *49*, 4682−4689.
(3) Casey, P. J.; Solski, P. A.; Der, C.; Buss, J. E. p21ras is modified by a farnesyl isoprenoid. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 8323−8327.
(4) Kato, K.; Cox, A. D.; Hisaka, M. M.; Graham, S. M.; Buss, J. E.; Der, C. J. Isoprenoid Addition to Ras Protein is the Critical Modification for its Membrane Association and Transforming Activity. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 6403−6407.
(5) Graham, S. L. Inhibitors of Protein Farnesylation; A New Approach to Cancer Chemotherapy. *Exp. Opin. Ther. Pat.* **1995**, *5*, 1269−1285.
(6) Reiss, Y.; Goldstein, J. L.; Seabra, M. C.; Casey P. C.; Brown M. S. Inhibition of Purified p21ras Farnesyl: Protein Transferase by Cys-AAX Tetrapeptides. *Cell* **1990**, *62*, 81−88.
(7) Goldstein, J. L.; Brown, M. S.; Stradley, S. J.; Reiss, Y.; Gierasch, L. M. Nonfarnesylated Tetrapeptide Inhibitors of Protein Farnesyltransferase. *J. Biol. Chem.* **1991**, *266*, 15575−15578.
(8) James, G. L.; Goldstein, J. L.; Brown, M. S.; Rawson, T. E.; Somers, T. C.; McDowell, R. S.; Crowley, C. W.; Lucas, B. K.; Levinson, A. D.; Marsters, J. C., Jr. Benzodiazepine Peptidomimetics: Potent Inhibitors of Ras Farnesylation in Animal Cells. *Science* **1993**, *260*, 1937−1942.
(9) Burns, C. J.; Guitton, J.-D.; Baudoin, B.; Lelièvre, Y.; Duchesne, M.; Parker, F.; Fromage, N.; Commerçon, A. Novel Conformationally Extended Naphthalene-Based Inhibitors of Farnesyltransferase. *J. Med. Chem.* **1997**, *40*, 1763−1767.
(10) Qian, Y.; Vogt, A.; Sebti, S. M.; Hamilton, A. D. Design and Synthesis of Non-Peptide Ras CAAX Mimetics as Potent Farnesyltransferase Inhibitors. *J. Med. Chem.* **1996**, *39*, 217−223.
(11) van der Pyl, D.; Cans, P.; Debernard, J. J.; Herman, F.; Lelièvre, Y.; Tahraoui, L. Vuilhorgne, M.; Leboul, J. RPR 113228, a Novel Farnesyl-Protein Transferase Inhibitor Produced by Chrysporium lobatum. *J. Antibiot.* **1995**, *48*, 736−737.
(12) Wallace, A.; Koblan, K. S.; Hamilton, K.; Marquis-Omer, D. J.; Miller, P. J.; Mosser, S. D.; Omer, C. A.; Schaber, M. D.; Cortese, R.; Oliff, A.; Gibbs, J. B.; Pessi, A. Selection of Potent Inhibitors of Farnesyl-protein Transferase from a Synthetic Tetrapeptide Combinatorial Library. *J. Biol. Chem.* **1996**, *271*, 31306−31311.
(13) Aoyama, T.; Satoh, T.; Yonemoto, M.; Shibata, J.; Nonoshita, K.; Arai, S.; Kawakami, K.; Iwasawa, Y.; Sano, H.; Tanaka, K.; Monden, Y.; Kodera, T.; Arakawa, H.; Suzuki-Takahashi, I.; Kamei, T.; Tomimoto, K. A New Class of Highly Potent Farnesyl Diphosphate-Competitive Inhibitors of Farnesyltransferase. *J. Med. Chem.* **1998**, *41*, 143−147.
(14) McNamara, D. J.; Dobrusin, E.; Leonard, D. M.; Shuler, K. R.; Kaltenbronn, J. S.; Quin, J., III; Bur, S.; Thomas, C. E.; Doherty, A. M.; Scholten, J. D.; Zimmerman, K. K.; Gibbs, B. S.; Gowan, R. C.; Latash, M. P.; Leopold, W. R.; Przybranowski, S. A.;

Sebolt-Leopold, J. S. C-Terminal Modifications of Histidyl-N-benzylglycinamides to Give Improved Inhibition of Ras FarnesylTransferase, Cellular Activity, and Anticancer Activity in Mice. *J. Med. Chem.* **1997**, *40*, 3319−3322.
(15) Vrignaud, P.; Bello, A.; Bissery, M. C.; Jenkins, R.; Hasnain, A.; Mailliet, P.; Lavelle, F. RPR 130401, a nonpeptidic farnesyltransferase inhibitor with in vivo activity. *Proc. Am. Assoc. Cancer Res. Meet.* **1998**, *39*, 270 #1846.
(16) Skagerberg, B.; Bonelli, D.; Clementi, S.; Cruciani, G.; Ebert, C. Principal Properties for Aromatic Substituents. A mutlivariate Approach for Design in QSAR. *Quant. Struct.-Act. Relat.* **1989**, *8*, 32−38.
(17) Baroni, M.; Clementi, S.; Cruciani, G.; Kettaneh-Wold, N.; Wold, S. D-Optimal Designs in QSAR. *Quant. Struct.-Act. Relat.* **1993**, *12*, 225−231.
(18) Wold, S.; Sjöström, M.; Carlson, R.; Lundstedt, T.; Hellberg, S.; Skagerberg, B.; Wikström, C.; Öhman, J. Multivariate Design. *Anal. Chim. Acta* **1986**, *191*, 17−32.
(19) Ericksson, L.; Andersson, P.; Johansson, E.; Tysklind, M.; Sandberg, M. and Wold, S. The Constrained Principal Property (CPP) Space in QSAR − Directional and Non-Directional Modelling Approaches. Proceedings for the 12th European Symposium on Structure−Activity Relationships − Molecular Modelling and Prediction of Bioactivity, Copenhagen, Denmark, 1998.
(20) Rhône-Poulenc Rorer S.A. WO 97/03050 and WO 98/29230.
(21) Mailliet, P.; et al. *J. Med. Chem.*, to be submitted.
(22) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters*; Wiley: New York, 1978. Austel, V. A manual method for systematic drug design. *Eur. J. Med. Chem. − Chim. Ther.* **1982**, *17*, 9−16.
(23) Austel, V. Selection of test compounds from a basic set of chemical structures. *Eur. J. Med. Chem. − Chim. Ther.* **1982**, *17*, 339−347.
(24) Clementi, S.; Cruciani, G.; Baroni, G.; Skagerberg, B. Selection of informative structures for QSAR studies. *Pharmacochem. Lib.* **1991**, *16*, 217−226.
(25) Box, G. E. P.; Draper, N. R. *Empirical Model Building and Response Surfaces*; Wiley: New York, 1986.
(26) Myers, R. H. *Response Surface Methodology*; Allyn and Bacon Inc.: Boston, 1971.
(27) SIMCA 6.01; Umetri AB, Box 7960, 907 19 Umea, Sweden; www.umetri.se.
(28) MODDE 4.0; Umetri AB, Box 7960, 907 19 Umea, Sweden; www.umetri.se.
(29) Pearlman, R. S.; Rusinko, A., III; Skell, J. M.; Balducci, R.; McGarity, C. M. CONCORD, version 4.01; Tripos Associates, Inc., 1969 S. Hanley Rd., Suite 303, St. Louis, MO 63944.
(30) Weiniger, D. SMILES, a Chemical Language and Information System 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(31) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.
(32) CERIUS^2, version 3.5; Molecular Simulations Inc., San Diego, CA.
(33) Daylight, release 4.51; Daylight Chemical Information Systems, Inc., 27401 Los Altos, Suite #370, Mission Viejo, CA 92691, and Santa Fe, NM 87501.
(34) Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar Molecular Surface Properties Predict the Instestinal Absorption of Drugs in Humans. *Pharm. Res.* **1997**, *14*, 568−571.
(35) Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; Beigi, F.; Lundahl, P.; Artursson P. Evaluation of Dynamic Polar Molecular Surface Area as Predictor of Drug Absorption: Comparison with Other Computational and Experimental Predictors. *J. Med. Chem.* **1998**, *41*, 5382−5392.
(36) Clark, D. E. Personal communication.
(37) de Aguiar, P. F.; Bourguignon, B.; Khots, M. S.; Massart, D. L.; Phan-Tan-Luu, R. D-optimal designs. *Chemomet. Intell. Lab. Syst.* **1995**, *30*, 99.
(38) Eriksson, L.; Johansson, E.; Müller, M.; Wold, S. Cluster-based design in environmental QSAR. *Quant. Struct.-Act. Relat.* **1997**, *16*, 383.

JM991166H