Predicting ligand binding to proteins by affinity fingerprinting

Lawrence M Kauvar[1*], Deborah L Higgins[1], Hugo O Villar[1], J Richard Sportsman[1], Åsa Engqvist-Goldstein[1], Robert Bukar[1], Karin E Bauer[1], Hara Dilley[1] and David M Rocke[2]

[1]Terrapin Technologies, Inc., 750-H Gateway Boulevard, South San Francisco, CA 94080, USA and [2]Graduate School of Management, University of California at Davis, Davis, CA 95616, USA

Background: There are many ways to represent a molecule's properties, including atomic-connectivity drawings, NMR spectra, and molecular orbital models. Prior methods for predicting the biological activity of compounds have largely depended on these physical representations. Measuring a compound's binding potency against a small reference panel of diverse proteins defines a very different representation of the molecule, which we call an affinity fingerprint. Statistical analysis of such fingerprints provides new insights into aspects of binding interactions that are shared among a wide variety of proteins. These analyses facilitate prediction of the binding properties of these compounds assayed against new proteins.

Results: Affinity fingerprints are reported for 122 structurally diverse compounds using a reference panel of eight proteins that collectively are able to generate unique fingerprints for about 75 % of the small organic compounds tested. Application of multivariate regression techniques to this database enables the creation of computational surrogates to represent new proteins that are surprisingly effective at predicting binding potencies. We illustrate this for two enzymes with no previously recognizable similarity to each other or to any of the reference proteins. Fitting of analogous computational surrogates to four other proteins confirms the generality of the method; when applied to a fingerprinted library of 5000 compounds, several sub-micromolar hits were correctly predicted.

Conclusions: An affinity fingerprint database, which provides a rich source of data defining operational similarities among proteins, can be used to test theories of cryptic homology unexpected from current understanding of protein structure. Practical applications to drug design include efficient pre-screening of large numbers of compounds against target proteins using fingerprint similarities, supplemented by a small number of empirical measurements, to select promising compounds for further study.

Chemistry & Biology February 1995, **2**:107–118

Key words: chemical classification, ligand binding, linear regression, target surrogate

Introduction

We describe here a strategy for identifying functional similarities in the binding sites of proteins that are unrelated by standard measures of structural homology. Since numerous drugs have been shown to cross-react to varying extents with proteins other than their ostensible target [1], contributing to unwanted side effects, identifying functional similarities among proteins is fundamentally important to drug design. A large database of cross-reactivities can be generated by choosing a diverse set of proteins that are all easy to assay, thereby enabling a variety of studies.

One way to analyze the data involves selection of a subset of compounds that have fingerprints that are all quite different from each other. Since the fingerprints reflect protein-binding properties, such a deliberately varied subset is useful as a core screening library. Using a small set of reference proteins, that were carefully selected to be representative of a much larger collection of proteins, we have fingerprinted over 5000 compounds; each fingerprint is the pattern of binding of a single compound to the reference panel of proteins. From this database, we selected a very small core screening library of 54 compounds based on the diversity in their fingerprints, which we designate as a training set. By physically assaying the training set of compounds for their binding to a new target protein, we obtain data that can be compared to the binding of the training set to the reference proteins. Mathematically, this comparison yields a computational surrogate of the target which describes the target in terms of its partial similarities to the various reference proteins. The surrogate created with only the data on the training compounds can then be used to predict the binding of all the other compounds in the database to the target.

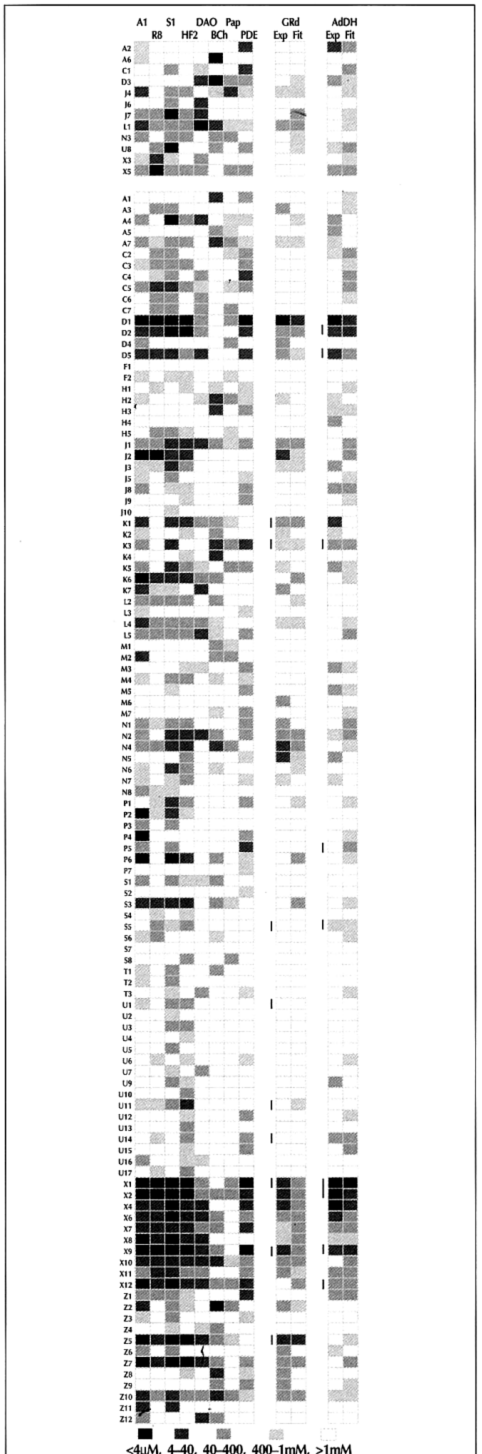
Formally, the new representation of molecules defined here as an affinity fingerprint consists of the collection of binding potencies (IC_{50}) against a panel of reference proteins (R_1, R_2, ... R_n), where the panel members have been empirically selected to provide binding sites

*Corresponding author.

Table 1. Compound library subjectively organized by prominent structural features.

Code	Compound Name
Amines	
A1	Butacaine
A2	4,4 -Diaminodiphenyl sulfone
A3	*p*-Dimethylamino-benzylaldehyde
A4	4,4'-Bis(dimethyl-amino)benzhydrol
A5	Dipyridamole
A6	Fendiline
A7	Glafenine
Cephalosporins	
C1	Cephaloglycin
C2	Cephapirin
C3	Cephalothin
C4	Cephradine
C5	Cephaloridine
C6	Cefoperazone
C7	Cefaclor
Dyes	
D1	Cibacron brilliant red 3BA
D2	Cibacron brilliant yellow 3GP
D3	Acridine orange hydrochloride hydrate
D4	Phenyl 9-acridinecarboxylate
D5	Pyrocatechol violet
Aliphatics	
F1	3-Hydroxy-1-methylpiperidine
F2	Fertilysin
Aromatic heterocyclics	
H1	6,7-Dimethyl-s,3-di-(2-pyridyl)-quinoxaline
H2	Harmaline
H3	Quinine
H4	8-Chlorotheophylline
H5	Murexide
Conjugated aromatics	
J1	2,2'-(1,3-Indenedi-formyl) dibenzoic acid
J2	Citrinin
J3	N-(2-Amino-4-chloro-phenyl)anthranilic acid
J4	Lasalocid
J5	Quinaldic acid
J6	Xanthurenic acid
J7	α-Cyano-3-hydroxycinnamic acid
J8	Flumequine
J9	Nalidixic acid
J10	Norfloxacin
Ketones	
K1	5,5'-Dibromosalicil
K2	4,5 Diphenyl-1,3-dioxolan-2-one
K3	Chalcone
K4	Ketotifen
K5	2-Hydroxy-3-(naphthyl)-1,4-napthoquinone
K6	1,2,3,4-Tetrafluoro-5,8-dihydroxy-anthraquinone
K7	Scopoletin
Phenols	
L1	Nordihydroguaiaretic acid
L2	Dienestrol
L3	Catechin
L4	Naringenin
L5	Hesperetin
Amides	
M1	α-(4-Chlorophenyl)-α(dihydro-oxo-pyridylmethyl)-imidazolyl benzyl alcohol
M2	1,3-Di-*p*-tolyl-2-thiourea
M3	Dansylamide
M4	Nimesulide
M5	Chloramphenicol base
M6	Colchicine
M7	Oxolamine
Nitro-aromatics	
N1	5-(4-Nitrophenyl)-2-furoic acid
N2	N-(4-Dimethylamino-3,5-dinitro phenyl)-maleimide
N3	4,5-Dichloro-2-nitroaniline
N4	2-(2,4-Dinitrostyryl) thiophene
N5	*tert*-Butyl 5-nitro-2-thiophene carboxylate
N6	4-Nitro-N-(2-thienyl-methylene)aniline
N7	N-(5-Nitro-2-pyridyl)-3,4,5,6-tetrachlorophthalamic acid
N8	N-(5-Nitro-3-pyridyl) phthalamic acid
Peptides	
P1	γ-Glu-S-hexyl Cys-Glu
P2	γ-Glu-S-hexyl Cys-Phe Gly
P3	γ-Glu-S-hexyl Cys-β-Ala
P4	γ-Glu-S-octyl Cys-Gly
P5	γ-Glu-S-butyl Cys-Gly
P6	γ-Glu-S-(β-methyl naphthyl)Cys-Gly
P7	Met-Leu-Phe
Steroids	
S1	5α-Androstane-3β, 17β-diol hydrate
S2	Cholic acid
S3	Lithocholic acid
S4	Deoxycholic acid
S5	Chenodeoxycholic acid
S6	Corticosterone
S7	Cymarin
S8	β-Escin
Triazines	
T1	2-Decanoyl-4,6-diamino-1,3,5-triazine
T2	Simazine
T3	4,6-Dihydroxy-1,3,5-triazine-2-acetic acid O-anisidide
Unconjugated aromatic acids	
U1	Bis(4-chlorophenoxy)acetic
U2	2-(4-Biphenyloxy)propionic
U3	2-(4-(Fluorosulfonyl) phenoxy)acetic
U4	2-(4-Benzyloxyphenoxy)-2-methylpropionic
U5	2-(4-*tert*-Butylphenoxy)acetic
U6	γ-Oxo-2-naphthalenebutyric
U7	2-(4-Aminophenoxy)acetic
U8	2-(4-Cinnamoylphenoxy)acetic
U9	2-(4-Formylphenoxy)acetic
U10	Ibuprofen
U11	Indomethacin
U12	Indoprofen
U13	Fenoprofen
U14	(S)-6-Methoxy-a-methyl-2-naphthaleneacetate
U15	Gemfibrozil
U16	Podocarpic acid
U17	Fenbufen
Xanthenes	
X1	Erythrosin B
X2	Phloxine B
X3	Fluoresceinamine, isomer II
X4	Pyrogallol red
X5	Fluorescein isothiocyanate, isomer I
X6	9-Phenyl-2,3,7-trihydroxy-6-fluorone
X7	4-(6-Hydroxy-3-oxo-3H-xanthen-9-yl)benzoic acid
X8	9-(4-(Dimethylamino)phenyl)-2,6,7-Trihydoxy-3H-xanthen-3-one sulfate
X9	6-Hydroxy-3-oxo-3H-xanthene-9-propionic acid
X10	9-(2,4-Dichlorphenoxymethyl)-6-hydroxy-3H-xanthen-3-one
X11	Dimethyl 4-(6-hydroxy-3-oxo-3H-xanthen-9yl) isophthalate
X12	2-(6-Hydroxy-3-oxo-3H-xanthen-9-yl)-cyclohexane-carboxylic acid
Miscellaneous	
Z1	1-Thio-β-D-glucose tetraacetate
Z2	Econazole
Z3	Taxol
Z4	Ajmaline
Z5	6-Chloro-3-nitro-2H-chromene
Z6	Cholecalciferol
Z7	1,1'-Dibenzoylferrocene
Z8	2,5-Diphenyloxazole
Z9	Ethaverine
Z10	Iodonitrotetrazolium Formazan
Z11	1-(Mesitylene-2-Sulfonyl) Imidazole
Z12	Olivetol

About 75 % of randomly chosen compounds, including the wide range of structures shown here, generate fingerprints with the panel detailed in Fig. 1.



which are well diversified with regard to interactions with small molecules. An important practical discovery is that the diversity of interactions attainable with a manageably-sized panel appears to account for most possible modes of interaction, at least to a first approximation. To explore the scope of utility of these data for predicting ligand binding, we attempted to model the binding potency of a compound (*i*) for a specified target protein (T) by a computation performed on the compound's affinity fingerprint. Mathematically [2,3], the computation consists of a linear combination of the compound's potencies against the reference panel of proteins. Thus, the affinity of *i* for T is represented in equation (1) as a summation of the binding potencies to the reference proteins, weighted by statistically-derived coefficients (C^{Rj}). These coefficients, calculated using data collected on the target's interaction with a small set of training compounds, provide the means to compare the target's binding properties to those of the reference proteins. A single set of coefficients, which constitutes the unique surrogate for each target, is used for predictions on all fingerprinted compounds. The coefficients are obtained by a standard method for model selection commonly referred to as multivariate stepwise linear regression.

$$\log (IC_{50})_{i,T} = \sum_{j=1}^{n} C^{Rj} \log (IC_{50})_{i,Rj} \qquad \text{Equation (1)}$$

For this approach to be useful in practice a small panel of reference proteins must suffice for fingerprinting large libraries of compounds. An analogous fingerprinting system is believed to underlie the olfactory system, which can distinguish among millions of different odorants using only a small panel of recognition proteins. In the case of one species of fish studied using recombinant DNA probes for the presumed olfactory receptors, the panel of olfactory receptors consists of only about 50 proteins [4]. For laboratory implementation, the number of reference proteins in the panel should be this small, or even smaller if possible. It is thus clear that the panel members must be selected for two properties aside from ease of assay. Each member must recognize a wide variety of compounds, ensuring high coverage of chemical types, but the variety should be quite different for

Fig. 1. Binding reactivity fingerprints. Semi-quantitative IC_{50} values (gray scale defined at bottom) of the compounds in Table 1 against eight reference proteins: A1, human glutathione S-transferase (GST) A1; R8, rat GST R8; S1, schistosome GST S1; HF2, housefly GST HF2; DAO, porcine D-amino acid oxidase; BCh, equine butyryl cholinesterase; Pap, papain; PDE, snake venom phosphodiesterase I. Experimental binding values (Exp) of the compounds to two different targets is compared to the predictions (Fit) made by the fitted computational surrogates described in Table 2. The two targets are yeast glutathione reductase (GRd) and aldehyde dehydrogenase (AdDH). The first 12 compounds are the first iteration training compounds used for both targets; vertical bars mark the target-specific ten additional compounds used for a second iteration fitting.

each panel member, thereby minimizing the redundancy of the information obtained.

A small set of proteins that meet these criteria have been used to fingerprint a structurally-diverse set of compounds. By applying equation (1) to data derived from testing target proteins against training sets of compounds, we have constructed surrogates of a variety of targets with no previously known homology to any of the reference proteins. The fidelity of such computational surrogates to the actual protein is sufficient to allow surprisingly reliable rank ordering of ligand binding potencies across several log units. By contrast, the fidelity of target models derived from X-ray crystallography allows only the enrichment of high affinity compounds over random selection, with no success at all in rank ordering. These results generalize fundamental questions about the uniqueness of protein binding sites [1,3], and suggest new approaches to defining the mechanisms by which proteins bind small molecule ligands. The results obtained so far suggest that the approach will have broad practical utility by assisting in the discovery of ligands that bind to proteins, for use in general biological research as well as in drug design efforts.

Results

Panel selection

Following preliminary testing of over 150 proteins from a variety of sources, chosen in part because each protein was expected to display broad cross-reactivity with small organic molecules, we selected a set of eight proteins to act as a reference panel for more detailed study. These eight proteins were chosen based on the diversity revealed in correlation tests comparing the binding pattern of each protein to the pattern seen for all of the others when assayed against ligands comprising a wide range of structures. Using this panel of eight proteins, affinity fingerprints can be obtained for compounds from numerous conventionally-defined chemical families. In a survey of several hundred small organic molecules of diverse structure, about 75 % show detectable binding to at least one of these reference proteins, and virtually none bind to more than four reference proteins, a distribution approaching the ideal defined by theoretical models for the olfactory receptor panel [6]. A representative sample of 122 structurally-varied compounds was used for this study (Table 1), and the fingerprints obtained are shown schematically in Fig. 1.

All of these reference proteins have easily assayable enzymatic activity. A compound's binding to each protein was quantified as the concentration needed to inhibit 50 % of that protein's activity (IC_{50}). Although the subsequent analyses were conducted on the actual numerical binding potencies, the precision of the assays used for rapid fingerprinting of the compounds is not substantially better than the gray scale in Fig. 1 indicates. The IC_{50} values obtained range over more than four log units, from 1000 μM to below 0.05 μM.

Training set assays

Once the reference panel was established, a subset of 12 compounds in the 122 compound fingerprint database was chosen, because these compounds exhibited the most diverse binding to the reference proteins. These 12 compounds constitute an initial set of training compounds. The training set was physically assayed for inhibitory activity against two enzymes of possible interest as targets for sensitizing tumors to cytotoxic chemotherapy: glutathione reductase (GRd), and aldehyde dehydrogenase (AdDH). These two proteins are unrelated to each other and to the reference proteins at the level of amino acid sequence, and their enzymatic functions are also distinct.

None of the training compounds is a particularly potent inhibitor against either of the targets. For each target, however, different members of the training set are a little better than the others, and these differences allow the fitting of a computational surrogate for the target, as defined by equation (1). This computer construct, which reflects the partial similarities of the various reference protein binding values to the target's binding values on the training set, can be tested for its ability to predict binding to the target by all other compounds fingerprinted using the same reference panel.

For each target protein (GRd and AdDH), a different computational surrogate was obtained. The surrogate was then used to choose for each target a second set of 10 new compounds that were predicted to be more evenly spread across the range of relevant potencies for inhibitors of the target, with each case generating a different set of 10. Following empirical measurement of the binding values for these second sets of compounds to their respective targets, a second iteration computational surrogate was calculated using the empirical data on all 22 compounds as a training set.

Predictions

For both GRd and AdDH, the particular reference proteins that were most heavily weighted in the first iteration continued to be prominent in the second iteration. These second iteration surrogates applied to the remaining 100 compounds in the database provided predictions which were then compared with the actual empirical values measured separately (Fig. 2); the data on the fitted and predicted values are also listed in Fig. 1. The associated statistical parameters, collated in Table 2, show substantial improvement between the first and second iteration predictions, as measured by standard statistical measures, especially the F-test [2].

Overall, the predictions provide a reliable rank ordering, with a good fit all across the observed span of three log units (from high to low micromolar potency), as indicated by dispersion factors (average scatter around the regression line) of less than 0.5. The dispersion factors partly reflect a known experimental error: in order to collect a complete data set quickly and inexpensively, the

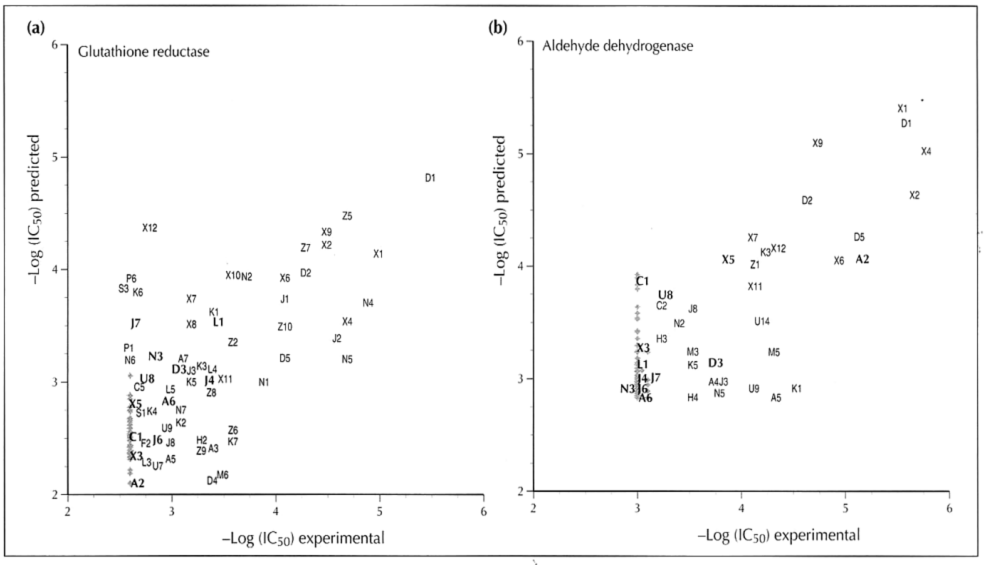


Fig. 2. Correlation of experimentally-measured target binding with predictions. **(a)** Glutathione reductase; **(b)** aldehyde dehydrogenase. Symbols are defined in Table 1; symbols for most of the compounds at the lower limit of measurable potency are replaced by gray diamonds for clarity; the initial training set compounds are in bold face.

majority of the IC_{50} values were calculated by fitting a line to single determinations of five serial dilution concentrations. We estimate from more extensive data collected on a few compounds that about half of the dispersion is attributable to assay error.

To test the dependence of the predictions on the empirically-measured training data, we fitted computational surrogates to a variety of random values. These control experiments yielded no valid correlation with the actual target data on the rest of the compounds. We also compared the actual binding data to that predicted by an arithmetic average of the reference proteins or to a more sophisticated statistical average, the first principal component. Again, there was no meaningful correlation. In short, the results cannot be attributed to any simple type of statistical fluctuation, a point reinforced by the fact that the overall predictive success was similar for two targets that differed markedly in all other aspects of the experiment.

Generality of surrogates

The fitting procedure defined by equation (1) represents a way to mimic an actual target protein by a combination of the reference proteins. Using an evolving set of reference proteins, with the most recent panel comprising 18 proteins, we have attempted to create analogous surrogates for ~40 different proteins. Regression coefficients in the initial fitting were greater than 0.7 for over 30 of them. At least some of the remaining cases are accounted for by proteins to which essentially none of the training compounds binds to any measurable degree.

To explore the generality of the surrogates when applied to a much larger library (over 5000 compounds fingerprinted

Table 2. Fitting procedures and statistical parameters of predictions.

Enzyme	Iteration	Regression equation	Statistical parameters			
			R_{fit}	σ_{fit}	σ_{pred}	F_{pred}
Glutathione reductase	I	0.11BCh+0.19HF2+1.79	0.72	0.22	0.59	4.7
	II	0.21BCh+0.72HF2+0.24S1-0.05	0.85	0.41	0.46	15.9
Aldehyde dehydrogenase	I	0.55PDE+1.35	0.64	0.51	0.46	6.9
	II	0.58PDE+0.25R8+0.43	0.86	0.50	0.48	27.4

First and second-iteration computational surrogates of two targets described in Fig. 2 are created by linear regression of data on training sets of compounds between the target and a panel of reference proteins (codes in Fig. 1). R_{fit} is the coefficient of multiple correlation for the fitted data; σ_{fit} (dispersion) is the residual standard deviation for the fitted equation; σ_{pred} is the standard deviation of the prediction errors when applying the fitted equation to the remainder of the data; F_{pred} measures the improvement of fit as the ratio of dispersion for the current fit compared to the previous iteration, using random data as the initial comparison.

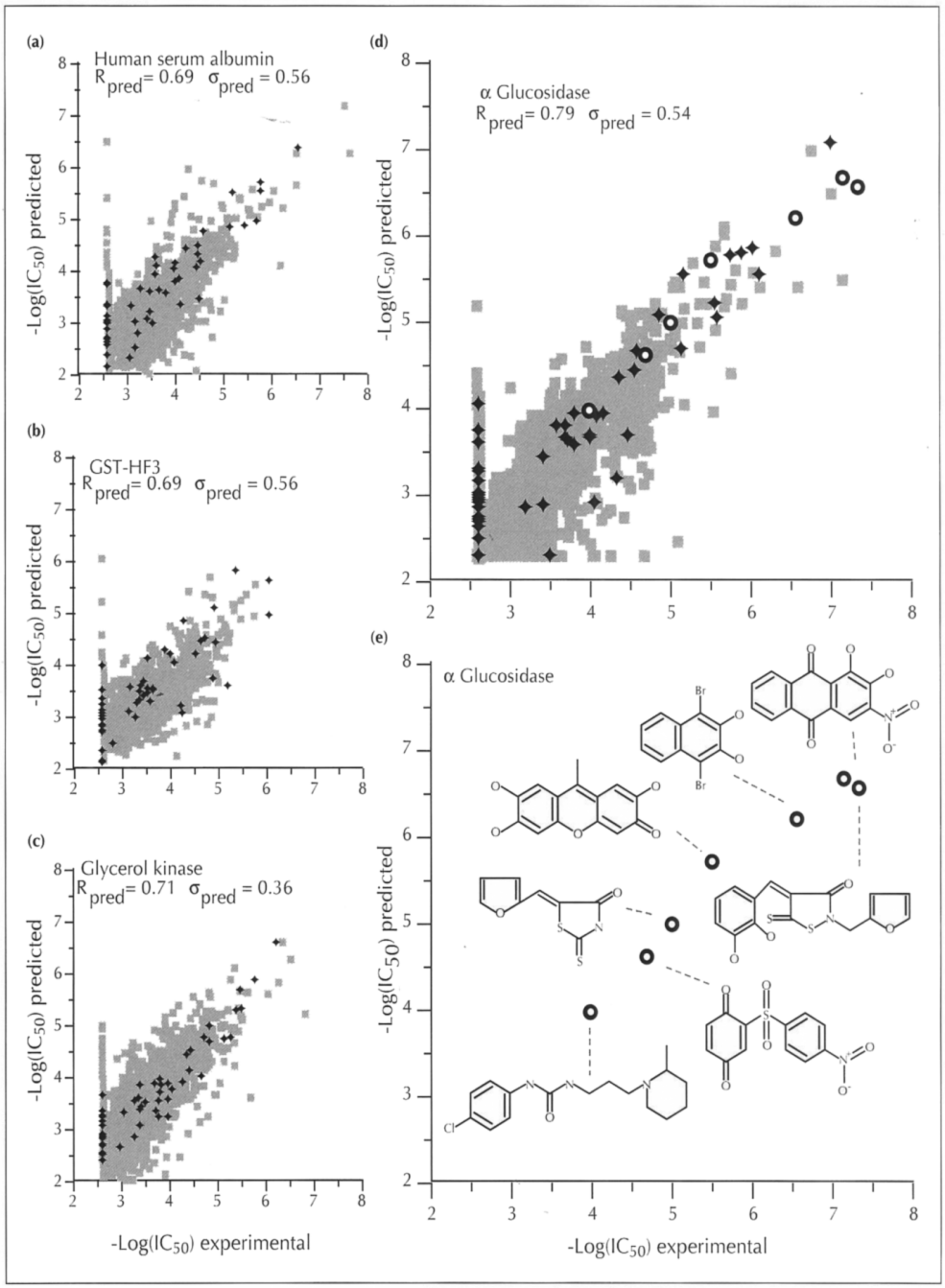


Fig. 3. Generality of computational surrogates. **(a–d)** IC_{50} data for 5000 compounds were collected for the four indicated target proteins. The correlation between a surrogate fit on data for 54 compounds (black diamonds) and the actual target on all compounds (gray squares) is illustrated, along with the statistical parameters R_{pred} and σ_{pred} which measure the overall correlation and dispersion respectively. Lower limit of experimental quantitation is 2.5. For yeast α glucosidase, the structures of compounds indicated by black circles in **(d)** are illustrated in **(e)** to indicate the diversity in structures for which binding was well predicted at several potency levels.

using a reference panel of 18 proteins), four proteins were chosen for further study using automated equipment to measure binding by competitive displacement of labeled ligand, rather than by enzyme activity as in the previous experiments. For each of these targets, data from a somewhat larger training set, consisting of 54 compounds, were used in fitting the surrogates, whose fidelity for predicting the binding to the rest of the 5000 compounds is illustrated in Fig. 3. Qualitatively, the results are the same as those in Fig. 2; the most striking quantitative difference is that the larger library includes compounds that bind several of the targets at potencies below 1 μM, extending the range over which the surrogate's fidelity has been confirmed. A further quantitative difference is that several more terms in the summation in equation (1) have non-zero coefficients for these examples; typically about 7 of the possible 18 terms were important, with the 3–5 most significant ones having roughly equal weightings.

These four examples illustrate applications that would be difficult to pursue without an efficient surrogate screening system. In the first case, a surrogate for a moderately high-affinity transporter site on serum albumin is calculated, demonstrating relevance of the technology to non-catalytic proteins. Transport proteins are important for determining biodistribution of compounds, which is difficult to predict at present. Other transporters, such as those postulated to affect oral absorption of drugs or their passage across the blood brain barrier, may also be accessible to study.

The surrogate for housefly type 3 glutathione S-transferase (GST) is expected to be useful for testing the ability of the technology to discriminate among a large family of isozymes. A variety of reagents are available for this study since we have previously used a focused approach to combinatorial chemistry [7] to develop selective inhibitors of key human GST isozymes [8], an approach which is proving useful in the development of novel cancer chemotherapeutic agents [9,10]. An even larger superfamily of isozymes are the protein kinases. As a first step in this area, we have calculated a surrogate for glycerol kinase. As we seek leads for kinases involved in signal transduction, we intend to use the glycerol kinase surrogate as a rapid predictor for cross-reactivity that would be expected to result in toxicity.

Finally, α glucosidase from yeast has been examined as a readily available prototype for the corresponding human isozyme, which is a target for development of antidiabetic compounds, with additional potential utility for antiviral and anticancer effects [11–13]. The structures of several compounds with varying potency against this target are drawn in Fig. 3, providing a further indication of the variety of structures that are effectively handled by the affinity fingerprinting system.

Characterizing chemical variety

In the experiments on 5000 compounds, the library's structural diversity is even larger than the set in Table 1, including ~300 known drugs and ~100 bioactive peptides. Trends observed in Fig. 1 persist in the larger database. Overall, each compound has a unique fingerprint, despite some degree of clumping not obviously related to common structural motifs. Conversely, compounds that appear to have very similar structures may nonetheless have very different fingerprints. In short, the fingerprints constitute a highly empirical data set, which is not easily explained by other criteria, confirming the widespread experience of other investigators that binding of ligands to proteins is very hard to predict. Having used a functionally diverse set of compounds to select the reference panel, it will be straightforward to reverse the logic and use the panel to assess the functional diversity of other compound libraries in an objective and quantitative manner.

To generate unique fingerprints for all of these compounds, it was necessary to expand the reference panel from 8 to 18 proteins, using similar criteria as before. As with the smaller panel, none of the 18 reference proteins used in the fingerprinting process has any previously recognizable similarity to the target protein. A convenient way to analyze the functional characteristics of the panels used for the experiments shown in Figs 2 and 3, in both relative and absolute terms, is to examine the number of principal components each panel provides. Principal components analysis is a method for extracting from large data matrices, such as that shown in Fig. 1, a minimal number of descriptors that can account for variance in the data [3]. By forming linear combinations of experimental descriptors, a new set of computed descriptors, the principal components, is formed. If the experimental descriptors are all highly correlated in their properties, then just one principal component may account for greater than 95 % of the variance in the data set; conversely, the more uncorrelated the individual experimental descriptors are, the larger the number of principal components that are needed to account for this high a percentage of the variance. In several other areas of research, this mathematical procedure has proven helpful in identifying mechanistically reasonable factors that account for the variance in the data.

For the panel of eight proteins in Fig. 1, 95 % of the variance in the 122 compound data set is accounted for by 6 principal components, compared to 14 principal components for the matrix of 18 proteins against 5000 compounds. For both panels, therefore, the proteins are almost completely uncorrelated in their binding properties, as intended. The most substantive difference between the panels is the absolute range of compounds yielding unique fingerprints, which is ~75 % for the smaller panel and ~95 % for the larger panel.

Additional insight into the functional characteristics of the panels can be obtained by plotting the increments of multi-dimensional variance accounted for by successive principal components (Fig. 4). In spite of the fact that the larger panel is able to fingerprint many more compounds, the overall pattern of principal components is similar in both cases, with the first principal component

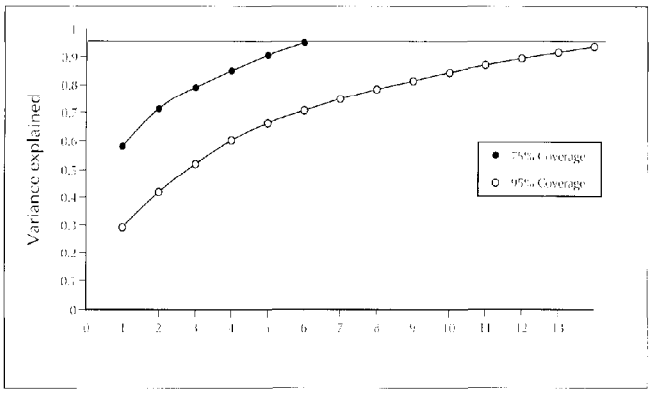


Fig. 4. Increments of multi-factorial variance accounted for by successive principal components derived from fingerprinting reference panels. The panels of proteins are those used for the experiments in Fig. 2 (•) and Fig. 3 (o), and differ primarily with respect to the range of compounds that can be fingerprinted, as indicated.

accounting for a substantial portion of the data. The factor described by this principal component presumably corresponds to some nearly-universal feature of small molecule ligands binding to proteins. The remaining principal components add smaller increments of information, which taper off smoothly. The small amount of variance accounted for by the last few principal components is consistent with the difficulty we have encountered in finding new proteins that are sufficiently independent of previously selected proteins to merit inclusion in an expanded panel. In all, over 300 proteins have now been evaluated to some degree as candidates for expanding the panel. As far as we can tell, including all of them in the panel would not materially change the number or character of the principal components. These results are also consistent with the surrogate-fitting results. Faithful mimics can only be calculated because new proteins tend to bind compounds in ways that are already accounted for by the reference panel.

Discussion

We have demonstrated that appropriate combinations of proteins from a reference panel selected for diversity in binding characteristics can provide reasonable mimics of other proteins. Such data had not previously been consistently collected for any extensive group of proteins, and thus no set suitable to act as a generic reference panel had ever been identified [14]. Since the predictions from the computational surrogates have proven effective across a wide range of potencies, for compounds that differ in a variety of characteristics, they cannot be a consequence of any known single feature, such as hydrophobicity, that is relevant to binding of all ligands to any protein. Hydrophobicity may, however, correspond to some part of the highly-dominant first principal component. These results, which are quite unexpected from comparisons of single proteins to each other, suggest that there are cryptic homologies among proteins that might be discernable statistically, even if they are not readily apparent in any single pair-wise comparison. Identifying mechanistic correlates for the major principal components provides a new, important, and quantitatively well-defined challenge for computational chemistry. The fact that only ~14 factors account for most of the data offers considerable hope for progress in this field.

Fundamental implications

Affinity fingerprinting is a phenomenological method that cannot by itself provide a means for interpreting the computational surrogates of proteins at the level of protein structure. Nonetheless, by providing a quantitative measure of similarity between proteins previously considered quite unrelated, it provides an empirical benchmark for evaluating new ways to compare protein structures. For example, evolutionary arguments imply that actual protein structures are constrained by history to have significant piecewise similarity [15]. Although such similarities may be in fragments too short to identify reliably through primary sequence alignment, they may be identifiable from analysis of secondary structures [16]. As such functionally-relevant molecular features become better understood, the structure-based approach to drug design pioneered by X-ray crystallographers [17] should become applicable to the much larger set of proteins whose detailed structures are not available.

Only a small minority of the proteins included in our reference panels have as yet been studied by crystallography. To obtain some preliminary insight into possible factors accounting for our results, we have therefore turned to an examination of known protein structures, beginning with a study of the amino-acid use in binding sites compared to bulk protein [5]. As is well known, the central core of most globular proteins contains a high frequency of hydrophobic amino acids, while the surface contains predominantly hydrophilic residues. In a survey of 50 diverse crystal structures, each of which has been solved with a bound ligand in place, we discovered that the amino acids in close proximity to the ligand are distinct from either of these patterns. Large residues, such as Arg and Trp, are substantially over-represented compared to their abundance in bulk protein, greatly reducing the number of possible permutations of amino acids in binding sites compared to that expected assuming all 20 were equally used. With only a few permutations, only a small reference panel should be needed to represent the most common motifs. We are now

studying patterns of pair-wise interactions in this reduced set in an effort to identify recurring patterns that may account for some of the principal components defined by our phenomenological data.

Computationally, the regression analysis used to predict binding is analogous to the traditional drug design technique of quantitative structure activity relationship (QSAR) [18], but applied to biochemical parameters that relate more to the target protein than the traditional parameters, which relate to physicochemical properties of the ligands. Conventional QSAR has established that general physical features of ligands are important for binding to proteins, but its utility has largely been limited to comparing compounds with quite similar structural backbones [19,20].

Practical consequences

Affinity fingerprinting, like other computer methods for quantitative chemical classification, is inherently an efficient approach to finding high-potency compounds. For comparison, direct physical screening, with thresholds set at 1 to 10 μM, yields compounds meriting further study at a rate of only about 0.01 %, or lower depending on the diversity of the library. Examination of Figs 2 and 3 show that in each of the fingerprinting examples, the rate for correctly predicting compounds at this potency level was about 70 %.

This result compares quite favorably to the rate of about 10 % in one of the best-studied cases of calculating binding energies by docking compounds into a crystal structure model [21]. Even with a high resolution structure available, our theoretical understanding of how ligands bind to proteins is currently too limited to allow predictions that produce any meaningful correlation coefficient in rank ordering of binding potencies. By contrast, the empirically-driven fingerprinting approach has yielded correlation coefficients in the range of 0.7 for a variety of proteins. The limitations of pure computational methods do not appear to be related to available computer size, which has grown exponentially for decades without a corresponding improvement in prediction accuracy. It seems more likely that what we know how to compute is, as yet, fundamentally incomplete.

In light of the theoretical efficiency of computational methods, the chief practical advantage of the more burdensome direct screening methods is their completeness, or low false negative rate. It is clear from inspection of Figs 2 and 3 that reducing the false negative rate, by lowering the threshold for selecting compounds, necessitates more confirmatory screening to overcome the resulting higher false positive rate. The trade-off between acceptable false positive and false negative rates can thus be adjusted based on the difficulty of follow-up secondary assays.

Few drugs are actually discovered in primary screening, however, since refinement of the leads is generally needed, if only to address issues of stability and manufacturing ease. Because comparison of leads with a diversity of structures is highly advantageous during optimization [22], the range of compounds that can be surveyed by affinity fingerprinting is a particularly attractive feature. Compounds successfully fingerprinted have come from conventional synthetic libraries, natural product libraries, and peptide combinatorial libraries, which have attracted interest as a source of numerous, although not necessarily diverse, chemicals [23]. Once a variety of leads with different structures are available, including both active and inactive analogs, computational methods designed to extract the key features of the ideal ligand (the pharmacophore) can be used to search structural formula databases for other useful ligands, lowering the false negative rate still further. Such three-dimensional database searches are of limited utility as a primary screening tool, however, since 15–20 high affinity compounds are typically needed to build a good pharmacophore model [19].

Practical implementation

Affinity fingerprinting operates in two phases. In the database collection phase, a suitable reference panel of proteins is assembled and assayed against a large library of compounds. From these data, a small subset of compounds with diverse fingerprints is chosen to act as a training set. In the second phase, the following steps are carried out for each new target protein. First, the training set is physically assayed against the target. The resulting pattern of IC_{50} data is then compared to the patterns of binding of the training compounds to the reference proteins, using stepwise linear regression software to calculate a mathematical surrogate of the target. The surrogate, an equation, is applied to the rest of the compounds in the database to yield estimated binding potencies against the target. Based on the predictions calculated in this manner, a collection of compounds is selected for a second round of physical assays. This set is chosen to test the accuracy of the surrogate across the full range of potencies as well as to pick out the most promising compounds for further study. The surrogate can then be revised iteratively.

By its nature, this search process is most efficiently implemented by centralizing the database collection and fingerprint analysis work, while leaving the target-specific assay work as a highly-decentralized function. The technology should thus be usable for exploring biological functions of numerous proteins, in much the same way as mutations are used today [24]. Aside from the commercial utility of providing lead compounds for drug development, this pharmacological approach to basic research offers more temporal control in regulating protein function than is normally possible by genetic methods, particularly in mammals for which conditional mutants are difficult to obtain.

Improvements

There are many ways in which it may be possible to improve the fingerprinting process further, such as using robotic techniques to increase throughput. We may also be able to increase the absolute range of the panel, for

example by using enzymes from the cellular toxic defense network, which often show broad cross-reactivity [25], or enzymatically silent binding proteins such as antibodies [26].

Initially we used extrapolation of binding data from moderate-quality training compounds from our initial set to choose a second group of training compounds. Additional iterations will allow us to use interpolation between the values for the high-quality training compounds discovered in the early rounds, which should be more accurate. Data smoothing techniques [27] may also help to overcome assay noise; recursive partitioning [28] and non-linear fitting [29] may also prove useful in striking a compromise between accuracy in rejecting low-affinity compounds and precision in predicting potency of high-affinity compounds. Deconvolution of multiple mechanisms of binding, such as at the active site and at an allosteric site, will benefit from more powerful techniques than the very simple linear regression methods used in this initial study.

It may be possible to increase the resolution of the cross-reactivity classification system by methodically adding binding proteins with overlapping specificities [30], which should have a considerable advantage over using individual antibodies as surrogates of drug targets, as has previously been suggested [31]. Selection of proteins with suitably-nested specificities should be feasible by probing recombinant libraries of proteins with appropriately-diversified training sets of compounds.

Significance

The commonly-used physical representations of a molecule's structure do not directly indicate its biological properties. For the most part, desirable pharmacological effects result from non-covalent binding to a target protein; unwanted side effects may arise from cross-reactivity with other proteins. A characteristic affinity fingerprint for a particular molecule can be generated by assaying its pattern of affinities towards a standardized panel of proteins, chosen to be highly independent in their binding properties. The fingerprint can be used to estimate cross-reactivity more generally, creating a new approach to estimating toxicity early in the drug design process.

In its simplest application, this approach to chemical classification provides an objective and quantitative means of assessing functional diversity of chemical libraries that is independent of current methods, which are based on analysis of structural formulae. It should therefore be useful in selecting well-diversified core screening sets from conventional chemical libraries, allowing the existing limited quantities that exist for most compounds to be conserved for follow-up screening. It also provides a means to guide combinatorial chemistry efforts towards construction of libraries that provide high diversity, not just large numbers.

Because the use of affinity fingerprints to construct computational surrogates of target proteins has proven useful for predicting binding of compounds with a very wide range of structures, it should be feasible to 'translate' products of combinatorial chemistry, including peptides, into small organic molecules with desirable properties for use as human therapeutics.

The inherent efficiency of affinity fingerprinting expands the scope of approachable drug targets by drastically reducing the number of direct assays of the target's biological activity needed to discover productive leads. This is particularly important when the target protein has not been purified to homogeneity, is unstable or is otherwise not available in adequate quantities for large scale screening, or when the assay procedure is complex and costly, as is the case for targets relevant to many of the major unmet medical needs.

Materials and methods

Reagents

Solutions were prepared using reagent grade materials purchased from several vendors. The compounds listed in Table 1 were purchased from Aldrich Chemical Co. (Milwaukee, WI), Sigma Chemical Co. (St. Louis, MO), or were synthesized at Terrapin Technologies [9]. For assay, the compounds were weighed, and dissolved in water if soluble, otherwise in 100 % DMSO, to make a stock solution of 5 mM which was then diluted for assay.

Reference panel proteins

All glutathione S-transferases (GST, E.C. 2.5.1.18) were recombinant homodimeric forms; provided by B. Mannervik (Univ. Uppsala): human A1, rat R8; provided by M. Syvanen (UC Davis): housefly HF2; from Pharmacia (as part of a fusion protein cloning vector): schistosome S1. D-Amino acid oxidase from porcine kidney (DAO, E.C. 1.4.3.3) was from Sigma, butyryl cholinesterase from horse serum (BCh, E.C. 3.1.1.8), papain (Pap, E.C. 3.4.22.2), and snake venom phosphodiesterase I from *Crotalus adamanteus* (PDE, E.C. 3.1.4.1) were from Worthington. The proteins used for the expanded panel experiments summarized in Fig. 3 are qualitatively similar to this core set, but include a variety of proprietary variants as well (to be described elsewhere).

Target proteins

Yeast glutathione reductase (GRd, E.C. 1.6.4.2), aldehyde dehydrogenase (AdDH, E.C. 1.2.1.5), and human serum albumin (Cohn fraction V) were from Sigma. Yeast α glucosidase (E.C. 3.2.1.23), and *Candida* glycerol kinase (E.C. 2.7.1.30) were from Boehringer Mannheim. Housefly HF3 GST (E.C. 2.5.1.18) was obtained from M. Syvanen (UC Davis); for fitting this enzyme, the HF2 GST was not included in the reference panel; sequence homology to mammalian GST enzymes, which contribute a small part to the fitting equation, is below 15 %.

Assays

GST activity was determined in a microplate assay using 1 mM each of GSH and 1-chloro-2,4-dintrobenzene (CDNB) in 200 mM sodium phosphate, pH 6.8 [8]. Five-fold serial dilutions of each compound, from 250 μM to 0.4 μM, were tested. The 50 % inhibition concentration (IC_{50}) was calculated from a line fitted to the data; for compounds with estimated IC_{50} below 0.4 μM, additional dilutions were tested until the true IC_{50} was bracketed. BCh was assayed in 50 mM sodium phosphate buffer, pH 7.2, containing 250 μM 5,5'-dithio-bis(2-nitrobenzoic acid) and 250 μM S-butyrylthiocholine chloride at 30 °C; the absorbance was monitored at 405 nm for 5 min. PDE activity was measured at 405 nm for 5 min. The assay mixture contained 200 μM *p*-nitrophenyl thymidine-5-phosphate in 0.11 M Tris–HCl, 0.11 M NaCl and 15 mM $MgCl_2$, pH 8.9. Papain was diluted in 1.25 mM EDTA, 0.07 mM 2-mercaptoethanol and 6.25 mM cysteine-HCl and assayed in 200 mM sodium phosphate, pH 6.8, containing 100 μM S-2302 (H-D-Pro-Phe-Arg-*p*-nitroanilide) from Pharmacia Hepar (Franklin, OH). Absorbance at 405 nm was measured for 5 min. DAO activity was measured in a coupled reaction with peroxidase at 405 nm. The assay contained 2 mM D-alanine, 0.25 mM *o*-dianisidine, and peroxidase in 0.02 M sodium phosphate, pH 8.3.

GRd activity was determined in microplates at 30 °C in 100 mM potassium phosphate buffer pH 7.4, containing 1 mM EDTA, 0.2 mM NADPH, and 0.5 mM oxidized glutathione (GSSG). The reaction was followed by monitoring the absorbance decrease at 340 nm as NADPH was consumed. AdDH activity was measured in 0.1 M sodium phosphate, 1.5 mM dithiothreitol, pH 7.5 containing 125 μM acetaldehyde and 10 mM NAD at 30 °C. Absorbance at 340 nm was read for 5 min.

Assays for the experiments in Fig. 3 were designed for a high-throughput robotic format. For each protein, a fluorescent tracer was chosen that competes for binding with known ligands for that protein. The IC_{50} in these cases represents displacement of the fluorescent tracer, as measured by determining fluorescence polarization; free tracer tumbles in solution faster than the fluorescent lifetime, causing a loss of polarization upon emission, while tracer bound to a much larger protein tumbles slowly enough to retain polarization [32]. A customized 96-well plate reader for measuring fluorescence polarization was purchased from Jolley Consulting and Research (Chicago, IL). Data from serial dilutions on each compound (50, 5, 0.5 and 0.05 mg ml^{-1}) was fitted to a four-parameter logistic function to estimate the IC_{50}.

The assays were performed in 0.1 M Na-phosphate buffer, pH 7.5. Each fluorescent tracer was diluted to give a signal-to-noise ratio in fluorescence units of at least 100, typically requiring a tracer concentration of about 50 nM. The appropriate protein was then added to a concentration sufficient to increase fluorescence polarization by at least 0.1 polarization unit over that of free tracer, typically requiring about 10 μg ml^{-1} protein.

Stepwise linear regression fitting

In this method, an initial model which uses no predictors is first established, and predictors from the candidate set (the reference proteins) are cyclically evaluated for inclusion in the model. Candidates are removed if the associated F test value is less than 4.0 (equivalent to a t-statistic of 2.0). Of the remaining candidates in each cycle, the reference protein with the largest F value (tightest overall clustering) is weighted into the model (equation 1) by assigning the slope of the regression line as that protein's coefficient. The process terminates when all variables in the equation, but none of the excluded variables, have F statistics of at least 4.0. The F statistic is a ratio of the dispersions for the predictor of interest compared to the best previous prediction, using random data for the initial step. The dispersion is a measure of how tightly the data cluster around the regression line comparing the test predictor to the actual data; clustering in turn is quantified as the average variance. Table 2 summarizes the statistical parameters, calculated using the S plus v. 3.1 software package (StatSci, Seattle, WA).

Acknowledgements: For guidance throughout the project, we thank John Tainer, Harel Weinstein, John Sedat, and Mike Bishop; for help in clarifying the presentation of the results, we thank Mike Venuti and Fred Cohen. We thank Christopher Berry for technical assistance in the initial experiments and Carol Topp for editorial and artistic assistance.

References

1. LaBella, F.S. (1991). Molecular basis for binding promiscuity of antagonist drugs. *Biochem. Pharm.* **42**, 51–58.
2. Montgomery, D.C. & Peck, E.A. (1992). *Introduction to Linear Regression Analysis*. Wiley, New York.
3. Massart, D., Vandeginste, B., Deming, S., Michotte, M. & Kaufman, L. (1988). *Chemometrics*. Elsevier, New York.
4. Ngai, J., Dowling, M.M., Buck, L., Axel, R. & Chess, A. (1993). The family of genes encoding odorant receptors in the channel catfish. *Cell* **72**, 657–666.
5. Villar, H.O. & Kauvar, L.M. (1994). Amino acid preferences at protein binding sites. *FEBS Lett.* **349**, 125–130.
6. Lancet, D., Sadovsky, E. & Seidemann, E. (1993). Probability model for molecular recognition in biological receptor repertoires: significance to the olfactory system. *Proc. Natl. Acad. Sci. USA* **90**, 3715–3719.
7. Kauvar, L.M. Panels of analyte-binding ligands. August 23, 1994. US Patent No. 5,340,474.
8. Flatgaard, J.E., Bauer, K.E. & Kauvar, L.M. (1993). Isozyme specificity of novel glutathione S-transferase inhibitors. *Cancer Chemother. Pharmacol.* **33**, 63–70.
9. Lyttle, M.H., *et al.*, & Bauer, K.E. (1994). Isozyme-specific glutathione S-transferase inhibitors: design and synthesis. *J. Med. Chem.* **37**, 189–194.
10. Lyttle, M.H., *et al.*, & Kauvar, L.M. (1994). Glutathione S-transferases activate a novel alkylating reagent. *J. Med. Chem.* **37**, 1501–1507.
11. Lebovitz, H.E. (1992). Oral antidiabetic agents: the emergence of alpha-glucosidase inhibitors. *Drugs* **44**, 21–28.
12. Elbein, A.D. (1991). Glycosidase inhibitors: inhibitors of N-linked oligosaccharide processing. *FASEB J.* **5**, 3055–3063.
13. Asano, N., Oseki, K., Kizu, H. & Matsui, K. (1994). Nitrogen in the ring pyranoses and furanoses: structural basis of inhibition of mammalian glycosidases. *J. Med. Chem.* **37**, 3701–3706.
14. Kauvar, L.M. (1993). Pharmaceutical targeting of GST isozymes. In *Structure and Function of Glutathione S-Transferase*. (Tew, K.D., Pickett, C.B., Mantle, T.J., Mannervik, B. & Hayes, J.D., eds), pp. 257–268, CRC Press, Boca Raton, FL.
15. Dorit, R.L., Schoenbach, L. & Gilbert, W. (1990). How big is the universe of exons? *Science* **250**, 1377–1382.
16. Prestrelski, S.J., Williams, A.L. Jr. & Liebman, M.N. (1992). Generation of a substructure library for the description of protein secondary structure. *Proteins* **14**, 430–440.
17. Cohen, N.C., Blaney, J.M., Humblet, C., Gund, P. & Barry, D.C. (1990). Molecular modeling software and methods for medicinal chemistry. *J. Med. Chem.* **33**, 883–894.
18. Hansch, C. & Klein, T.E. (1986). Molecular graphics and QSAR in the study of enzyme–ligand interactions. On the definition of bioreceptors. *Accounts Chem. Res.* **19**, 392–400.
19. Loew, G.H., Villar, H.O. & Alkorta, I. (1993). Strategies for indirect computer-aided drug design. *Pharm. Res.* **10**, 475–486.
20. Franke, R. (1984). *Theoretical Drug Design Methods*. Akademie Verlag, Berlin.
21. Schoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D. & Perry, K.M. (1993). Structure-based discovery of inhibitors of thymidylate synthase. *Science* **259**, 1445–1450.

22. Andrews, P.R., Craik, D.J. & Martin, J.L. (1984). Functional group contributions to drug–receptor interactions. *J. Med. Chem.* **27**, 1648–1657.
23. Kenan, D.J., Tsai, D.I. & Keene, J.D. (1994). Exploring molecular diversity with combinatorial shape libraries. *Trends Biochem. Sci.* **19**, 57–64.
24. Mitchison, T.J. (1994). Towards a pharmacological genetics. *Chemistry & Biology* **1**, 3–6.
25. Jakoby, W.B. & D.M. Ziegler, D.M. (1990). The enzymes of detoxification. *J. Biol. Chem.* **265**, 20715–20718.
26. Marks, J.D., Hoogenboom, H.R., Griffiths, A.D. & Winter, G. (1992). Molecular evolution of proteins on filamentous phage. *J. Biol. Chem.* **267**, 16007–16010.
27. Koorwinder, T.H., ed. (1993). *Wavelets: an elementary treatment of theory and applications*. World Scientific, River Edge, NJ.
28. Ciampio, A., Chang, C.H., Hogg, S., & McKinney, S. (1987). Recursive partitioning: a versatile method for exploratory data analysis in biostatistics. In *Biostatistics*. (MacNeil, I.B. & Humprey, G.J., eds), pp. 211-228, D. Reidel Publishing, New York, NY.
29. Bates, D.M. & Watts, D.G. (1988). *Nonlinear regression analysis and its applications*. J. Wiley, New York, NY.
30. Cheung, P.Y.K., Kauvar, L.M., Engqvist-Goldstein, A.E., Ambler, S.M., Karu, A.E. & Ramos, L.S. (1993). Harnessing immunochemical cross-reactivity: use of pattern recognition to classify molecular analogs. *Anal. Chim. Acta* **282**, 181-191.
31. Wolff, M.E. & McPherson, A. (1990). Antibody-directed drug discovery. *Nature* **345**, 365–366.
32. Dandliker, W.B., Hsu, M.L., Levin, J. & Rao, B.R. (1981). Equilibrium and kinetic inhibition assays based upon fluorescence polarization. *Methods Enzymol.* **74**, 3–28.

Received: **22 Dec 1994**; revisions requested: **19 Jan 1995**; revisions received: **6 Feb 1995**. Accepted: **6 Feb 1995**.