WizePairZ: A Novel Algorithm to Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry

Daniel J. Warner,*,† Edward J. Griffen,‡ and Stephen A. St-Gallay†

Department of Chemistry, AstraZeneca R&D Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RH, United Kingdom, and Cancer and Infection Research, AstraZeneca R&D Alderley, Alderley Park, Macclesfield, Cheshire. SK10 4TG, United Kingdom

Received March 1, 2010

An algorithm to automatically identify and extract matched molecular pairs from a collection of compounds has been developed, allowing the learning associated with each molecular transformation to be readily exploited in drug discovery projects. Here, we present the application to an example data set of 11 histone deacetylase inhibitors. The matched pairs were identified, and corresponding differences in activity and lipophilicity were recorded. These property differences were associated with the chemical transformations encoded in the SMIRKS reaction notation. The transformations identified a subseries with the optimal balance of these two parameters. Enumeration of a virtual library of compounds using the extracted transformations identified two additional compounds initially excluded from the analysis with an accurate estimation of their biological activity. We describe how the WizePairZ system can be used to archive and apply medicinal chemistry knowledge from one drug discovery project to another as well as identify common bioisosteres.

INTRODUCTION

Matched Molecular Pair Analysis. The concept of a matched pair of compounds, i.e., two compounds that differ from each other by just one or a small, localized group of atoms is one that is familiar to medicinal chemists. Matched pairs are commonly used to help analyze structure activity relationships (SAR) by associating each pair of molecules with a corresponding change in physical or biological properties. It is well-known, for example, that the presence of specific structural motifs such as sterically unhindered N-containing heterocycles is likely to result in cytochrome P450 (CYP) 3A4 inhibition.[1] Furthermore, it is possible to reduce this problem through steric effects, or electronic substitution that disfavors interactions between the nitrogen atom and the CYP heme. In this case, an example of a matched pair of compounds would be one containing the unsubstituted heterocycle and an identical compound with either a bulky substituent hindering the nitrogen atom, an electron withdrawing group at an alternative position, or the nitrogen atom located at a more sheltered position within the ring. Matched molecular pairs can be used across a range of molecular properties. Recent studies have reported how molecular matched pair analysis (MMPA) can facilitate lead optimization against a range of parameters including: aqueous solubility,[2] plasma protein binding,[2] oral exposure,[2] primary potency,[3] and liver microsome stability.[4] The technique resembles Free−Wilson analysis,[5] in which the contribution toward a particular property is estimated for a collection of substituents, yet MMPA is believed to benefit from only requiring calculation of the change in a given property resulting from a single substitution, as opposed to the absolute value for the entire molecule by analogy with free energy perturbation.[3,6]

Automation and Matched Molecular Pairs. The fastest and most common approach to the identification of molecular matched pairs is based on Murcko style core and R-group fragmentation of the molecular structures.[7,8] Using this kind of approach, structures are fragmented according to precise definitions of rings, linkers, side chains, and scaffolds, which are limited to the breakage of acyclic bonds between heavy atoms. Such fragment definitions have been used in the identification of matched molecular pairs by a number of groups.[9−11] Cases where just a single fragment differs between two compounds are considered a matched pair, and the frequency with which fragments are replaced by one another is recorded. The transformation and the frequency with which it is observed is then associated with either the absence of any change to the biological activity in the identification of potential bioisosteric replacements[10,12,13] or differences in measured properties such as biological activity, solubility, plasma protein binding, and log D.[9,11]

While the approach described in the papers cited above has obvious utility, the reliance on the cleavage of single bonds to define molecular cores and identify side chains introduces a number of limitations. For example, a common tactic for reducing metabolic clearance is to replace aryl hydrogen atoms with fluorine, in order to block likely sites of metabolism.[14] While an acyclic bond breakage approach would identify 4-fluoro-phenyl as a favorable replacement for phenyl, a medicinal chemist might consider aryl fluorination to be the same regardless of whether the substitution occurred specifically on a phenyl ring or on a number of other aromatic heterocycles. In this example, the key for capturing the chemical knowledge required to achieve the desired outcome is not only that hydrogen must be replaced by fluorine but that the nature of the local environment (in

* Corresponding author e-mail: dan.warner@astrazeneca.com.
† AstraZeneca R&D Charnwood.
‡ AstraZeneca R&D Alderley.

10.1021/ci100084s
Published on Web 08/06/2010

this case an aryl group) must also be included. The precise nature of the ring containing the aryl carbon in this case is not necessarily relevant. It should be noted that, in contrast to conventional MMPA, WizePairZ does not rely on the use of hydrogen as a "reference" substituent, which can result in structural relationships being missed.

CYP 3A4 activity demonstrates another limitation of the single bond fragmentation approach. For a nitrogen-containing heterocycle, the key knowledge to be captured is movement of the nitrogen to a more sterically hindered position, regardless of whether the exposed atom lies in, for example, a pyridyl or isoqunolinyl ring. An approach to incorporating information of this kind has been described by Sheridan et al. where a substructure descriptor difference vector is used in the definition of molecular transformations.[15] The inclusion of atom type descriptors in this way requires that in order to reduce metabolism, a fluorine atom must be added, while at the same time a carbon atom with two non-hydrogen neighbors and one π electron (type C21) must be converted to a carbon with three non-hydrogen neighbors and one π electron (type C31). Information about the longer range connectivity required for the CYP 3A4 example is introduced with the use of atom pair[16] and topological torsion descriptors.[17] As with single bond fragmentation schemes, maximum common substructure (MCS) has also been applied to the identification of the most common structural transformations in extensive collections of drug-like compounds.[18,19]

In 2006, Abbott Laboratories published Drug-Guru,[20] a medicinal chemistry "expert system" that contained a collection of molecular transformations compiled from the literature and medicinal chemists' experience. The system required a user to provide a molecular structure as input, to which stored transformations were applied where applicable to generate a population of novel compounds. The output structures may then have been evaluated for suitability as chemical targets on the basis of synthetic tractability and their predicted properties according to QSAR/QSPR models. The transformations were stored using SMIRKS,[21] an extension of SMILES notation that stores molecular transformations as text strings. SMIRKS strings can store transformations that include the addition and removal of atoms from a structure within a specific local molecular environment. Furthermore, they can be directly interpreted using visualization software such as Accelrys' Accord plug-in for Excel.[22]

In this paper, we present WizePairZ, a collection of algorithms that combine the key advantages of the methods described above. First, we use MCS searches to identify cores common to both members of a putative matched molecular pair. From this we can determine not only the nature of the molecular transformation translating one molecule into the other, but also the chemical environment within which the transformation occurs. Our method captures different levels of the local environment, from the immediate neighborhood of the transformation up to four bonds out. At each level of the specificity hierarchy, an individual SMIRKS string is encoded and stored in a database, along with any known experimental information associated with transformations of that type. The advantage of storing the extracted knowledge using SMIRKS is that it can be stored in massive, yet rapidly searchable databases, retrieved using SQL statements,[23] and used to suggest possible solutions to medicinal chemistry problems. This is in keeping with one of the principles of expert systems that knowledge should be extracted and stored in the format in which it will be used.

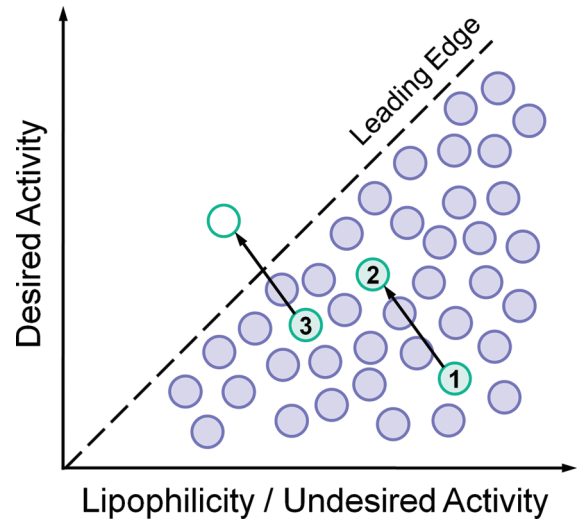
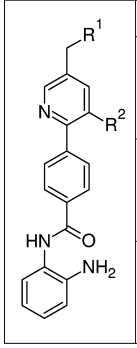
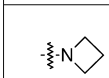
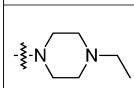
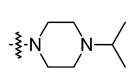


Figure 1. Schematic representation of the medicinal chemistry challenge to push the leading edge. The desired activity, such as primary potency, often tracks with an undesired activity, such as hERG, through a cocorrelation with lipophilicity. A structural transformation between two pairs of compounds, 1 and 2, results in a change in both activities. Applying this transformation to compound 3 predicts a virtual compound that has pushed the leading edge.

One area where this methodology is likely to be of use is in the simultaneous optimization of a compound's physicochemical properties and intended biological effects. Achieving the balance of properties in a molecule that will give the necessary pharmacokinetic, pharmacodynamic, and safety profile is a medicinal chemistry challenge that frequently lies on a knife's edge. Improvements to the desired activity commonly results in the introduction of unwanted characteristics such as insolubility, hepatic clearance, CYP inhibition, or hERG activity, with the root cause of these relationships often being a dependence on lipophilicity.[24,25] Consequently, the challenge is to identify compounds that present an improvement in the desired activity while simultaneously reducing lipophilicity (or any undesired other activity).

Fortunately for medicinal chemists, the properties of molecules are not solely governed by lipophilicity but also depend on the molecular structure. The potency/lipophilicity relationship in Figure 1 can be thought of as a multiobjective optimization problem, where the dotted line marks the "leading edge",[25] also referred to as the Pareto frontier.[26] From Figure 1, a pair of compounds in this data set (numbers 1 and 2) has an activity change (defined by the arrow) and an associated structural transformation. The same transformation can be applied to compound 3, resulting in a suggested compound that is predicted to push the leading edge, i.e., a compound that has an increase in the desired activity with a corresponding drop in the undesired activity. The steps of this approach—identifying matched pairs, associating structural modifications with a change in activity, and then applying them to another compound—are a familiar medicinal chemistry process. WizePairZ uses this approach to automatically consider every possible pair of compounds

Table 1. Data Set Used in This Study (Compounds 13c and 14d Were Excluded from the Matched Pair Analysis)[a]

Structure	R^1	R^2	Compound	logD @ pH7.4	HDAC pIC_{50}*
	-OH	CN	3	1.58	7.77 ± 0.69
		CH_3	13a	1.18	7.67 ± 0.14
		CN	13b	1.22	8.01 ± 0.35
		Cl	13c	NV	7.73 ± 0.06
		F	13d	1.69	7.30 ± 0.24
		CH_3	14a	1.20	7.73 ± 0.40
		CN	14b	1.14	8.01 ± 0.18
		Cl	14c	1.81	7.79 ± 0.25
		F	14d	NV	7.42 ± 0.35
		CH_3	15a	1.44	7.72 ± 0.16
		CN	15b	1.28	7.98 ± 0.07
		Cl	15c	1.68	7.88 ± 0.79
		F	15d	1.60	7.48 ± 0.08

[a] Taken from Andrews et al.[27] The ± values denote 95% confidence intervals for the mean. The intervals are calculated on the basis of multiple measurements of the activity of each compound in JMP.[30]

in a data set (without core definition) including knowledge gleaned from areas that may have been discounted during analysis. Such areas might include regions of relatively high lipophilicity and low potency, which nevertheless may still contain transformations that have utility in the optimization process. The approach may also reinforce evidence supporting individual target compounds by providing multiple transformations from different starting points that propose the same target.

METHODS

Step 1: Identification of Potential Matched Pairs. A subset of the literature data set of 13 histone deacetylase (HDAC) inhibitors, containing a number of matched molecular pairs, was prepared as a table of SMILES strings, pIC_{50}, and log D values (Table 1).[27] Compounds 13c and 14d were excluded from the analysis due to the absence of measured log D values and subsequently used for validation. A matrix of all possible molecular pairs was created such that every compound could be compared with every other compound in the data set, bar itself. The following steps were implemented with the use of the OEChem toolkit[28] and Python.[29]

For each pair of molecules, the two SMILES strings were converted into OpenEye molecular graphs, assigning aromaticity, ring membership, and implicit hydrogen counts. An MCS search was performed on the two molecules, using the default atom and bond expression options, which match atoms with the same atomic number, aromaticity, and formal charge and bonds with the same order and aromaticity (Figure 2). As formal charges were not assigned at the time of graph creation, the formal charge parameter is redundant in this process. The matches with the largest number of mapped atoms were scored most favorably. In the event of a tie, any ring bonds in the template molecule that were not mapped to rings in the target incurred a penalty. A matched pair was defined as having no more than 10% of the structure differing between the two molecules in the initial substructure search. Thus, the minimum number of atoms considered as a match was set to 90% of the number of atoms in the largest

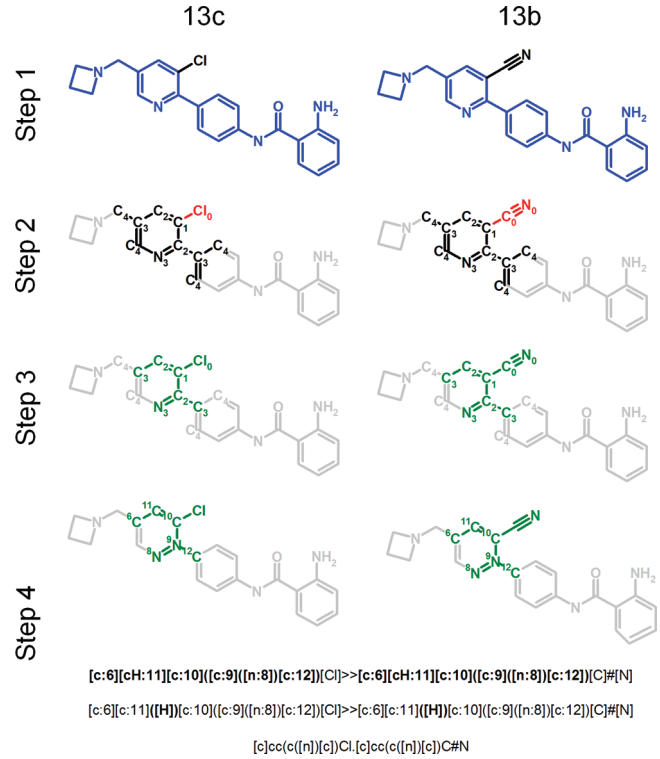


Figure 2. Visual representation of the method for encoding the SMIRKS. MCS is shown in blue, with what remains after elimination of the MCS shown in red. Atomic radii are marked as subscript characters, whereas atom mappings are marked as superscripts. Green denotes the fragment that remains following the deletion of atoms with atomic radii > 3.

of the two molecules. The exhaustive search option was employed in all cases.

Step 2: Matched Pair Verification. To generate a valid molecular transformation, the modification between the matched pairs had to be located at a single site—not spread across the molecules—so all atoms that were not part of the common substructure were "marked" (by setting the atomic radii to zero, Figure 2). In this context, the atomic radii are used simply as a placeholder to mark the distance from the site of modification and do not imply the creation of any 3D molecular conformations. A second iteration over the common substructure identified atoms where either the valence state, implicit hydrogen count, ring membership status, or ring size differed between the two molecules. Atoms where any of these properties differed were assigned an atomic radius of 1. For the remaining atoms, the minimum number of bonds to a "marked" atom was counted and radii assigned as to the number of bonds separating that atom from the noncommon structure. All atoms with a bond count greater than four were deleted to leave an extended version of Sheridan's "remainder after elimination of common substructure" (RECS).[15]

If the RECS for each molecule contained just a single fragment, the molecular transformation was located at a single site and a valid transformation had been identified. Fragment counts of greater than one indicated molecular pairs where the locations were spread across the molecule, and these transformations were not processed.

Step 3: Definition of Local Environment Shells. For valid transformations, a SMIRKS string was encoded (see below), and the outer shell of atoms (i.e., those three bonds

out from the structural modification) was pruned. This process was repeated, decreasing the number of bonds out from the structural modification by one each time until either none of the remaining atoms were part of the common substructure or the number of fragments remaining in either RECS was no longer one, i.e., not a valid transformation. Thus, every molecular transformation might be encoded by up to four SMIRKS strings, each encoding different amounts of the transformation's local structural environment.

Step 4: Encoding the SMIRKS String. SMIRKS strings define atom mappings that relate "reactant" atoms to their counterparts in a "product". These mappings were created by iteration over the RECS and setting the atom map indices for the remaining atoms on the template (reactant) graph to be the same as their atomic indices. The corresponding target (product) atoms were then identified such that their atom map indices could be set to the same values. The atom map indices were all incremented by 1, as the atom indexing in an OpenEye molecular graph begins at 0, while the SMILES creation function will fail to output mappings where the atom map index is less than 1.

At this point, the graphs were converted back into SMILES strings, one for each fragment, now including appropriate atom mappings separated from their respective atomic symbols by a colon. The penultimate step of the encoding procedure was to transfer the hydrogen atoms from within the square brackets to outside (e.g., [C**H2**:8] is converted to [C:8]**([H])([H])**), a manipulation which was performed on the text string directly using regular expressions. While this results in what may appear to be overly verbose fragment descriptions, it was found to be important in guarding against misapplication of the transformation when it was eventually put to use. For example, when methylating a tertiary amine, not only must a methyl group be added to a molecule but hydrogen must also be removed to prevent the transformation from further methylating quaternary amines. Similarly, aryl substitution transformations require hydrogen removal to prevent creation of an sp^3 center and ensure that aromaticity is preserved. Finally, the two SMILES strings were concatenated with "≫" in the center to form the complete SMIRKS string.

Step 5: Aggregation of Transformations. To group together particular transformations with other transformations of the same type, a common identifier was created after the SMIRKS string was encoded. For example, [c:7]([H])≫[c:7][F] clearly encodes the same structural modification as [c:18]([H])≫[c:18][F], yet a common identifier must be defined in order for them to be grouped together. These identifiers were created by removal of all atom mappings from the RECS, and a second call to the SMILES creation function. The two new strings were concatenated with a dot separator, e.g., cH.cF in the case of the transformation above. All instances of transformations with the same identifier could then be aggregated, while only one example of a SMIRKS string need be retained. For each transformation, the mean change in HDAC pIC_{50} and log D were calculated, along with the percentage of observations where each property either increased or decreased.

Step 6: Application of Extracted Transformations. WizePairZ can apply the transformations directly in the enumeration of virtual libraries. As a means of validating the methodology, all SMIRKS strings created at the three-bond cleavage level were used to create a collection of reactions using the OpenEye library generator. Hydrogen atoms were treated by the generator in an explicit fashion with valence correction disabled. The reactions were subsequently applied to each of the 11 seed compounds listed above in turn. After ensuring that only a single compound was generated by the transformation, a SMILES string for the product was reported, alongside an estimated value for the HDAC activity, by summing the measured pIC_{50} for the seed compound and the mean change in activity for that transformation.

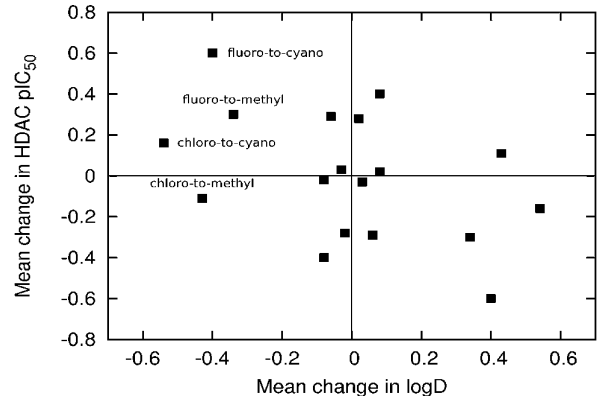


Figure 3. Mean changes in HDAC potency with respect to lipophilicity for the 18 unique molecular transformations identified by WizePairZ. Halo-to-cyano and halo-to-methyl transformations are highlighted in the top left of the plot.

Table 2. Summary of the Individual Matched Pairs from Which the Transformation [c]C.[c]C#N (Methyl to Cyano) Was Derived

compound A	compound B	ΔHDAC pIC_{50}	Δlog D @ pH 7.4
13a	13b	+0.34	+0.04
14a	14b	+0.28	−0.06
15a	15b	+0.26	−0.16
mean change for [c]C.[c]C#N(methyl to cyano)		+0.29	−0.06

RESULTS

Matched Pair Identification. The protocol described above generated 18 unique transformations, covering a total of 36 matched molecular pairs (Figure 3). The two most frequently occurring were the replacement of ethyl with isopropyl at R1 and the replacement of methyl with cyano at R2, with each transformation (plus its reverse) occurring three times. Neither transformation had a dramatic impact on the measured log D of the compounds; however, while the HDAC activities of the groups of ethyl/isopropyl matched pairs were largely unaffected, the replacement of methyl with cyano at R2 brought about an increase in potency across three distinct matched pairs, with an average gain of 0.29 log units (Table 2).

Compound 14a (ethyl piperazine) was identified as being a matched pair of 13a (azetidine), whereas 15a (isopropyl piperazine) was not. This disparity highlighted the significance of the 90% common substructure requirement in the definition of a molecular matched pair. Compound 14a contains 32 heavy atoms, which forced any matched molecules to contain at least 28 (28.8 rounded down) heavy atoms in the common substructure. Compound 15a on the other hand contains 33 heavy atoms, and therefore 13a would

Table 3. Summary of the Individual Matched Pairs from which the Transformations [c]F.[c]C#N (Fluoro to Cyano) and [c]Cl.[c]C#N (Chloro to Cyano) Were Derived

compound A	compound B	ΔHDAC pIC_{50}	Δlog D @ pH 7.4
13d	13b	+0.71	−0.47
15d	15b	+0.50	−0.32
mean change for [c]F.[c]C#N(fluoro to cyano)		+0.60	−0.40
14c	14b	+0.22	−0.67
15c	15b	+0.10	−0.40
mean change for [c]Cl.[c]C#N(chloro to cyano)		+0.16	−0.54

require 29 atoms in the common substructure to be considered paired—one atom more than its actual count. While omission of the azetidine to isopropyl piperazine transformation may not have been a desirable outcome, the 90% substructure cutoff represents an empirically derived compromise between completeness, performance, and utility. Reducing the cutoff to 80% would ensure the encoding of this transformation, but the evaluation of additional putative MCS matches at the initial search stage (step 1) would incur further computational cost. While this increase in cost is likely to be problem-specific, our benchmarks indicate that in comparison to a 90% substructure limit, an increase in time of around 50% can be expected for an 80% substructure requirement and a 100% increase for 70% common substructure. Another consideration was the potential applicability of the transformations encoded at the 80% level; as such transformations generate enumerated structures with greater diversity from the initial seed molecule, and the concept of a matched molecular pair is thus challenged.

Figure 3 summarizes the transformations identified by WizePairZ from this data set. The transformation featuring the largest increase on HDAC potency was the switch from fluorine at R2 to cyano, with a mean increase of +0.60 log units. In addition, it was chloro to cyano substitution that exhibited the largest reduction in log D. These transformations in the top left quadrant represent the most interesting SAR, as they correspond to an increase in potency with a decrease in lipophilicity. As 13c and 14d were excluded from the analysis, the algorithm identified just two examples in each of these sets of matched pairs (Table 3). These findings were consistent with the original report on this data set, which states: "*the chloro and flouro-substituded pyridines are consistently less potent than their methyl and cyano counterparts. Interestingly, the enhanced potency of the cyano series does not appear to arise as a consequence of increased lipohilicity*".

The collection of points surrounding the origin in Figure 3 represented transformations that have little or no effect on either the activity or lipophilicity of the compounds. These transformations included the inter conversion between azetidine and ethyl piperazine, and ethyl piperazine and isopropyl piperazine at R1.

Variation in the extent of the common substructure included with each transformation resulted in up to a possible four SMIRKS strings being deposited for each molecular pair, with each encompassing an ever greater degree of the local environment surrounding the structural change. There were a total of 18 unique transformations, 16 of which encoded up to four degrees of local environment and two encoded up to two degrees, resulting in a total of 68 recorded individual SMIRKS strings. The transformations with fewer than four SMIRKS strings had extended transformations in more than one part of the molecule yet were linked together by fewer than four bonds; an example can be seen in Figure 4.

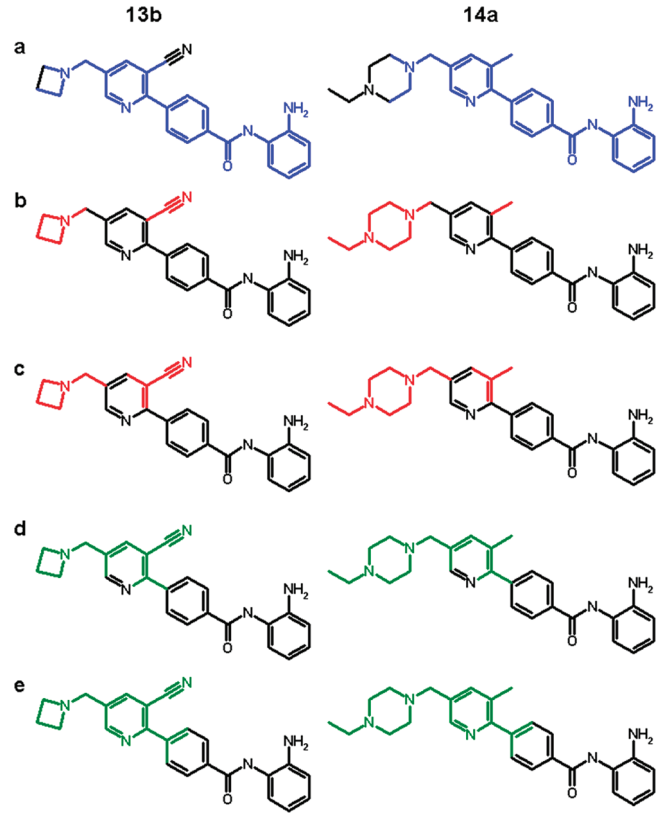


Figure 4. Encoding the transformation of 13b into 14a (and vice versa). (a) Original MCS match encompassing 90% of the larger of the two molecules, shown in blue. (b) First degree of local environment; atoms within one bond of the structural change are marked in red. As these are discontinuous, no valid SMIRKS existed. (c) Second degree of local environment; atoms within two bonds of the structural change are marked in red. Again, no valid SMIRKS existed. (d) Third degree of local environment; atoms within three bonds of the structural change are marked in green. As this was a continuous structure, it generated a valid SMIRKS. (e) Fourth degree of local environment; atoms within four bonds of the structural change are marked in green. Again, a valid SMIRKS was generated.

Validation of the Methodology. All three-bond transformations were applied to the 11 compounds on which the analysis was performed, resulting in the generation of 15 unique chemical structures. Twelve were replicas of existing compounds in Table 1, but three were novel compounds—the methyl, fluoro, and chloro derivatives of the initial cyanopyridine lead 3. As 13c and 14d were excluded from the data-mining steps reported above, these "virtual" compounds were used to validate the methodology.

The structure of 13c was accessed by four different routes: methyl, cyano, and flouro to chloro substitution of 13a, b, and d, respectively, and from ethyl piperazine to azetidine substitution of 14c (Table 4). Each of these routes provided its own estimate for the HDAC activity of 13c, based on the addition of the measured pIC_{50} value of the seed compound to the mean change in activity for a given transformation. These four estimates provided a mean activity of 7.78 with

Table 4. Routes to the Enumeration of Compound 13c Based on the Application of Three-Bond Transformations to Four Distinct Seed Compounds[a]

seed compound	transformation	HDAC pIC_{50} (seed)	ΔHDAC pIC_{50}	HDAC pIC_{50} (13c, estimated)
13a	[c]cc(c([n])[c])C.[c]cc(c([n])[c])Cl (methyl pyridine to chloro pyridine)	7.67	+0.11	7.78
13b	[c]cc(c([n])[c])C#N.[c]cc(c([n])[c])Cl (cyano pyridine to chloro pyridine)	8.01	−0.16	7.85
13d	[c]cc(c([n])[c])F.[c]cc(c([n])[c])Cl (fluoro pyridine to chloro pyridine)	7.30	+0.40	7.70
14c	[c]C[NH]1CC[NH](CC1)CC.[c]C[NH]1CCC1 (ethyl piperazine to azetidine)	7.79	−0.03	7.76

[a] Estimated pIC_{50} values were calculated from the sum of the mean change in activity for a given transformation and the seed compound activity.

Table 5. Routes to the Enumeration of Compound 14d Based on the Application of Three-Bond Transformations to Five Distinct Seed Compounds[a]

seed compound	transformation	HDAC pIC_{50} (seed)	ΔHDAC pIC_{50}	HDAC pIC_{50} (14d, estimated)
13d	[c]C[NH]1CCC1.[c]C[NH]1CC[NH](CC1)CC (azetidine to ethyl piperazine)	7.30	0.03	7.33
14a	[c]cc(c([n])[c])C.[c]cc(c([n])[c])F (methyl pyridine to fluoro pyridine)	7.73	−0.30	7.43
14b	[c]cc(c([n])[c])C#N.[c]cc(c([n])[c])F (cyano pyridine to fluoro pyridine)	8.01	−0.60	7.41
14c	[c]cc(c([n])[c])Cl.[c]cc(c([n])[c])F (chloro pyridine to fluoro pyridine)	7.79	−0.40	7.39
15d	[CH2][NH]([CH2])C(C)C.[CH2][NH]([CH2])CC (isopropyl piperazine to ethyl piperazine)	7.48	−0.02	7.46

[a] Estimated pIC_{50} values were calculated from the sum of the seed compound activity and the mean change in activity for a given transformation.

a tight spread, which compared very favorably with the measured pIC_{50} value of 7.73.

The other compound omitted from the initial analysis was 14d, which was accessed from a total of five seed compounds. Methyl, cyano, and chloro were substituted for fluoro in 14a, b, and c, while azetidine and isopropyl piperazine were replaced by ethyl piperazine in 13 and 15d, respectively (Table 5). As with the example above, a tight range of estimates gave a very good approximation of the measured activity of 7.42, with a mean estimate of 7.40.

The three novel compounds were all neutral compounds: the methyl, chloro, and fluoro analogues of compound 3 with predicted HDAC pIC_{50}'s of 7.48, 7.61, and 7.17, respectively.

DISCUSSION

Performance and Utility. The protocol for identifying matched pairs and encoding SMIRKS strings is robust and effective and can be used to guide optimization of compounds against a wide range of medicinal chemistry criteria. The initial data sets analyzed do not have to be complete, as the net change in each parameter is simply collected together for a given transformation at the aggregation stage. While there might be five examples of matched pairs where the transformation affects activity against one receptor, there might be only one example where the effect at a second receptor is known. This transformation may still be used to optimize both parameters simultaneously, albeit with the caveat that the evidence supporting the predicted effect on the second receptor is weaker than the first.

If the user plans to draw any conclusions from the mean changes in a parameter then the data must be expressed in an appropriate manner; e.g., inhibitory concentrations should be expressed as pIC_{50} values. If only the qualitative effect is of interest, i.e., the percentage of times a transformation caused a given property to increase or decrease, then the only real consideration for the user is the time that the analysis is likely to take. The matched pair identification algorithm scales almost linearly with the number of processors used, allowing data sets containing tens of thousands of compounds to be analyzed; as a rule of thumb, each potential matched pair evaluation (with SMIRKS encoding when appropriate) takes approximately 15 ms. The only other major factor governing the speed of execution of the protocol is the size and diversity of the structures undergoing analysis. Highly disparate collections of compounds taken from a range of projects can be analyzed in a short space of time, as the number of common substructures to be evaluated for each molecular pair is far fewer than for data sets containing larger compounds featuring the same molecular core.

While the data set analyzed in this paper is only small, we have applied this methodology *in house* to a single data set containing over 35 000 compounds (data not shown). In line with the estimation above, this analysis took around 3 days when distributed across 36 2.8 GHz dual core AMD Opteron 8220 processors (~5 000 CPU hours). The process generated approximately 35 000 unique transformations at the one-bond level, 90 000 at the two-bond level, 140 000 at the three-bond level, and just over 200 000 four-bond transformations, all coupled with at least qualitative information regarding how the transformation affects that property. These structural modifications are now at the disposal of

medicinal chemists, providing a summary of the 300 000 matched pair relationships within that data set.

The value in the WizePairZ approach can be broken down into three main areas. It is more thorough than the Murcko type rule based fragmentation: the transformations themselves (and also the pairs of compounds from which the information has been extracted to help understand the SAR) may be easily visualized, and the application of prestored transformations to a seed molecule can be used to improve the properties of lead compounds. The program can be applied to small data sets typical of individual drug discovery projects to solve specific issues, or much larger data sets of more diverse compounds to identify general fixes for common medicinal chemistry issues.

Subseries Selection. Inspection of the transformations described above highlighted a number of cases where substitution at R2 gave rise to an increase in HDAC activity without much effect on the overall lipophilicity of the compounds. In contrast, there is little evidence of much HDAC SAR surrounding the amine position at R1. How this information is utilized is likely to depend on the circumstances of the medicinal chemist. If an additional gain in primary potency is required, then one may wish to explore further possibilities for substitution at R2, as untested substituents could further optimize interactions with the receptor.

In an alternative situation, this same information may be used in a more direct sense. Having identified the R2-cyano subseries as the most polar and potent, this feature could be fixed while manipulating the pharmacokinetic (PK) profile through adjustment of log P and pK_a at R1. The evidence presented here would suggest that the SAR in this region is essentially flat, as none of the transformations encoded for this part of the molecule had a mean effect on HDAC activity of greater than ± 0.03 log units. From simple plots of the type presented in Figure 3, a user can quickly identify which series of compounds are situated on the leading edge (in this case, the methyl and cyano series), thus providing the most potency with a minimal amount of lipophilicity. Subseries such as these are more likely to be progressible due to the likelihood of avoiding undesirable activity at other receptors.

Application of Encoded Knowledge. We have shown here how it is possible to fill gaps in data sets containing pIC_{50} data, which provides a valuable sanity check when reviewing whether all readily identifiable compounds have been considered for synthesis within a project. However, the WizePairZ approach has far wider utility than the example presented in this study. As the matched molecular pairs identified by the algorithm are stored in a database as text strings, they can be easily retrieved and applied to solve medicinal chemistry problems wherever they may be required. This is of considerable use when facing situations where a drug discovery project is handed from one team to another; while being familiar with a handful of the most potent and well profiled examples, the new team taking on the development of a chemical series may not be aware of learning obtained from structural transformations derived from the less active members of the data set. Stored transformations could be applied to the best leads to suggest potential chemistry targets that are not detrimental to activity but may offer benefits in terms of selectivity, pharmacokinetic profiles or other properties in need of optimization. A more traditional Free−Wilson analysis may be applicable in this situation in terms of potency analysis, but the incorporation of physicochemical and pharmacokinetic parameters from series which do not share the same scaffold highlights the advantages of an MMP approach.[3]

With the SAR surrounding a collection of receptors deposited in the database, it becomes possible to enrich the knowledge base further with the identification of bioisosteres. In the HDAC data set presented in this paper, replacement of a chlorine atom with a methyl group brought about a reduction of activity of just 0.11 log units (mean of two examples). In large corporate collections, there is likely to be information describing chloro/methyl matched pairs across a whole range of receptors, and aggregation of these data becomes simply a matter of informatics. It has been shown from previous analysis of large data sets that the net effect of a transformation on any given activity is most likely to be zero.[11] However, in a subset of cases where there is no net effect, the removal of either of the two functional groups being transformed will also lead to a significant reduction in potency. Here, it is possible to claim the identification of a bioisostere and, once flagged, may be applied to chemical series where the SAR is yet to be determined.

While analyzing primary potency data with WizePairZ can be interesting and helpful within an individual drug discovery project, the greatest value comes from the analysis of data sets with applicability across a range of programs. For example, chemists working on a series of compounds where it is expected that the compounds will interact with a receptor implicated in toxicological side effects would be expected to assess the activity of a considerable number of compounds from that series. In so doing, those chemists build up knowledge of functional group transformations with the ability to manipulate activity at that receptor, hence limiting the risk of detecting toxicity *in vivo*. Meanwhile, or even sometime well into the future, another chemistry team might detect activity at the same receptor on a program where it had not been anticipated. Rather than repeat the expensive and time-consuming process of back-screening a large number of compounds from the new series, the original data set can simply be mined using WizePairZ to identify molecular modifications that previously brought about a reduction in activity and applied to the new lead. The suggested structures could then be submitted for screening where samples of the compound are available or prioritized for synthesis where they are not. In this way, historical knowledge can be stored and shared across projects and applied to common medicinal chemistry problems wherever they may occur. We have already published a "vision" of how such a system could be developed in the future and the effects that this approach to generation of an autocurating medicinal chemistry knowledge base could have in transforming drug-hunting.[31]

CONCLUSIONS

WizePairZ is an automated expert system for medicinal chemistry based on matched molecular pairs analysis. The example presented in this paper illustrates how it is possible, given a simple data set containing structures, biological activity, and lipophilicity information, to extract key SARs and use the relationships to propose further chemical targets

for synthesis in an informed and intelligent fashion. The software successfully identified a total of 18 unique transformations from a collection of 11 compounds with associated changes to the properties under investigation. From these data, it was possible to identify the cyano subseries as offering the optimal balance of potency and lipophilicity, as was highlighted in the original study.[27] Modifications to the molecular structures were recorded as SMIRKS strings, applied back to the seed molecules, and used to enumerate the two compounds omitted from the analysis with accurate estimates of their biological activity.

While the example presented in this paper is only small, WizePairZ can be applied to larger data sets containing tens or even hundreds of thousands of molecules. Matched pairs where there is an indication that a key recognition feature for a given receptor is being introduced or removed can then be validated across a whole range of chemical series. Any medicinal chemist who is unfamiliar with the SAR surrounding a particular property can provide their lead compound as a seed structure to the system and instantly get an indication of what impact specific structural modifications are likely to have on their compound. By this mechanism, it is possible to store and transfer historic medicinal chemistry experience from team to team and across large organizations.

ACKNOWLEDGMENT

The authors would like to thank the members of the Department of Medicinal Chemistry at AstraZeneca R&D Charnwood and, in particular, Nick Tomkinson and Hitesh Sanganee for discussions contributing to the development of WizePairZ. Thanks also go to Dave Cosgrove and Pete Kenny for proof reading and comments on this manuscript.

REFERENCES AND NOTES

(1) Riley, R. J.; Parker, A. J.; Trigg, S.; Manners, C. N. Development of a Generalized, Quantitative Physicochemical Model of CYP3A4 Inhibition for Use in Early Drug Discovery. *Pharm. Res.* **2001**, *18*, 652–655.
(2) Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *J. Med. Chem.* **2006**, *49*, 6672–6682.
(3) Birch, A. M.; Kenny, P. W.; Simpson, I.; Whittamore, P. R. O. Matched molecular pair analysis of activity and properties of glycogen phosphorylase inhibitors. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 850–853.
(4) Lewis, M. L.; Cucurull-Sanchez, L. Structural pairwise comparisons of HLM stability of phenyl derivatives: Introduction of the Pfizer metabolism index (PMI) and metabolism-lipophilicity efficiency (MLE). *J. Comput.-Aided Mol. Des.* **2008**, *23*, 97–103.
(5) Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *7*, 395–399.
(6) Zwanzig, R. W. High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys.* **1954**, *22*, 1420–1426.
(7) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
(8) Bemis, G. W.; Murcko, M. A. Properties of known drugs. 2. Side chains. *J. Med. Chem.* **1999**, *42*, 5095–5099.
(9) Haubertin, D. Y.; Bruneau, P. A database of historically-observed chemical replacements. *J. Chem. Inf. Model.* **2007**, *47*, 1294–1302.
(10) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339–348.
(11) Hajduk, P. J.; Sauer, D. R. Statistical analysis of the effects of common chemical substituents on ligand potency. *J. Med. Chem.* **2008**, *51*, 553–564.
(12) Sheridan, R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103–108.
(13) Ertl, P. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374–380.
(14) Tandon, M.; O'Donnell, M. M.; Porte, A.; Vensel, D.; Yang, D.; Palma, R.; Beresford, A.; Ashwell, M. A. The design and preparation of metabolically protected new arylpiperazine 5-HT1A ligands. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 1709–1712.
(15) Sheridan, R. P.; Hunt, P.; Culbersons, J. C. Molecular transformations as a way of finding and exploiting consistent local QSAR. *J. Chem. Inf. Model.* **2006**, *46*, 180–192.
(16) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
(17) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82–85.
(18) Southall, N. T. Kinase Patent Space Visualization Using Chemical Replacements. *J. Med. Chem.* **2006**, *49*, 2103–2109.
(19) Raymond, J. W.; Watson, I. A.; Mahoui, A. Rationalizing Lead Optimization by Associating Quantitative Relevance with Molecular Structure Modification. *J. Chem. Inf. Model.* **2009**, *49*, 1952–1962.
(20) Stewart, K. D.; Shiroda, M.; James, C. A. Drug Guru: A computer software program for drug design using medicinal chemistry rules. *Bioorg. Med. Chem.* **2006**, *14*, 7011–7022.
(21) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(22) *Accord For Excel*, version 7.0; Accelrys Software Inc.: San Diego, CA, 2008.
(23) Chamberlin, D. D.; Boyce, R. F. In *SEQUEL: A structured English query language*; ACM New York: New York, 1974; pp 249−264.
(24) Kubinyi, H. Lipophilicity and drug activity. *Progr. Drug Res.* **1979**, *23*, 97.
(25) Shamovsky, I.; Connolly, S.; David, L.; Ivanova, S.; Nordén, B.; Springthorpe, B.; Urbahns, K. Overcoming undesirable HERG potency of chemokine receptor antagonists using baseline lipophilicity relationships. *J. Med. Chem.* **2008**, *51*, 1162–1178.
(26) Pareto, V. Cours d'Economie Politique, volume I and II. *F. Rouge, Lausanne* **1896**, 250.
(27) Andrews, D. M.; Gibson, K. M.; Graham, M. A.; Matusiak, Z. S.; Roberts, C. A.; Stokes, E. S. E.; Brady, M. C.; Chresta, C. M. Design and campaign synthesis of pyridine-based histone deacetylase inhibitors. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 2525–2529.
(28) *OEChem*, version 1.6.1; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2008.
(29) *Python*, version 2.5.2; Python Software Foundation: Wolfeboro Falls, NH, 2008.
(30) *JMP*, version 7; SAS Institute Inc.: Cary, NC, 2007.
(31) Griffen, E. The rise of the intelligent machines in drug hunting. *Future Med. Chem.* **2009**, *1*, 405–408.

CI100084S