ELSEVIER

Oncology exploration: charting cancer medicinal chemistry space

David G. Lloyd[1], Georgia Golfis[1], Andrew J.S. Knox[2], Darren Fayne[1], Mary J. Meegan[2] and Tudor I. Oprea[3]

[1]Molecular Design Group, School of Biochemistry and Immunology, Trinity College Dublin, Dublin 2, Ireland
[2]School of Pharmacy and Pharmaceutical Sciences, Trinity College Dublin, Dublin 2, Ireland
[3]Division of Biocomputing, MSC11 6145, University of New Mexico School of Medicine, Albuquerque, NM 87131, USA

Approaches for the experimental determination of protein–ligand molecular interactions are reliant on the quality of the compounds being tested. The application of large, randomly designed combinatorial libraries has given way to the creation of more-focused 'drug-like' libraries. Prior to synthesis, we wish to screen the potential compounds to remove undesired chemical moieties and to be within a required range of physiochemical properties. We have used a principal-component analysis (PCA) computational approach to analyze the 3D descriptor space of active and non-active (hit-like) cancer medicinal chemistry compounds. We define hit-like those molecules passing the unmodified OpenEye FILTER program. Our analysis indicates that these compounds occupy quite different regions in space. Cancer-active compounds exist in a much greater volume of space than generic hit-like space and most of them fail the commonly applied filters for orally bioavailable drugs. This is of great significance when designing orally bioavailable cancer target drugs.

The successful application of the processes of virtual and physical screening for active ligands is totally reliant on the quality of the molecules being screened. In the simplest terms, if there are no hits in the database or compound library, there is no point in performing the screen. Recent years have seen great advances in our understanding of what makes a molecule drug- or lead-like and cheminformatic treatment of screening collections has focused the attention of discovery research towards drug and lead chemical space. However, when dealing with oncology, the applied and trusted rules of engagement do not always apply.

Partitioning and classifying cancer medicinal chemistry space is not straightforward. A multitude of active cancer ligands, containing different molecular scaffolds, have been identified for the relatively small number of cancer targets. The past two decades have witnessed a tremendous increase in our understanding of the pathology and molecular biology of human cancers [1]. Although enormous progress has been made in the development and identification of new molecular medicines and targets in this area, many of the current clinical treatments for cancers have limitations with respect to efficacy, resistance and toxicity in the patient [2]. There is much scope for the exploitation of the many new molecular targets to develop new cancer treatments with improved specificity, toxicology profiles and efficacy [3].

Cancer chemotherapeutics

The bulk of existing cancer chemotherapeutic drugs causes cell death by several different mechanisms – the majority by non-selectively targeting the cellular processes that cancers utilize to rapidly grow and divide (i.e. the ability to replicate their DNA) [4]. In general, dividing tumor cells have lost the ability to respond to environmental cues, which in normal cells would control physiological functions such as cell division. Normal cells exhibit a higher basal resistance to chemotherapeutic drugs, whereas rapidly dividing cells, such as bone marrow and intestinal mucosa, are highly susceptible to them and severe toxic side effects are common [5]. The discovery of defects in oncogenes has allowed the development of exciting anticancer therapeutics that can selectively target tumor cells and specific tumor biochemical processes,

Corresponding author: Lloyd, D.G. (lloyddg@tcd.ie).

1359-6446/06/$ – see front matter PII: S1359-6446(05)03688-3

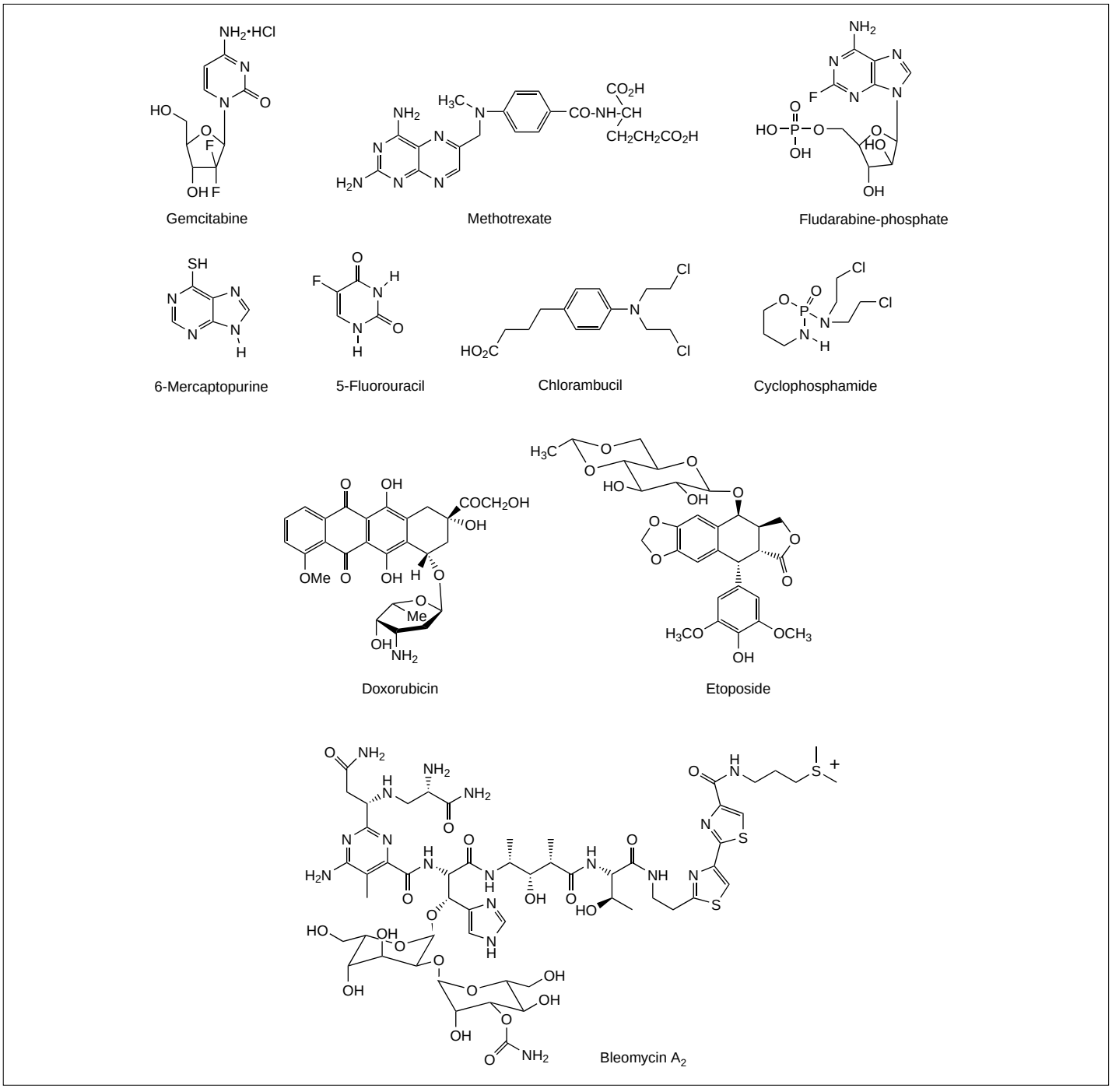


FIGURE 1

Chemical structures of non-selective chemotherapy drugs acting through different mechanisms. A selection of genotoxic and antimetabolic drugs are depicted, illustrating the great variety in structural form of anticancer ligands.

avoiding the cellular toxicity associated with conventional cancer chemotherapeutic drugs [6]. Knowledge of the specific biochemical differences between normal and cancer cells is growing at an exponential rate and this can be potentially exploited for cancer chemotherapy. The recent development of some FDA-approved treatments that target cancer-specific cellular processes demonstrate the utility of these novel approaches [7].

To appreciate the breadth of cancer-related chemical space, we first examine some of the major mechanistic groupings within this space by looking at various classes of oncology therapeutics, with a view to mapping their relative locations in terms of the medicinal chemical space they occupy.

Antimetabolites

Antimetabolites, such as methotrexate, 6-mercaptopurine and gemcitabine (Figure 1), were developed to interfere with specific enzymatic steps in nucleotide biosynthesis in tumor cells. Methotrexate is an inhibitor of dihydrofolate reductase (DHFR) and, as such, limits the formation of nucleotides and thus inhibits DNA replication. It has a significant role in the treatment of breast

cancer, acute lymphocyte leukemia and lymphomas [8]. Purine antagonists, such as 6-mercaptopurine and fludarabine-phosphate (Figure 1), function by inhibiting DNA production either by blocking the production of the required purines or by incorporating themselves into the growing DNA as false nucleotides through DNA polymerase. 6-Mercaptopurine can be given orally or intravenously and is useful in the treatment of acute lymphocytic and myelogenous leukaemias [9]. Similarly, pyrimidine antimetabolites [e.g. 5-fluorouracil and gemcitabine (Figure 1)] act by preventing the biosynthesis of the required pyrimidine nucleoside, hence inhibiting DNA replication. Gemcitabine is used clinically for the treatment of non-small cell lung cancer, whereas 5-fluorouracil is useful for a range of malignancies in tissues such as breast, colon, liver and skin [10].

Genotoxic drugs

Genotoxic drugs modify DNA and thus prevent accurate nucleic acid replication. Examples of widely used genotoxic chemotherapeutic drugs include alkylating agents [e.g. chlorambucil and cyclophosphamide that is activated by hepatic metabolism (Figure 1)], which alkylate and crosslink guanine bases, thus preventing DNA replication. Chlorambucil is used to treat a variety of cancers, including chronic lymphocytic leukemia, lymphomas and Hodgkin's disease. Cyclophosphamide is useful in the treatment of bladder, bone, cervical and lung cancers [11]. The platinum-containing coordination complexes, cisplatin, carboplatin and ormaplatin, act in a similar manner, by intra-strand crosslinking of the guanine bases of DNA, and are useful alone or as part of a combination regimen to treat a wide variety of neoplastic diseases [12].

Intercalating agents, such as doxorubicin and epirubicin (Figure 1), insert themselves in the minor groove between the nitrogen base pairs in the DNA double helix, causing a distortion in the helix shape and therefore interfering with DNA and RNA replication and transcription processes. Doxorubicin also inhibits topoisomerase type II, reducing the ability of the cell to repair breaks in the DNA. It is clinically useful for the treatment of a wide variety of cancers, such as breast and ovarian cancer, Hodgkin's and non-Hodgkin's lymphomas, testicular, bladder and lung cancers. Like doxorubicin, the natural product etoposide affects the DNA by specifically inhibiting the action of the enzyme topoisomerase II, which allows unwinding of the DNA in the normal process of replication by causing a break in the DNA strands. It is clinically useful for small cell lung cancer [13]. The cytotoxic antibiotic bleomycin A_2 (Figure 1) inhibits DNA synthesis by a mechanism of DNA chain cutting and so prevents cell replication. It is a useful agent for Hodgkin's and non-Hodgkin's lymphomas, together with squamous cell carcinomas.

Mechanism-based cancer therapies

The chemotherapeutic drugs outlined above are effective but relatively non-selective because they target fundamental biochemical processes, such as DNA and protein production. This indiscriminate action is one of the reasons for their severe toxic side effects. In recent years, there has been significant progress in the identification of cancer-specific cellular drug targets and in the design of drugs that selectively target cancer-specific cellular processes. These treatments should exploit the biochemical differences between normal and cancerous cells, resulting in drugs with greater potency and less toxic side effects [14].

Several oncogenes have been identified (e.g. *TP53*, *SRC*, *RAS*, *BCR-ABL*) which affect tumor growth and development, and new methods for oncogene identification are continually progressed [15]. Overexpression of specific gene products, such as epidermal growth factor (EGF) receptors and human epidermal growth factor receptor 2 (HER-2), can also be linked to the progression of some cancers. Significant developments in targeted treatments have been made in the areas of protein kinase inhibitors, farnesyl transferase inhibitors and matrix metalloproteinase inhibitors, as well as for agents targeting nuclear receptors implicated in disease progression.

Kinase inhibitors

The kinases represent a group of enzymes that regulate protein and cellular activity by phosphorylation. Intracellular signaling pathways that stimulate cell proliferation are frequently controlled by kinases [16]. In chronic myeloid leukemia, a specific chromosome (the Philadelphia chromosome) produces a damaged kinase fusion receptor protein BCR-ABL, which signals uncontrolled proliferation. Imatinib (Figure 2) is a tyrosine kinase inhibitor that binds to the BCR-ABL receptor, preventing ATP binding and resultant kinase activity [17]. It is approved by the FDA for the treatment of chronic myeloid leukemia and also for gastrointestinal tumors. EGF receptors are overexpressed on the surface of some lung and colon cancer cells. Activation of these receptors by EGF and TGF-α (transforming growth factor α) is crucial for tumor cell proliferation. Gefitinib (Figure 2) is a tyrosine kinase inhibitor that binds to the EGF receptor and inhibits the phosphorylation of the tyrosine residue by the tyrosine kinase enzyme. It is used clinically in the treatment of refractive non-small cell lung cancer. However, a recent randomized clinical trial, comparing the efficacy of gefitinib versus placebo after chemotherapy and radiation in patients with non-small cell lung cancer, which had spread only to nearby tissues or lymph nodes, was terminated by researchers when an interim data review indicated that gefitinib treatment would not improve survival (http://www.cancer.gov/newscenter/pressreleases/gefitinibNSCLC).

Proteasome inhibitors

Proteasomes function by controlling the processing of cellular proteins; for example they regulate proteins that control cell cycle and are involved in the degradation of damaged proteins. They are composed of several proteolytic enzymes. Inhibition of proteasome function can cause cell-cycle arrest and cell death [18]. Cancer cells are very susceptible to proteasome inhibitors because they divide rapidly and their normal control mechanisms do not function. Bortezomib (Figure 2) is used clinically in the treatment of multiple myeloma. It interacts with a threonine residue at the catalytic site of the proteasome, limiting metastasis and apoptosis, which are both regulated by proteasomal mechanisms.

Antimitotic agents

Antimitotic drugs interfere with the normal process of cell division. In mitosis, tubulin polymerizes to form spindle microtubules, which facilitate the separation of the two replicated chromosomes. Some drugs bind to tubulin monomers and inhibit the formation of microtubules (e.g. vincristine), whereas paclitaxel and epothilones have the opposite effect in that they stabilize the microtubules and

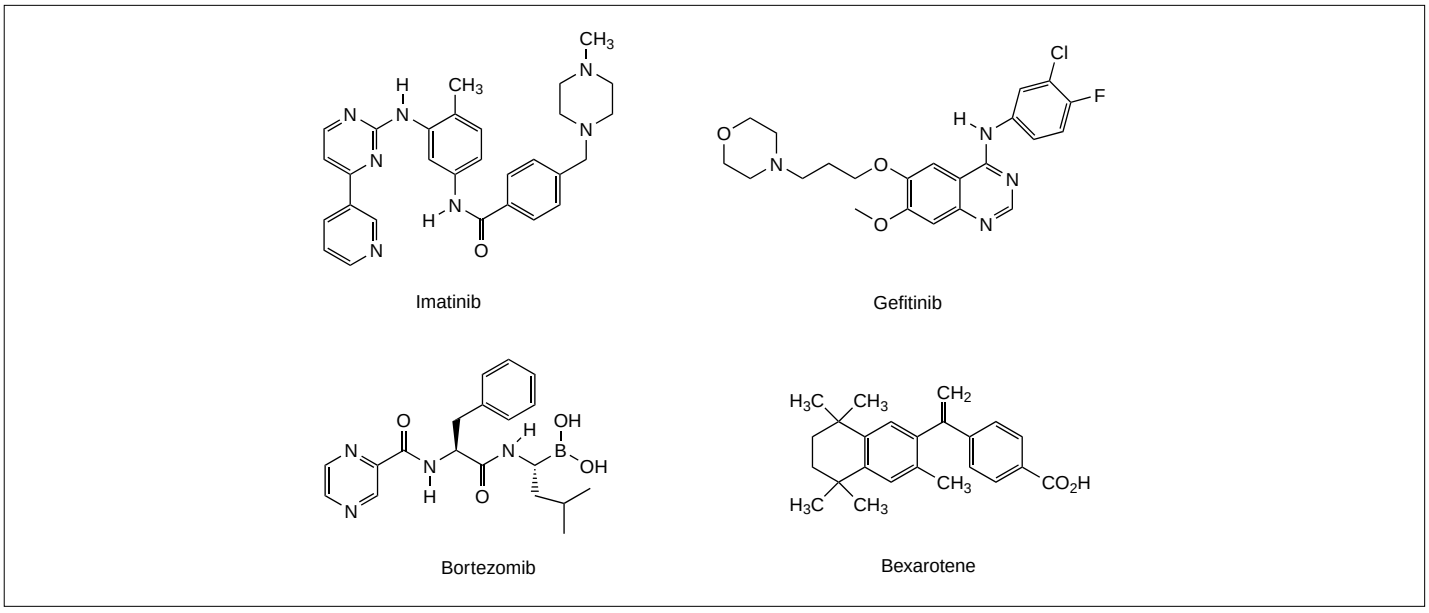


FIGURE 2

Chemical structures of small molecule cancer target selective inhibitors. Kinase and proteasome inhibitors bind to specific protein targets rather than to proteins involved in basic biochemical processes.

prevent the normal cell division process to be completed (Figure 3) [19]. Paclitaxel has several indications for various cancer chemotherapies, including combination regimens with cisplatin for lung and ovarian cancers and with the monoclonal antibody trastuzumab for breast cancer.

Hormonal agents

The proliferation of various breast, ovarian and prostate cancers are known to be dependent on signaling hormones, such as estrogen or testosterone; many of the drugs used in the adjuvant treatments of these cancers are effective by blocking either the biosynthetic production of the hormone or the activity of the hormone in the target cell [20].

Selective estrogen-receptor modulators

Drugs such as tamoxifen and toremiphene (Figure 3) act as estrogen receptor antagonists in breast tumor cells. Binding of these drugs causes a change in the ligand-binding domain conformation, clearly observed in X-ray crystallography data, that effectively blocks estrogen action in the cell, disrupting cell signaling processes and preventing cancer cell proliferation. Tamoxifen is used clinically for the adjuvant treatment of pre- and post-menopausal breast cancer and also for prevention in 'at risk' patients. The related selective estrogen-receptor modulator, raloxifene, is currently indicated for osteoporosis treatment. The compound was originally developed as keoxifene, for the treatment of tamoxifen-resistant breast cancer patients. Although effective in the treatment of estrogen receptor (ER)-dependent breast cancers, keoxifene was unable to meet clinical endpoints for the resistant indication. In contrast to tamoxifen, raloxifene has differential intrinsic actions at its target organs, acting as agonist at bone and as antagonist at the uterus.

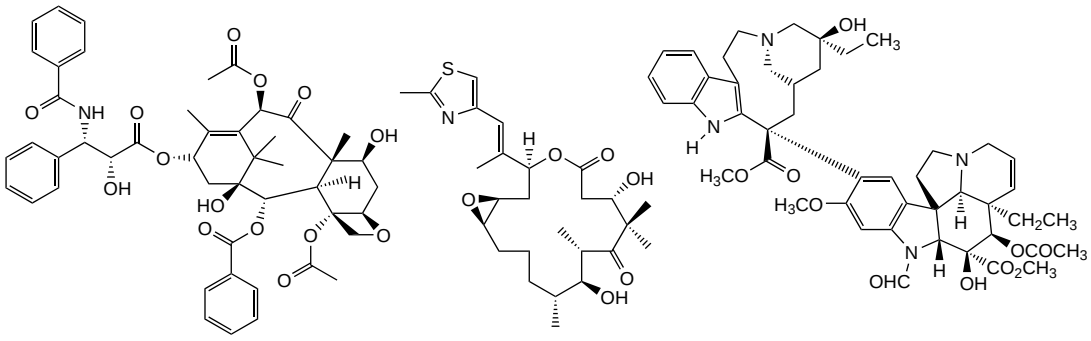
Fulvestrant is an estrogen receptor antagonist (pure antiestrogen) that binds reversibly to the ER and is used clinically to treat ER-positive metastatic postmenopausal breast cancers, with much continued interest in the development of structurally constrained analogues of this drug [21]. Megestrol is a synthetic analogue of progesterone and inhibits proliferation of hormonally sensitive metastatic breast and endometrial cancers (Figure 3). Recent studies have also highlighted the utility of this class of compounds in targeting GPR30, a transmembrane estrogen receptor identified in the endoplasmic reticulum and novel target in estrogen-dependent cancers. Although the physiological and pathological implications of GPR30 remain to be clarified, it is important to note that known antiestrogen agents such as tamoxifen and raloxifene are not estrogen-receptor-selective drugs because they act as agonists on GPR30 [22].

Selective androgen-receptor modulators

The androgen hormones, testosterone and dihydroxytestosterone, play a signaling role in the normal growth of prostate cells. These androgens also bind to the androgen receptors in prostate cancer cells, resulting in the proliferation of cancer cells. The antiandrogens flutamide, bicalutamide and nilutamide (Figure 3) block the action of testosterone by preventing the binding of testosterone to its receptors and therefore control the growth of prostate cancer cells [23]. Leutenising hormone releasing hormone analogues, such as the peptides goserelin and buserelin, which suppress the production of the steroid hormone testosterone, are routinely used in prostate cancer treatments and also for endometriosis [24].

Aromatase inhibitors

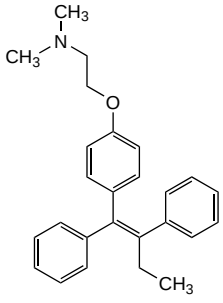
The biosynthesis of estrogen from androgen precursors can be prevented by blocking the action of the enzyme aromatase [25]. There are two types of aromatase inhibitors used clinically for postmenopausal ER-positive breast cancer, the steroidal aromatase inhibitor examestane and the nonsteroidal aromatase inhibitors, anastrazole and letrozole (Figure 3). Recent clinical trials with anastrozole suggest an increased role of this drug in the adjuvant treatment of the disease [26].



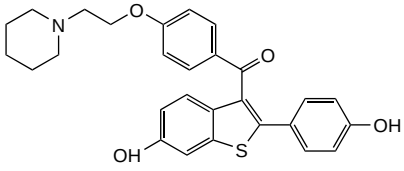
Paclitaxel

Epothilone A

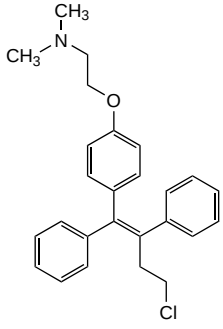
Vincristine



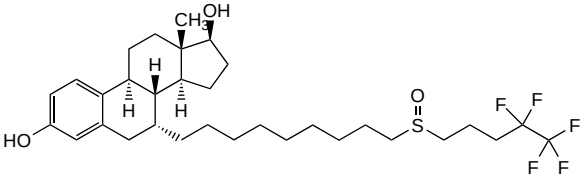
Tamoxifen



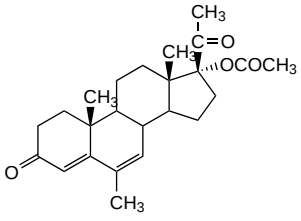
Raloxifene



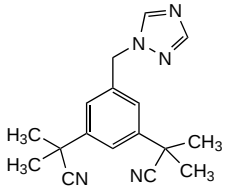
Toremiphene



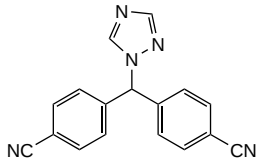
Fulvestrant, ICI 182780



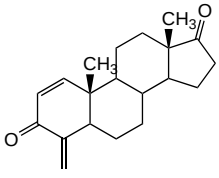
Megestrol



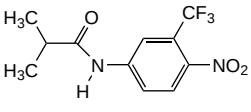
Anastrozole



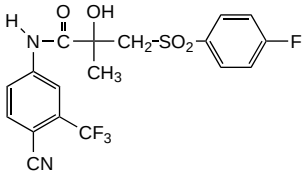
Letrozole



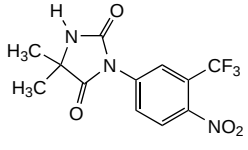
Exemestane



Flutamide



Bicalutamide



Nilutamide

FIGURE 3

Examples of antimitotic, hormonal, SERM, SARM and aromatase targeting ligands. These drugs specifically target different classes of cancer-related proteins.

Exploring chemical space: where are we looking, and what are we looking at?

Although it is often useful from a review perspective to group molecules based on target or activity, such groupings are not necessarily useful in advancing the identification of novel ligands. Cancer-related chemical space has been described as being intrinsically different from general drug-like space [27]. The extent to which oncology compounds move away from traditional medicinal chemistry space is related to properties that, most importantly, can be identified using computational filters, commonly applied at the earliest stages of the drug discovery process. For example, if significantly lower toxicity (e.g. moving away from alkylating agents) or improved oral bioavailability are desired (e.g. aromatase inhibitors, antiestrogens and antiandrogens), filters can be incorporated to a selection protocol to converge on molecules that occupy the appropriate portion of the chemical space. To attain a 'big picture' perspective on the distribution of cancer-targeting compounds in the broader medicinal chemical space, we have examined the regions of medicinal chemical space occupied by cancer-targeted molecules and known classes of oncology therapeutics through the application of the principal-component analysis (PCA) [28].

Virtual screening and cheminformatic filtering

An *in silico* or virtual screening (VS) approach helps to converge on possible active molecules from large molecular libraries and focus physical assaying on a smaller subset of compounds [29]. A developing area in VS is the use of computational methods that filter a molecular library before docking towards compounds with favorable pharmacokinetics, optimum oral bioavailability, compatibility with some types of metabolisms and consequently low toxicity [30]. As with all human endeavors, the urge to impose rules and structures on process is firmly engrained in the early phase of drug discovery. Framing rational discovery are a set of guiding rules that describe cheminformatic properties desirable in lead- or drug-like chemical scaffolds.

The 'rule of five'

Based on a study of the properties of orally available drugs, Lipinski's analysis of the reasons why compounds fail in progression highlighted the necessity to consider pharmacokinetic properties in compound library design [31]. This work furnished drug designers with the 'rule of five' (RO5), which essentially directed library creation and our consideration of screening collection diversity towards orally available drug-like space. Application of the RO5 would enforce the following properties on a screening collection:

(i) molecular weight (MW) ≤500 Da;
(ii) hydrophobicity (log *P*) ≤5;
(iii) number of H-bond donors ≤5 and acceptors ≤10.

The 'rule of three'

This rule set was proposed by Congreve *et al.* [32] and is designed for compliance in fragment-based drug discovery [33]. Application of the rule of three would enforce the following properties on a screening collection:

(i) MW <300 Da;
(ii) log *P* <3;
(iii) number of H-bond donors and acceptors <3;
(iv) flexible bonds <3.

Cheminformatic filters

Cheminformatic treatment of computational representations of screening collections allows the filtering of the collections according to the criteria of the designer, using calculable properties of the compounds to prescribe discriminating parameters for grouping, excluding or considering those subsets of the dataset that can be advanced to *in silico* or *in vitro* HTS studies. Much effort has been expended in the database creation to ensure that maximum chemical structural diversity is in-built, while adhering to the guiding rules of drug- and lead-like chemical properties [34,35].

Popular software utilities, such as FILTER (OpenEye), MOE (Chemical Computing Group) and the Daylight Toolkit (Daylight Chemical Information Systems) readily enable users to apply rapid partitioning of compound databases using tunable cheminformatic parameters based on the 'classical' rules, often enhanced with functionality to recognize and remove compounds with toxic properties [36]. These utilities are frequently employed in internal discovery efforts [37], as well as in the presentation and marketing of focused commercial screening libraries and non-targeted HTS collections.

Rational screening: looking in the right places

In drug discovery, it's fairly well established that a biological HTS campaign, which looks at ~1,000,000 compounds, will deliver viable lead structures that are ultimately progressed to furnish one successful drug compound [38]. Approaches to the discovery of novel chemotypes that can exert a desired therapeutic effect have lain in the hands of large compound database screening campaigns. Although useful in the earliest phases of discovery, testing every available compound against all disease states using HTS is not particularly efficient. One aim of rational drug design is to focus and limit the breadth of screening that must be undertaken to reach this goal. Effective preselection of screening candidates is crucial, usually employing computation means, namely VS [39] and cheminformatic filtering [40], to direct the search towards areas in chemical space where the actives are expected to be.

The overarching trend when commencing a discovery research program with a new target is to work within a so-called drug-like compound set for hit identification and then to work back to lead-like space to advance a patentable novel chemotype. Therefore, the design of compound libraries will often (correctly) take filtering and the design rules into account when creating a diverse screening set for general application in HTS [41]. However, when we consider the challenges associated with active identification within the cancer medicinal chemical space, having too narrow a focus in the hit identification stages might not be ideal or appropriate.

Lead-like versus drug-like compounds

Broadly speaking, compounds sought after in the drug discovery process can be split into two categories, drug-like and lead-like. The concept of lead-likeness implies a physicochemical profile in chemical libraries where the members have reduced complexity (e.g. MW <400) and other more-restricted properties than those deemed drug-like. This leaves room for chemical modification in lead optimization rounds, which subsequently modify the properties

towards drug-likeness [42]. Various authors have proposed different concepts of what constitutes drug-like compounds. These models refer to having molecular similarities to known drugs [43] or acceptable absorption, distribution, metabolism, excretion and toxicology (ADME–Tox) properties [44]. Drug-likeness is often entirely dependent on the mode of administration. Other researchers have utilized the application of hard and fast cheminformatic rule systems to partition chemical space into drug-like and non-drug-like bins. We have previously illustrated that RO5 compliance alone does not necessarily imbue drug-likeness on the post-partitioned cohort [40].

It is generally accepted that the initial active molecules discovered in VS and HTS are often far removed from the drugs into which they might evolve through optimization. In such practices the best we can hope for is the identification of screening hits. For the purpose of this study, we are conferring the loose description of 'hit-likeness' to those small molecule compounds that pass the unmodified OpenEye FILTER (which uses XLOGP [45] as a measure of the hydrophobicity partition coefficient with a maximum cut-off value of six) cheminformatic criteria including a strict application of the RO5. These criteria are now widely adopted in database filtering applied in the early stages of VS.

Describing cancer medicinal chemistry space

Douglas Adams pointed out to would-be intergalactic hitch-hikers that space is mind-bogglingly big [46]. Similarly, chemical space is vast, so vast that it can be conceptually advantageous to notionally partition it into smaller, more-manageable sections. From a drug design perspective, these areas could include biologically relevant or medicinal chemical space, which would encompass smaller areas of theoretical chemical space and synthetically accessible space. These subsets further contain places such as drug space and lead space. The challenge for drug discovery is to explore areas of value in medicinal chemical space and, within these areas, retrieve bioactive, workable molecules. The navigational use of 'chemography' (GPS for drug discovery), mapping compounds onto chemical descriptors, allows us to find our way among the vast number of chemical entities (10^{60}) that exist *in potentia* [47]. Paradoxically, this landscape is so large that we can never hope to explore it fully; however, the biologically relevant and valuable subset of hit space is much smaller than chemical space as a whole but elements of hit space are dispersed within the larger system.

To describe accurately what we term cancer medicinal chemistry space, it was first necessary to amass a set of compounds known to be active against cancer (clinical and preclinical compounds), as well as quality data on compounds purported to be anticancer active but later shown to be inactive in assay. A study set of clinical cancer compounds (34 ligands) and literature-claimed active (4026 ligands tested *in vitro* with a minimum activity of $-\log IC_{50} > 6$) and verified inactive (4285) anticancer compounds were selected from the WOMBAT database (the 2004.2 release of this database contains chemical and biological data from 4773 papers published in medicinal chemistry journals between 1975 and 2005) [48]. The new NCI activity database (41,086 compounds) was also used to identify a subset of anticancer active (8688 ligands with minimum log $GI_{50} > 6$ in assay in NCI screens) and verified inactive ligands (32,398 ligands with minimum log $GI_{50} < 6$ in NCI assay) from compounds that had undergone general anticancer screening at NCI.

Finally, to contextualize cancer space in comparison with general non-cancer chemical space, it was necessary to describe and populate generic hit-like space. To achieve this, a pre-filtered subset of the ZINC database [49] was used, containing 109,432 commercially available molecules, with no specific activity claimed. ZINC was our choice of database because it is pre-partitioned into various regions (e.g. drug-like and lead-like). ZINC has been widely adopted by the VS community because of the quality of its content and the availability of compounds therein for validation studies, and for this reason it has been used as a benchmark set. Once uniqueness of the data had been verified by cross-correlating all dataset members, the study set of medicinal chemical space equated to a total of 158,863 compounds, of which 12,714 are known cancer actives, 36,683 are known cancer inactives and the remainder can be described within our study context as hit-like ligands in nature.

Mapping methods

To present the relationship of cancer medicinal chemistry space in the context of wider chemical space in a meaningful and accessible way, it is necessary to construct graphical distributions of the dataset in the 3D space. It was also useful for us to quantify the subsets within our data that would conform to our working definition of hit-like (i.e. those compounds which would pass an application of the FILTER software protocol including RO5 compliance). To partition the dataset on this basis, it was necessary to first calculate 2D descriptors for all members in the MOE and to identify those descriptors which related to adherence to the RO5. A total of 48 2D molecular descriptors were identified, describing atomic nature, molecular size, polarity, lipophilicity and flexibility. PCA is a relatively easy way to transform an *n*-descriptor space into a more-manageable 3D space. In our analysis, we transformed the 48 vectors space into a 3D space described by 3 principal component vectors, where each of the 3 vectors is a combination of the 48 weighted descriptors [50]. These operations facilitated the creation of graphical representations of the 3D space spanned by the compound set.

TABLE 1

Breakdown of cancer compound hit-like nature

Dataset (source)	Active	Inactive	Total members	Hit-like	Rejected[a]
Clinical drugs (WOMBAT-PK)	36	0	36	10	26 (72%)
Designed compounds (WOMBAT)	4026	4285	8311	4150 50% actives, 50% inactives	4161 (50%)
NCI assayed compounds (NCI)	8688	32,398	41,086	13,042 17% actives, 83% inactives	28044 (68%)
Anticancer active (WOMBAT and NCI)	12,714	0	12,714	4248	8466 (66.6%)

[a]Rejected: fails on application of cheminformatic tool FILTER, which takes into account RO5 fails and also the presence of 'toxic' or undesirable reactive functionalities (OpenEye Scientific Software).

Analysis

Table 1 illustrates the nature of cancer compound space considered in this study. In a total of 12,714 verified-active anticancer compounds, only 33.4% pass a cheminformatic hit-like filter, positioning almost two-thirds of cancer actives outside what is accepted as hit-like chemical space. One would expect a modicum of attrition in such a process when utilities such as FILTER are designed to remove not only RO5 fails but also specific molecules containing toxic and reactive functionalities by removing staples such as alkylating agents, nitrogen mustards that make up a large proportion of our anticancer arsenal. In our analysis, however, the vast majority of compound failures stemmed from a lack of MW, log *P* and H-bond acceptor compliance with regard to the RO5. The actual level of attrition in these circumstances is significant when the diversity of the active molecules is considered in the 3D space; this is not simply a matter of 'nasty' groups removing cancer actives

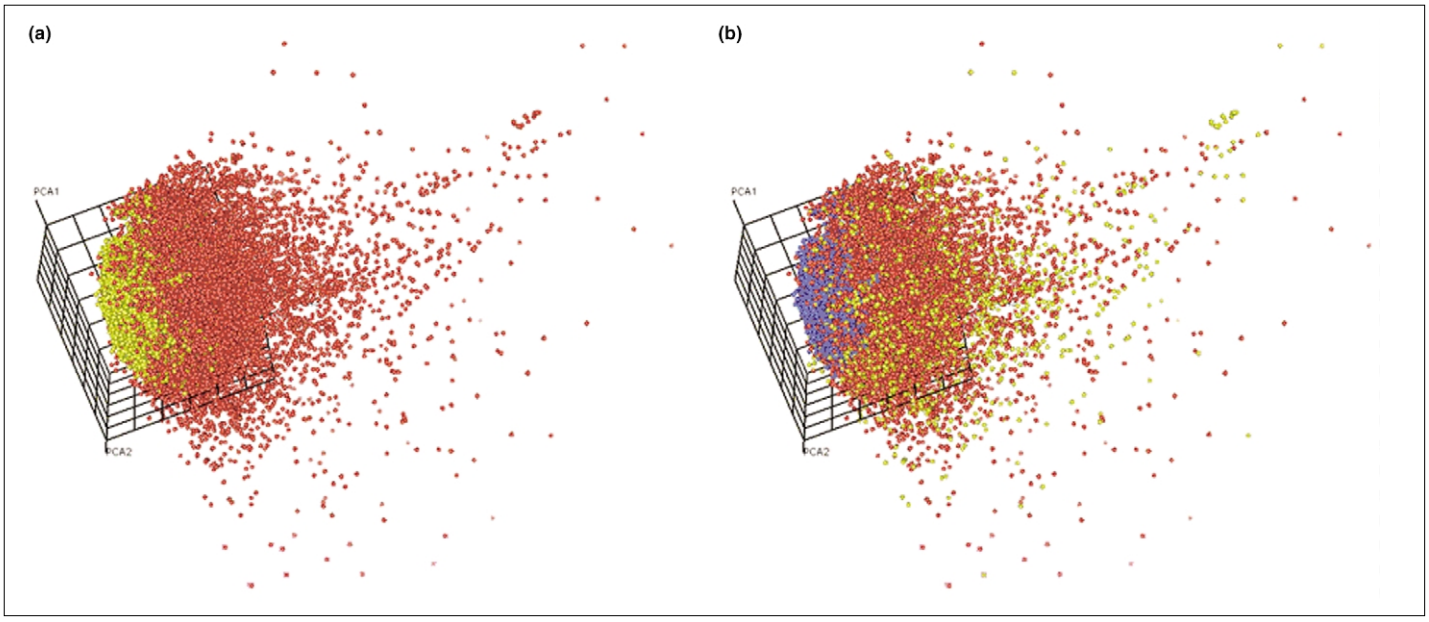


FIGURE 4

Charting cancer medicinal chemistry space. The magnitude of cancer-active medicinal chemistry space, as compared with generic drug-like space, is vast. This illustrates the need to focus in on target related areas of chemical space when seeking new selective drugs. **(a)** Key: yellow sphere, generic hit space; red spheres, medicinal chemistry space. **(b)** Key: blue spheres, generic hit space; red spheres, cancer-inactive medicinal chemistry space; yellow spheres, cancer-active medicinal chemistry space.

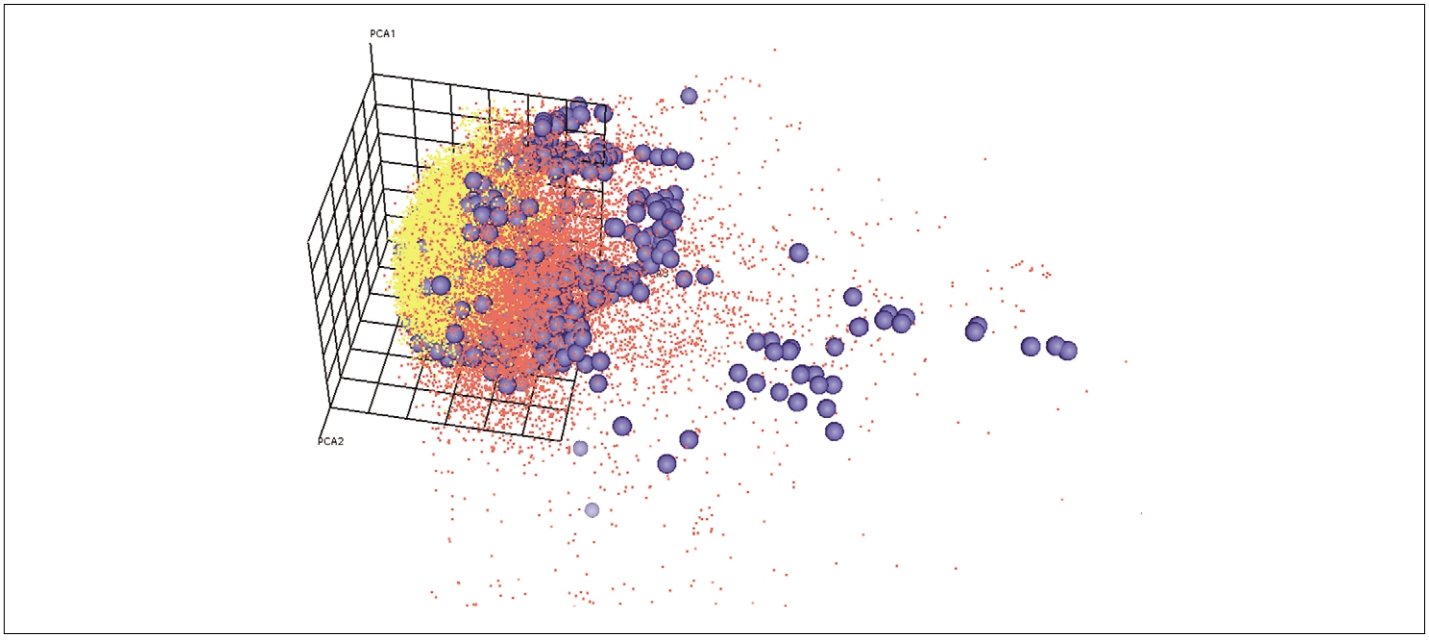


FIGURE 5

Anticancer kinase-targeted space. When a specific cancer target is studied, it is obvious that the majority of its active ligands populate non-generic drug like space. Thus, care needs to be exercised when applying general drug-like filtering criteria to cancer ligand selection procedures. Key: blue spheres, active ligands; yellow cloud, generic hit space; red cloud, non-kinase targeted medicinal chemistry space.

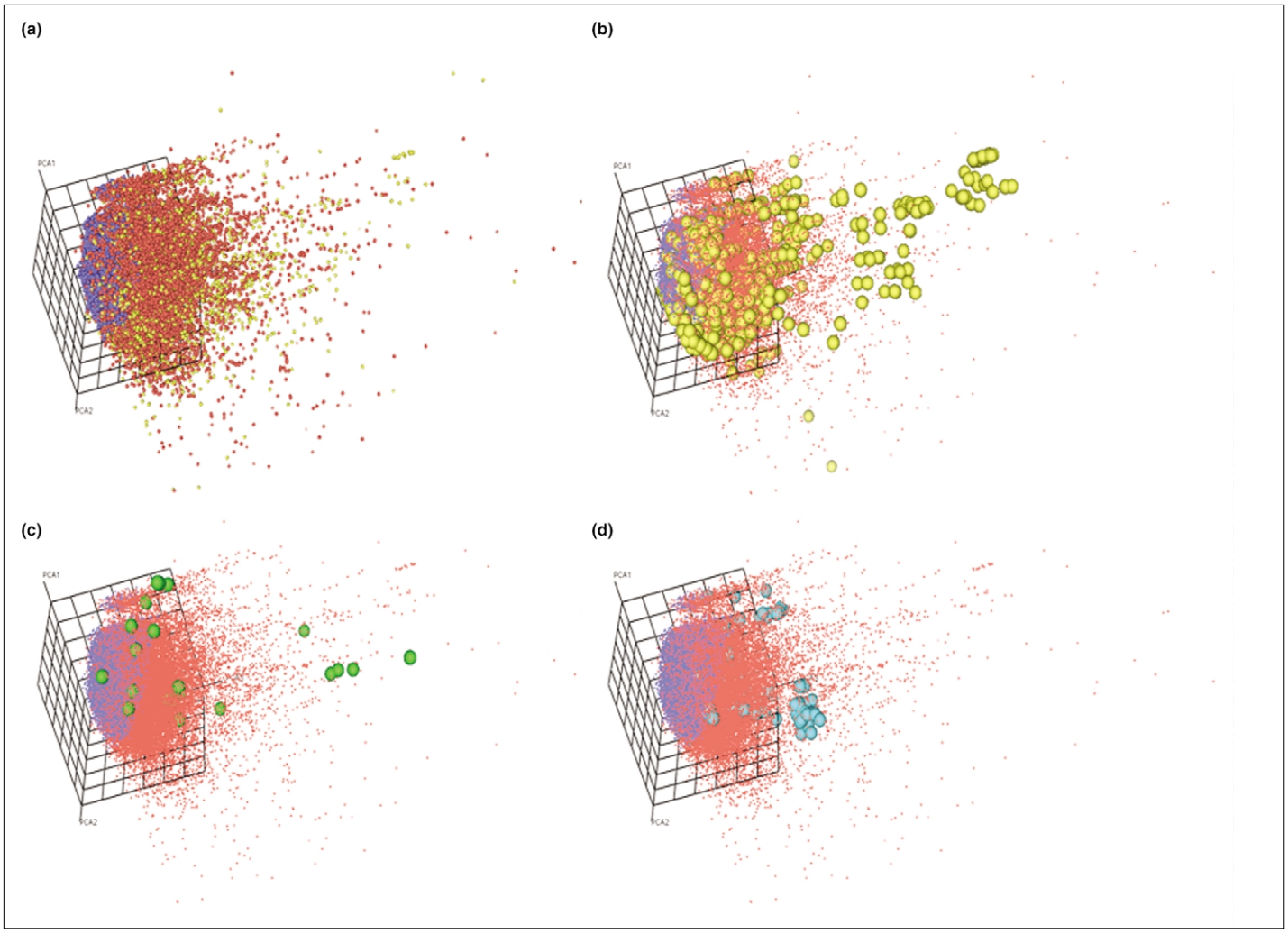


FIGURE 6

Charting cancer medicinal chemistry space. An illustration of the regions of chemical space occupied by different cancer targets shows that filters need to be carefully applied even when considering cancer-active ligands. Drugs binding to different cancer targets do not necessarily reside in closely related areas of chemical space. **(a)** Key: yellow spheres, all cancer actives; red spheres, cancer inactives; blue spheres, traditional hit-like space (indeterminate activity – passes FILTER). **(b)** Key: yellow spheres, cancer actives known to target kinases; red cloud, cancer medicinal chemistry space; blue cloud, traditional hit-like space. **(c)** Key: green spheres, clinically used anticancer compounds; red cloud, cancer medicinal chemistry space; blue cloud, traditional hit-like space. **(d)** Key: cyan spheres, cancer actives known to target tubulin; red cloud, cancer medicinal chemistry space; blue cloud, traditional hit-like space.

from the search space. The implication when looking for cancer actives in prefiltered compound collections is immediately clear.

We can learn a similar lesson when we examine the graphical distribution of the cancer medicinal chemistry space (all compounds assayed *in vitro*) in relation to the wider chemical space (entire study population). The overall distribution of active anticancer compounds spans an area of medicinal chemistry space far beyond that described by our hit-like definition (Figure 4).

Figure 4a contextualizes the nature of medicinal chemistry space considered in this study. The yellow spheres are nonspecific hit-like compounds taken from the filtered ZINC database, illustrating the relatively compact nature of the RO5 space when compared with the wider medicinal chemical space. The red spheres are compounds which were claimed, assayed or demonstrated as anticancer agents from WOMBAT and the NCI databases. These compounds represent charted cancer medicinal chemical space. Figure 4b illustrates the distribution of active anticancer compounds (yellow) in comparison with inactive cancer medicinal chemical space (red) and in relation to generic hit-like compound space (blue).

Even when examining familial distributions, as for the cancer kinome [17], the breadth of cancer space spanned by actives is considerable (Figure 5). Figure 5 shows a view of the distribution of anticancer kinase-targeting compounds in medicinal chemistry space: from a total of 915 active compounds examined (blue spheres), only 156 (15%) lie within our defined hit-like space (yellow cloud), whereas 759 (83%) lie outside (red cloud) and do not pass application of FILTER incorporating RO5 compliance.

Such a general spatial distribution of actives outside the boundaries of the traditional hit-space precludes the creation of all-encompassing cancer-generic filtering rules for database pre-processing. By adopting a class-by-class focus on targeted compound sets (e.g. antitubulin or anti-EGF receptor), compounds could be used to craft tailored cheminformatic filters biased to the target of study for the creation of more rationally focused screening collections. These filters can be utilized in the exploration of target-relevant chemical space. The caveat here is the need for unambiguous

target information for selecting the regions of space that are to be explored.

In Figure 6 the comparative distribution of targeted actives is presented, with reference to the wide spatial distribution of clinically used oncology compounds. It is clear that clusters exist in targeted medicinal chemistry space, in some instances these clusters are not far removed from hit-space and they could potentially be optimized into orally available drug-like space through design.

Conclusion

Cancer medicinal chemical space is far broader than just hit space or orally available drug space and, although it shares common areas to these spaces, it has unique untapped pockets still ripe for exploration. To explore cancer space, drug designers must bear in mind that cancer medicinal chemistry space is not simply a subset of hit- or drug-like space and application of ubiquitous rules and generic filters in these instances will seriously limit the realm of exploration, particularly when dealing with novel targets in the earliest phases of discovery, perhaps to the detriment of the discovery program underway. We have shown that application of the most commonly used cheminformatic filters to bestow hit-likeness on a screening collection results in spatial partitions that are not generally occupied by oncology therapeutics. Particular attention must be given to MW, log *P* and H-bond acceptor parameters in the available filters, as these are primarily responsible for the removal of potential cancer clinical candidate compounds in such filtering processes. It is crucial to think where one wants to be and to take the best route to get there, rather than discarding the avenues available because of a conditioning to follow rules which don't always need to be applied.

Acknowledgement

D. G. Lloyd acknowledges support for this research from the Higher Education Authority Programme for Research in Third Level Institutes IITAC project, Science Foundation Ireland and the Irish Health Research Board. Part of this work was funded by the NCI grant number CA118100 for the UNM Cancer Research and Treatment Center (TIO).

References

1 Desany, B. and Zhang, Z. (2004) Bioinformatics and cancer target discovery. *Drug Discov. Today* 9, 795–802
2 Wong, C.F. *et al*. (2005) Exploiting novel cell cycle targets in the development of anticancer agents. *Curr. Cancer Drug Targets* 5, 85–102
3 Cozzi, P. *et al*. (2004) Recent anticancer cytotoxic agents. *Curr. Med. Chem. Anti-Canc. Agents* 4, 93–121
4 Martinez, R. and Chacon-Garcia, L. (2005) The search of DNA-intercalators as antitumoral drugs: what it worked and what did not work. *Curr. Med. Chem.* 12, 127–151
5 Piazza, F. and Semenzato, G. (2004) Molecular therapeutic approaches to acute myeloid leukemia: targeting aberrant chromatin dynamics and signal transduction. *Expert Rev. Anticancer Ther.* 4, 387–400
6 Benvenuti, S. *et al*. (2005) Identification of cancer genes by mutational profiling of tumor genomes. *FEBS Lett.* 579, 1884–1890
7 Chen, E.X. and Siu, L.L. (2005) Development of molecular targeted anticancer agents: successes, failures and future directions. *Curr. Pharm. Des.* 11, 265–272
8 McGuire, J.J. (2003) Anticancer antifolates: current status and future directions. *Curr. Pharm. Des.* 9, 2593–2613
9 Brathe, A. *et al*. (2003) Cytotoxic activity of 6-alkynyl- and 6-alkenylpurines. *Bioorg. Med. Chem. Lett.* 13, 877–880
10 Malet-Martino, M. *et al*. (2002) The prodrugs of 5-fluorouracil. *Curr. Med. Chem. Anti-Canc. Agents* 2, 267–310
11 Izbicka, E. and Tolcher, A.W. (2004) Development of novel alkylating drugs as anticancer agents. *Curr. Opin. Investig. Drugs* 5, 587–591
12 Brana, M.F. *et al*. (2001) Intercalators as anticancer drugs. *Curr. Pharm. Des.* 7, 1745–1780
13 Larsen, A.K. *et al*. (2003) Catalytic topoisomerase II inhibitors in cancer therapy. *Pharmacol. Ther.* 99, 167–181
14 Saglio, G. *et al*. (2004) Rational approaches to the design of therapeutics targeting molecular markers: the case of chronic myelogenous leukemia. *Ann. N. Y. Acad. Sci.* 1028, 423–431
15 Collier, L.S. *et al*. (2005) Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature* 436, 272–276
16 Vieth, M. *et al*. (2005) Kinomics: characterizing the therapeutically validated kinase space. *Drug Discov. Today* 10, 839–846
17 Druker, B.J. (2004) Imatinib as a paradigm of targeted therapies. *Adv. Cancer Res.* 91, 1–30
18 Mitsiades, C.S. *et al*. (2005) Proteasome inhibition as a therapeutic strategy for hematologic malignancies. *Expert Rev. Anticancer Ther.* 5, 465–476
19 Jordan, M.A. (2002) Mechanism of action of antitumor drugs that interact with microtubules and tubulin. *Curr. Med. Chem. Anti-Canc. Agents* 2, 1–17
20 Meegan, M.J. and Lloyd, D.G. (2003) Advances in the science of estrogen receptor modulation. *Curr. Med. Chem.* 10, 181–210
21 Lloyd, D.G. *et al*. (2004) Benzoxepin-derived estrogen receptor modulators: a novel molecular scaffold for the estrogen receptor. *J. Med. Chem.* 47, 5612–5615
22 Revankar, C.M. *et al*. (2005) A transmembrane intracellular estrogen receptor mediates rapid cell signaling. *Science* 307, 1625–1630
23 Balk, S.P. (2002) Androgen receptor as a target in androgen-independent prostate cancer. *Urology* 60(*Suppl. 1*), 132–138
24 Gaillard-Moguilewsky, M. (1991) Pharmacology of antiandrogens and value of combining androgen suppression with antiandrogen therapy. *Urology* 37(*Suppl. 2*), 5–12
25 Narashimamurthy, J. *et al*. (2004) Aromatase inhibitors: a new paradigm in breast cancer treatment. *Curr. Med. Chem. Anti-Canc. Agents* 4, 523–534
26 Ingle, J.N. and Suman, V.J. (2005) Aromatase inhibitors for therapy of advanced breast cancer. *J. Steroid Biochem. Mol. Biol.* 95, 113–119
27 Wang, J. and Ramnarayan, K. (1999) Toward designing drug-like libraries: a novel computational approach for prediction of drug feasibility of compounds. *J. Comb. Chem.* 1, 524–533
28 Jackson, J.E. (1991) *A Users Guide to Principal Components*, Wiley
29 Mestres, J. (2002) Virtual screening: a real screening complement to high-throughput screening. *Biochem. Soc. Trans.* 30, 797–799
30 Hou, T. and Xu, X. (2004) Recent development and application of virtual screening in drug discovery: an overview. *Curr. Pharm. Des.* 10, 1011–1033
31 Lipinski, C.A. *et al*. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25
32 Congreve, M. *et al*. (2003) A 'Rule of Three' for fragment-based lead discovery? *Drug Discov. Today* 8, 876–877
33 Fattori, D. (2004) Molecular recognition: The fragment approach in lead generation. *Drug Discov. Today* 9, 229–238
34 Leach, A.R. and Hann, M.M. (2000) The *in silico* world of virtual libraries. *Drug Discov. Today* 5, 326–336
35 Shuttleworth, S.J. *et al*. (2005) Design and synthesis of protein superfamily-targeted chemical libraries for lead identification and optimization. *Curr. Med. Chem.* 12, 1239–1281
36 Olah, M.M. *et al*. (2004) Strategies for compound selection. *Curr. Drug Disc. Tech.* 1, 211–220
37 Chen, G. *et al*. (2005) Focused combinatorial library design based on structural diversity, druglikeness and binding affinity score. *J. Comb. Chem.* 7, 398–406
38 Oprea, T.I. (2002) Lead structure searching: are we looking at the appropriate property? *J. Comput. Aided Mol. Des.* 16, 325–334
39 Lyne, P.D. (2002) Structure-based virtual screening: an overview. *Drug Discov. Today* 7, 1047–1055
40 Oprea, T.I. (2000) Property distribution of drug-related chemical databases. *J. Comput. Aided Mol. Des.* 14, 251–264
41 Harper, G. *et al*. (2004) Design of a compound screening collection for use in high-throughput screening. *Comb. Chem. High Throughput Screen.* 7, 63–70
42 Hann, M.M. and Oprea, T.I. (2004) Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* 8, 255–263
43 Walters, W.P. and Murcko, M.A. (2002) Prediction of "drug-likeness". *Adv. Drug Deliv. Rev.* 54, 255–271
44 Lipinski, C.A. (2000) Drug-like properties and the cause of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* 44, 235–249

45 Wang, R. *et al.* (1997) A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Comput. Sci.* 37, 615–621
46 Adams, D. (1979) *The hitch-hiker's guide to the galaxy*, Pan Books
47 Oprea, T.I. and Gottfries, J. (2001) Chemography: the art of navigating in chemical space. *J. Comb. Chem.* 3, 157–166
48 Olah, M. *et al.* (2004) WOMBAT: world of molecular bioactivity. In *Chemoinformatics in drug discovery* (Oprea, T.I. ed.), pp. 223–239, Wiley-VCH
49 Irwin, J.J. and Shoichet, B.K. (2005) ZINC – a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* 45, 177–182
50 Haggarty, S.J. *et al.* (2004) Mapping chemical space using molecular descriptors and chemical genetics: deacetylase inhibitors. *Comb. Chem. High Throughput Screen.* 7, 669–676