OPEN ACCESS Freely available online

PLoS one

A Kernelisation Approach for Multiple d-Hitting Set and Its Application in Optimal Multi-Drug Therapeutic Combinations

Drew Mellor[1,2], Elena Prieto[1,2], Luke Mathieson[1,2], Pablo Moscato[1,2]*

1 Centre for Bioinformatics, Biomarker Discovery and Information Based Medicine, The University of Newcastle, Newcastle, Australia, **2** Information Based Medicine Program, Hunter Medical Research Institute, Newcastle, Australia

Abstract

Therapies consisting of a combination of agents are an attractive proposition, especially in the context of diseases such as cancer, which can manifest with a variety of tumor types in a single case. However uncovering usable drug combinations is expensive both financially and temporally. By employing computational methods to identify candidate combinations with a greater likelihood of success we can avoid these problems, even when the amount of data is prohibitively large. Hitting Set is a combinatorial problem that has useful application across many fields, however as it is NP-complete it is traditionally considered hard to solve exactly. We introduce a more general version of the problem (α,β,d)-Hitting Set, which allows more precise control over how and what the hitting set targets. Employing the framework of Parameterized Complexity we show that despite being NP-complete, the (α,β,d)-Hitting Set problem is fixed-parameter tractable with a kernel of size $O(\alpha dk^d)$ when we parameterize by the size k of the hitting set and the maximum number α of the minimum number of hits, and taking the maximum degree d of the target sets as a constant. We demonstrate the application of this problem to multiple drug selection for cancer therapy, showing the flexibility of the problem in tailoring such drug sets. The fixed-parameter tractability result indicates that for low values of the parameters the problem can be solved quickly using exact methods. We also demonstrate that the problem is indeed practical, with computation times on the order of 5 seconds, as compared to previous Hitting Set applications using the same dataset which exhibited times on the order of 1 day, even with relatively relaxed notions for what constitutes a low value for the parameters. Furthermore the existence of a kernelization for (α,β,d)-Hitting Set indicates that the problem is readily scalable to large datasets.

Citation: Mellor D, Prieto E, Mathieson L, Moscato P (2010) A Kernelisation Approach for Multiple *d*-Hitting Set and Its Application in Optimal Multi-Drug Therapeutic Combinations. PLoS ONE 5(10): e13055. doi:10.1371/journal.pone.0013055

Editor: Maria A. Deli, Hungarian Academy of Sciences, Hungary

Received June 15, 2010; **Accepted** August 19, 2010; **Published** October 18, 2010

Funding: The authors acknowledge the support of the Hunter Medical Research Institute, The University of Newcastle, and ARC Discovery Project DP0773279 (Application of novel exact combinatorial optimisation techniques and metaheuristic methods for problems in cancer research). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pablo.moscato@newcastle.edu.au

Introduction

Typically the selection of a drug therapy for a disease is limited to a single drug, however diseases such as cancer may present as a heterogeneous mix of subtypes of the general disease. In cases such as these multi-drug therapies may prove more effective than single drug therapies, and many trials have been conducted to this end [1–3]. Furthermore combinations of drugs may allow a more targeted approach for a selection of subtypes of a disease, while minimizing effects on unaffected cells. Unfortunately with the abundance of compounds available for the treatment of many conditions of interest, the time and expense in testing even all two drug combinations may be prohibitive. Therefore a smarter approach is needed. Vazquez [4] introduces the Hitting Set problem for this task in the context of oncological drug therapy. The Hitting Set problem is a combinatorial problem that proves extremely useful in modeling a large variety of problems in many domains including protein network discovery [5], metabolic network analysis [6], diagnostics [7–9], gene ontology [10] and gene expression analysis [11,12].

The Hitting Set Problem

Hitting Set is a combinatorial problem that models the problem of selecting a small group of elements to represent or cover a collection of sets. Such a group that covers every set in the collection is called a hitting set. Finding such a set without any constraint is simple, however if we required that the size of the hitting set be relatively small, the problem becomes computationally challenging (*NP*-complete in a formal sense). This difficulty in obtaining solutions with desirable qualities thus requires more thoughtful approaches.

We now give some technical details and formal definitions of the problems of interest.

Hitting Set is equivalent to the Set Cover problem [13], and when otherwise unrestricted, is equivalent to the Red/Blue Dominating Set [14] problem and is related to the *k*-Feature Set [15] problem.

The decision version of the Hitting Set problem is defined as follows:

Hitting Set

Instance: A set S and a collection $C \subseteq 2^S$ and an integer k.
Question: Is there a set $S' \subseteq S$ with $|S'| \leq k$ such that for every $c \in C$ we have $c \cap S' \neq \phi$?

The set S' is called a *hitting set for* C, or simply a *hitting set*. For an element $s \in S'$ and an element $C \in C$ if $s \in c$ we say that s *hits* c. This problem is *NP*-complete even when the maximum size of each element of C is two (by equivalence with Vertex Cover [13]) and $W[2]$-complete for parameter k; Cotta and Moscato [16] give a parameterized proof via k-Feature Set and Paz and Moran [17] give a proof which along with the equivalence of Hitting Set and Set Cover leads to the same result, though predates the parameterized complexity framework. However if we restrict the cardinality of the elements of C to d the problem, while remaining *NP*-complete, becomes fixed-parameter tractable where d is a constant and the parameter is k [18]. In this case the problem is known as the Hitting Set for Sets of Size d or d-Hitting Set problem. We note that Hitting Set has several equivalent formulations, in particular we choose to use the bipartite graph representation where S and C form the two partite vertex sets of the graph and an edge sc corresponds to the element $s \in S$ being an element of $c \in C$. This allows us to employ some simplifying graph theoretic terminology and techniques. We generalize this problem to include the case where we may want the elements of C to be hit more than once. In particular this includes the case where we ask if all the sets of C can be hit α times, but extends to the case where the elements of C can be hit up to α times. We encode this by the use of a hitting function η. Our problem then becomes the α-Multiple d-Hitting Set (or (α,d)-Hitting Set):

(α,d)-Hitting Set
Instance: A bipartite graph $G=(S \uplus C,E)$ where for all $c \in C$ we have $d(c) \leq d$, a hitting function $\eta : C \rightarrow [0,\alpha]$ and an integer k.
Question: Is there a set $S' \subseteq S$ with $|S'| \leq k$ such that for every $c \in C$ we have $|N(c) \cap S'| \geq \eta(c)$?

When $\eta(c)=\alpha$ for all $c \in C$, (α,d)-Hitting Set can be $(1+\ln d)$-approximated in time $O(\alpha \cdot |C| \cdot |S|)$ [19], but cannot be approximated with a factor of $(1-\varepsilon)\ln n$ for any $\varepsilon \in (0,1)$ unless $NP \subseteq \mathrm{DTIME}(n^{\log \log n})$ [20].

Results and Discussion

The Fixed-Parameter Tractability of (α,d)-Hitting Set

As we prove in the *Materials and Methods* section, the (α,d)-Hitting Set problem is fixed-parameter tractable, and indeed a more general variant the (α,β,d)-Hitting Set problem is also fixed parameter tractable when we take the maximum degree d of the class vertices C as a constant and the size k of the hitting set and the maximum desired coverage α as a joint parameter. Though the problem is formally hard - which would normally give the intuition that an exact solution would be too expensive to compute - the fixed-parameter tractability indicates that it is likely that we can obtain an exact solution efficiently. Armed with this knowledge we proceed with the experiments of the following section, where we use the drug response data of the NCI60 anti-tumor drug screening program to determine a sets of drugs that hit cancerous cell lines multiple times. These drug sets are than mathematically supportable candidates for combination chemotherapies. Moreover we are able to tune the nature of the hitting sets via the numbers k, α and β, which allows us to control which cell lines are targetted (and which are specifically not) and how much each cell line is hit in the solution.

A Comparative Application

The NCI60 human tumor anti-cancer drug screen dataset [21] was established in the 1980s as an enabling tool for anti-cancer drug development. Included in this dataset is response data for over 40,000 drugs against the 60 cell lines of the dataset. Vazquez [4] highlights the utility of a hitting set approach in developing multi-drug therapies for heterogeneous malignancies; given the plethora of available compounds, testing multi-drug combinations exhaustively is prohibitive if not impossible. Applying hitting set to efficacy data measured on an individual basis for each compound allows us to determine possible drug combinations that would provide the best chance of efficacy against many cancer types. Using the GI50 response NCI60 dataset (available from the DTP website [22]) Vazquez uncovers a minimum hitting set with three compounds that cumulatively gives a good response with all cell lines in the dataset, where a response is considered good if it is more than two standard deviations above the mean of the z-transformed response data. Vazquez uses first a greedy highest-degree-first approach to give an estimate of the maximum size of a minimum hitting set, followed by either an exhaustive search or simulated annealing, depending on the size of the hitting set. Vazquez reports times for such approaches on the order of one day on a desktop computer.

We revisit Vasquez's experiment, using data reduction (though it is not necessary to employ the more complex rules given in the kernelization proof) with IBM ILOG CPLEX [23] as the kernel solver by framing the problem as a integer programming problem. We use the same threshold for the z-transformation to identify significant response levels. Using this approach we reduce the time to solve the instance to less than 5 seconds, where most of the time is spent loading and reducing the data, with CPLEX solving the integer programming instance in approximately 0.08 milliseconds. Furthermore this approach guarantees optimality in the size of the hitting set.

From here we employ more a more recent version of the NCI60 dataset (2009 as compared to Vazquez's 2006). At the time of writing, the latest NCI60 dataset includes 14 additional cell lines, however we remove these, as there is insufficient response data in the dataset, leading to inflated hitting set sizes. The latest data also includes a further 2281 compounds. We note that employing the new GI50 response data we are able to uncover 3 element hitting sets involving compounds not available in the earlier dataset (an example is given in Table 1 and Figure 1), in particular Everolimus (NSC 733504) a drug now used for the treatment of advanced renal cancer which is also giving positive results in phase II trials for metastatic melanoma [24,25]. However there have recently been some concerns over the provenance of some of the cell lines in the NCI60 dataset. In particular Lorenzi *et al.* [26] suggested that the MDA-N cell line, nominally a breast cancer cell line is in fact similar the M14 and MDA-MB-435 cell lines, and thus should be is in fact a melanoma cell line. Chambers [27] however suggests that although M14 and MDA-MB-435 are identical cell lines, they may not in fact be melanoma cell lines. We do not attempt to resolve this dispute, however with regard to this, and as a indication of the flexibility of the method we employ we consider both the case where MDA-N is a breast cancer cell line and the the case where MDA-N is a melanoma cell line.

Employing the (α,β,d)-Hitting Set model gives more flexibility in what kind of therapy we would like to pursue. For instance, by choosing $\eta_1 = 2$ for all vertices, we are able to find a hitting set that hits every cell line at least twice (see Table 2). However the size of

Table 1. Minimal hitting set using 2009 NCI60 data.

NSC Number	Compound Name
174121	Methotrexate Derivative
691039	(S)-7-Hydroxy-1,2,3-trimethoxy-10-methylsulfanyl-6,7-dihydro-5H-benzo[a]heptalen-9-one
733504	Everolimus/Afinitor

Minimal hitting set for NCI60 GI50 response data from 2009.
doi:10.1371/journal.pone.0013055.t001

this hitting set is 6, which is likely to be beyond the point where the trade off between anti-cancer efficacy and side effects is acceptable. Fortunately we can exploit (α,β,d)-Hitting Set more intelligently. For example we may wish to find a hitting set that specifically targets breast cancer cell lines – for which we set all breast cancer cell line vertices to have $\eta_1 = 1$ and all other cell lines to have $\eta_2 = 0$. This gives a hitting set that hits *only* breast cancer cell lines, which may be useful in minimizing unwanted peripheral damage to non-breast cancer cells. This gives a hitting set with three elements. In the case where we considered MDA-N to be a breast cancer cell line (see Table 3 and Figure 2) this set includes the compound deoxypodophyllotoxin, which is known to induce apoptosis [28]. If we consider MDA-N as a melanoma cell line we obtain a different hitting set (see Table 4 and Figure 3). If we relax our requirements an allow other cell lines to be hit at most once we can obtain a hitting set that hits the breast cancer cell lines more (Table 5 and Figure 4). The results when we set η_1 to 2 for all breast cancer lines are given in Table 6 and Figure 5 (including MDA-N) and Table 7 and Figure 6 (excluding MDA-N). We note particularly that in the case where MDA-N is included, the optimal hitting set uncovered includes Docetaxel, a well known anti-cancer agent [29] for several cancer types including breast cancer. Interestingly Docetaxel is also currently included in several clinical trials examining its potential as part of a multi-drug therapy [30–34].

In another example, we may wish to target melanoma cell lines exclusively, and furthermore, we may wish to attack each cell line with at least two drugs at once. However in this case (where $\eta_1 = 2$ for melanoma cell lines and $\eta_2 = 0$ for all others) the minimal hitting set size is 6 (or 5 if MDA-N is included as a melanoma cell line – Table 8 and Figures 7 & 8). Considering that a therapeutic cocktail involving 6 compounds may have excessive side effects, we can relax the requirements, and allow $\eta_2 = 1$ for non-melanoma cell lines. In this case we find that the smallest hitting set is of size 3. By altering the focus when solving the kernel by fixing the hitting set size (k) at 3 and maximizing the total degree of the vertices in the hitting set, subject to the η_1 and η_2 constraints, we can obtain the minimal size hitting set that hits our targets as much as possible, within the bounds given by the constraints. This results in the hitting sets in Tables 9 & 10 and Figures 9 & 10. Of note is AZD6244, which is currently involved in 21 anti-cancer drug trials [35] and has been identified as a potent kinase inhibitor [36,37].

Conclusion

Given the size of modern datasets, and the expectation that they will only get larger, it is clear that we require efficient approaches to solving important computational biology problems. The first phase of any such approach is simply defining the problem at hand. Unfortunately once clearly stated, many such problems are *NP*-hard or worse. However this need not mean that we must resort to inexact or approximate approaches, which could be undesirable in a field such as drug selection. Parameterized Complexity provides a toolkit for dealing with nominally hard problems, and identifying cases where despite super-polynomial running times, we may still expect good performance.

The drug selection problem as examined here is one such problem. It is modeled well by the d-Hitting Set problem, which is fixed-parameter tractable when parameterized by the maximum size of the hitting set. Therefore we can expect that despite being *NP*-complete, it would be relatively quick to solve when these parameters are small. However we demonstrate that the much more flexible variant (α,β,d)-Hitting Set is also fixed-parameter tractable, with only the addition of a single parameter - the maximum of the minimum number of times any vertex should be hit. With (α,β,d)-Hitting Set we are able to better control the nature of the hitting set uncovered, and thus tailor any such hitting set to a useful set of constraints, such as limits on which cell lines are to be hit, the maximum any of these can be hit and of course the minimum number of times any cell line should be hit. Moreover we can solve this problem quickly, and guarantee optimality - without any notable restrictions on the parameters and constants. This allows the quick generation of possible drug combinations for testing, with guarantees of a certain baseline performance, eliminating the need to exhaustively test all possible combinations, which would be financially and temporally prohibitive.

In brief this paper provides a robust and flexible methodology for multiple drug selection, which can easily be applied to other domains that are modeled by the d-Hitting Set problem, with a sound theoretical background as to why and how the problem can be solved efficiently, despite its *NP*-completeness. Moreover the existence of a kernelization for (α,β,d)-Hitting Set indicates that even without using a specialized commercial solver such as CPLEX, the problem is readily scalable to large datasets. Given the speed at which we are able to solve instances with on the order of 40,000 vertices, we can expect that much larger datasets are also solvable in a reasonable time.

A future extension that may be of interest would be to somehow encode in the problem the notion that some hitting vertices are incompatible, e.g., two compound may have severe adverse interactions, and thus can never be used together as a therapy, regardless of their individual usefulness.

Materials and Methods

Dataset and Computational Method

The dataset primarily employed is the NCI60 DTP Human Tumor Cell Line Screen, available from [22]. We use the version released in October 2009, and downloaded in April 2010. The raw dataset is presented as a series of cell line and compound pairs, along with the GI50 response measurement (the method for producing the measurements is also detailed by [22]) for that pair plus concentration information and statistical information. Where there are multiple entries for the same compound-cell line pair, we

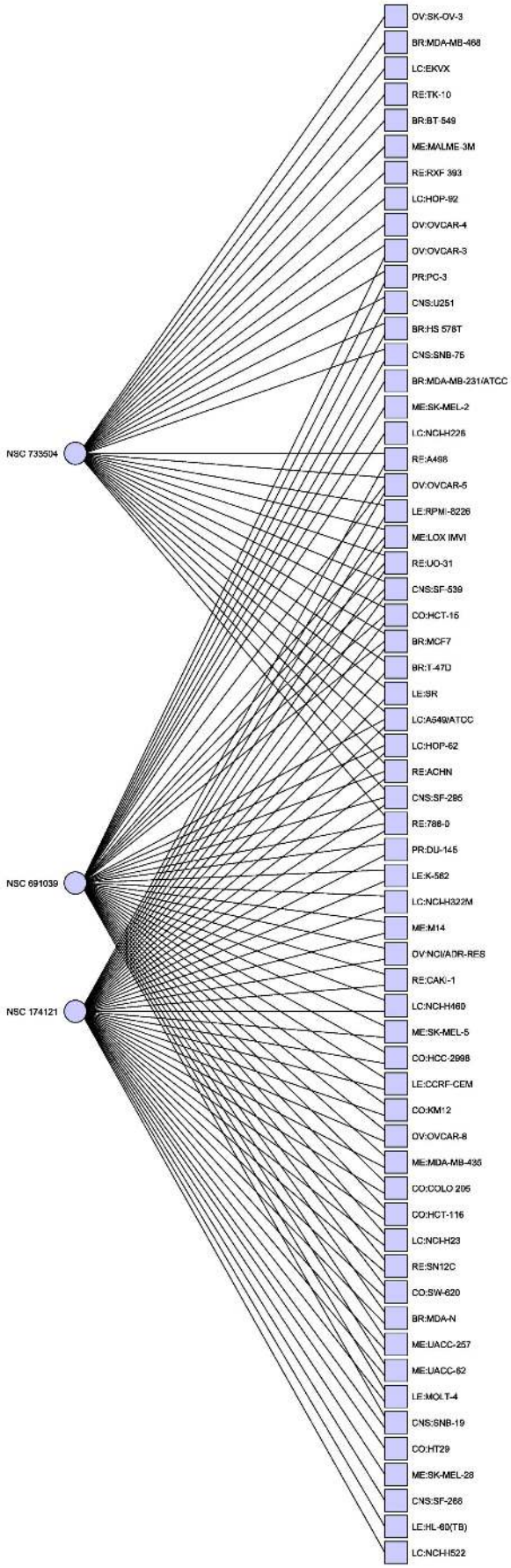

Figure 1. Minimal hitting set hitting for the NCI60 dataset. This hitting set hits all cell lines at least once, but is further optimized to hit all target cell lines the maximal number of times. Of particular note are NSC 174121, a methotrexate derivative and NSC733504, Everolimus/Afinitor, both known anti-cancer agents.
doi:10.1371/journal.pone.0013055.g001

select the entry resulting from the experiment using the highest concentration of the compound. We extract this data into a matrix cross indexed by the NSC number of the compound and the name of the cell line. Where an entry does not exist for a given compound-cell line pair, we enter "NA" for that entry in the matrix.

Once the data is in this matrix format we threshold the data according to the method used by Vazquez [4] whereby the raw data is subject to a z-transformation over a logarithmic scale and then any value above a certain threshold expressed in terms of the standard deviation to 1, and anything below, including "NA" values, to 0. In line with Vazquez we choose two standard deviations as our particular threshold for this paper, though this is adjustable.

We then construct a graph for the hitting set instance using the Java Universal Network/Graph Framework (JUNG) [38] with the SetHypergraph class, representing each compound with a vertex and each cell line with a (hyper)edge which carries a weight indicating the number of times that edge is to be hit. This graph is then reduced to remove vertices of zero degree, edges with no incident vertices (which are noted as technically this would indicate a no instance unless that edge does not require hitting) and vertices that are only adjacent to edges that require zero hits. This basic reduction alone typically reduces the number of vertices significantly, bringing the graph within a reasonable size for immediate processing. From a theoretical standpoint the constant d is of importance, for the graph constructed as stated, $d=4741$ (as we allow the natural value, rather than imposing an external limit). In practice a d value of this magnitude proves perfectly workable, and returning to the theoretical viewpoint indicates that the instance is in a sense already kernelized.

Once the graph is reduced, we construct an integer programming instance equivalent of the problem given the graph, and pass this instance to CPLEX [23] (version 11.200) and search for an optimal solution to one of two objective functions, given the constraints of the number of hits for each cell line (given by the η_1 value). The first objective function simply minimizes the size of the hitting set (k), for the second objective function we fix the size of the hitting set, and maximize the number of hits on vertices where no maximum number of hits has been set (the η_2 value). As part of this search CPLEX may apply some unspecified proprietary reduction process.

The figures were created using yEd Graph Editor [39].

The computer hardware employed is a Dell PowerEdge III Dual Xeon 5550 server with 32Gb of RAM, operating Red Hat Linux 64 bit EL 4 Server.

Theoretical Background and Kernelization Proof

Graph Theory and Notation. A *(simple undirected) graph* consists of a set V (the vertices), and a set E of two element subsets of V (the edges). A *bipartite graph* is a graph where the vertices are partitioned into two partite sets, where all edges have one endpoint in one set and the other endpoint in the other set, i.e., $V=V_1 \uplus V_2$ and $E \subseteq V_1 \times V_2$.

Given a graph $G=(V,E)$ and two vertices $u,v \in V$, we denote the edge between u and v by uv or equivalently vu. Given two vertices u,v in V, if there is an edge $uv \in E$ we say that u and v are

Table 2. Minimal double hitting set.

NSC Number	Compound Name
147340	Anisomycin hydrochloride
174121	Methotrexate derivate
314018	Ansamitocin derivate TN-006
691039	(7S)-7-hydroxy-1,2,3-trimethoxy-10-methylsulfanyl-6, 7-dihydro-5H-benzo[a]heptalen-9-one
712807	Capecitabine
733504	Everolimus/Afinitor

Minimal hitting set hitting each cell line at least twice.
doi:10.1371/journal.pone.0013055.t002

Table 3. Minimal hitting set targeting only breast cancer.

NSC Number	Compound Name
403148	Deoxypodophyllotoxin
697188	2-(4-methoxyphenyl)-5-[8-[5-(4-methoxyphenyl)-1,3,4-oxadiazol-2-yl]octyl]-1,3,4-oxadiazole
732011	21-(2-N,N-Diethylaminoethyl)oxy-7.alpha.-methyl-19-norpregna-1,3,5(10)-triene-3-O-sulfamate

Minimal hitting set hitting breast cancer cell lines at least once, and all other cell lines zero times.
doi:10.1371/journal.pone.0013055.t003

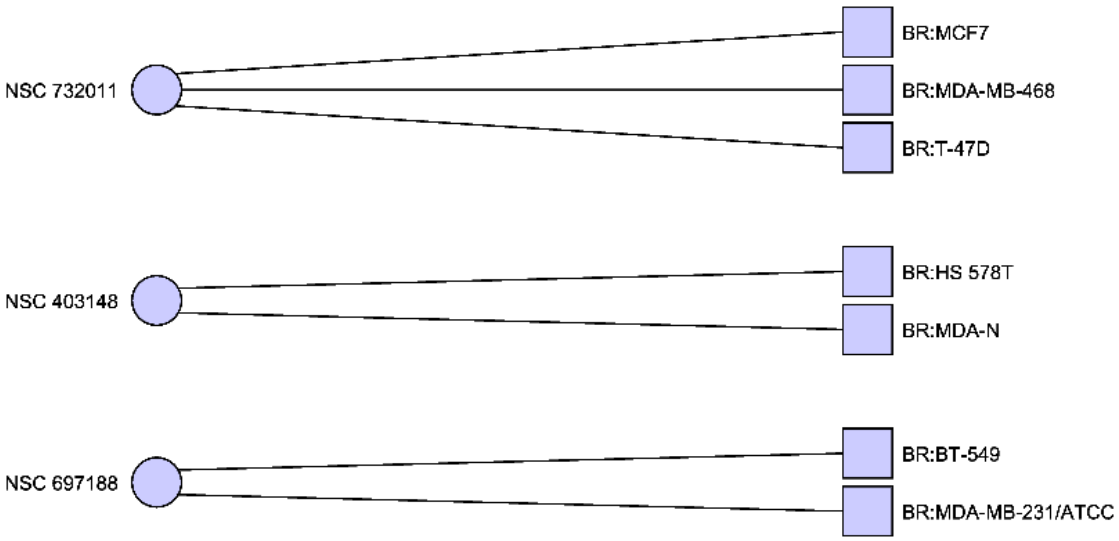

Figure 2. Minimal hitting set hitting only breast cancer cell lines. Including the disputed MDA-N cell line. This hitting set also reveals additional structure with each drug targeting a specific, disjoint subset of the breast cancer cell lines. Only cell lines with at least one adjacent compound are shown.
doi:10.1371/journal.pone.0013055.g002

Table 4. Minimal hitting set targeting only breast cancer without MDA-N.

NSC Number	Compound Name
630678	Streptomyces antibiotic
732011	21-(2-N,N-Diethylaminoethyl)oxy-7.alpha.-methyl-19-norpregna-1,3,5(10)-triene-3-O-sulfamate
734235	isoindolo[1,2-a]quinoxalin-4(5H)-one

Minimal hitting set hitting breast cancer cell lines at least once, and all other cell lines zero times.
doi:10.1371/journal.pone.0013055.t004

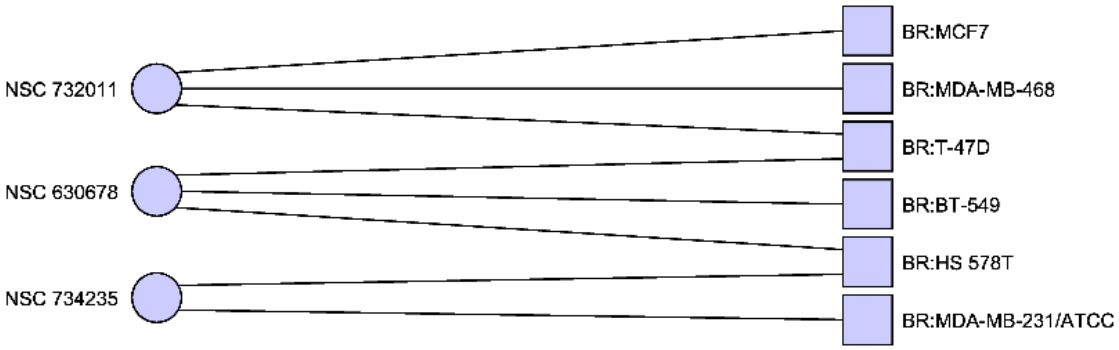

Figure 3. Minimal hitting set hitting only breast cancer cell lines. Excluding the disputed MDA-N cell line. In this case the hitting set is much less clearly separated, though two of the cell lines are now hit twice. Only cell lines with at least one adjacent compound are shown.
doi:10.1371/journal.pone.0013055.g003

adjacent and the u and v are *incident* on uv. Given a vertex $u \in V$, the set $N(u)$ is the *(open) neighborhood* of u and consists off all vertices adjacent to u in G, we extend this notion in the natural way to sets of vertices.

Parameterized Complexity. A *parameterized (decision) problem* is a formally defined computational problem consisting of three components; the input, a special part of the input called the parameter, and the question. Following Flum and Grohe's [40] definition we may assume that the parameter is derived from a polynomial time computable mapping from the input to the natural numbers. A parameterized problem Π is *fixed-parameter tractable* if there is an algorithm $\mathcal{A}$ such that for every instance (x,k) where x is the input, k is the parameter and $|x|=n$, $\mathcal{A}$ correctly answers YES or NO in time bounded by $f(k)p(n)$ where p is a polynomial and f is a computable function.

A *polynomial time kernelization* (or just *kernelization*) is a polynomial time mapping that given an instance (x,k) of a parameterized problem produces a new instance (x',k') of the problem such that:

1. x is a YES-instance if and only if x' is a YES-instance,
2. $k' \leq k$ and
3. $|x'| \leq g(k')$ for some computable function g.

It is easy to see that if a problem has kernelization, then it is fixed-parameter tractable. It is also easy to prove that if a problem is fixed-parameter tractable, then it has a kernelization [41].

Parameterized complexity has a fully developed theory for determining when a problem is unlikely to be fixed-parameter tractable, but as this is not necessary for this work, we refer the reader to the monographs of Flum and Grohe [40] and Downey and Fellows [42] for full discussion, and simply state that if a problem is $W[t]$-hard or $W[t]$-complete for any $t \in \mathbb{N}^+$, then the problem is not fixed-parameter tractable unless certain complexity theoretic assumptions are false, which seems unlikely.

Table 5. Minimal hitting set targeting breast cancer but allowing other cell lines to be hit.

NSC Number	Compound Name
652903	Saframycin AR1(AH2)
685006	2-imino-8-methoxy-N-phenylchromene-3-carboxamide
733504	Everolimus/Afinitor

Minimal hitting set hitting breast cancer cell lines at least once, and all other cell lines zero times.
doi:10.1371/journal.pone.0013055.t005

The Fixed-Parameter Tractability of (α,d)-Hitting Set

Our kernelization for (α,d)-HITTING SET follows the basic format of Abu-Khzam's kernelization for d-HITTING SET [18].

Let (G,k) be an instance of (α,d)-HITTING SET which we assume to have been preprocessed for nonsense input such as vertices $c \in C$ with $d(c) > d$ or $d(c) < \eta(c)$. Therefore we may assume that for all $c \in C$ we have $\eta(c) \leq d(c) \leq d$ and that for all vertices $s \in S$ we have $d(s) \geq 0$.

We first apply Reduction Rules 1 to 3 exhaustively, before applying Rules 4 and 5.:

Reduction Rule 1: If there is a vertex $c \in C$ with $d(c) = \eta(c)$ then for every vertex $s \in N(c)$ for every vertex $b \in N(s)$ reduce $\eta(b)$ by 1, delete s from G and reduce k by 1. Finally, delete c from G.

Lemma 1 *Reduction Rule 1 is sound.*

Proof. If such a vertex c exists, then all its neighbors in S must be in the hitting set, and we can remove them from the graph after suitably noting the effect for the vertices of $N(N(c))$.

Note in particular that this rule effectively allows us to assume that m is at most $d-1$. This will be used implicitly in Reduction Rule 4.

Reduction Rule 2: If there is a vertex $c \in C$ with $\eta(c) = 0$, delete c from G.

Lemma 2 *Reduction Rule 2 is sound.*

Proof. Clearly c requires no vertices to hit it, so may be ignored.

Reduction Rule 3: If there are two vertices $c,b \in C$ such that $N(c) \subseteq N(b)$ and $\eta(c) \geq \eta(b)$, delete b from G.

Lemma 3 *Reduction Rule 3 is sound.*

Proof. If two such vertices c and b exist, then any hitting set that hits c at least $\eta(c)$ times will hit b at least $\eta(b) < \eta(c)$ times.

Let $B \subseteq S$ be a set of size $d-1$ vertices such that B is the pairwise intersection of the neighborhoods of a vertex set $N \subseteq C$. Let $N_i = \{n \in N | \eta(n) = i\}$.

Reduction Rule 4: Let $B \subseteq S$ and $N \subseteq C$ be vertex sets as described. For each $i \in [1,\alpha]$ such that $|N_i| > k$ add a vertex c to C with $\eta(c) = i$ and edges such that $N(c) = B$ and delete N_i from G.

Lemma 4 *Reduction Rule 4 is sound.*

Proof. Let (G,k) be a YES-instance of (α,d)-HITTING SET. Then there is a set $A \subseteq S$ with $|A| \leq k$ that hits each element c of C at least $\eta(c)$ times. Assume that there are sets B and N as described in the reduction rule and that for some $i \in [1,\alpha]$ we have that $|N_i| > k$. Let A_{N_i} be the subset of A that hits N_i. Assume further that $A_{N_i} \not\subseteq B$, then for each $n \in N_i$ there is at least one other vertex in

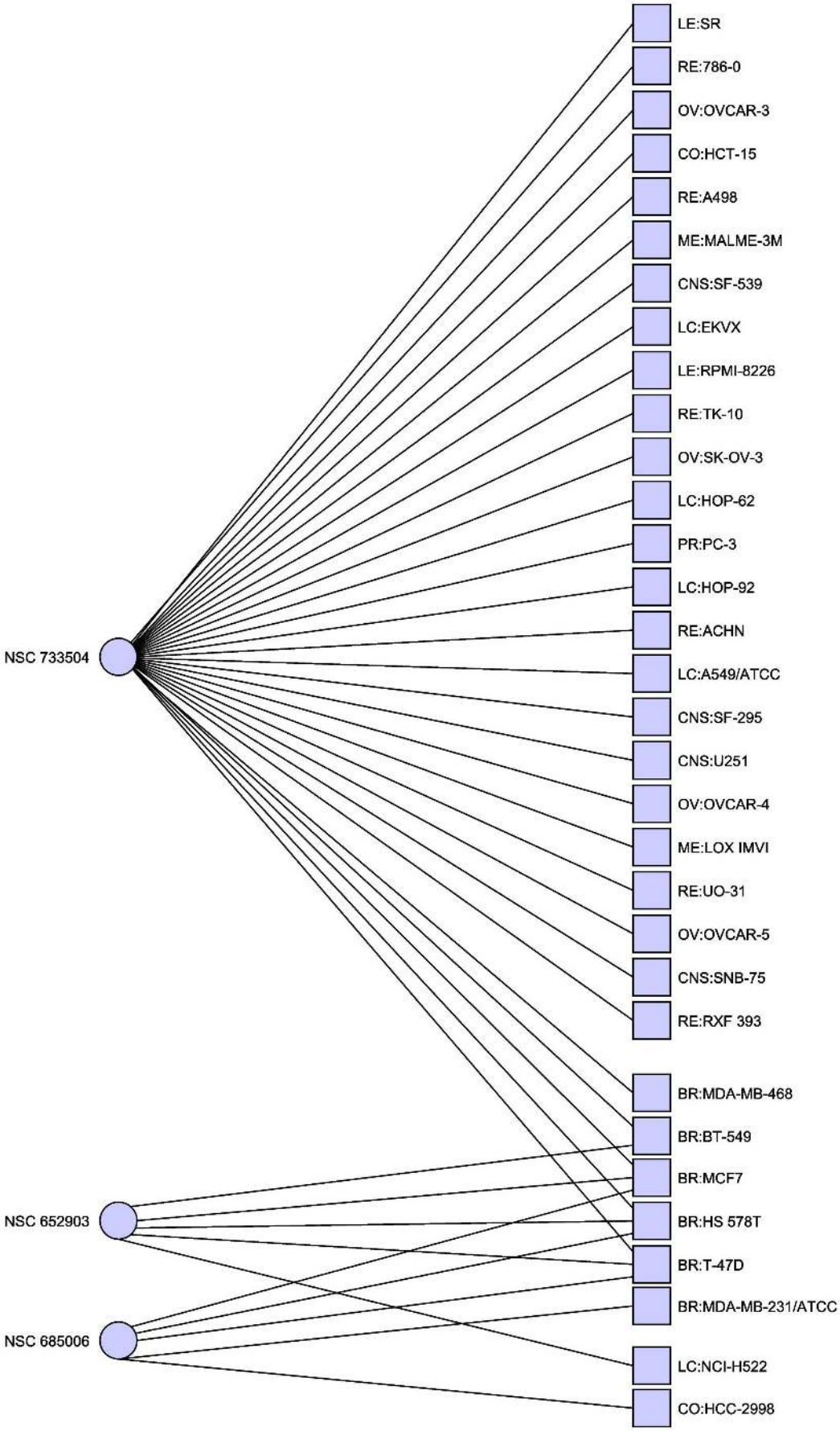
LE:SR
RE:786-0
OV:OVCAR-3
CO:HCT-15
RE:A498
ME:MALME-3M
CNS:SF-539
LC:EKVX
LE:RPMI-8226
RE:TK-10
OV:SK-OV-3
LC:HOP-62
PR:PC-3
LC:HOP-92
RE:ACHN
LC:A549/ATCC
CNS:SF-295
CNS:U251
OV:OVCAR-4
ME:LOX IMVI
RE:UO-31
OV:OVCAR-5
CNS:SNB-75
RE:RXF 393
BR:MDA-MB-468
BR:BT-549
BR:MCF7
BR:HS 578T
BR:T-47D
BR:MDA-MB-231/ATCC
LC:NCI-H522
CO:HCC-2998
NSC 733504
NSC 652903
NSC 685006

Figure 4. Minimal hitting set hitting only breast cancer cell lines. Excluding the disputed MDA-N cell line. In this case we allow non-breast cancer cell lines to be hit at most once. By relaxing the restriction on hitting non-breast cancer cell lines, we obtain a hitting set which hits more of the breast cancer cell lines repeatedly. The trade-off being that other cell lines are also affected, increasingly the likelihood that non-cancerous cells are also affected by the treatment, as the compounds are less specific to a particular genetic signature. Only cell lines with at least one adjacent compound are shown.
doi:10.1371/journal.pone.0013055.g004

Table 6. Minimal hitting set hitting breast cancer twice, and no others, with MDA-N.

70929	Hedamycin
156565	1-hydroxy-4-[4-(2-hydroxyethyl)anilino]anthracene-9,10-dione
628503	Docetaxel
697188	2-(4-methoxyphenyl)-5-[8-[5-(4-methoxyphenyl)-1,3,4-oxadiazol-2-yl]octyl]-1,3,4-oxadiazole
732011	21-(2-N,N-Diethylaminoethyl)oxy-7.alpha.-methyl-19-norpregna-1,3,5(10)-triene-3-O-sulfamate
734235	isoindolo[1,2-a]quinoxalin-4(5H)-one

Minimal hitting set hitting breast cancer cell lines at least once, and all other cell lines zero times.
doi:10.1371/journal.pone.0013055.t006

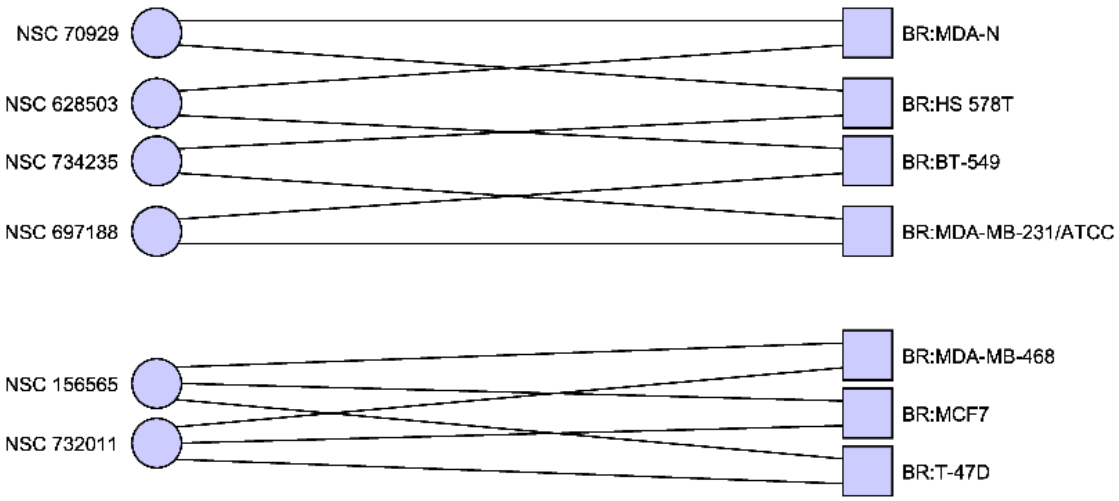

Figure 5. Minimal hitting set hitting breast cancer cell lines twice. Including the disputed MDA-N cell line. In this case the breast cancer cell lines separate neatly into two groups, with the first group forming a cycle and the second group forming a complete bipartite graph. Only cell lines with at least one adjacent compound are shown.
doi:10.1371/journal.pone.0013055.g005

Table 7. Minimal hitting set hitting breast cancer twice, and no others, without MDA-N.

156565	1-hydroxy-4-[4-(2-hydroxyethyl)anilino]anthracene-9,10-dione
630678	Streptomyces antibiotic
697188	2-(4-methoxyphenyl)-5-[8-[5-(4-methoxyphenyl)-1,3,4-oxadiazol-2-yl]octyl]-1,3,4-oxadiazole
698400	5-(1,3-benzodioxol-5-yl)-1,2,3,4-tetrahydrobenzo[a]phenanthridine
732011	21-(2-N,N-Diethylaminoethyl)oxy-7.alpha.-methyl-19-norpregna-1,3,5(10)-triene-3-O-sulfamate

Minimal hitting set hitting breast cancer cell lines at least once, and all other cell lines zero times.
doi:10.1371/journal.pone.0013055.t007

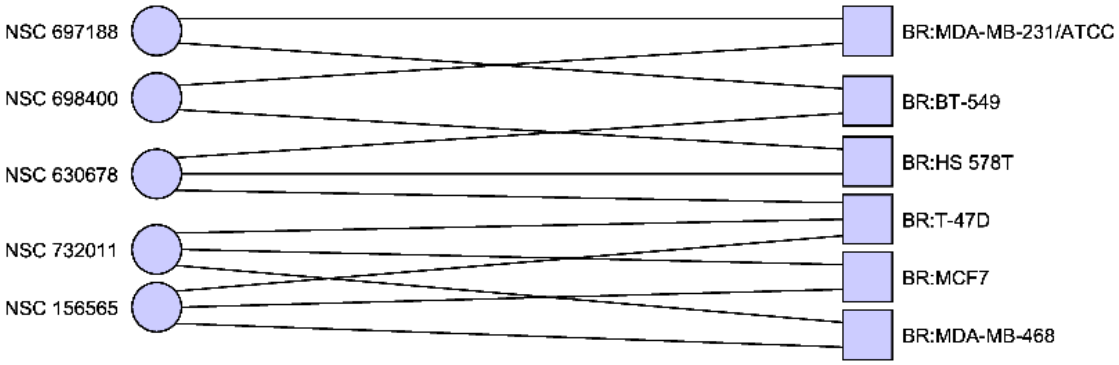

Figure 6. Minimal hitting set hitting breast cancer cell lines twice. Excluding the disputed MDA-N cell line. Without the MDA-N cell line, the breast cancer cell lines do not separate, although the complete bipartite component is a subgraph of this graph, however we gain a greater number of hits per cell line in this case. Only cell lines with at least one adjacent compound are shown.
doi:10.1371/journal.pone.0013055.g006

Table 8. Minimal hitting set targeting melanoma twice, without MDA-N.

624206	N-[2-[(4-chlorophenyl)methyldisulfanyl]ethyl]decan-1-amine hydrochloride
646807	2-(2-Isonicotinoylhydrazino)-N-(3-methyl-1,4-dioxo-1,4-dihydro-2-naphthalenyl)-2-oxoacetamide
674092	2-phenyl-N-[3-[4-[3-[(2-phenylquinoline-4-carbonyl)amino]propyl]piperazin-1-yl]propyl]quinoline-4-carboxamide hydrochloride
677944	6-[2-(4-hydroxy-3-methoxyphenyl)ethylamino]quinoline-5,8-dione
697989	dicopper 2-acetyloxy-3,5-di(propan-2-yl)benzoate
708559	2-(3,4-dichlorophenyl)-N-methyl-N-[3-[methyl(3-pyrrolidin-1-ylpropyl)amino]propyl]acetamide

Minimal hitting set hitting melanoma cell lines at least twice and no others. This result does not include MDA-N as a melanoma cell line.
doi:10.1371/journal.pone.0013055.t008

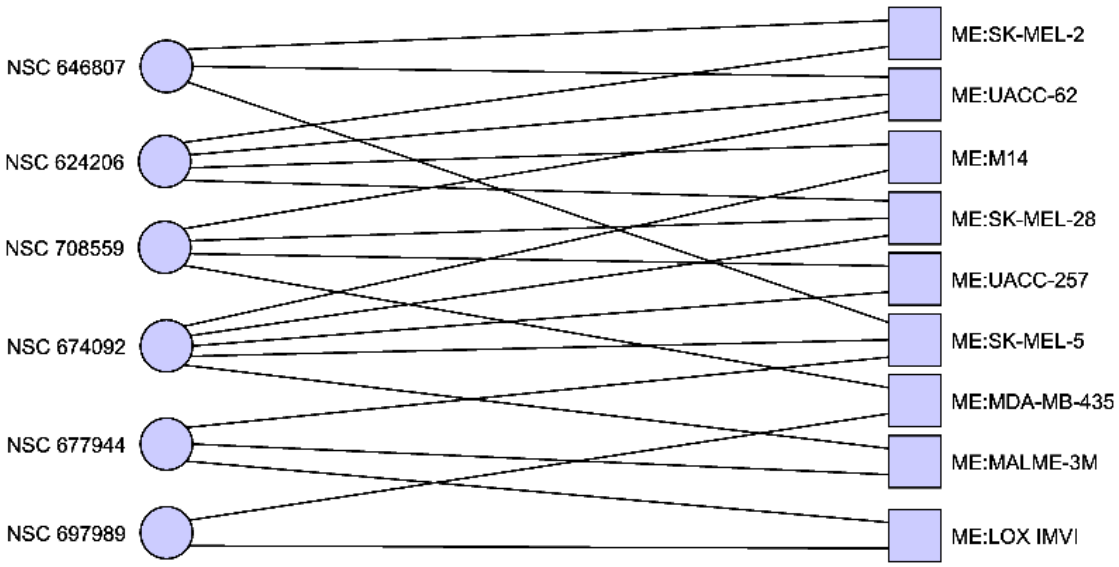

Figure 7. Minimal hitting set hitting melanoma cell lines at least 2 and no other cell lines. This hitting set also maximizes the number of hits on the melanoma cell lines. Only cell lines with at least one adjacent compound are shown.
doi:10.1371/journal.pone.0013055.g007

A_{N_i}, but then $|A| \geq |A_{N_i}| > k$, which contradicts the assumption that (G,k) is a YES-instance.

Therefore the set N_i must be hit by B, so we may restrict our search to the intersection. .

Lemma 5 *Reduction Rule 4 can be computed in polynomial time.*

Proof. Given a set of vertices $B \subseteq N(s)$ for some $s \in S$ with $|B| = d-1$, we construct an auxiliary graph G' by taking for each i the subgraph of G induced by the vertices $N(N_i) \setminus B$. If there is a

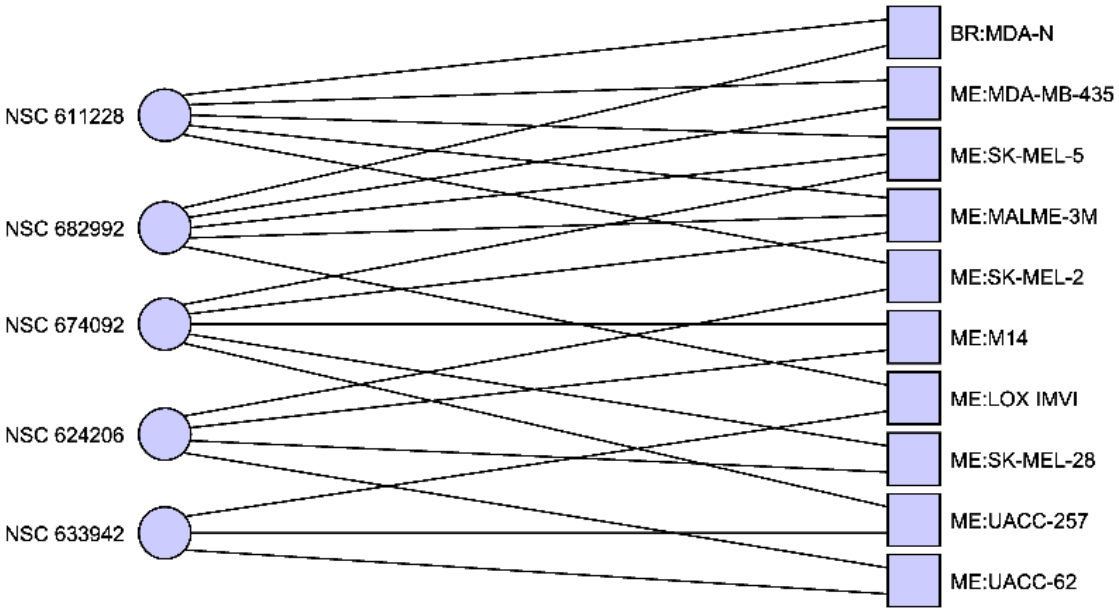

Figure 8. Minimal hitting set hitting melanoma cell lines at least 2 and no other cell lines. Including the disputed MDA-N cell line. It is interesting to note that including MDA-N as a melanoma cell line rather than a breast cancer cell line reduces the size of the minimal hitting set from 6 to 5. This hitting set also maximizes the number of hits on the melanoma cell lines. Only cell lines with at least one adjacent compound are shown.
doi:10.1371/journal.pone.0013055.g008

Table 9. Minimal hitting set targeting melanoma, without MDA-N.

NSC Number	Compound Name
646807	2-(2-Isonicotinoylhydrazino)-N-(3-methyl-1,4-dioxo-1,4-dihydro-2-naphthalenyl)-2-oxoacetamide
656238	2-Methyl-4,8-dihydrobenzo[1,2-b:5,4-b′]dithiophene-4,8-dione
741078	AZD6244 (ARRY-142886)

Minimal hitting set hitting melanoma cell lines at least twice, all others at most once, maximizing the degree of the melanoma cell line vertices.
doi:10.1371/journal.pone.0013055.t009

Table 10. Minimal hitting set targeting melanoma, with MDA-N.

NSC Number	Compound Name
361127	Destruxin E
624206	N-[2-[(4-chlorophenyl)methyldisulfanyl]ethyl]decan-1-amine hydrochloride
656238	2-Methyl-4,8-dihydrobenzo[1,2-b:5,4-b′]dithiophene-4,8-dione

Minimal hitting set hitting melanoma cell lines at least twice, all others at most once, maximizing the degree of the melanoma cell line vertices.
doi:10.1371/journal.pone.0013055.t010

maximum matching in G' of size greater than k, then the matched vertices from S form the required set with pairwise neighbohood intersection B.

As d is a constant, we can iterate over all sets of vertices of size $d-1$ in time $O(|S|\cdot d)$. The matchings can be computed in time $O(\alpha\cdot|N(B)\cup(N(N(B)\setminus B))|^{3/2})$.

Definition 6 (Weakly Related Vertices) Given two vertices $s,t\in C$, s and t are *weakly related* if $|N(s)\cap N(t)|\leq d-1$, and both $N(s)\not\subseteq N(t)$ and $N(t)\not\subseteq N(s)$.

Let $W\subseteq C$ be a maximal set of pairwise weakly related vertices. Let $B\subseteq S$ be a set of vertices, and denote by W_B the set of vertices of W whose neighborhood is a superset of B. Further denote by $W_{B,i}$ the subset of W_B where for each $v\in W_{B,i}$ we have $\eta(v)=i$.

Reduction Rule 5: Compute a maximal collection W of pairwise weakly related vertices. If $|W|>\alpha k^d$ apply the following algorithm:

for $j=d-1$ **downto** 1 **do**
  for $t=\alpha$ **downto** 1 **do**
    for each set $B\subseteq N(v)$ where $v\in W$ and $|B|=j$ **do**
      if $|W_{B,t}|>k^{d-j}$ **then**
        Add a vertex c to C, edges such that $N(c)='$ and set $\eta(c)=t$.
        Delete $W_{B,t}$ from G.

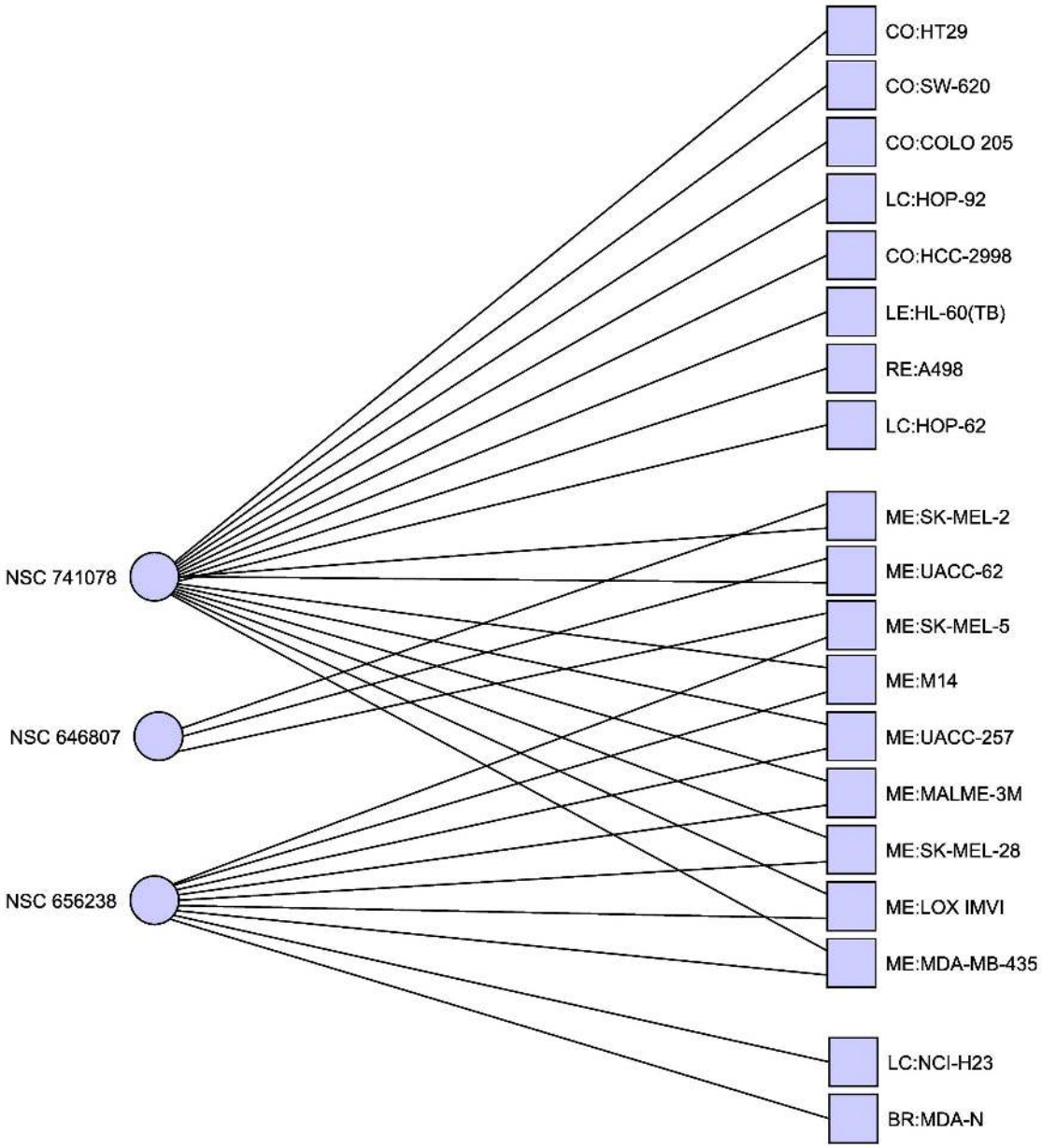

Figure 9. Minimal hitting set hitting melanoma cell lines at least 2 and all other cell lines at most once. For this we consider MDA-N as a non-melanoma cell line, however it is also hit by the hitting set, though only once. This hitting set also maximizes the number of hits on the melanoma cell lines. Only cell lines with at least one adjacent compound are shown.
doi:10.1371/journal.pone.0013055.g009

Lemma 7 *Reduction Rule 5 is sound.*

Proof. We defer the proof of the bound on the size of W until the proof of Lemma 8.

Let (G,k) be a YES-instance of (α,d)-HITTING SET. Then there is a set $A \subseteq S$ that hits S sufficiently. For sets of size $d-1$, Reduction Rule 4 proves the soundness of the first iteration of the outer loop.

For each other iteration, assume that the iteration for sets of size j holds, then let B be set of size $j-1$ where $|W_{B,t}| > k^{d-j+1}$ for some t. If $|A \cap B| < t$ then by the pigeon hole principle there is some vertex $v \in A$ that is in at least k^{d-j} neighborhoods of vertices in $W_{B,t}$, but then $B \cup \{v\}$ is a set that is the intersection of at least k^{d-j} neighborhoods of vertices in some subset of W, contradicting the correctness of the previous iteration. Therefore the entire set of vertices hitting each $W_{B,t}$ vertex is contained within B if $|W_{B,t}| > k^{d-j}$, so we may replace $W_{B,t}$ with a single vertex.

Note also that for each element of W there is at most 2^d sets B, so we may iterate through all sets in time $O(\alpha d 2^d \cdot |W|)$, so we can perform the replacements in polynomial time.

Lemma 8 *If (G,k) is a* YES*-instance of (α,d)-*HITTING SET*, reduced under Reduction Rules 1 to 5, then $|V(G)| \leq (d+1)\alpha k^d$.*

Proof. If (G,k) is a YES-instance of (α,d)-HITTING SET, then there is a set $A \subseteq S$ such that for every $s \in S$ we have $|N(s) \cap A| \geq \eta(s)$ with $|A| \leq k$.

Claim 9 $C = W$.

By construction, every vertex in C with degree at most $d-1$ is in W. Assume there is some $c \in C$ with $d(c) = d$ and $c \notin W$, then

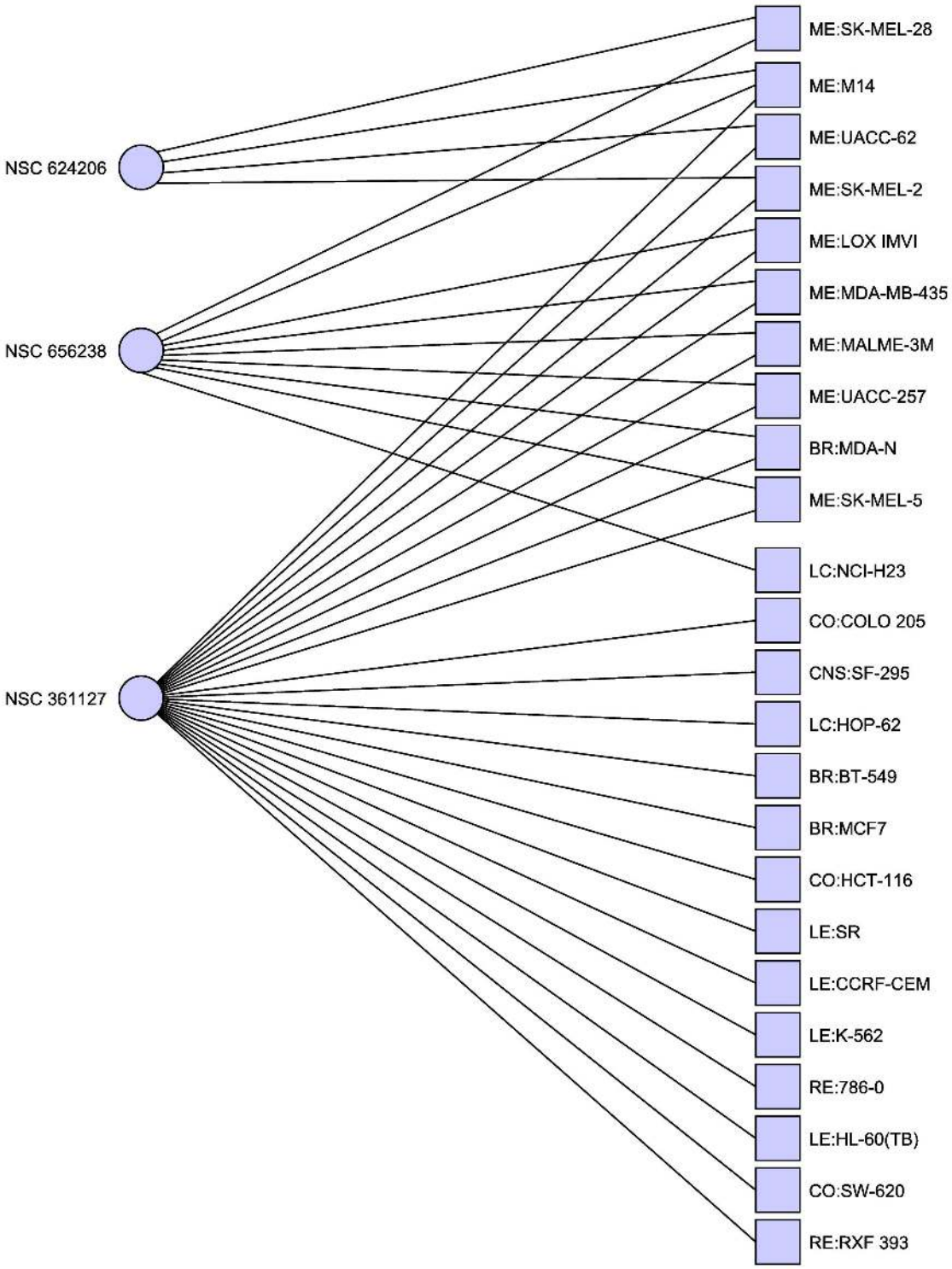

Figure 10. Minimal hitting set hitting melanoma cell lines at least 2 and all other cell lines at most once. Including MDA-N as a melanoma cell line. The key difference with the case where we consider MDA-N to be a non-melanoma cell line is that in this case we obtain a hitting set that hits the melanoma cell lines slightly more. Only cell lines with at least one adjacent compound are shown.
doi:10.1371/journal.pone.0013055.g010

there must be some vertex $c' \in C$ such that $|N(c) \cap N(c')| > d-1$, but then as the degree of any vertex in C is at most d, $N(c) = N(c')$, and Reduction Rule 3 would apply. Therefore there are no vertices from C not in W.

Claim 10 $|W| \leq \alpha k^d$.

As A hits each vertex of W at least once, by Reduction Rule 5 each element of A as a singleton is in the neighborhood of at most αk^{d-1} vertices from C. Therefore $|W| \leq |A| \cdot \alpha k^{d-1} \leq \alpha k^d$.

Combining Claims 9 and 10 we have $|C| \leq \alpha k^d$. As each vertex of C has degree at most d, there are at most αdk^d vertices in S, and the bound follows.

Theorem 11 (α,d)-HITTING SET *is fixed-parameter tractable with parameter k and has a kernel of size at most αdk^d.*

We note that although d must be a constant to obtain a polynomial time kernelization, α may be alternatively given as an additional parameter, without change to the kernelization.

This kernelization may be extended to an even more general version of the problem, where we not only specify lower bounds for the number of hits, but also upper bounds:

(α,β,d)-HITTING SET

Instance: A bipartite graph $G=(S \uplus C,E)$ where for all $c \in C$ we have $d(c) \leq d$, two hitting functions $\eta_1 : C \rightarrow [0,\alpha]$ and $\eta_2 : C \rightarrow [0,\beta]$ and an integer k.
Question: Is there a set $S' \subseteq S$ with $|S'| \leq k$ such that for every $c \in C$ we have $\eta_2(c) \geq |N(c) \cap S'| \geq \eta_1(c)$?

Corollary 12 (α,β,d)-HITTING SET *is fixed-parameter tractable with parameter k and has a kernel of size at most αdk^d.*

Author Contributions

Conceived and designed the experiments: DM LM PM. Performed the experiments: DM LM PM. Analyzed the data: LM PM. Contributed reagents/materials/analysis tools: DM EP LM PM. Wrote the paper: DM EP LM PM.

References

1. Albain KS, Crowley JJ, LeBlanc M, Livingston RB (1990) Determinants of improved outcome in small-cell lung cancer: an analysis of the 2,580-patient southwest oncology group data base. Journal of Clinical Oncology 8: 1563–1574.
2. Flamant F, Schwartz L, Delons E, Caillaud JM, Hartmann O, et al. (1984) Nonseminomatous malignant germ cell tumors in children. Multidrug therapy in stages III and IV. Cancer 54: 1687–1691.
3. Fu KK, Silverberg IJ, Phillips TL, Friedman MA (1979) Combined radiotherapy and multidrug chemotherapy for advanced head and neck cancer: results of a radiation therapy oncology group pilot study. Cancer Treatment Reports 63: 351–357.
4. Vazquez A (2009) Optimal drug combinations and minimal hitting sets. BMC Systems Biology 3: 81–86.
5. Berman P, DasGupta B, Sontag ED (2007) Randomized approximation algorithms for set multicover problems with applications to reverse engineering of protein and gene networks. Discrete Applied Mathematics 155: 733–749.
6. Haus UU, Klamt S, Stephen T (2008) Computing knock-out strategies in metabolic networks. Journal of Computational Biology 15: 259–268.
7. de Kleer J, Mackworth AK, Reiter R (1992) Characterizing diagnoses and systems. Artificial Intelligence 56: 197–222.
8. Leipins GE, Potter WD (1991) A genetic algorithm approach to multiple-fault diagnosis. In: Davis L, ed. Handbook of Genetic Algorithms, Van Nostrand Reinhold Company. pp 237–250.
9. Reiter R (1987) A theory of diagnosis from first principles. Artificial Intelligence 32: 57–95.
10. Hvidsten TR, Lægreid A, Komorowski HJ (2003) Learning rule-based models of biological process from gene expression time profiles using gene ontology. Bioinformatics 19: 1116–1123.
11. Ruchkys D, Song S (2002) A parallel approximation hitting set algorithm for gene expression analysis. In: 14th Symposium on Computer Architecture and High Performance Computing (SBAC-PAD 2002). Vitoria, Espirito Santo, Brazil. pp 75–81.
12. Vinterbo SA, Kim EY, Ohno-Machado L (2005) Small, fuzzy and interpretable gene expression based classifiers. Bioinformatics 21: 1964–1970.
13. Garey MR, Johnson DS (1979) Computers and Intractability: A Guide to the Theory of NP-Completeness. New York: W. H. Freeman & Co.
14. Fernau H (2005) Parameterized Algorithmics: A Graph-Theoretic Approach.
15. Davies S, Russell S (1994) NP-completeness of searches for smallest possible feature sets. In: Greiner R, Subramanian D, eds. AAAI Symposium on Intelligent Relevance. New Orleans, 41–43.
16. Cotta C, Moscato P (2003) The *k*-feature set problem is W[2]-complete. Journal of Computer and System Sciences 67: 686–690.
17. Paz A, Moran S (1981) Non deterministic polynomial optimization problems and their approximations. Theoretical Computer Science 15: 251–277.
18. Abu-Khzam FN (2010) A kernelization algorithm for d-hitting set. Journal of Computer and Systems Sciences 76: 524–531.
19. Vazirani V (2001) Approximation Algorithms. Berlin: Springer-Verlag. 380 p.
20. Feige U (1998) A threshold of ln *n* for approximating set cover. Journal of the ACM 45: 634–652.
21. Shoemaker RH (2006) The NCI60 human tumour cell line anticancer drug screen. Nature Reviews Cancer 6: 813–823.
22. NCI/NIH Website (Accessed 2010). Developmental Theraputics Program. http://dtp.nci.nih.gov/.
23. IBM Website (Accessed 2010). ILOG CPLEX. http://www-01.ibm.com/software/integration/optimization/cplex/.
24. Rao RD, Windschitl HE, Allred JB, Lowe VJ, Maples WJ, et al. (2006) Phase II trial of the mTOR inhibitor everolimus (RAD-001) in metastatic melanoma. Journal of Clinical Oncology 24: 8043.
25. Peyton JD, Spigel DR, Burris HA, Lane C, Rubin M, et al. (2009) Phase II trial of bevacizumab and everolimus in the treatment of patients with metastatic melanoma: Preliminary results. Journal of Clinical Oncology 27: 9027.
26. Lorenzi PL, Reinhold WC, Varma S, Hutchinson AA, Pommier Y, et al. (2009) DNA fingerprinting of the NCI-60 cell line panel. Molecular Cancer Therapeutics 8: 713–724.
27. Chambers AF (2009) MDA-MB-435 and M14 cell lines: Identical but not M14 melanoma? Cancer Research 69: 5292–5293.
28. Shin SY, Yong Y, Kim CG, Lee YH, Lim Y (2010) Deoxypodophyllotoxin induces G2/M cell cycle arrest and apoptosis in HeLa cells. Cancer Letters 287: 231–239.
29. Lyseng-Williamson KA, Fenton C (2005) Docetaxel: A review of its use in metastatic breast cancer. Drugs 65: 2513–2531.
30. Slamon D, Eiermann W, Robert N, Pienkowski T, Martin M, et al. (2009) Phase III randomized trial comparing doxorubicin and cyclophosphamide followed by docetaxel (AC-¿T) with doxorubicin and cyclophosphamide followed by docetaxel and trastuzumab (AC-¿TH) with docetaxel, carboplatin and trastuzumab (TCH) in Her2neu positive early breast cancer patients: BCIRG 006 study. Cancer Research 69: 62.
31. Perez EA, Hillman DW, Dentchev T, Le-Lindqwister NA, Geeraerts LH, et al. (2010) North central cancer treatment group (NCCTG) N0432: phase II trial of docetaxel with capecitabine and bevacizumab as first-line chemotherapy for patients with metastatic breast cancer. Annals of Oncology 21: 269–274.
32. Polyzos A, Malamos N, Boukovinas I, Adamou A, Ziras N, et al. (2010) FEC versus sequential docetaxel followed by epirubicin/cyclophosphamide as adjuvant chemotherapy in women with axillary node-positive early breast cancer: a randomized study of the Hellenic Oncology Research Group (HORG). Breast Cancer Research and Treatment 119: 95–104.
33. Joensuu H, Bono P, Kataja V, Alanko T, Kokko R, et al. (2009) Fluorouracil, Epirubicin, and Cyclophosphamide With Either Docetaxel or Vinorelbine, With or Without Trastuzumab, As Adjuvant Treatments of Breast Cancer: Final Results of the FinHer Trial. Journal of Clinical Oncology 27: 5685–5692.
34. Sparano JA, Makhson AN, Semiglazov VF, Tjulandin SA, Balashova OI, et al. (2009) Pegylated Liposomal Doxorubicin Plus Docetaxel Significantly Improves Time to Progression Without Additive Cardiotoxicity Compared With Docetaxel Monotherapy in Patients With Advanced Breast Cancer Previously Treated With Neoadjuvant-Adjuvant Anthracycline Therapy: Results From a Randomized Phase III Study. J Clin Oncol 27: 4522–4529.
35. ClinicalTrialsgov Website (Accessed 2010). U.S. clinical trial registry. http://clinicaltrials.gov/ct2/home.
36. Davies B, Logie A, McKay JS, Martin P, Steele S, et al. (2007) AZD6244 (ARRY-142886), a potent inhibitor of mitogen-activated protein kinase/extracellular signal-regulated kinase kinase 1/2 kinases: mechanism of action in vivo, pharmacokinetic/pharmacodynamic relationship, and potential for combination in preclinical models. Molecular Cancer Theraputics 6: 2209–2219.
37. Yeh TC, Marsh V, Bernat BA, Ballard J, Colwell H, et al. (2007) Biological characterization of ARRY-142886 (AZD6244), a potent, highly selective mitogen-activated protein kinase kinase 1/2 inhibitor. Clinical Cancer Research 13: 1576.
38. Java Universal Network/Graph Framework Website (Accessed 2010). JUNG. http://jung.sourceforge.net/.
39. yWorks Website (Accessed 2010). yEd. http://www.yworks.com/en/products_yed_about.html.
40. Flum J, Grohe M (2006) Parameterized Complexity Theory. Berlin: Springer. 493 p.
41. Niedermeier R (2006) Invitation to Fixed-Parameter Algorithms. Oxford: Oxford University Press. 316 p.
42. Downey RG, Fellows MR (1999) Parameterized Complexity. Berlin: Springer. 533 p.