BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

A graph theoretic approach to utilizing protein structure to identify non-random somatic mutations

Gregory A Ryslik[1*], Yuwei Cheng[2], Kei-Hoi Cheung[2,3], Yorgo Modis[4] and Hongyu Zhao[1,2*]

Abstract

Background: It is well known that the development of cancer is caused by the accumulation of somatic mutations within the genome. For oncogenes specifically, current research suggests that there is a small set of "driver" mutations that are primarily responsible for tumorigenesis. Further, due to recent pharmacological successes in treating these driver mutations and their resulting tumors, a variety of approaches have been developed to identify potential driver mutations using methods such as machine learning and mutational clustering. We propose a novel methodology that increases our power to identify mutational clusters by taking into account protein tertiary structure via a graph theoretical approach.

Results: We have designed and implemented *GraphPAC* (**G**raph **P**rotein **A**mino acid **C**lustering) to identify mutational clustering while considering protein spatial structure. Using *GraphPAC*, we are able to detect novel clusters in proteins that are known to exhibit mutation clustering as well as identify clusters in proteins without evidence of prior clustering based on current methods. Specifically, by utilizing the spatial information available in the Protein Data Bank (PDB) along with the mutational data in the Catalogue of Somatic Mutations in Cancer (COSMIC), *GraphPAC* identifies new mutational clusters in well known oncogenes such as EGFR and KRAS. Further, by utilizing graph theory to account for the tertiary structure, *GraphPAC* discovers clusters in DPP4, NRP1 and other proteins not identified by existing methods. The R package is available at: http://bioconductor.org/packages/release/bioc/html/GraphPAC.html.

Conclusion: *GraphPAC* provides an alternative to *iPAC* and an extension to current methodology when identifying potential activating driver mutations by utilizing a graph theoretic approach when considering protein tertiary structure.

Background

Cancer, one of the most widespread and heterogeneous diseases, is at its most fundamental level a disease brought on by the accumulation of somatic mutations [1]. These mutations typically occur in either tumor suppressors or oncogenes. While oncogenic mutations either tend to deregulate or up-regulate the resulting protein behavior, mutations within tumor suppressors typically lower the activity of genes that prevent cancer. Pharmacological intervention has shown to be more effective with inhibiting activating oncogenes than with restoring functionality of tumor suppressing genes. Combined with the theory of "oncogene addiction", that many cancers are dependent upon a small set of key genes to drive their rapid cellular multiplication with the rest of the mutations simply being passenger mutations [2,3], the identification of driver oncogenic mutations has become of critical importance in cancer research.

Due to the importance of this problem, several approaches have been proposed to detect naturally selected regions in which activating mutations occur. One general approach postulates that driver mutations will have a higher non-synonymous mutation rate as compared to the background level after normalizing for the length of the

*Correspondence: gregory.ryslik@yale.edu; hongyu.zhao@yale.edu
[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA
[2]Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA
Full list of author information is available at the end of the article

gene [4-6]. Similarly, assuming that the neutral rate of nucleotide substitution is surpassed when positive selection is acting on a specific region, one can check if the ratio of nonsynonymous (K_a) to synonymous (K_s) mutations per site is greater than 1 [7]. Relatedly, Ye *et al.* [8] and Ryslik *et al.* [9] showed that mutational clusters can be indicative of activating mutations and that finding such clusters is a way to reduce the driver mutation search space needing to be analyzed. An alternative approach relies on creating classifiers to categorize mutations. Machine learning algorithms such as *Polyphen-2* [10], which predicts whether a missense mutation is damaging, and *CHASM* [11-13], which discriminates between known driver mutations and a set of passenger mutations, rely upon a set of rules developed using a variety of machine learning techniques such as Random Forests [14] and Support Vector Machines [15]. These rules can be used to calculate a score for each mutation based upon both sequence and non-sequence-based features such as evolutionary conservation, size and polarity of the substituted residue as well as accessible surface area [16]. Other classifiers, such as *SIFT* [17], use only a subset of these features, e.g. evolutionary conservation, for prediction.

While the methods based upon background mutational rates have had some success in identifying regions of positive selections or driver mutations, they nonetheless suffer from several shortcomings. First, many of these methods rely upon calculating the difference between synonymous and non-synonymous mutations but do not take into account that selection can act upon minute regions of the gene. Thus, when the mutations rates are averaged over the entire gene, the signal may be lost. Second, the methods proposed by Kreitman [7] and Wang [4] do not differentiate between activating gain-of-function mutations and inactivating loss-of-function non-synonymous mutations. Third, many of the machine learning methods require an extensive rule set that must first be trained using a well annotated database that is still limited. Until the requisite literature and information is developed, the machine learning algorithm is unable to create a well-performing classifier. Furthermore, the rules must be updated periodically to reflect updated knowledge and information. For a recent review of several popular methods that attempt to discern missense substitution effect on protein function see Gnad *et al.* [18] and Gonzalez-Perez *et al.* [19].

Building on the work of Bardelli *et al.* [5] and Torkamani and Schork [20], which stipulated that only a small number of specific mutations can activate a protein, Ye *et al.* [8] developed Non-Random Mutational Clustering (*NMC*) to identify potential activating mutations. *NMC* works on the hypothesis that absent any previously known mutational hotspot, a mutational cluster is indicative of a possible activating mutation. This is based on the observation that most amino acid substitutions are either neutral or incompatible with protein function, resulting in a concentration of activating mutations within a small subset of protein residues and domains [8]. For the null hypothesis that mutation locations are random in the candidate protein when represented in linear form, *NMC* identifies clustering by evaluating whether there is statistical evidence of mutations occurring closer together on the line than expected by chance. While *NMC* is able to implicate some cancer related genes, it is limited by the fact that it considers the protein as a linear sequence and does not take into account the tertiary protein structure. To account for protein structure information, Ryslik *et al.* [9] developed *iPAC* (**i**dentification of **P**rotein **A**mino acid **C**lustering), which reorganizes the protein into a one dimensional space that preserves, as best as possible, the three dimensional amino acid pairwise distances using Multidimensional Scaling (MDS) [21]. As described by Ryslik *et al.* [9], utilizing the tertiary information is critical when identifying clustering as mutations that occur far apart when the protein is considered linearly can be very close together once the protein is folded in 3D space. The 3D proximity of such mutations might thus yield novel clusters. While it was shown that *iPAC* provides an improvement over *NMC*, the reliance upon a global method like MDS can potentially result in a distorted rearrangement of the protein, since distant residues will nevertheless have an impact on each other's final position in one dimensional space.

In this manuscript, we provide an alternative method to *iPAC* by remapping the protein into one dimensional space via a graph theoretic approach. This approach allows for a more natural consideration of the protein, one that is sensitive to protein domains and linkers. We show that our methodology is effective in identifying proteins with mutational clustering that are missed by both *iPAC* and *NMC* such as NRP1 and MAPK24. We also show that for some proteins, *GraphPAC* identifies fewer clusters than inferred by both *iPAC* and *NMC* while for other proteins *GraphPAC* identifies more clusters than the other two methods. While both *GraphPAC* and *iPAC* are an improvement over *NMC* since they account for tertiary structure, the differences between *GraphPAC* and *iPAC* point to the fact that different rearrangements of the protein must be considered in order to better understand the mutational clustering landscape. We show that many of the clusters identified by *GraphPAC* are also classified as damaging by *Polyphen-2* and as an activating mutation by *CHASM*. By providing a more complete picture of mutational clustering than *iPAC* or *NMC* individually, *GraphPAC* allows us to obtain a more accurate landscape of where potential activating mutations may occur on the protein.

Methods

GraphPAC uses a four step approach to identifying mutational clusters. The first step, as described in Sections 'Obtaining mutational data' and 'Obtaining the 3D structural data', retrieves mutational and positional data from COSMIC [22] and the PDB [23], respectively. After reconciling the mutational and positional databases (Section 'Reconciling the structural and mutational data'), the residues are realized as a connected graph where each residue is a vertex whereupon the traveling salesman problem is heuristically solved in order to find the shortest path through the protein (Section 'Traveling salesman approach'). Once the shortest path has been identified, the protein residues are reordered along this path providing a one dimensional ordering of the protein. The linear *NMC* algorithm is then used to calculate which mutations are closer together than expected by chance. Lastly, the clusters are unmapped back into the original space and the results reported back to the user. We detail each of the steps in the sections below.

Obtaining mutational data

The mutational positions were obtained from the 58th version of the COSMIC database that was downloaded via the following ftp site: ftp.sanger.ac.uk/pub/CGP/cosmic. The database was implemented locally using Oracle 11g. Only missense mutations that were classified as "Confirmed somatic variant" or "Reported in another cancer sample as somatic" were selected, with nonsense and synonymous mutations excluded. Moreover, we only considered mutations originating from studies that were classified as whole gene screens. Next, since multiple studies can report mutational data from the same cell line, mutational redundancies were removed to avoid double counting the mutations. Lastly, only the proteins with a UniProt Accession Number [24] were kept in order to correctly match the mutational and positional data, resulting in 777 proteins. See "COSMIC query" in *Additional file 1* for the SQL code required to generate the mutational data.

Obtaining the 3D structural data

The PDB web interface was used to obtain the protein tertiary information for each of the 777 proteins described in Section 'Obtaining mutational data'. Since multiple structures are often available for the same protein, all structures with a matching UniProt Accession Number were used and an appropriate multiple comparisons adjustment (see Section 'Multiple comparison adjustment for structures') was performed afterwards. For proteins where the resolution provided alternative conformations, the first conformation listed in the file was used. Similarly, for structures where more than one polypeptide chain with a matching Uniprot Accession Number was available, the first matching chain listed in the file was used (typically chain A). Finally, after the chain and conformation were selected, the cartesian coordinates of all the α-carbon atoms were used to represent the tertiary backbone structure of the protein. While we only used the α-carbon location to represent the residue location in this paper, our methods are robust if any of the other backbone atoms are used including the amide nitrogen, main chain carbonyl carbon or the main chain carbonyl oxygen.

Also, while X-ray crystallography was used to determine many of the tertiary structures in the PDB, we note that molecular dynamics (MD) could in principle be used to model the protein structure in solution. However, taking into account the time complexity of such an approach for larger proteins as well as the number of structures that we consider, such a task is beyond the scope of this paper [25]. Further, as crystal structures are almost always representative of the correctly folded protein, using the current structural information is more than sufficient until MD simulations can be applied on much faster time scales. See "Structure Files" in *Additional file 2* for a full listing of all the 1,904 structure/chain combinations used.

Reconciling the structural and mutational data

In order to reference the same residue in the COSMIC and PDB databases, an alignment was performed to accommodate their different numbering systems. Like *iPAC*, *GraphPAC* allows two such reconciliations. The first is based upon a pairwise alignment as described in Pages *et al.* [26] while the second is based upon a numerical reconstruction from the structural information available in the PDB file. Due to the fact that the PDB file structure potentially changes depending upon the structure release date along with other technical complications, pairwise alignment was used for all the analysis described in this paper unless specifically noted. For further information on the alignment please see the documentation in the *GraphPAC* package available on Bioconductor. Protein/structure/chain combinations that resulted in only one mutation or no mutations on the residues for which tertiary information was available were dropped. Similar to *iPAC*, a successful alignment of the tertiary and mutational data was obtained for 140 proteins corresponding to 1100 unique structure/chain combinations. See "Structure Files" in *Additional file 2* for a full listing and description.

Traveling salesman approach

Since the *NMC* algorithm requires order statistics to identify clustering (see Section 'NMC'), we need to map the protein from a three dimensional to a one dimensional space so that order statistics may be constructed. Contrary to *iPAC*, which employed MDS, a graph theoretic approach is used by *GraphPAC*. As discussed above, one

major limitation of MDS is that the minimization of the stress function:

$$\sigma_1 = \sqrt{\frac{\sum_{i,j}[f(\delta_{i,j}) - d_{i,j}(\mathbf{X})]^2}{\sum_{i,j} d_{i,j}^2(\mathbf{X})}} \quad (1)$$

results in every residue having an effect on the final position of every other residue. In Equation 1, δ_{ij} represents the Euclidean distance between residues i and j in the original higher-dimensional space while $d_{i,j}(\mathbf{X})$ represents the distance between them in the lower dimensional space $\mathbf{X}$. Lastly, f, is used to account for situations where the proximity measures $\delta_{i,j}$ do not come from a true metric space. Since in our case, $\delta_{i,j} \in \mathbb{R}$, f is the identity function. Minimization of σ_1 may not capture that a protein is typically comprised of several domains and that only residues within a specific domain should influence each other's final position in linear space (see Figure 1).

Under the *GraphPAC* algorithm, we first construct a complete graph with each residue represented by a vertex[a]. We then create a linear ordering of the protein by finding a Hamiltonian[b] path through the graph. As the number of distinct Hamiltonian paths on a complete graph with N vertices is equal to $\frac{(N-1)!}{2}$, a direct consideration of all possible paths is computationally unfeasible. Further, selective pruning of the edges based upon edge distance is also often impractical due to the domain structure where many residues are close to each other. Because of these factors, we use a heuristic algorithm that solves the Traveling Salesman Problem (TSP) [27,28] to find a linear path that is approximate of the shortest path through the protein. We then use this path as a representative reordering of the protein into one dimensional space. Unlike *iPAC*, which is based on a global remapping, this methodology takes into account only locally neighboring residues to remap the protein to one dimensional space.

While there are many heuristic solutions for the TSP (see Gutin and Punnen [29]), we consider three of the most common insertion methods [30]: cheapest insertion, farthest insertion and nearest insertion as described below. Specifically, the objective of the TSP is to find a cyclic permutation π of $\{1, 2, 3, \ldots, n\}$ that minimizes the total tour distance, namely:

$$\min_{\pi} \sum_{i=1}^{n} d(i, \pi(i))$$

Here, $d(i, j)$ represents the distance between residues i and j (with $d(i, i) = 0$) and $\pi(i)$ represents the residue that follows residue i on the tour. The difference between the three insertion methods rests on how the next residue k is selected for insertion. Under cheapest insertion, the next k to be inserted into the tour is chosen such that the increase in tour length is minimal. Under nearest insertion, at each iteration, the k that is closest to a residue already on the tour is selected. Finally, under farthest insertion, the k that is farthest away from any residue already on the tour is selected.

These algorithms have different upper bounds on their tour lengths. For example, the farthest insertion algorithm creates tours that approach $\frac{3}{2}$ of the shortest length while the nearest and cheapest insertion algorithms can be linked to the minimal spanning tree algorithm and thus have an upper bound of twice the shortest tour length when distances satisfy the triangular inequality [28]. Due to the varied nature of these methods and that there is no biological justification to favor one over the other, we consider all three methods when identifying clusters and then perform an appropriate multiple comparison adjustment to infer the statistical evidence of mutation clusters (see Section 'Multiple comparison adjustment for structures').

As can be seen from Figure 2, all the rearrangement options present a positive skew and are mostly consistent with each other. For the majority of the proteins, all three insertion approaches as well as the MDS approach result in little rearrangement. However, if one method results in radical rearrangement when the protein is mapped to 1D space, the other methods do so as well. This makes selection of a specific insertion method less critical and for the rest of this manuscript, unless otherwise specified, we use the insertion method with the most significant cluster for analysis. Please see "Distribution Summary" in

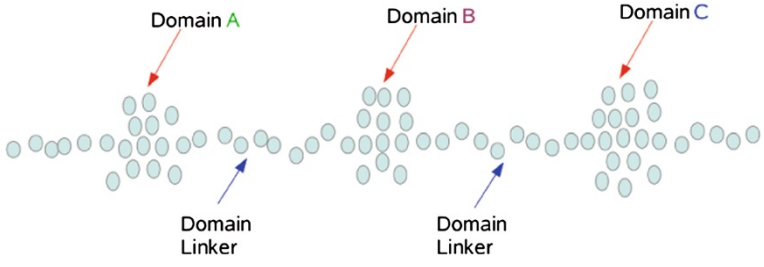

Figure 1 An example protein with three different domains. Under *iPAC*, the Domain A residues will influence the final positions of Domain C residues and vice versa, a result that is undesirable if the three domains are independent of each other. The residues in Domain A and Domain C will have no effect on each other's final position via the graph theoretic approach.

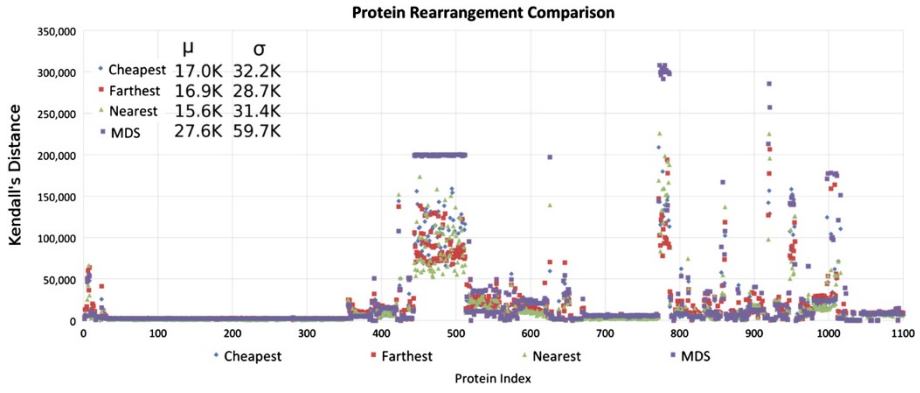

Figure 2 The amount of rearrangement performed under each of the three insertion methods described as well as MDS. Each column on the x-axis represents one of the 1100 structures considered, with structures from the same protein adjacent to one another and the protein order determined lexicographically by protein name. The y-axis shows the Kendall Tau distance, which is equivalent to the number of swaps required to sort the protein back into {1,2,3,...,..} order using bubble sort. The proteins with at least one rearrangement higher than 150,000 represent the DPP4, F5, IDE, MET, PIK3Cα, SEC23A and TF proteins, from left to right, respectively.

Additional file 3 for a full listing of each structure's Kendall Tau distance, protein index and a high resolution plot.

Path lengths between nearby and distant residues are statistically different

We employed a statistical test to verify that the TSP algorithm yields a shorter path between residues that are close together in 3D space versus residues that are far apart. First, we selected 200 random protein structures from our data set. For each structure we then selected 100 random amino acids and categorized them as "close" versus "far" in 3D space (see Table 1 for more information on the classification). For structure i, we then calculated the average path distance between all pairwise close residues, denoted $\bar{c}_i$, and the average path distance between all pairwise far residues, denoted $\bar{f}_i$. Next, we calculated the average close and far path distance over all structures: $\bar{c} = (\sum_{i=1}^{200} \bar{c}_i)/200$ and $\bar{f} = (\sum_{i=1}^{200} \bar{f}_i)/200$. Finally, as $\bar{c}$ and $\bar{f}$ are averages, we applied the central limit theorem and performed a t-test with $H_0 : \bar{c} = \bar{f}$ vs $H_a : \bar{f} > \bar{c}$. This test was performed for each of the insertion methods described in Section 'Traveling salesman approach' and at various classifications of close and far.

As shown in Table 1, the p-value is ≈ 0 for each combination of distance and insertion method, allowing us to conclude that the average path distance between residues that are far apart in 3D space is larger than the average path distance between residues that are close together in 3D space.

NMC

The *NMC* algorithm as described by Ye *et al.* [8], and briefly reviewed here, was used to find the mutational clusters once the protein was remapped to 1D space. To begin, suppose we had m samples of a protein that was N residues long and that there were a total of n mutations over all m proteins. As shown in Figure 3, by collapsing over the m samples, we can construct order statistics for every mutation. Then, given order statistics $X_{(k)}$ and $X_{(i)}$ where $i < k$, we define a cluster to exist if $Pr(C_{ki} = X_{(k)} - X(i)) \leq \alpha$, for some predetermined significance level α. As shown in Ye *et al.* [8], while a closed form calculation of the above probability is possible, it often becomes

Table 1 This table shows the p values when testing the difference in path length between residues that are close versus far apart

Close vs. Far	Method		
Residue distance	GraphPAC- Cheapest	GraphPAC- Farthest	GraphPAC- Nearest
$<$ 5 Å vs $>$ 25Å	9.50 E-143	2.10 E-141	1.38 E-145
$<$ 10 Å vs $>$ 30Å	9.32 E-147	5.82 E-145	2.81 E-149
$<$ 15 Å vs $>$ 35Å	2.91 E-151	1.01 E-148	1.33 E-153
$<$ 20 Å vs $>$ 40Å	1.42 E-155	1.63 E-151	7.15 E-158

The left column shows the requirement to label two residues as close or far apart. For instance, "$<$ 5Å versus $>$ 25Å" signifies that residues that are less than $<$ 5Å apart are labeled as close while residues that are $>$ 25Å apart are considered far apart. The p-values for each method are shown in columns 2–4.

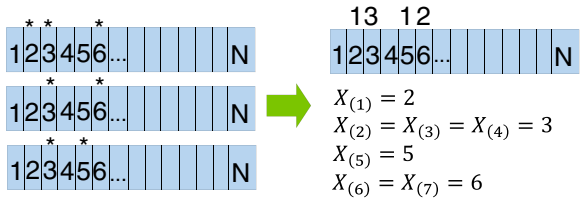

Figure 3 An example constructing order statistics over 3 samples with 7 total mutations. The number inside the box indicates the residue number. A "*" above a residue signifies a non-synonymous missense substitution mutation for that residue. Figure from Ryslik *et al.* [9].

computationally costly. To overcome this, we calculate $\frac{C_{ki}}{N}$ and assume that the statistic is uniform on $(0, 1)$. Then in the limit, it can be shown that:

$$\begin{aligned} &Pr\left(\frac{C_{ki}}{N} = \frac{X_{(k)} - X_{(i)}}{N} \leq c\right) \\ &= \int_0^c \frac{n!}{(k-i-1)!\,(i+n-k)!} y^{k-i-1}(1-y)^{i+n-k} dy \\ &= Pr(Beta(k-i, i+n-k+1) \leq c) \end{aligned} \tag{2}$$

The above calculation is then performed on all pairwise mutations and an appropriate multiple comparison adjustment is then applied. For the remainder of this study, we use the more conservative Bonferroni correction [31,32] to adjust for the intra-protein cluster p-values. See Section 'Multiple comparison adjustment for structures' for a description of how we account for the inter-protein multiple comparisons. Lastly, it is important to mention that the structural information obtained for each protein does not always contain the (x, y, z) coordinates for every residue in the protein. In such cases, in order to compare *GraphPAC*, *iPAC* and *NMC* on an equal basis, these missing residues are removed from the protein.

We also note that since we obtained our mutational data from COSMIC, some tissue types are more represented than others in the database. However, this scenario results in our analysis being more conservative and our findings even more significant. Assuming that mutations occur in different parts of the protein for different tissue types, when collapsing over all tissues a larger value of n is obtained while the values of i and k (as seen in Equation 2) for two specific mutations are not changed. This results in a larger p-value signifying that clusters found when collapsing over tissue types would be even more significant if only a unique tissue type was analyzed.

Multiple comparison adjustment for structures

In addition to the Bonferroni adjustment performed to account for multiple testing within a specific structure, we perform a second multiple comparison adjustment to account for testing all 1,100 structures. Since a single protein can have many structures that are similar to each other, a second Bonferroni adjustment is too conservative and an integrated Bonferroni-FDR approach was performed. Specifically, for a given protein, the Bonferrroni adjusted p-value of each cluster was multiplied by $\frac{n(n-1)}{2}$ to calculate p^*. Thus, p^* could be compared directly to an α-level of 0.05 in order to determine the cluster's significance. Next, a rFDR[33] approach, which is a good approximation for the standard FDR method when there are a large number of independent or positively correlated tests, was used. Under this method, the expected value of α is estimated over all k tests and then used as the significance threshold. Setting k as the total number of structures under all three insertion methods, the mean alpha can be approximated by:

$$rFDR = \alpha\left(\frac{k+1}{2k}\right)$$

where $k = 3 \times 1100 = 3300$. Using $\alpha = 0.05$, $rFDR$ is calculated to be ≈ 0.025007. Rounding down, all the clusters for which $p^* \leq 0.025$ were deemed to be significant. To avoid confusion in the rest of the paper, we only report the p-value (with the exception of Table 2). However, each cluster discussed in Section 'Results' is significant after the Bonferroni-FDR multiple comparison adjustment described here.

Results

In this section we compare the results between *GraphPAC*, *iPAC* and *NMC* in terms of the number of structures found (Section 'Method comparison') and describe the new proteins identified by *GraphPAC* (Section '*GraphPAC* identifies novel proteins with significant clustering'). We also show the results of our method in comparison to two machine learning methods along with a descriptions of whether our results overlap biologically relevant structures (Section 'Cluster localization in relevant sites and performance evaluation').

Method comparison

Using the *GraphPAC* algorithm, out of the 140 proteins analyzed, 9, 10 and 12 proteins with statistically significant clusters were found under the cheapest, nearest and farthest insertion methods, respectively. This corresponded to 223, 225 and 226 significant structures (out of the 1100 total structures considered) under the three methods. It is important to note that failure to utilize the tertiary information results in either an over or an underestimation of the number of clusters in approximately 70% of the structures analyzed (see Figure 4). Hence, failure to account for the protein structure provides either an overly

Table 2 A comparison of the 15 proteins that were found to contain significant clustering via *GraphPAC*, *iPAC* or *NMC*

	GraphPAC		*iPAC*		*NMC*	
Protein	**p-value**	**p***	**p-value**	**p***	**p-value**	**p***
KRAS	4.21 E-233	4.33 E-229	6.17 E-185	6.35 E-181	4.39 E-233	4.52 E-229
TP53	4.05 E-152	4.48 E-147	5.23 E-128	6.11 E-123	4.37 E-086	5.30 E-081
BRAF	3.84 E-130	1.04 E-126	3.73 E-130	1.01 E-126	3.84 E-130	1.04 E-126
PIK3CA	8.20 E-084	3.58 E-080	8.20 E-084	3.58 E-080	8.20 E-084	3.58 E-080
NRAS	8.26 E-029	9.91 E-027	5.38 E-026	6.46 E-024	8.26 E-029	9.91 E-027
HRAS	1.54 E-014	6.94 E-013	1.23 E-010	5.54 E-009	5.61 E-010	8.42 E-009
AKT1	2.47 E-005	2.47 E-004	1.18 E-005	7.08 E-005	2.47 E-005	7.41 E-005
IDE	1.56 E-003	4.67 E-003	2.20 E-005	6.60 E-005	1.56 E-003	4.67 E-003
EGFR	9.04 E-004	9.04 E-003	1.35 E-004	1.35 E-003	-	-
DPP4	3.17 E-003	3.63 E-002	-	-	-	-
MAP2K4	1.21E-003	1.21E-002	-	-	-	-
NRP1	1.58E-002	1.58E-002	-	-	-	-
PCSK9	5.61 E-003	1.68E-002	-	-	-	-
HAO1	-	-	7.95 E-003	2.39 E-002	-	-
EIF2AK2	-	-	2.45 E-003	7.36 E-003	-	-

If a specific method did not find a particular protein to contain significant clustering, a "-" is shown. The p* calculation is described in Section 'Multiple comparison adjustment for structures'. The smallest p-value from all of the insertion methods was selected.

complicated or overly simplified view of the mutational orientation.

On the protein level, as shown in Table 2, eight proteins were identified as having significant clusters by *GraphPAC*, *NMC* and *iPAC* while 7 proteins were identified as having significant clusters by only a subset of these methods. We note that of these seven proteins, four of them were only identified via the *GraphPAC* methodology while two of them were identified only via the *iPAC* methodology. We further note that *GraphPAC* identifies the largest number of proteins with significant clustering at the same false discovery rate, thereby showing an increased power to detect mutational clustering. We also observe that there were no proteins found to have significant clustering under the linear *NMC* algorithm that were subsequently missed by the *GraphPAC* algorithm. See Section

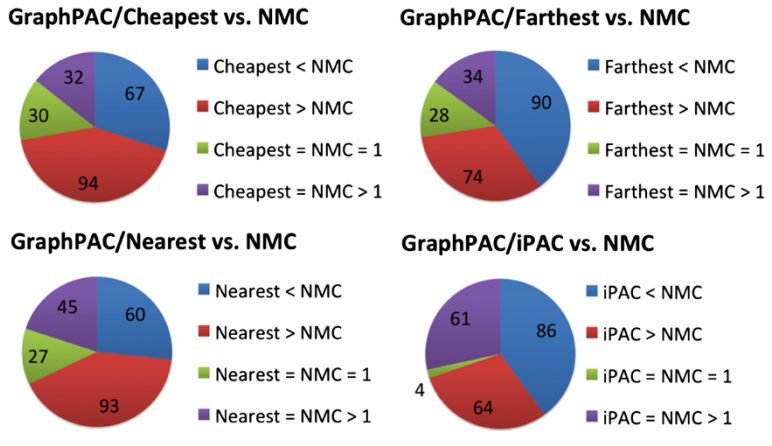

Figure 4 A comparison of *GraphPAC*, *iPAC* and *NMC* over all the structures that were found to be significant. Each of the 3D methods are considered: all three *GraphPAC* insertion methods and *iPAC*. The size of each colored block represents the number of structures with the relationship described. For instance, out of the 223 structures with significant clusters found under the cheapest insertion method of *GraphPAC* (top left), 94 structures had more clusters identified under the *GraphPAC* approach as compared to the *NMC* approach. Green is used to designate structures where the 3D and *NMC* methods identified 1 cluster while purple is used to designate structures where the 3D and *NMC* methods identified more than 1 cluster.

'Cluster localization in relevant sites and performance evaluation' for a summary of cluster overlap with active biological sites along with a performance evaluation via machine learning methods.

GraphPAC identifies novel proteins with significant clustering

GraphPAC identified four proteins with clustering that are missed by the *iPAC* algorithm: DPP4, MAP2K4, NRP1, and PSCK9. DPP4 is a serine protease that can modify tumor cell behavior and is a potential cancer therapeutic target [34]. Both MAP2K4 and NRP1 are well known to be associated with lung cancer [35,36]. Finally, while PCSK9 mutations are well known in causing hypercholesterolemia [37], recent research shows that absence of PCSK9 can provide a protective benefit against melanoma due to lower circulating LDLc. This allows for a potential additional cancer therapy via PCSK9 inhibitors [38]. [38]. For a full listing of which structure-protein combinations were found significant, see "Results Summary" in *Additional file 4*. Please see Sections '*GraphPAC* finds novel proteins compared to *iPAC* and *NMC*', '*GraphPAC* identifies additional clusters compared to *iPAC* and *NMC*' and '*GraphPAC* finds fewer clusters compared to *NMC*' for an in-depth review of selected protein-structure combinations.

Cluster localization in relevant sites and performance evaluation

We note that 9 of the 13 proteins that *GraphPAC* identified as having significant clustering have their most significant cluster overlap a binding site, catalytic domain or kinase domain. Out of the remaining four proteins, three proteins have their most significant cluster fall within a previously identified biologically relevant region. For instance, IDE's most significant cluster is located on residues 684–698, a denaturation-resistant epitope region [39]. For NRP1, which plays roles in angiogenesis [40] and axon guidance [41], the most significant cluster directly overlaps the F5/8 type C 1 domain - a domain in many blood coagulation factors. Finally, for PIK3C-α, the most significant cluster overlaps residue 1047 which has been shown to potentially increase the substrate turnover rate, a common oncogenic behavior [42]. For further detail on relevant biological site information, please see "Relevant Sites" in *Additional file 5*.

Further, we evaluated the performance of *GraphPAC* via two well-known machine learning algorithms: *CHASM* and *PolyPhen-2*. It is critical to first note however, that the machine learning algorithms utilize a much more detailed set of features when evaluating the mutation. Thus these algorithms may identify mutations as significant while *GraphPAC* would not. Nevertheless, of all the mutations that fall within significant clusters identified by *GraphPAC*, 93% and 91% of them were also identified as significant (FDR $\leq$ 20%) by *CHASM* and *PolyPhen-2* (respectively). We note that *GraphPAC* is only able to determine statistically significant clustering and not whether a mutation is truly damaging and/or activating. However, given the high percentages described above, the evidence supports the hypothesis that clustering is in fact indicative of potential driver mutations. Thus, via *GraphPAC*, the researcher has a fast and easily available tool to identify potential driver mutations for further study. The benefit of *GraphPAC* is that it is able to be executed with far less prior information as compared to the machine learning approaches. For further details, see "Performance Evaluation" in *Additional file 6*.

Finally, we note that while *GraphPAC* provides an improvement in cluster identification compared to prior work, the algorithm is unable to distinguish between mutations that increase or decrease kinase activity nor between gain-of-function (GOF) or loss-of-function (LOF) mutations. As described by Lapenna and Giordano [43], Brognard *et al.* [44], Geiger *et al.* [45], Ahn *et al.* [36], Lisabeth *et al.* [46] and Linka *et al.* [47], a large body of literature suggests that inactivating loss-of-function mutations are more common than previously thought and often occur in regions that regulate kinase activation. Nevertheless, as described above, many of the clusters identified by *GraphPAC* contain mutations that are classified as driver and/or damaging by common machine learning algorithms. As such, *GraphPAC* provides a fast and easy method to identify such potential mutations, which can then be verified and analyzed via additional approaches. These approaches can range from the aforementioned machine learning algorithms to experimental approaches that test for GOF mutations as described by Fawdar *et al.* [48].

Discussion

In this section we discuss in depth some of the clustering results presented in Section 'Results'. Specifically, we review in detail three situations: 1) *GraphPAC* identifies novel proteins (Section '*GraphPAC* finds novel proteins compared to *iPAC* and *NMC*'), 2) *GraphPAC* finds additional clusters in proteins identified to contain clustering by other methods (Section '*GraphPAC* identifies additional clusters compared to *iPAC* and *NMC*') and 3) *GraphPAC* finds fewer clusters compared to other methods (Section '*GraphPAC* finds fewer clusters compared to NMC'). In each of these sections, we discuss the biological relevance of our findings.

GraphPAC finds novel proteins compared to *iPAC* and *NMC*

As shown in Table 2, *GraphPAC* identified five additional proteins as compared to the linear *NMC* algorithm. In this section we will consider two of these proteins, both of

which are directly related to cancer: EGFR, which is also identified by *iPAC*, and NRP1, which is not identified by *iPAC*.

EGFR is a cell-surface receptor for ligands in the epidermal growth factor family [49] and is present in a wide range of diseases such as glioblastoma multiforme [50], lung adenocarcinoma [51] and colorectal cancer [52]. The most significant cluster found was in the 2ITX structure [53] between residues 719–768 (see Figure 5) with a corresponding p-value of 0.0009. This cluster contains mutations G719S, T751I and S768I which are all found in non-small cell lung carcinomas (NSCLC) [54-56] with mutation G719S well known for increased kinase activity [57]. It is also interesting to note that all three mutations within this cluster, which was identified purely through statistical clustering analysis, show a beneficial clinical response to either Erlotinib or Getfinib [55,58,59]. Exclusion of the tertiary information would have resulted in this cluster being missed. Finally, it is worth noting that recent research has shown that signaling by EGFR is dependent upon an allosteric interaction between two kinase domains in an asymmetric dimer as opposed to phosphorylation. As the formation of this asymmetric dimer is believed to activate all EGFR family members, it is likely that oncogenic activation of EGFR may differ from other protein kinases [60,61].

We now consider the NRP-1 protein, a coreceptor for the vascular endothelial growth factor (VEGF) which is upregulated in a large variety of cancers including lung tumors [35], gastrointestinal metasteses [62] and pancreatic carcinomas [63]. In NSCLC patients, it has been shown to be an independent predictor of cancer relapse and reduced survival as well as a cancer invasion enhancer [64]. Moreover, research has shown that NRP-1 inhibitors provide an additive effect to anti-VEGF therapy in reducing tumor progression. Monoclonal antibodies that attach to the b1-b2 domains, the domains responsible for VEGF binding, have already been created [65]. The b1 domain, which spans residues 275–424 almost exactly overlaps the most significant cluster found by *GraphPAC*, which consists of residues 277–432 (p-value 0.0158) in the 2QQI [66] structure (Figure 6). Finally, it is worth noting that mutations on residues 297 and 320 were recently found that completely disrupt VEGF binding, both of which also fall within the *GraphPAC* identified cluster of 277-432 in the 2QQI structure.

GraphPAC identifies additional clusters compared to *iPAC* and *NMC*

A representative example where *GraphPAC* identifies additional clusters as compared to *NMC* and *iPAC* is in the KRAS protein for the 3GFT structure[c] [67] (Figure 7). KRAS, a GTPase, is one of the most pervasively activated oncogenes, with some estimates stating that between 17–25% of all human tumors contain an activating mutation of the gene [68]. Due to the large number of samples with mutations in this gene and the resulting strong statistical signal, *GraphPAC*, *iPAC* and *NMC* all identify that KRAS contains highly statistically significant mutational clusters. Nevertheless, *GraphPAC* identifies several novel clusters that are missed by *iPAC* and NMC. While all three methods identify clustering at residues 12–13, 12–61 and 12–146, only *iPAC* and *GraphPAC* identify two additional clusters at 1) 61–117 and 2) 117–146.

Moreover, only *GraphPAC* (under the cheapest and nearest insertion methods) identifies a statistically significant cluster for residues 12–23 and 23–61 as shown in Table 3. Considering the 12–23 cluster, we see that a sub-cluster of 12–13 is identified as well. This follows biological function as mutations on residues 12 and 13 appear in a large variety of cancers, such as breast, lung,

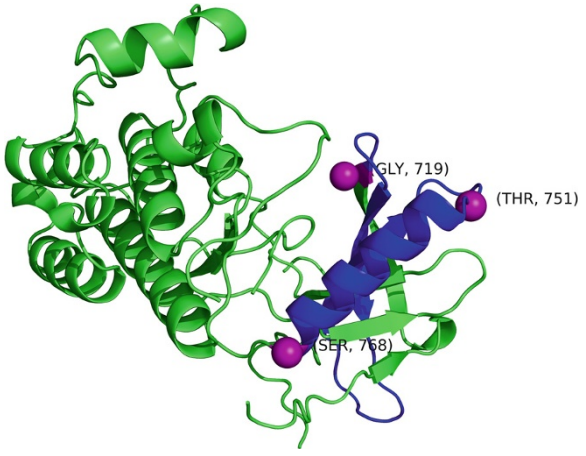

Figure 5 The EGFR ectodomain fragment structure (PDB ID 2ITX) where the 719–768 cluster is colored in blue. The three mutations, 719, 751 and 768 are displayed as purple spheres.

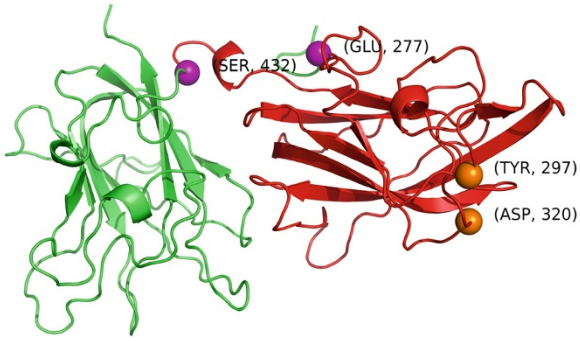

Figure 6 The NRP-1 structure (PDB ID 2QQI) where the 277–432 cluster is colored in red. The mutations that disrupt VEGF binding, 297 and 320 are shown as orange spheres while the end-points of the cluster, 277 and 432, are shown as purple spheres.

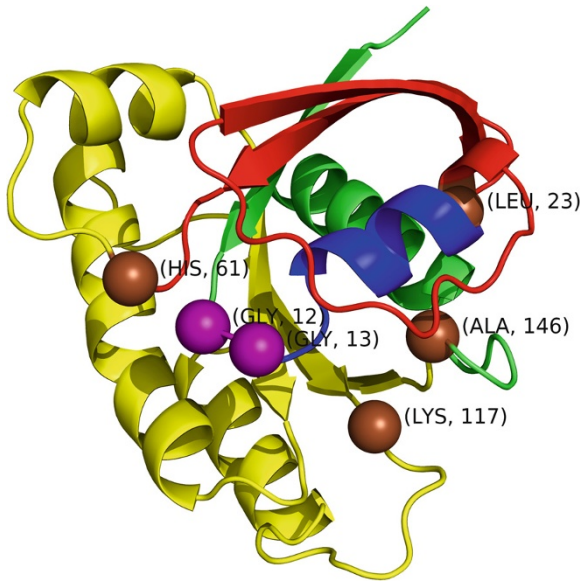

Figure 7 The KRAS structure (PDB ID 3GFT) color coded by region: amino acids 13–22 are blue, 24–60 are red and 62–145 are yellow. Residues 12 and 13 which make up the most significant cluster are shown as purple spheres, while residues 23, 61, 117 and 146 are shown as brown spheres.

bladder, pancreas and colon [6,69,70] while mutations on residues 22 and 23 appeared in colorectal/large intestine tissue samples in our data. It is interesting to note that germline mutations on residue 22 often result in developmental disorders such as Noonan Syndrome Type 3 (NS3) as well as Cardiofaciocutaneous Syndrome (CFC) [71,72].

Finally, the majority of mutations in cluster 61–146 also segregate along pathological lines with all the mutations in our data either occurring in lung or gastrointestinal tract/large intestine carcinomas. Specifically, residue 61 mutations are typically found in colorectal and lung cancer [6,73] while mutations K117N and A146T are found in colorectal cancer [6].

Table 3 P-value comparison of the three algorithms for several significant clusters

Residues	*NMC*	*iPAC*	*GraphPAC*
12–13	9.45 E-229	3.91 E-165	8.95 E-229
12–23	-	-	1.31 E-99
12–61	4.34 E-65	2.38E E-87	5.49 E-164
12–146	3.85 E-13	3.81 E-90	2.87 E-16
23–61	-	-	1.01 E-105
61–146	-	3.01 E-106	4.35 E-31
117–146	-	1.66 E-102	-

A "-" signifies that the method did not find that cluster to be significant. For *GraphPAC*, the cheapest insertion results are reported here.

GraphPAC finds fewer clusters compared to *NMC*

As seen from Figure 4, between 25%–40% of the structures identified with significant clustering had fewer clusters under the *GraphPAC* methodology as compared to the linear *NMC* algorithm with the vast majority of these structures corresponding to BRAF, HRAS and TP53. Here we consider a representative example, the 4E26 structure [74] for BRAF when analyzed using the farthest insertion method (Figure 8). As *iPAC* identified even more clusters than NMC, we compare *GraphPAC* to NMC in this section when showing that fewer mutational clusters is of benefit. Further, as V600 is well known to be the most likely mutated position in BRAF, the most significant cluster identified by *GraphPAC*, *iPAC* and *NMC* is located only on that residue with a p-value of 2.12×10^{-129} under all three methods. In all, *GraphPAC* identifies 16 clusters while *NMC* identifies 22, with the differences shown in Table 4.

Although it is outside the scope of this manuscript to consider every difference between Tables 4a and 4b, we observe that three of the longest clusters 464–671, 466–671 and 469–671 are dropped by *GraphPAC*. Since after alignment of the protein structural data to the mutational data (see Section 'Obtaining the 3D structural data'), tertiary information was available on residues 448–603 and 610–723, these clusters cover 77.0%, 76.3% and 75.2% of all the available residues, respectively. By considering the 3D structure via *GraphPAC*, the longest clusters are dropped and the remaining overlapping clusters focus almost exclusively on residues 464–600.

After structure and mutation alignment, the residue substitutions in significant clusters include: G464V, G466V, G469V, G469A, N581S, G596R, L597V, LV597R,

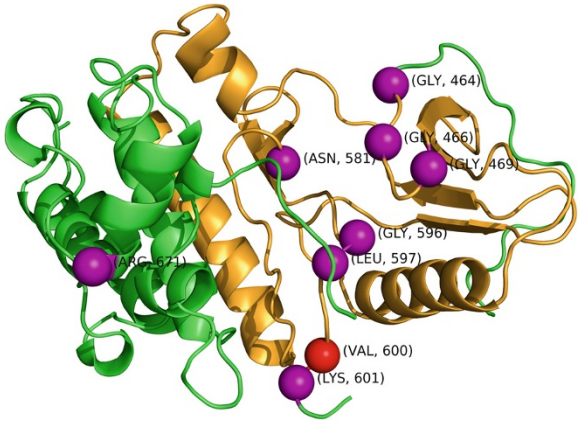

Figure 8 The BRAF structure (PDB ID 4E26) color coded by segment: I) amino acids 464–599 are orange 2) amino acids 601–671 are green. The α-carbons of the mutated residues 464, 466, 469, 581, 596, 597 , 601 and 671 are shown as purple spheres. Residue 600 is shown as a red sphere.

Table 4 A comparison of GraphPAC and NMC identified clusters for the BRAF structure

(a) Clusters found by *GraphPAC*				
			p-value	
Start	**End**	**# Muts.**	***GraphPAC***	***NMC***
600	600	60	2.12E-129	2.12E-129
597	600	62	1.49E-104	1.49E-104
600	601	62	1.49E-104	9.22E-117
596	600	64	7.16E-102	7.16E-102
596	601	66	3.37E-91	1.16E-100
597	601	64	8.07E-91	7.16E-102
601	671	3	5.85E-38	-
600	671	63	8.30E-37	7.08E-26
469	601	72	2.59E-22	5.92E-17
581	601	68	1.23E-21	1.33E-65
581	600	66	2.94E-20	3.13E-63
469	600	70	3.98E-20	4.91E-15
466	601	74	2.15E-17	9.69E-19
466	600	72	7.01E-16	1.60E-16
464	601	75	1.15E-15	1.12E-19
464	600	73	2.33E-14	2.97E-17

(b) Clusters found by *NMC* and dropped by *GraphPAC*			
Start	**End**	**# Muts.**	***NMC* Pvalue**
596	671	67	4.12E-29
597	671	65	4.79E-27
581	671	69	3.33E-26
464	671	76	5.92E-09
466	671	75	3.32E-08
469	671	73	8.11E-07

A "-" for the *NMC* value signifies that cluster was not identified under the linear algorithm.

V600E, V600K, K601N and R671Q. Since mutation R671Q does not have extensive literature and comes from a non-specified tissue sample in the COSMIC database, it will no longer be considered here. Thus, by considering the tertiary structure, we significantly narrow the window of which residues to consider for potential driver mutations and can partition the protein into three segments: I) 464–599 and II) 600 and III) 601. Segment I is primarily associated with lung and colorectal cancer as shown in [3,75-77]. Segment II represents the two most common mutations in BRAF, V600E and V600K. Overall, 95% of BRAF mutations occur on V600, with some studies showing that V600E occurs within 73% to 79% of patients while V600K occurs within 12% to 19% of patients [78,79]. Mutations at this position result in the oncogene being constitutively activated with increased kinase activity and have been found in a wide range of cancers such as metastatic melanoma [80], ovarian serous carcinoma [81] and hairy cell leukemia [82]. Furthermore, recent inhibitors, such as Vemurafenib and GSK2118436 specifically target the V600E and V600E/K mutations (respectively), supporting the hypothesis that somatic clusters can provide pharmacological targets [83]. Lastly, segment III is comprised of the much less common K601N mutation which has been observed in myeloma cases along with V600E. Since these patients share the more common BRAF mutations as well, they may also potentially benefit from BRAF inhibitors [84].

Further, as shown in Section 'Results' and described above, *GraphPAC* finds fewer clusters for a significant percentage of the structures analyzed. Overall, the reduction in total clusters identified can result from two sources: the removal of some residues because no tertiary data was available or the cluster is no longer significant when using the traveling salesman algorithm to account for 3D structure. The first case, which is already rare, will become increasingly more so as additional studies result in more complete and detailed structural information. For the second case, if a cluster is not found to be significant under *GraphPAC* when compared to NMC, a near or overlapping cluster is usually found (see Tables 4a and 4b). For BRAF specifically, under every type of graph insertion method (cheapest, nearest and farthest), every "probably damaging" or "possibly damaging" mutation (as classified by *PolyPhen-2*) was still identified in at least one significant cluster for the structure. For a complete analysis, see "Potential Driver Loss" in *Additional file 7*.

Conclusion

In this manuscript we provide an alternative method to utilize protein tertiary structure when identifying somatic mutation clusters. By employing a graph theoretic approach to restructuring the protein order, we identify both new clusters in proteins previously shown to have clustering as well as proteins that were not previously shown to have clustering. We have also provided several examples where we are able to identify clusters of mutations that may benefit from pharmacological treatment. Moreover, as *GraphPAC* uses the *NMC* algorithm to identify clusters rather than a fixed window size, we are able to detect clusters of varying lengths. Finally, the methodology is fast and robust with the overwhelming majority of structure/protein combinations taking under 10 minutes each to analyze on a consumer desktop.

The *GraphPAC* algorithm, while presenting a viable alternative to the MDS restriction of *iPAC* and an improvement over NMC, nevertheless contains several limitations. First, while no longer bound to the MDS requirement of *iPAC*, there is no closed form solution to the shortest path problem and our algorithm must appeal to heuristic approximations. Second, to satisfy the

uniformity assumption, the mutation status of all residues must be known ahead of time. With the growth of high-throughput sequencing however, this issue is temporary. Next, unequal rates of mutagenesis along with hyper-mutability of specific genomic regions may violate the assumption that every residue has a uniform probability of mutation. To help ensure that this assumption holds, we only consider single residue missense substitutions and have removed insertions and deletions from the analysis since they tend to be sequence dependent. Further, research has shown that CpG dinucleotides may have a mutational frequency ten times or higher compared to other dinucleotides [85]. However, in the analyses presented in Sections '*GraphPAC* finds novel proteins compared to *iPAC* and *NMC*', '*GraphPAC* identifies additional clusters compared to *iPAC* and *NMC*', '*GraphPAC* finds fewer clusters compared to NMC', only approximately 13% of the mutations used to identify clustering occurred in CpG sites. Relatedly, colorectal carcinomas [86] contain more transition mutations while cigarette use results in more transversion mutations in lung carcinomas [8]. Still, when considering KRAS, the overwhelming majority of substitutions occur on residues 12, 13, and 61 for both colorectal and lung cancer, implying that while the mutational landscape may vary, it does not have a significant effect on mutation location and thus would not violate the uniformity assumption. Hence, while this analysis is influenced by a variety of factors, as are previous studies, it nevertheless appears that the primary cause of clustering is selection for a cancer phenotype.

Several areas for future research are also directly evident. First, an approach that considers the protein directly in 3D space via simulation may be employed. However, such an approach would not be able to use the order statistic methodology to identify clustering and thus might not be as sensitive for small mutation counts. Moreover, while we only consider distance when finding the shortest path through the graph, future research can incorporate the physico-chemical properties of the specific residues or domains by appropriately increasing or decreasing edge length. The potential additive effect of multiple cancer mutations in the same protein, as discussed in the case of EGFR by Hashimoto *et al.* [87], can also be incorporated via additional refinement of the edge weights. Additional research is required in this area in order to incorporate these improvements.

Overall however, *GraphPAC* utilizes protein tertiary structure via a graph theoretic approach in identifying mutational clustering. We show that this method identifies new clusters that are otherwise missed and that in some cases, pharmaceutical targets for mutations in these clusters have already been found and therapies created. Specifically, Erlotinib and Getfinib are used to target mutations in EGFR significant clusters (see Section '*GraphPAC* finds novel proteins compared to *iPAC* and *NMC*') while Vemurafenib is used to target mutations that occur within BRAF significant clusters (see Section '*GraphPAC* finds fewer clusters compared to NMC'). This helps confirm the hypothesis that mutational clustering may be indicative of driver mutations and as new protein structures become available, *GraphPAC* can provide a rapid methodology to identify such potential mutations.

Availability and requirements

Project Name: *GraphPAC*: Identification of Mutational Clusters in Proteins via a Graph Theoretical Approach.
Project Home Page: http://www.bioconductor.org/packages/release/bioc/html/GraphPAC.html
Operating system(s): Platform independent
Programming Language: R
Other Requirements: R $\geq$ 2.15 (see homepage for R package requirements).
License: GPL-2

Endnotes

[a]Under a complete graph, every vertex is connected to every other vertex. The length of the edge between vertices i and j is set to be equal to the length between amino acids i and j in $\mathbb{R}^3$.

[b]A Hamiltonian path is a walk through the graph that visits every vertex once and only once.

[c]For this analysis, a manual reconstruction was performed in order to include residue 61 which is listed as a histidine under isoform 2B in the Uniprot Database and a glutamine in the COSMIC database. As the substitution of one amino acid in the structure would not have a significant impact on the spatial structure of the protein, and residue 61 is a highly mutated position, the residue was kept in the analysis. As a result, amino acids 1–167 are used.

Additional files

Additional file 1: Cosmic Query. The SQL query used to extract the mutations from COSMIC.

Additional file 2: Structure Files. A detailed list of which protein-structure combinations were used and what chains were selected.

Additional file 3: Distribution Summary. A detailed list of Kendall's Distance for each structure under each of the methods.

Additional file 4: Results Summary. A summary of each structure's most significant p-value for both *iPAC* and NMC.

Additional file 5: Relevant Sites. A review showing which of the *iPAC* clusters fall within structurally relevant sites.

Additional file 6: Performance Evaluation. In-depth results validating the *iPAC* results using *PolyPhen-2* and *CHASM*.

Additional file 7: Potential Driver Loss. An analysis of whether any potential driver mutations are lost when *iPAC* finds fewer clusters than NMC.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
GR and HZ developed the *GraphPAC* methology. KC was responsible for obtaining the mutation data from the COSMIC database. GR and YC executed the methodology on the protein structures. GR drafted the original manuscript while KC, YC, YM and HZ were responsible for revisions. HZ finalized the manuscript. All authors have read and approved the final text.
This work was supported in part by NSF Grant DMS 1106738 (GR, HZ), NIH Grants GM59507 and CA154295 (HZ) as well as a Fellowship from the Yale World Scholars Program of the China Scholarship Council (YC).

Acknowledgements
We thank Drs. Robert Bjornson and Nicholas Carriero for their time and help in discussing this methodology.

Author details
[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA. [2]Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. [3]Yale Center for Medical Informatics, Yale School of Medicine, New Haven, CT, USA. [4]Department of Molecular Biophysics & Biochemistry, Yale University, New Haven, CT, USA.

Received: 9 July 2013 Accepted: 11 March 2014
Published: 26 March 2014

References

1. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control.** *Nat Med* 2004, **10**(8):789–799.
2. Weinstein IB, Joe AK: **Mechanisms of disease: Oncogene addiction–a rationale for molecular targeting in cancer therapy.** *Nat Clin Pract Oncol* 2006, **3**(8):448–457.
3. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, et al.: **Patterns of somatic mutation in human cancer genomes.** *Nature* 2007, **446**(7132):153–158.
4. Wang T: **Prevalence of somatic alterations in the colorectal cancer cell genome.** *Proc Natl Acad Sci* 2002, **99**(5):3076–3080.
5. Bardelli A, Parsons DW, Silliman N, Ptak J, Szabo S, Saha S, Markowitz S, Willson JKV, Parmigiani G, Kinzler KW, Vogelstein B, Velculescu VE: **Mutational analysis of the tyrosine kinome in colorectal cancers.** *Science* 2003, **300**(5621):949.
6. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JKV, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE: **The consensus coding sequences of human breast and colorectal cancers.** *Science* 2006, **314**(5797):268–274.
7. Kreitman M: **Methods to detect selection in populations with applications to the human.** *Annu Rev Genomics Hum Genet* 2000, **1:**539–559.
8. Ye J, Pavlicek A, Lunney EA, Rejto PA, Teng C: **Statistical method on nonrandom clustering with application to somatic mutations in cancer.** *BMC Bioinformatics* 2010, **11:**11.
9. Ryslik GA, Cheng Y, Cheung KH, Modis Y, Zhao H: **Utilizing protein structure to identify non-random somatic mutations.** *BMC Bioinformatics* 2013, **14:**190.
10. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7**(4):248–249.
11. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R: **Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations.** *Cancer Res* 2009, **69**(16):6660–6667.
12. Carter H, Samayoa J, Hruban RH, Karchin R: **Prioritization of driver mutations in pancreatic cancer using cancer-specific high-throughput annotation of somatic mutations (CHASM).** *Cancer Biol Ther* 2010, **10**(6):582–587.
13. Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, Cooper DN, Ryan M, Karchin R: **CRAVAT: cancer-related analysis of variants toolkit.** *Bioinformatics* 2013, **29**(5):647–648.
14. Breiman L: **Random forests.** *Mach Learn* 2001, **45:**5–32.
15. Cortes C, Vapnik V: **Support-vector networks.** *Mach Learn* 1995, **20**(3):273–297.
16. Reva B, Antipin Y, Sander C: **Predicting the functional impact of protein mutations: application to cancer genomics.** *Nucleic Acids Res* 2011, **39**(17):e118–e118.
17. Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions.** *Genome Res* 2001, **11**(5):863–874.
18. Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z: **Assessment of computational methods for predicting the effects of missense mutations in human cancers.** *BMC Genomics* 2013, **14**(Suppl 3):S7.
19. Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GRS, Creixell P, Karchin R, Vazquez M, Fink JL, Kassahn KS, Pearson JV, Bader GD, Boutros PC, Muthuswamy L, Ouellette BFF, Reimand J, Linding R, Shibata T, Valencia A, Butler A, Dronov S, Flicek P, Shannon NB, Carter H, Ding L, Sander C, Stuart JM, Stein LD, Lopez-Bigas N: **International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group: Computational approaches to identify functional genetic variants in cancer genomes.** *Nature methods* 2013, **10**(8):723–729.
20. Torkamani A, Schork NJ: **Prediction of cancer driver mutations in protein kinases.** *Cancer Res* 2008, **68**(6):1675–1682.
21. Borg I, Groenen PJF: *Modern Multidimensional Scaling : Theory and Applications*. New York: Springer; 1997.
22. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR: **The Catalogue of Somatic Mutations in Cancer (COSMIC).** *Curr Protoc Hum Genet* 2008, **Chapter 10:**Unit 10.11.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P: **The protein data bank.** *Nucleic Acids Res* 2000, **28:**235–242.
24. The UniProt Consortium: **Reorganizing the protein space at the Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2011, **40**(D1):D71–D75.
25. Durrant J, McCammon JA: **Molecular dynamics simulations and drug discovery.** *BMC Biology* 2011, **9:**71.
26. Pages H, Aboyoun P, Gentleman R, DebRoy S: **Biostrings: String objects representing biological sequences, and matching algorithms.** 2012. [R package version 2.24.1]. www.bioconductor.org/packages/release/bioc/html/Biostrings.html.
27. Applegate DL: *The Traveling Salesman Problem: A Computational Study*. Princeton: Princeton University Press; 2006. Princeton series in applied mathematics.
28. Hahsler M, Hornik K: **TSP—Infrastructure for the Traveling Salesperson Problem.** *J Stat Software* 2007, **23**(2):1–21.
29. Gutin G, Punnen AP: *The Traveling Salesman Problem and its Variations*. New York: Springer; 2007. No. 12 in Combinatorial optimization.
30. Rosenkrantz DJ, Stearns RE, Lewis PM II: **An analysis of several heuristics for the traveling salesman problem.** *SIAM J Comput* 1977, **6**(3):563–581.
31. Dunn OJ: **Confidence intervals for the means of dependent, normally distributed variables.** *J Am Stat Assoc* 1959, **54**(287):613–621.
32. Dunn OJ: **Multiple comparisons among means.** *J Am Stat Assoc* 1961, **56**(293):52–64.
33. Gong Y, Kakihara Y, Krogan N, Greenblatt J, Emili A, Zhang Z, Houry WA: **An atlas of chaperone–protein interactions in Saccharomyces cerevisiae: implications to protein folding pathways in the cell.** *Mol Syst Biol* 2009, **5:**. doi:10.1038/msb.2009.26.
34. Kelly T: **Fibroblast activation protein-α and dipeptidyl peptidase IV (CD26): Cell-surface proteases that activate cell signaling and are potential targets for cancer therapy.** *Drug Resist Updat* 2005, **8**(1–2):51–58.
35. Lantuéjoul S, Constantin B, Drabkin H, Brambilla C, Roche J, Brambilla E: **Expression of VEGF, semaphorin SEMA3F, and their common receptors neuropilins NP1 and NP2 in preinvasive bronchial lesions, lung tumours, and cell lines.** *J Pathol* 2003, **200**(3):336–347.
36. Ahn YH, Yang Y, Gibbons DL, Creighton CJ, Yang F, Wistuba II, Lin W, Thilaganathan N, Alvarez CA, Roybal J, Goldsmith EJ, Tournier C, Kurie JM: **Map2k4 functions as a tumor suppressor in lung adenocarcinoma and inhibits tumor cell invasion by decreasing peroxisome**

proliferator-activated receptor 2 expression. *Mol Cell Biol* 2011, **31**(21):4270–4285.
37. Abifadel M, Varret M, Rabès JP, Allard D, Ouguerram K, Devillers M, Cruaud C, Benjannet S, Wickham L, Erlich D, Derré A, Villéger L, Farnier M, Beucler I, Bruckert E, Chambaz J, Chanu B, Lecerf JM, Luc G, Moulin P, Weissenbach J, Prat A, Krempf M, Junien C, Seidah NG, Boileau C: **Mutations in PCSK9 cause autosomal dominant hypercholesterolemia.** *Nat Genet* 2003, **34**(2):154–156.
38. Seidah NG: **Proprotein Convertase Subtilisin Kexin 9 (PCSK9) Inhibitors in the treatment of hypercholesterolemia and other pathologies.** *Curr Pharm Des* 2013, **19**(17):3161–3172.
39. Cavender JF, Mummert C, Tevethia MJ: **Transactivation of a ribosomal gene by simian virus 40 large-T antigen requires at least three activities of the protein.** *J Virol* 1999, **73:**214–224.
40. Jubb AM, Strickland LA, Liu SD, Mak J, Schmidt M, Koeppen H: **Neuropilin-1 expression in cancer and development.** *J Pathol* 2012, **226:**50–60.
41. Maden CH, Gomes J, Schwarz Q, Davidson K, Tinker A, Ruhrberg C: **NRP1 and NRP2 cooperate to regulate gangliogenesis, axon guidance and target innervation in the sympathetic nervous system.** *Dev Biol* 2012, **369**(2):277–285.
42. Mankoo PK, Sukumar S, Karchin R: **PIK3CA somatic mutations in breast cancer: Mechanistic insights from Langevin dynamics simulations.** *Proteins: Struct, Funct, Bioinf* 2009, **75**(2):499–508.
43. Lapenna S, Giordano A: **Cell cycle kinases as therapeutic targets for cancer.** *Nat Rev Drug Discovery* 2009, **8**(7):547–566.
44. Brognard J, Zhang YW, Puto LA, Hunter T: **Cancer-associated loss-of-function mutations implicate DAPK3 as a tumor-suppressing kinase.** *Cancer Res* 2011, **71**(8):3152–3161.
45. Geiger TR, Song JY, Rosado A, Peeper DS: **Functional characterization of human cancer-derived TRKB mutations.** *PLoS ONE* 2011, **6**(2):e16871.
46. Lisabeth EM, Fernandez C, Pasquale EB: **Cancer somatic mutations disrupt functions of the EphA3 receptor tyrosine kinase through multiple mechanisms.** *Biochemistry* 2012, **51**(7):1464–1475.
47. Linka RM, Risse SL, Bienemann K, Werner M, Linka Y, Krux F, Synaeve C, Deenen R, Ginzel S, Dvorsky R, Gombert M, Halenius A, Hartig R, Helminen M, Fischer A, Stepensky P, Vettenranta K, Köhrer K, Ahmadian MR, Laws HJ, Fleckenstein B, Jumaa H, Latour S, Schraven B, Borkhardt A: **Loss-of-function mutations within the IL-2 inducible kinase ITK in patients with EBV-associated lymphoproliferative diseases.** *Leukemia* 2012, **26**(5):963–971.
48. Fawdar S, Trotter EW, Li Y, Stephenson NL, Hanke F, Marusiak AA, Edwards ZC, Ientile S, Waszkowycz B, Miller CJ, Brognard J: **Targeted genetic dependency screen facilitates identification of actionable mutations in FGFR4, MAP3K9, and PAK5 in lung cancer.** *Proc Natl Acad Sci* 2013, **110**(30):12426–12431.
49. Herbst RS: **Review of epidermal growth factor receptor biology.** *Int J Radiat Oncol Biol Phys* 2004, **59**(2, Supplement):S21–S26.
50. Heimberger AB, Hlatky R, Suki D, Yang D, Weinberg J, Gilbert M, Sawaya R, Aldape K: **Prognostic effect of epidermal growth factor receptor and EGFRvIII in glioblastoma multiforme patients.** *Clin Cancer Res* 2005, **11**(4):1462–1466.
51. Ladanyi M, Pao W: **Lung adenocarcinoma: guiding EGFR-targeted therapy and beyond.** *Mod Pathol* 2008, **21:**S16–S22.
52. Markman B, Javier Ramos F, Capdevila J, Tabernero J: **EGFR and KRAS in colorectal cancer.** *Adv Clin Chem* 2010, **51:**71–119.
53. Yun CH, Boggon TJ, Li Y, Woo MS, Greulich H, Meyerson M, Eck MJ: **Structures of lung cancer-derived egfr mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity.** *Cancer Cell* 2007, **11**(3):217–227.
54. Simonetti S, Molina M, Queralt C, de Aguirre I, Mayo C, Bertran-Alamillo J, Sanchez J, Gonzalez-Larriba J, Jimenez U, Isla D, Moran T, Viteri S, Camps C, Garcia-Campelo R, Massuti B, Benlloch S, y Cajal S, Taron M, Rosell R: **Detection of EGFR mutations with mutation-specific antibodies in stage IV non-small-cell lung cancer.** *J Transl Med* 2010, **8:**135.
55. Masago K, Fujita S, Irisa K, Kim YH, Ichikawa M, Mio T, Mishima M: **Good clinical response to gefitinib in a non-small cell lung cancer patient harboring a rare somatic epidermal growth factor gene point mutation; codon 768 AGC > ATC in exon 20 (S768I).** *Jpn J Clin Oncol* 2010, **40**(11):1105–1109.
56. Yoshikawa S, Kukimoto-Niino M, Parker L, Handa N, Terada T, Fujimoto T, Terazawa Y, Wakiyama M, Sato M, Sano S, Kobayashi T, Tanaka T, Chen L, Liu ZJ, Wang BC, Shirouzu M, Kawa S, Semba K, Yamamoto T, Yokoyama S: **Structural basis for the altered drug sensitivities of non-small cell lung cancer-associated mutants of human epidermal growth factor receptor.** *Oncogene* 2012, **32:**27–38.
57. Yun CH, Boggon TJ, Li Y, Woo MS, Greulich H, Meyerson M, Eck MJ: **Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity.** *Cancer Cell* 2007, **11**(3):217–227.
58. Kancha RK, Peschel C, Duyster J: **The epidermal growth factor receptor-L861Q mutation increases kinase activity without leading to enhanced sensitivity toward epidermal growth factor receptor kinase inhibitors.** *J Thorac Oncol* 2011, **6**(2):387–392.
59. Peraldo-Neia C, Migliardi G, Mello-Grand M, Montemurro F, Segir R, Pignochino Y, Cavalloni G, Torchio B, Mosso L, Chiorino G, Aglietta M: **Epidermal Growth Factor Receptor (EGFR) mutation analysis, gene expression profiling and EGFR protein expression in primary prostate cancer.** *BMC Cancer* 2011, **11:**31.
60. Zhang X, Pickin KA, Bose R, Jura N, Cole PA, Kuriyan J: **Inhibition of the EGF receptor by binding of MIG6 to an activating kinase domain interface.** *Nature* 2007, **450**(7170):741–744.
61. Jura N, Endres NF, Engel K, Deindl S, Das R, Lamers MH, Wemmer DE, Zhang X, Kuriyan J: **Mechanism for activation of the EGF receptor catalytic domain by the juxtamembrane segment.** *Cell* 2009, **137**(7):1293–1307.
62. Hansel DE, Wilentz RE, Yeo CJ, Schulick RD, Montgomery E, Maitra A: **Expression of neuropilin-1 in high-grade dysplasia, invasive cancer, and metastases of the human gastrointestinal tract.** *Am J Surg Pathol* 2004, **28**(3):347–356.
63. Parikh AA, Liu WB, Fan F, Stoeltzing O, Reinmuth N, Bruns CJ, Bucana CD, Evans DB, Ellis LM: **Expression and regulation of the novel vascular endothelial growth factor receptor neuropilin-1 by epidermal growth factor in human pancreatic carcinoma.** *Cancer* 2003, **98**(4):720–729.
64. Hong TM, Chen YL, Wu YY, Yuan A, Chao YC, Chung YC, Wu MH, Yang SC, Pan SH, Shih JY, Chan WK, Yang PC: **Targeting neuropilin 1 as an antitumor strategy in lung cancer.** *Clin Cancer Res* 2007, **13**(16):4759–4768.
65. Pan Q, Chanthery Y, Liang WC, Stawicki S, Mak J, Rathore N, Tong RK, Kowalski J, Yee SF, Pacheco G, Ross S, Cheng Z, Le Couter J, Plowman G, Peale F, Koch AW, Wu Y, Bagri A, Tessier-Lavigne M, Watts RJ: **Blocking neuropilin-1 function has an additive effect with anti-VEGF to inhibit tumor growth.** *Cancer Cell* 2007, **11:**53–67.
66. Appleton BA, Wu P, Maloney J, Yin J, Liang WC, Stawicki S, Mortara K, Bowman KK, Elliott JM, Desmarais W, Bazan JF, Bagri A, Tessier-Lavigne M, Koch AW, Wu Y, Watts RJ, Wiesmann C: **Structural studies of neuropilin/antibody complexes provide insights into semaphorin and VEGF binding.** *EMBO J* 2007, **26**(23):4902–4912. [PDB ID: 2QQI].
67. Tong Y, Tempel W, Shen L, Arrowsmith C, Edwards A, Sundstrom M, Weigelt J, Bockharev A, Park H: **Human K-Ras in complex with a GTP analogue.** 2009. [http://www.rcsb.org/pdb/explore.do?structureId=3GFT] [PDB ID: 3GFT].
68. Kranenburg O: **The KRAS oncogene: past, present, and future.** *Biochim Biophys Acta Rev Canc* 2005, **1756**(2):81–82.
69. McCoy MS, Bargmann CI, Weinberg RA: **Human colon carcinoma Ki-ras2 oncogene and its corresponding proto-oncogene.** *Mol Cell Biol* 1984, **4**(8):1577–1582.
70. Motojima K, Urano T, Nagata Y, Shiku H, Tsurifune T, Kanematsu T: **Detection of point mutations in the Kirsten-ras oncogene provides evidence for the multicentricity of pancreatic carcinoma.** *Ann Surg* 1993, **217**(2):138–143.
71. Zenker M, Lehmann K, Schulz AL, Barth H, Hansmann D, Koenig R, Korinthenberg R, Kreiss-Nachtsheim M, Meinecke P, Morlot S, Mundlos S, Quante AS, Raskin S, Schnabel D, Wehner LE, Kratz CP, Horn D, Kutsche K: **Expansion of the genotypic and phenotypic spectrum in patients with KRAS germline mutations.** *J Med Genet* 2007, **44**(2):131–135.
72. Gremer L, Merbitz-Zahradnik T, Dvorsky R, Cirstea IC, Kratz CP, Zenker M, Wittinghofer A, Ahmadian MR: **Germline KRAS mutations cause aberrant biochemical and physical properties leading to developmental disorders.** *Hum Mutat* 2011, **32:**33–43.
73. Tam IYS, Chung LP, Suen WS, Wang E, Wong MCM, Ho KK, Lam WK, Chiu SW, Girard L, Minna JD, Gazdar AF, Wong MP: **Distinct epidermal**

growth factor receptor and KRAS mutation patterns in non-small cell lung cancer patients with different tobacco exposure and clinicopathologic features. *Clin Cancer Res* 2006, **12**(5):1647–1653.
74. Qin J, Xie P, Ventocilla C, Zhou G, Vultur A, Chen Q, Liu Q, Herlyn M, Winkler J, Marmorstein R: **Identification of a Novel family of BRAF V600E inhibitors.** *J Med Chem* 2012, **55**(11):5220–5230. PDB ID: 4E26.
75. Naoki K, Chen TH, Richards WG, Sugarbaker DJ, Meyerson M: **Missense mutations of the BRAF gene in human lung adenocarcinoma.** *Cancer Res* 2002, **62**(23):7001–7003.
76. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, Davis N, Dicks E, Ewing R, Floyd Y, Gray K, Hall S, Hawes R, Hughes J, Kosmidou V, Menzies A, Mould C, Parker A, Stevens C, Watt S, Hooper S, Wilson R, Jayatilake H, Gusterson BA, Cooper C, Shipley J, et al.: **Mutations of the BRAF gene in human cancer.** *Nature* 2002, **417**(6892):949–954.
77. Gandhi J, Zhang J, Xie Y, Soh J, Shigematsu H, Zhang W, Yamamoto H, Peyton M, Girard L, Lockwood WW, Lam WL, Varella-Garcia M, Minna JD, Gazdar AF: **Alterations in genes of the EGFR signaling pathway and their relationship to EGFR tyrosine kinase inhibitor sensitivity in lung cancer cell lines.** *PLoS ONE* 2009, **4**(2):e4576.
78. Lovly CM, Dahlman KB, Fohn LE, Su Z, Dias-Santagata D, Hicks DJ, Hucks D, Berry E, Terry C, Duke M, Su Y, Sobolik-Delmaire T, Richmond A, Kelley MC, Vnencak-Jones CL, Iafrate AJ, Sosman J, Pao W: **Routine multiplex mutational profiling of melanomas enables enrollment in genotype-driven therapeutic trials.** *PloS one* 2012, **7**(4):e35309.
79. Menzies AM, Haydu LE, Visintin L, Carlino MS, Howle JR, Thompson JF, Kefford RF, Scolyer RA, Long GV: **Distinguishing clinicopathologic features of patients with V600E and V600K BRAF-mutant metastatic melanoma.** *Clin Cancer Res* 2012, **18**(12):3242–3249.
80. Sosman JA, Kim KB, Schuchter L, Gonzalez R, Pavlick AC, Weber JS, McArthur GA, Hutson TE, Moschos SJ, Flaherty KT, Hersey P, Kefford R, Lawrence D, Puzanov I, Lewis KD, Amaravadi RK, Chmielowski B, Lawrence HJ, Shyr Y, Ye F, Li J, Nolop KB, Lee RJ, Joe AK, Ribas A: **Survival in BRAF V600–mutant advanced melanoma treated with vemurafenib.** *N Engl J Med* 2012, **366**(8):707–714.
81. Grisham RN, Iyer G, Garg K, DeLair D, Hyman DM, Zhou Q, Iasonos A, Berger MF, Dao F, Spriggs DR, Levine DA, Aghajanian C, Solit DB: **BRAF Mutation is associated with early stage disease and improved outcome in patients with low-grade serous ovarian cancer.** *Cancer* 2013, **119**(3):548–554.
82. Ewalt M, Nandula S, Phillips A, Alobeid B, Murty VV, Mansukhani MM, Bhagat G: **Real-time PCR-based analysis of BRAF V600E mutation in low and intermediate grade lymphomas confirms frequent occurrence in hairy cell leukaemia.** *Hematol Oncol* 2012, **30**(4):190–193.
83. Lemech C, Infante J, Arkenau HT: **The potential for BRAF V600 inhibitors in advanced cutaneous melanoma: rationale and latest evidence.** *Ther Adv Med Oncol* 2011, **4**(2):61–73.
84. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet JP, Ahmann GJ, Adli M, Anderson KC, Ardlie KG, Auclair D, Baker A, Bergsagel PL, Bernstein BE, Drier Y, Fonseca R, Gabriel SB, Hofmeister CC, Jagannath S, Jakubowiak AJ, Krishnan A, Levy J, Liefeld T, Lonial S, Mahan S, Mfuko B, Monti S, Perkins LM, et al.: **Initial genome sequencing and analysis of multiple myeloma.** *Nature* 2011, **471**(7339):467–472.
85. Sved J, Bird A: **The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model.** *Proc Natl Acad Sci* 1990, **87**(12):4692–4696.
86. Hollstein M, Sidransky D, Vogelstein B, Harris CC: **p53 mutations in human cancers.** *Science* 1991, **253**(5015):49–53.
87. Hashimoto K, Rogozin IB, Panchenko AR: **Oncogenic potential is related to activating effect of cancer single and double somatic mutations in receptor tyrosine kinases.** *Hum Mutat* 2012, **33**(11):1566–1575.

doi:10.1186/1471-2105-15-86
Cite this article as: Ryslik *et al.*: A graph theoretic approach to utilizing protein structure to identify non-random somatic mutations. *BMC Bioinformatics* 2014 **15**:86.