Gain and loss of phosphorylation sites in human cancer

Predrag Radivojac1,†, Peter H. Baenziger2,†, MarceI G. Kann3, Matthew E. Mort2, Matthew W. Hahn1,4 and Sean D. Mooney2,5

1School of Informatics, Indiana University, 901 East Tenth Street, Bloomington, IN 47408, 2Center for Computational Biology and Bioinformatics, Department of Medical and Molecular Genetics, Indiana University School of Medicine, 410 West Tenth Street, Suite 5000, Indianapolis, IN 46222, 3Department of Biological Sciences, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250 and 4Department of Biology, Indiana University, 1001 East Third Street, Bloomington, IN 47405, USA

ABSTRACT

Motivation: Coding-region mutations in human genes are responsible for a diverse spectrum of diseases and phenotypes. Among lesions that have been studied extensively, there are insights into several of the biochemical functions disrupted by disease-causing mutations. Currently, there are more than 60 000 coding-region mutations associated with inherited disease catalogued in the Human Gene Mutation Database (HGMD, August 2007) and more than 70 000 polymorphic amino acid substitutions recorded in dbSNP (dbSNP, build 127). Understanding the mechanism and contribution these variants make to a clinical phenotype is a formidable problem.

Results: In this study, we investigate the role of phosphorylation in somatic cancer mutations and inherited diseases. Somatic cancer mutation datasets were shown to have a significant enrichment for mutations that cause gain or loss of phosphorylation when compared to our control datasets (putatively neutral nsSNPs and random substitutions). Of the somatic cancer mutations, those in kinase genes represent the most enriched set of mutations that disrupt phosphorylation sites, suggesting phosphorylation target site mutation is an active cause of phosphorylation deregulation. Overall, this evidence suggests both gain and loss of a phosphorylation site in a target protein may be important features for predicting cancer-causing mutations and may represent a molecular cause of disease for a number of inherited and somatic mutations.

Contact: sdmooney@iupui.edu

1 INTRODUCTION

Mutations in cancers likely alter a great number of molecular events. One of these events is protein phosphorylation, which when altered may result in system-wide disruption and deregulation of signal transduction (Lim, 2005). Indeed, because of their involvement in cancer, kinases remain important drug targets in several classes of human cancers. Currently, kinase inhibitors Gleevec and Herceptin represent the most powerful therapies (Garber, 2006; Lim, 2005; Moasser, 2007).

There are many known spontaneous or somatic amino acid substitutions (Shimizu et al., 2007) and it is likely that some of these will have profound effects on protein function (Kaminker et al., 2007). In addition, two large-scale studies recently identified amino acid substitutions linked to cancer (Greenman et al., 2007; Sjoblom et al., 2006) and others have developed computational methods to predict certain types of cancer mutation sites (Kaminker et al., 2007). Within these sets of mutations, we expect to observe loss-of-function mutations that turn-off normal molecular function. We also expect to observe mutations that cause a gain of function; that is, mutations that cause a molecular function to have deregulated activation when compared to normal function. Finally, we expect to observe many mutations that do not participate in neoplastic development and progression (Sjoblom et al., 2006). These so-called passenger mutations, in fact, may comprise an overwhelming majority of cancer-associated mutations (Futreal et al., 2005). As we begin to discover inherited and spontaneous amino acid substituting mutations in cancer, we have an opportunity to hypothesize their effects and classify as loss-of-function, gain-of-function and no-contributing-function (passengers).

Phosphorylation of amino acid residues serine (S), threonine (T) and tyrosine (Y) is common in cancer-associated proteins (Ikonen et al., 2004) and known to be deregulated in cancer (Lim, 2005). It is well understood that changes in phosphorylation signaling can be due to deregulation of kinase and phosphatase function, usually detected through altered gene expression (Blumes-Jensen and Hunter, 2001; Stephens et al., 2005). This could include amino acid substitutions on kinases or phosphatases that directly interrupt the stability and/or the function of the kinase or phosphatase, resulting in changes in target phosphorylation. Effects of kinase or phosphatase regulators can also lead to altered phosphorylation. Here, we hypothesize that kinase and phosphatase substrates, too, could be subject to the effects of mutation (Fig. 1). A test of this hypothesis is performed to understand how phosphorylation sites in proteins are altered in cancer.

There is evidence in the literature that disruptions of phosphorylation sites are associated with cancer, for instance, mutations of T286 in cyclin D1 (CCND1). Phosphorylation of T286 by GSK3B in the wild type form of cyclin D1 initiates its nuclear export and subsequent degradation in the cytoplasm (Alt et al., 2000; Diehl et al., 1998), while the loss of phosphorylation is causatively implicated in nuclear accumulation of cyclin D1 in esophageal cancer and generally increased oncocgenic potential (Benzeno et al., 2006). Interestingly, while overexpression of cyclin D1 has been linked to a wide range of human cancers, it is becoming increasingly evident that overexpression of its wild type form is not sufficient for oncogenesis (Benzeno et al., 2006), further emphasizing the importance of understanding the functional effects of protein mutations.

© The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
prominent pathway having proteins with mutations likely to disrupt phosphorylation.

2 METHODS

Datasets. Mutation data included inherited disease amino acid substitutions from the Swiss-Prot database (Borkmann et al., 2003), two cancer tumor sequencing studies (Greenman et al., 2007; Sjoblom et al., 2006) and the SeattleSNPs resequencing project. The two cancer tumor sequencing projects and the Swiss-Prot dataset provide our "disease" datasets. The breast and colorectal cancer dataset (herein referred to as the 'B&C Cancer set') originates from a sequencing project involving breast and colorectal cancer tumors (Sjoblom et al., 2006). A subset of B&C Cancer limited to mutations in genes statistically overmutated is referred to as 'B&C Cancer—CAN'. The 'Kinase' dataset originates from a project that sequenced kinase genes from over 200 individual cancer tumors (Greenman et al., 2007). For the analysis of Swiss-Prot, only disease-associated variants were included, while variants noted as 'in dbnp' or having reference to collagen were excluded. We removed collagen variants due to overrepresentation in the set of inherited disease-associated mutations and the regularity of the collagen sequence that can bias the prediction process. To obtain population-relevant polymorphism data, we obtained all amino acid substitutions from the SeattleSNPs website. In total, 661 amino acid substitutions were analyzed in 189 genes and allele frequencies were determined from the SeattleSNPs project resequencing of African and European descent populations. For the purposes of this study, the wild type allele was defined as the most common allele for polymorphisms.

Additionally, several control sets were created. First, a model of 'neutral' mutations was constructed by collecting amino acid differences between human protein sequences and orthologous sequences from Ensembl's Compara database (Hubbard et al., 2007), version 41. Neutral mutations were considered to be those where the orthologous sequence differed from both the wild type human protein and all considered mutants. They were limited to five closely related species: Bos taurus, Canis familiaris, Macaca mulatta, Oryctolagus cuniculus and Rattus norvegicus. Second, a random set of amino acid substitutions was constructed using Swiss-Prot data (version 45), where mutations were generated by randomly substituting the wild type amino acid by any of the remaining amino acids at a random position in wild type protein sequence. This set serves as a model of a computationally random comparison set. Lastly, two codon-based, random sets of mutations were created. These datasets were created by mutating the wild type sequence in our previously described cancer datasets (B&C Cancer and Kinase) and are therefore referred to as the 'B&C Cancer Control' set and the 'Kinase Control' set. The mutations were generated at the codon level and took into account measures of amino acid difference, codon frequency, transition-transversion ratio and gene variability (Goldman and Yang, 1994). Parameters for this mutating algorithm include: µ (mutation rate per site per generation), µ (mutation rate per site per generation), and γ (transition–transversion ratio, set to 1.45, as suggested by Goldman and Yang) and γ (gene variability, set to 43.99, as suggested by Goldman and Yang). Goldman and Yang provide a set of equations to calculate the probability that a given codon will mutate to the nine alternatives, limiting mutations to only one of the three nucleotides, as

\[ Q_{ij} = \frac{\mu \pi_{i} e^{-d_{ij}/\mu}}{\sum_{k} \pi_{i} e^{-d_{ij}/\mu}} \] (for transversions)

\[ Q_{ij} = \frac{\mu \pi_{i} e^{-d_{ij}/\mu}}{\sum_{k} \pi_{i} e^{-d_{ij}/\mu}} \] (for transitions)

where \( Q_{ij} \) is the probability that codon \( i \) will mutate to codon \( j \) and \( d_{ij} \) is the distance between amino acids \( i \) and \( j \) according to Li et al. (1985).

To summarize the dataset information, we list the number of mutations and proteins in each of the datasets: B&C Cancer (1099 mutations; 847 proteins), B&C Cancer—CAN (150; 171), B&C Cancer Control (7658; 847), Ensembl Orthologs (1181; 498), Kinase Cancer (695; 312), Kinase Cancer Control
because of the loss of phosphorylation will be about 0.3. If the two probabilities are 0.9 and 0.1, the probability of a loss of phosphorylation at residue \( s \) is:

\[
P_{\text{loss}}(\text{at } s) = \frac{P_{\text{loss}}(s)}{P_{\text{loss}}(s)} = \frac{1}{1} = 1.
\]

This definition is practically useful because it dictates that the DisPhos predictions have a 5–10% false positive rate (depending on the residue type) and that 10–20% of all predicted positives are false identifications (Iakoucheva et al., 2004). In reality, we believe that a lower percentage of all ‘false positive’ identifications are indeed false, because the set of ‘negative’ phosphorylation sites used to estimate false positive rate is believed to contain a number of actual phosphorylation sites that are yet unidentified or not stored in major databases.

### 3 RESULTS

#### 3.1 Prevalence of phosphorylation site predictions in training set sequences

To verify the datasets were similar in the number of mutations involving a phosphorylatable amino acid mutating to a non-phosphorylatable amino acid or vice versa, we compared percentages of each dataset representing these types of mutations. Figure 2 shows that the datasets are closely gathered around an average of 12.2% of the data involving mutations from S, T or Y to another amino acid (Fig. 2; dark grey bars), and around 13.5% for the mutations to a phosphorylatable amino acid (Fig. 2; light grey bars). Interestingly, both datasets created with the biologically intelligent mutating algorithm have fewer mutations to phosphorylatable amino acids (Fig. 2; light grey bars). This likely occurs because of the consideration the algorithm gives to evolutionary preferred amino acid substitutions (mostly neutral).

Prediction was completed on 53,690 mutations from 5,390 unique protein sequences. All insertions and deletions, all substitutions...
containing discrepancies between the sequence and mutations, and all protein sequences containing symbols not in A were excluded from the analysis, therefore resulting in smaller mutation datasets than those originally published. From the starting set of 53,690 mutations, 14,844 had changes of phosphorylation prediction (>0.001) on the mutation site. However, only 1033 of these mutations were considered high-confidence loss/gain-of-phosphorylation sites. The original set of 53,690 mutations also caused phosphorylation score changes (>0.001) on 612,435 non-mutation sites, yet none of the non-mutation sites were of high confidence. Only 20 such mutations yielded a loss/gain-of-phosphorylation score >0.5. This is expected from two standpoints: (i) functionally, a mutation around a phosphorylation site is more likely to change the efficiency of enzymatic reaction instead of fully disrupting it and (ii) statistically, a single change away from the modification site is less likely to be critically important due to an implicit assumption of statistical inference methods that similar sequence patterns should generally result in similar prediction scores, especially for robust inference models such as DisPhos.

3.2 Analysis of phosphorylation sites in germline polymorphic positions in two populations

In order to understand phosphorylation in the context of mutational frequency, we studied mutations from the SeattleSNPs resequencing project. We partitioned these amino acid substitutions into allele frequency ranges (low: frequency <0.05; medium: 0.25 < frequency <0.75; and high: frequency >0.95) based on data from the SeattleSNPs project and averaged the allele DisPhos scores. As Figure 3 shows, mutations with low allele frequency have a positive correlation with the propensity for phosphorylation. This result seems to support the recent finding that rare alleles tend to be mildly deleterious (Kryukov et al., 2007).

3.3 Predicted gain and loss of phosphorylation sites in cancer compared to other datasets

In Figure 4A, we illustrate that the three cancer mutation sets show a greater rate of gain of phosphorylation than the non-disease-associated sets. For the highly confident predictions, the Kinase and B&C Cancer sets show gain of phosphorylation occurring on 1.87% (n = 13) and 1.91% (n = 21) of the mutations, respectively. These percentages are significantly larger than those in the control sets (Swiss-Prot Random: 0.81% (n = 194), B&C Cancer Control: 0.86% (n = 66), Kinase Control: 0.88% (n = 48) and Swiss-Prot: 0.89% (n = 112)). The originating authors of the B&C Cancer set used more stringent statistical filters to isolate genes showing significantly higher mutation rates than expected. Limiting the B&C Cancer set to mutations from the overly mutated genes shows a gain of phosphorylation predicted on 2.29% (n = 8) of the dataset (B&C Cancer—CAN). Finally, while the inclusion of data from orthologous sequences seems to have provided a good neutral model, we note that there are multiple examples where disease-associated mutations in humans correspond to the wild type sequences even in closely related species (Ostegdaard et al., 2007; Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007).

Figure 4B shows the rates of high-confidence loss of phosphorylation in each dataset. Again, cancer datasets show the greatest loss of phosphorylation and the Kinase dataset is particularly interesting because of its large percentage of loss of phosphorylation sites: 3.17% (n = 22). As mentioned earlier and shown in Figure 2, these results are not biased by the number of mutations to or from phosphorylatable amino acids.

3.4 Motif and pathway analysis of proteins affected by change of phosphorylation

Disease sets showing higher gain and/or loss of phosphorylation were analyzed to identify possible overrepresentation in known phosphorylation site motifs. We used Scansite (Obenauer et al., 2003) to identify the frequency of motifs present in each dataset and compared highly disrupted motifs in the cancer datasets to the motifs disrupted in the cancer control sets. Several motifs showed large differences: Akt Kinase, Clk2 Kinase, 14-3-3 Mode 1, Nick 2, Cortactin SH3 and Grb2 SH3, Ctk SH3, Amphilysin SH3, Erk1 Kinase, DNA-activated Protein Kinase, Casein Kinase 1 and 2 and ATM Kinase. However, due to the further fragmentation of disease datasets, we were able to identify only two motifs that are disrupted in the cancer data with statistical significance (Bonferonni-corrected t-test): Cortactin SH3 and Grb2 SH3.

For pathway analysis, we created several subsets of proteins from the breast and colorectal cancer mutation set. By comparing the percent of each of these sets classified into specific pathways, we observe that gain of phosphorylation-associated mutations are enriched for several pathways (Fig. 5). Specifically, we see that these proteins are associated with signaling pathways such as Wnt/β-catenin, which has been shown to be associated with cancer and tumorigenesis (Mori et al., 1992; Spink et al., 2000).

3.5 Evidence for our predictions in the literature

A review of the literature provides evidence in support of our predictions. For example, mutation of T286, a known phosphorylation site, in cyclin D1, a known cancer-associated
Gain and loss of phosphorylation sites in human cancer

Fig. 4. Mutations that cause loss and gain of phosphorylation target sites are enriched in cancer. Relative frequencies of high-confidence gain-of-phosphorylation sites (A) and high-confidence loss-of-phosphorylation sites (B) in various datasets. P-values were calculated using a t-test and statistical significance determined using a Bonferroni-corrected threshold of 0.05.

Fig. 5. IPA of proteins hosting mutations altering phosphorylation sites. After running IPA, the percent of proteins from each dataset classified into the pathways above was graphed for each subset of the B&C Cancer dataset.

Our results also complement those found by studying individual phosphorylation site mutations. In a review of EGFR and ErbB4's roles in tumorigenesis and signal transduction, several examples of phosphorylation site mutations are given which through creation of new docking sites lead to changes in regulation (Schulze et al., 2005; Yarden and Sliwkowski, 2001). Some interesting sites show gain and loss of phosphorylation, but have not been experimentally validated. In AXIN-1 (Spink et al., 2000) and BARD-1 (GenBank gi|20532378

4 DISCUSSION

4.1 Phosphorylation target site mutation is likely a mechanism in cancer

Based on the results of this study, we believe both gain and loss of phosphorylation are an important mechanism causing deregulation of phosphorylation mediated signal transduction. We find that kinases in cancer are twice as likely to have mutations disrupting phosphorylation sites as compared to a kinase control set, but the increase compared to the set of Swiss-Prot mutations, human variation data (SeattleSNPs) and orthologous sequences is 4-, 5- and 8-fold, respectively. Similarly, gain of phosphorylation sites in cancer-associated mutations is about 2-fold as compared to Swiss-Prot and human variation data. This is an intriguing finding because it suggests that a mechanism of signal transduction deregulation in cancer is mediated by either removal or creation of phosphorylation sites thereby causing either a loss or a gain of phosphorylation function, depending on the role of the phosphorylated residue. This is further supported by the fact that an increase in number
of high-confidence phosphorylation sites is correlated with low allele frequencies in two human populations, suggesting purifying evolutionary selection against these mutations.

Perhaps most interesting of all is that cancer-associated mutations are predicted to be enriched in gain of phosphorylation sites. It can be speculated that mutations which create phosphorylation sites could destabilize proteins, interrupt protein interactions, or disrupt enzyme catalysis or other normal protein functions. They may also recruit kinases or phosphatases necessary for other cellular processes causing system-wide deregulation. When we examine the signaling pathways in which we observe phosphorylation disrupting mutations, we find that several pathways are enriched in mutations. Notably, the Wnt/β-Catenin pathway is enriched in both gain and loss of phosphorylation sites. This pathway is well known in cancer and several reviews convey its importance (Bienz and Clevers, 2000; Spink et al., 2000).

It is tempting to speculate on the percentages of known mutations likely causing gain or loss of phosphorylation sites. Using the high-confidence scores only, we predict that about 2% of disease-associated mutations in Swiss-Prot are caused by changes in phosphorylation patterns. For a less stringent threshold of 0.5 for the prediction of phosphorylation sites, our prediction rises to about 7%. However, these estimates should be interpreted with caution given by the recent analysis of Care et al. (2007).

The findings of our study that gain and loss of phosphorylation sites are linked to human cancer add to the increasing breadth of literature addressing post-translational modifications and disease, e.g. the role of glycosylation in inherited mutations (Vogt et al., 2005, 2007).

4.2 Machine learning methods that predict functional sites can be used to predict mutation disruption

One of the conclusions of this study is that residue functional prediction methods can give insight into the molecular mechanism of specific disease-causing mutations. Currently, prediction methods focus on utilizing sequence, protein structure and evolutionary features to predict whether an amino acid substitution is likely to affect protein function or be involved in disease (Bromberg and Rost, 2007; Ng and Henikoff, 2003; Ramensky et al., 2005, 2007).


Gain and loss of phosphorylation sites in human cancer


