Systematic Analyses and Prediction of Human Drug Side Effect Associated Proteins from the Perspective of Protein Evolution

Tina Begum[1], Tapash Chandra Ghosh[2], and Surajit Basak[1,3,*]

[1]Bioinformatics Centre, Tripura University, Suryamaninagar, Tripura, India

[2]Bioinformatics Centre, Bose Institute, C.I.T. Scheme VII M, Kolkata, India

[3]Department of Molecular Biology & Bioinformatics, Tripura University, Suryamaninagar, Tripura, India

*Corresponding author: E-mail: basaksurajit@gmail.com.

Accepted: February 16, 2017

Abstract

Identification of various factors involved in adverse drug reactions in target proteins to develop therapeutic drugs with minimal/no side effect is very important. In this context, we have performed a comparative evolutionary rate analyses between the genes exhibiting drug side-effect(s) (SET) and genes showing no side effect (NSET) with an aim to increase the prediction accuracy of SET/NSET proteins using evolutionary rate determinants. We found that SET proteins are more conserved than the NSET proteins. The rates of evolution between SET and NSET protein primarily depend upon their noncomplex (protein complex association number = 0) forming nature, phylogenetic age, multifunctionality, membrane localization, and transmembrane helix content irrespective of their essentiality, total druggability (total number of drugs/target), m-RNA expression level, and tissue expression breadth. We also introduced two novel terms—killer druggability (number of drugs with killing side effect(s)/target), essential druggability (number of drugs targeting essential proteins/target) to explain the evolutionary rate variation between SET and NSET proteins. Interestingly, we noticed that SET proteins are younger than NSET proteins and multifunctional younger SET proteins are candidates of acquiring killing side effects. We provide evidence that higher killer druggability, multifunctionality, and transmembrane helices support the conservation of SET proteins over NSET proteins in spite of their recent origin. By employing all these entities, our Support Vector Machine model predicts human SET/NSET proteins to a high degree of accuracy (~86%).

Key words: side effect associated drug target (SET), nonside effect associated drug target (NSET), killer druggability, essential druggability, protein evolutionary rates, support vector machine (SVM).

Introduction

The extents to which evolutionary changes have affected drug side effects among human targets remain unanswered till date. The present study intends to elucidate the grounds of unwanted toxic side effects in drug targets. A handful databases (Kuhn et al. 2013; Juan-Blanco et al. 2015; Zhou et al. 2015) for human proteome are of immense importance in this regard. Based on one of such databases (SIDER 2 of Kuhn et al. [2013]) with actual experimental drug adverse data, Wang et al. (2013a) worked on drug target proteins from a network perspective to explore the biological factors leveraging drug side effects in human. They (Wang et al. 2013a) discovered that protein essentiality and centrality largely drive side effects of the reported drugs. Focusing on human essential targets, a recent study (Liu and Altman 2015) has highlighted the influence of low affinity binding to "off-target" proteins as a potential cause of drug side effect(s). Binding to multiple targets could be another plausible reason of producing drug side effects in human (Wang et al. 2013b).

Earlier, we have (Begum and Ghosh 2014) emphasized that interactome network perturbation (drug-induced perturbation) strongly influences target drug side effects. In 2015, Perez-Lopez et al. (2015) found that drug side effect associated targets (SET) are better spreaders of network perturbations than nonside effect associated targets (NSET). Using two different interactome network perturbation models (edgetic perturbation model and drug-induced perturbation model), we have previously established that proteins prone to network perturbations are conserved in nature. All the aforementioned studies stress on the fact that evolutionary rates should vary between SET and NSET proteins. Therefore, in the present

study, we want to identify various factors responsible for between groups (SET vs. NSET) evolutionary rate variation. A wide range of publications have already disentangled determinants of protein evolutionary rates in terms of nonsynonymous amino acid changing substitution rates (*d*N) and/or using the ratio of nonsynonymous to synonymous substitution rates (*d*N/*d*S) (Hirsh and Fraser 2001; Fraser et al. 2002; Alba and Castresana 2005; Drummond et al. 2006; Pal et al. 2006; Cai et al. 2009; Wolf et al. 2009a; Begum and Ghosh 2010; Panda et al. 2012; Zhang and Yang 2015). A number of determinants are yet to be identified to explain total variability of protein evolutionary rates. In this perspective, another objective of our present study is to find out some novel biological attributes which we can use along with previously known evolutionary rate determinants to increase the prediction accuracy of side/non side effect associated proteins in human.

Using high-coverage genome sequence data from thirteen vertebrate species (including human, mouse, cow etc.) to evaluate the consistency of *d*N and *d*S estimates using eight commonly-used methods, Wang et al. (2011) has established that all the methods yielded a nearly uniform result when estimating *d*N, but not *d*S (or *d*N/*d*S). Along with Wang et al. (2011), a couple of other researchers used *d*N rather than *d*N/*d*S to compute protein evolutionary rates efficiently (Alba and Castresana 2005; Drummond et al. 2005; Wolf et al. 2009b; Wang et al. 2011). Hence, in this article, we emphasized on commonly used human–mouse orthologous pair (Liao and Zhang 2006; Gharib and Robinson-Rechavi 2011; Georgi et al. 2013) to define fast evolving and slow evolving protein coding genes in SET and NSET groups on the basis of their *d*N data. Using the druggable subset (drug targeted proteins) of human genome, our study demonstrates that NSET proteins evolve faster than SET proteins, mainly within noncomplex forming group. The trend of our result does not change after controlling two major determinants of protein evolutionary rates—gene expression level and tissue expression breadth (Drummond et al. 2005, 2006; Pal et al. 2006; Cai et al. 2009; Park and Choi 2010; Zhang and Yang 2015). Although, target essentiality (the target protein is essential in nature) is claimed to play lead role in drug side effects (Wang et al. 2013a, 2013b; Liu and Altman 2015), we observed that the variation in evolutionary rates between SET and NSET proteins does not solely depend on target essentiality. We also used three druggability measures of targets: (i) total druggability, essential druggability, and killer druggability to explain the rate discrepancy of SET vs. NSET proteins. Among the three factors, killer druggability can partially explain the conservation of comparatively recently emerged SET proteins over older NSET proteins. In the context of protein evolution, our study establishes relationship between protein age, killer druggability, protein multifunctionality, and subcellular localization to explain the variation of evolutionary rates between the SET and NSET proteins. We observed that recently emerged SET proteins are conserved if they are membrane localized, more multifunctional, have higher killer druggability and elevated transmembrane helix content. Finally, Support Vector Machine (SVM) (Chang and Lin 2011) approach was adopted to predict SET/NSET proteins considering all the evolutionary rate attributes (genomic, structural, and functional) those differ significantly between the SET and NSET proteins. Our machine learning approach can predict SET/NSET proteins to a high degree of accuracy (~86%) with 94% precision level.

Materials and Methods

Collection of Protein Evolutionary Rate Data of Human SET and NSET Proteins

DrugBank v.4.3 (Law et al. 2014) was used to retrieve a catalogue of total 2426 human drug targets and drug related information. We chose SIDER v.2 (http://sideeffects.embl.de/) database to identify drugs with 996 experimental side effect information (Kuhn et al. 2013; Zhou et al. 2015). For a target protein, more than one drug may have association to side effect(s). We, therefore, considered a protein as side effect associated (SET) if its target drug(s) have at least one known side effect. Nonside effect associated proteins (NSET) are those whose drug targets are not associated with any side effect. BioMart interface of Ensembl v.82 (Flicek et al. 2013) was utilized to obtain evolutionary rate (*d*N) data using human–mouse orthologous pairs with $dS < 3$ (Tang and Epstein 2007; Begum and Ghosh 2014; Acharya and Ghosh 2016). However, proteins with $dS = 0$ were not considered for this study (although the result remained unchanged when we considered them). The strength of selection was inferred based on the value of *d*N/*d*S (Drummond et al. 2006; Pal et al. 2006; Cai et al. 2009; Wang et al. 2011; Zhang and Yang 2015). Mapping of those evolutionary rate data to SET and NSET proteins finally yielded 388 SET and 1488 NSET proteins (supplementary table S1, Supplementary Material online) with accessible *d*N (and *d*N/*d*S) data for further analyses.

Acquiring Essential Proteins and Computing Target Essential Druggability

Target essentiality data for 2472 human proteins were obtained from the study of Georgi et al. (2013). After matching these essential proteins with our data set, we finally attained target essentiality information of 120 SET and 361 NSET proteins.

Considering all 2472 human essential proteins, we first catalogued the drugs that targets essential proteins using DrugBank v.4.3 (Law et al. 2014) data. Next, we counted the number of such drugs against each protein target of our data set to get the essential druggability of that target. By this way, we got essential druggability of 1382 proteins out of total 1876 proteins.

Estimation of Target Killer Druggability

To estimate killer druggability of a target protein, we listed the drugs for which killing/toxic side effect(s) currently exist in SIDER2 database (Kuhn et al. 2013). Zhou et al. (2015) considered death, sudden death, sudden cardiac death, cardiac death, cancer, hemorrhagic strokes, heart failure, and congestive heart failure as killing side effects. However, we considered death, sudden death, cancer, metastasis, myocardial infarction, heart failure, congestive heart failure, coma, completed suicide, stroke, and cardiac arrest as potentially toxic/killing side effects for our study. Next, we counted the number of drug(s) exhibiting killing side effects against each protein target to assign the killer druggability of the target. It is a property of SET proteins and we found the killer druggability of 215 (out of total 388) SET proteins for our investigation.

Retrieval and Analyses of Protein Complex Data

CORUM database (http://mips.helmholtz-muenchen.de/genre/proj/corum/) was browsed to collect human protein complex assembly data (Ruepp et al. 2010). The number of complexes in which a target protein participates represents the complex number of that target (Chakraborty et al. 2010). Thus, we perceived 385 complex forming proteins for further investigation.

To reconfirm our result, we further considered large protein complexes (size ≥ 5), since, very small protein complexes of CORUM could produce high hit rate fluctuations (Bin Goh et al. 2015). Following this method, protein complex data for SET/NSET proteins get reduced from 385 to 178.

Estimation of Gene Expression Level and Tissue Expression Breadth

We fetched http://genes.mit.edu/burgelab/mrna-seq/ to retrieve human mRNA-seq data to evaluate gene expression levels (Wang et al. 2008). Following Huang et al. (2013), we call a gene is expressed in a tissue if its expression value is greater than $M + 2 \times MAD$, where M and MAD are determined by $M = \text{median}(x)$; and x indicates the average expression values for the corresponding gene among all tissues. For each gene, we counted the number of expressed tissues to indicate its tissue expression breadth (Begum and Ghosh 2014). To represent the expression level of a gene, we took the average expression value in the tissues where it is found to be expressed (Begum and Ghosh 2014). By this way, we obtained expression data of total 1411 genes of our data set.

Dating Protein Age

Two different approaches were taken to quantify the phylogenetic emergence of a protein. As a first measure, we collected the list of human gene encoded proteins along with their phylostrata of origin from the supplementary data given by Domazet-Lošo and Tautz (2008). We considered a protein as "old" if it falls in ps1 or ps2 (before fungi and up to eukaryotes) group and as "new" if it belongs to any other ps group (Domazet-Loso and Tautz 2008; Nagaraj et al. 2010). We then mapped the age data to our data set of SET/NSET proteins which finally provided 365 "new" and 1459 "old" proteins for further analyses. As an alternative to the first method (which provides categorical age information), we estimated the evolutionary origin of our proteins of interest (in Ma: million years ago) using a phylogenetic approach implemented by ProteinHistorian tool (Capra et al. 2012). By this way, we obtained numerical age data of 1836 proteins.

Determining Pleiotropic Index of a Protein

We contemplated on Gene Ontology (GO) annotation for the "biological process" from Ensembl Genome Browser v.82 (He and Zhang 2006; Chakraborty and Ghosh 2013; Flicek et al. 2013) to calculate the multifunctionality/pleiotropic index of a protein.

Analyzing Protein Subcellular Localization

Even though a couple of databases are available to predict subcellular localization of a protein, we deliberately used the data provided by Wang et al. (2013b) to identify the subcellular localization of proteins of our data set. Following this method, we spotted the subcellular localizations of 1004 SET/NSET proteins.

However, to predict the transmembrane helices of a protein, we relied on hidden Markov model based TMHMM v.2.0 Server (http://www.cbs.dtu.dk/services/TMHMM/), the best performing transmembrane prediction program (Huang et al. 2002).

Statistical Analyses

All statistical analyses were done with the help of SPSS v.13. We considered nonparametric Spearman rank correlation coefficient and two-tailed Mann–Whitney U test (MWT) to calculate correlation and difference between two data sets, respectively. To create randomized data set for performing simulation study, we used R package v. 2.13.1 (http://www.r-project.org). For randomization, if n1 and n2 be the sample sizes respectively ($n1 > n2$); we drew n2 points from the first sample of size n1 at random to compute the mean of each subset (Necsulea et al. 2009). We did n2 times of such simulations. Finally, we used R v 2.13.1 package to compare the mean of two groups (Necsulea et al. 2009; Hesterberg et al. 2010). The results were considered to be statistically significant if the P value is less than 0.05.

Implementation of Support Vector Machine

The SVM algorithm implemented in the open access LIBSVM v.2.84 package (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) (Chang and Lin 2011) was adopted to perform prediction of

SET/NSET proteins using different evolutionary rate attributes. In our data set of 1876 proteins, 388 SET proteins were classified as positives and rest of the 1488 NSET proteins was treated as negative cases. Using "subset.py" function of LIBSVM (Chang and Lin 2011), 70% of the total data set was randomly chosen as the training data set and the remaining 30% was used as test set. All the attributes in the training and test data sets were scaled in the range of −1 to 1. For our study, we considered most widely used C-SVC type SVM and Radial Basis Function (RBF) kernel for better performance of our model (Dey et al. 2012). The penalty parameter C and the RBF kernel parameter g were optimized using grid search, a built in function of LIBSVM (Chang and Lin 2011). In addition to 5-fold cross-validation, we checked the average performance of our model on 10 different randomly generated training and test sets.

Performance Assessment

Considering TP, FP, TN, and FN as the number of true positives, false positives, true negatives and false negatives respectively; we estimated the prediction accuracy $\left[\frac{(TP+TN)}{(TP+FN+TN+FP)} \times 100\right]$ of our model (Dey et al. 2012; Kisslov et al. 2014). In addition, we relied on a balanced measure "Matthews correlation coefficient (MCC)" during cross-validation, as the positive to negative case ratio was not one (Dey et al. 2012). The MCC value was determined using the following formula:

$$MCC = \frac{[(TP \times TN) - (FP \times FN)]}{\sqrt{[(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)]}}$$

Where, the values +1, 0, and −1 signify a perfect, a random and an inverse prediction, respectively (Dey et al. 2012). We also determined sensitivity or recall $\left[\frac{TP}{(TP+FN)} \times 100\right]$, specificity $\left[\frac{TN}{(TN+FP)} \times 100\right]$, positive predictive value (PPV) or precision $\left[\frac{TP}{(TP+FP)} \times 100\right]$, negative predictive value (NPV) $\left[\frac{TN}{(TN+FN)} \times 100\right]$, and F-measure $\left[\frac{2 \times \text{Precision} \times \text{Sensitivity}}{(\text{Precision} + \text{Sensitivity})}\right]$ to evaluate the prediction performance of our model (Dey et al. 2012; Kisslov et al. 2014).

Results

Side Effect Associated Drug Targets Are Conserved and Exhibit Higher Druggability

In this study, we collected evolutionary rates (dN) data from public resources for the druggable subset of human SET and NSET proteins ($n = 1876$). We observed that SET proteins evolve slower than the NSET proteins (average $d\text{N}_{\text{SET}} = 0.062$, average $d\text{N}_{\text{NSET}} = 0.083$, $P_{\text{MWT}} = 2.81 \times 10^{-5}$). Similar trend was observed using $d\text{N}/d\text{S}$ data (average $d\text{N}/d\text{S}_{\text{SET}} = 0.092$, average $d\text{N}/d\text{S}_{\text{NSET}} = 0.126$, $P_{\text{MWT}} = 2.08 \times 10^{-6}$). Interestingly, we did not find any difference in dS between SET and NSET proteins (average $d\text{S}_{\text{SET}} = 0.676$, average $d\text{S}_{\text{NSET}} = 0.669$, $P_{\text{MWT}} = 3.65 \times 10^{-1}$). This analysis suggests that the difference in $d\text{N}/d\text{S}$ value between NSET and SET proteins mainly depends upon the difference in dN. To assess that the significance is not due to the sample size difference of the proteins of our interest, we performed nonparametric test with randomized data set of SET and NSET proteins (10,000 simulated replicates). Thereby, a significant difference (Wilcoxon rank sum test: $P < 2.20 \times 10^{-16}$) in dN between SET and NSET strengthens the fact that our observation is free from sample size bias.

We observed an inverse correlation between total druggability (total number of drugs/target) and protein evolutionary rates ($\rho_{\text{Number of drugs/target}}{}^{d\text{N}} = -0.060$, $P = 9.00 \times 10^{-3}$, $n = 1876$) which is in agreement with the earlier observation by Wang et al. (2013b). To understand the evolutionary rate heterogeneity between human SET and NSET proteins, we pooled all the proteins into two bins according to their target drug number(s) (bin 1 [number: 1–3]: Low, bin 2 [number ≥4]: High). Interestingly, we found that the difference in dN between SET and NSET proteins exists only in bin 2 (fig. 1), whereas the total druggability differs between SET and NSET proteins in both bin 1 ($P_{\text{MWT}} = 1.39 \times 10^{-8}$) and bin 2 ($P_{\text{MWT}} = 5.58 \times 10^{-16}$). Similar result was obtained using $d\text{N}/d\text{S}$ (data not shown). This study affirms that total druggability may not be a crucial factor for obtaining evolutionary rate heterogeneity between human SET and NSET proteins.

Recently, Zhou et al. (2015) have claimed that 51% of FDA approved drugs have higher killing index, which usually denotes the potentiality of a drug having harmful side effects. In our analysis, we computed the number of drugs exhibiting killing side effects for each target protein and termed them as "killer druggability" of that target protein. When we correlated dN and the killer druggability of proteins, killer druggability has been found to have only a weak but significant negative correlation with protein evolutionary rates (Number of killer drugs/target $\rho^{d\text{N}} = -0.085$, $P = 2.16 \times 10^{-4}$, $n = 1876$). This observation suggests that killer druggability may influence evolutionary rates of SET/NSET proteins. Based on protein evolutionary rates, equally populated bins (i.e., having the identical number of proteins in each of the two bins: bin 1 [rate: 0.0006–0.0576; slow evolving], bin 2 [rate: 0.0576–0.5762; fast evolving]) were constructed to better understand the distribution pattern of killer druggable proteins (proteins with killer druggability > 0). The pattern is found to be consistent with our correlative study since slowly evolving group (bin 1) contains more proteins (58.60%) with higher killer druggability (killer druggability > 0) than that of other proteins (48.89%) with killer druggability (killer druggability = 0) (Two sided Fisher's exact test: $P = 4.10 \times 10^{-2}$). To comprehend the relative influences of total druggability and killer druggability on the rates of

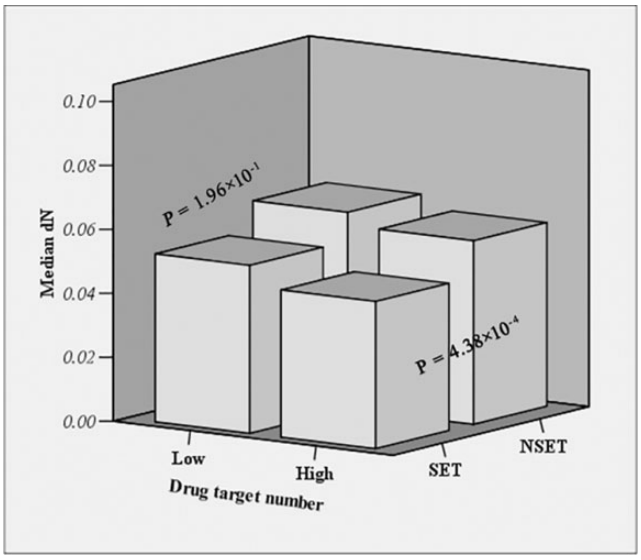

Fig. 1.—The relationships between protein evolutionary rates (*d*N) and target drug numbers of human SET vs. NSET proteins in different bins. $P_{MWT} < 0.05$ denotes significant difference between groups.

molecular evolution, we performed a linear regression analysis considering *d*N as dependent variable and total druggability, killer druggability as independent variables. We observed that killer druggability is a better predictor of protein evolutionary rates ($F = 6.979$, $P = 1.00 \times 10^{-3}$) than total druggability ($\beta_{\text{total druggability/target}} = -0.021$, $P = 5.26 \times 10^{-1}$; $\beta_{\text{killer druggability/target}} = -0.070$, $P = 3.40 \times 10^{-2}$).

Target Essentiality and Target Essential Druggability Are Not Sole Determinants of SET/NSET Protein Evolutionary Rates

It was observed that essential genes evolved slowly than nonessential genes (Kimura and Ohta 1974; Chakraborty and Ghosh 2013; Zhang and Yang 2015) and essential druggable targets were shown to have strong affinities towards drug side-effects (Liu and Altman 2015). We mapped human essential genes [provided by Georgi et al. (2013)] to our data set and found that essential proteins evolve slower than nonessential proteins (average $d\text{N}_{\text{Essential proteins }[n=481]} = 0.059$, average $d\text{N}_{\text{Nonessential proteins }[n=1395]} = 0.085$, $P_{MWT} = 4.84 \times 10^{-16}$). This finding led us to presume that the conservation of SET proteins over the NSET group may largely attributable to their difference in essential protein contents. Despite the fact that the proportions of essential proteins are significantly higher in SET (30.93%; 120/388) compared to the NSET proteins (24.26%; 361/1488) (SET vs. NSET: *Z* score = 2.679, $P = 7.40 \times 10^{-3}$), no significant difference in the rate of protein evolution was detected between SET and NSET proteins within the essential gene set (average $d\text{N}_{\text{SET}} = 0.055$, average $d\text{N}_{\text{NSET}} = 0.061$, $P_{MWT} = 7.74 \times 10^{-1}$). From this study, we infer that the difference in rate of evolution between SET and NSET proteins is independent of target essentiality.

In other words, the number of essential proteins that bind to a drug is a key predictor of drug side-effects (Wang et al. 2013a; Liu and Altman 2015). To investigate, we introduced a new term "essential druggability" of a target protein which actually denotes the number of drug(s) having one or more essential partner(s) for each target protein (irrespective of the essential/nonessential nature of the target). We hypothesize that increasing essential druggability of a target will intensify its sequence conservation rate and indeed we observed the same trend (Number of drugs that binds to essential proteins/target $\rho^{dN} = -0.147$, $P = 1.00 \times 10^{-6}$, $n = 1382$). We also found that essential druggability differs between SET and NSET protein targets (average essential $\text{druggability}_{\text{SET}\ [n=360]} = 9.328$, average essential $\text{druggability}_{\text{NSET}\ [n=1022]} = 2.208$, $P_{MWT} = 1.21 \times 10^{-62}$). We thus infer that essential druggability could be a potential evolutionary rate determinant. To better understand, proteins are pooled into a number of bins by means of their essential druggability (bin 1 [number: 1–2]: Low, bin 2 [number: 3–4]: Medium, bin 3 [number $\geq$5]: High). Thereby, we found that the difference in *d*N does not exist between SET and NSET proteins in any of the three bins (P_{MWT}: bin $1 = 2.87 \times 10^{-1}$, bin $2 = 2.08 \times 10^{-1}$, bin $3 = 4.16 \times 10^{-1}$) while target druggability differs in bin 1 and bin 3 (P_{MWT}: bin $1 = 2.00 \times 10^{-3}$, bin $2 = 8.90 \times 10^{-2}$, bin $3 = 2.42 \times 10^{-5}$). The result does not change using *d*N/*d*S data (data not shown). All these analyses indicate that essential druggability may not be a sole crucial factor for obtaining evolutionary rate heterogeneity between SET and NSET proteins.

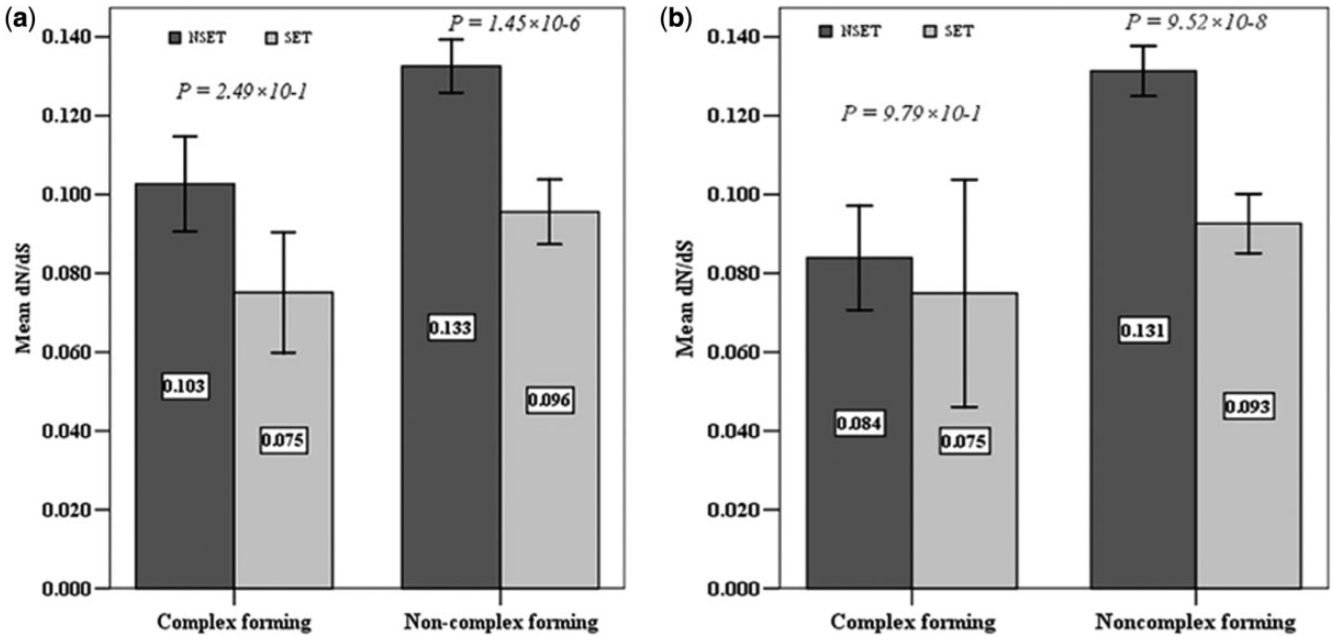

Fig. 2.—The impact of complex forming ability on *d*N/*d*S of SET and NSET proteins. The bar graphs demonstrate the difference in the distribution of *d*N/*d*S between SET and NSET proteins in complex/noncomplex forming groups considering (*a*) all protein complexes (*b*) large protein complexes (size ≥ 5) of CORUM database. $P_{MWT} < 0.05$ between groups was used to represent a statistically significant difference. Error bars signify 95% confidence interval.

Noncomplex Forming Ability Plays Key Role in Evolutionary Rate Disparity between SET and NSET Proteins

Protein complexes where two or more proteins integrate to carry out and regulate a variety of cellular functions and essential genes are more likely to be components of the protein complexes (Wang et al. 2009; Chakraborty and Ghosh 2013; White et al. 2013). Such complex forming proteins endure strong selective pressure due to their functional constraints (Teichmann 2002; Mintseris and Weng 2005; Chakraborty et al. 2010). Even though, complex forming proteins are found to be slow evolving than noncomplex forming protein (average $d\mathrm{N}_{\text{complex forming }[n=385]} = 0.063$, average $d\mathrm{N}_{\text{noncomplex forming }[n=1491]} = 0.083$, $P_{MWT} = 7.45 \times 10^{-13}$), we did not find any difference in protein complex number (number of complexes a protein belongs to) between SET and NSET protein sets (average complex $\text{number}_{\text{SET }[n=388]} = 0.771$, average complex $\text{number}_{\text{NSET }[n=1488]} = 0.795$, $P_{MWT} = 4.95 \times 10^{-1}$). We, thus, infer that complex number may not be an evolutionary rate determinant of SET and NSET proteins.

We splitted our data set into (i) complex forming (complex number > 0) and (ii) noncomplex forming (complex number $= 0$) groups. Subsequent analysis demonstrates that SET and NSET proteins are under equal selective constraint (*d*N/*d*S) within complex forming group (fig. 2*a*) plausibly due to their comparable functional utility. However, the difference in *d*N (within complex difference: $P_{MWT} = 5.15 \times 10^{-1}$; within noncomplex difference: $P_{MWT} = 9.11 \times 10^{-6}$) and *d*N/*d*S between SET and NSET proteins persist in noncomplex forming group (fig. 2*a*), where almost equal proportions of SET (80.67%; 313/388) and NSET (79.17%; 1178/1488) proteins are present (SET vs. NSET: Z score $= 0.653$, $P = 5.14 \times 10^{-1}$). From this analysis, we conclude that the observed difference in evolutionary rates between SET and NSET proteins may be guided by protein noncomplex forming ability irrespective of their sample size bias.

A recent article has emphasized that some protein complexes those are very small in size may prompt high hit rate variations (Bin Goh et al. 2015). To reconfirm our result, we considered protein complexes with size ≥ 5 (Bin Goh et al. 2015). Further analyses revealed an identical trend i.e. the *d*N (within complex difference: $P_{MWT} = 6.11 \times 10^{-1}$; within noncomplex difference: $P_{MWT} = 6.83 \times 10^{-7}$) or *d*N/*d*S difference between SET and NSET proteins remains insignificant only within the complex forming group (fig. 2*b*). This study affirms that noncomplex forming ability acts as an important evolutionary rate determinant of SET and NSET proteins, independent of protein complex sizes.

Gene Expression Level and Tissue Expression Breadth Are Inadequate to Explain Evolutionary Rate Variation between SET and NSET Proteins

Proteins belonging to multiple complexes (higher protein complex association number) are comparatively highly expressed than proteins those have subunits of fewer protein complexes or those do not form any complex (Chakraborty and Ghosh

2013). In addition, gene expression level is known as primary determinant of evolutionary rate because highly expressed genes reduce mistranslation induced misfolding cost by experiencing higher selective constraints (Drummond et al. 2005, 2006; Pal et al. 2006; Wolf et al. 2009b; Zhang and Yang 2015). Therefore, it is necessary to determine the influence of gene expression levels on the rates of evolution of SET/NSET proteins to draw any further inference. We did not find any significant correlation between gene expression level and *d*N; although a weak negative correlation exists between gene expression level and *d*N/*d*S ($_{\text{Expression level}}\rho^{dN} = -0.024$, $P = 3.60 \times 10^{-1}$, $n = 1411$; $_{\text{Expression level}}\rho^{dN/dS} = -0.077$, $P = 4.00 \times 10^{-3}$, $n = 1411$). In contrast, we obtained a lower gene expression level in SET group of genes compared to the NSET group (average gene expression $\text{level}_{\text{SET}\ [n=385]} = 48.740$, average gene expression $\text{level}_{\text{NSET}\ [n=1491]} = 92.624$, $P_{\text{MWT}} = 6.22 \times 10^{-5}$). We also estimated expression levels of SET and NSET genes separately within complex and noncomplex forming groups. SET genes are found to be lowly expressed than NSET genes, whereas difference in *d*N or *d*N/*d*S between SET and NSET does not exist (*complex forming group*: average gene expression $\text{level}_{\text{SET}\ [n=55]}$ vs. $_{\text{NSET}\ [n=268]} = 44.411$ vs. 82.126, $P_{\text{MWT}} = 3.00 \times 10^{-3}$; *noncomplex forming group*: average gene expression level $_{\text{SET}\ [n=194]}$ vs. $_{\text{NSET}\ [n=894]} = 49.967$ vs. 95.771, $P_{\text{MWT}} = 3.00 \times 10^{-3}$). This finding highlights the fact that gene expression level does not have any influence in explain the evolutionary rate variations between SET and NSET group.

It has been argued that tissue expression breadth has more influence on protein evolutionary rate than gene expression level and such tissue-specific genes evolve faster than broadly expressed housekeeping genes (Duret and Mouchiroud 2000; Zhang and Li 2004; Park and Choi 2010). In this context, it is expected that conserved SET proteins are comparatively broadly expressed than the NSET proteins. Surprisingly, we noticed a lower tissue expression breadth of SET group of proteins (average expression breadth ≈ 8, $n = 249$) compared to NSET (average expression breadth ≈ 12, $n = 1162$, and P_{MWT} for SET vs. NSET tissue expression breadth $= 1.80 \times 10^{-12}$) proteins, implies that side effect associated genes tend to be expressed in limited number of tissues in spite of their lower evolutionary rates. Similar trends were observed in complex (SET vs. NSET: $P_{\text{MWT}} = 8.48 \times 10^{-5}$) and noncomplex (SET vs. NSET: $P_{\text{MWT}} = 2.40 \times 10^{-9}$) forming groups.

Protein Age, Functionality, and Killer Druggability Simultaneously Influence the Conservation of SET Proteins over NSET Group

The time of origin of a protein strongly influences its rate of molecular evolution since protein cites some "memory" of its age (Vishnoi et al. 2010). It has been found that younger proteins evolve faster than older proteins (Alba and Castresana 2005; Wolf et al. 2009b). Therefore, it is expected that SET proteins have to be comparatively older than NSET proteins. Using categorical gene age data of Domazet-Lošo and Tautz (2008), we found that younger proteins are significantly overrepresented in SET group than random expectation (odds ratio = 1.325, Z score = 2.233, $P < 2.55 \times 10^{-2}$). Similar trend was observed when we carried another study with publicly available tool: "ProteinHistorian" (Capra et al. 2012) (average $\text{age}_{\text{SET}\ [n=382]} \approx 949$ Ma, average $\text{age}_{\text{NSET}\ [n=1454]} \approx 1172$ Ma, $P_{\text{MWT}} = 4.00 \times 10^{-3}$), although there exists an inverse correlation between protein age and evolutionary rates ($_{\text{Age of proteins}}\rho^{dN} = -0.364$, $P = 1.00 \times 10^{-6}$, $n = 1836$). In view of the fact that tissue specific proteins are comparatively younger than housekeeping proteins (Duret and Mouchiroud 2000; Zhang and Li 2004; Wolf et al. 2009a, 2009b; Park and Choi 2010), we performed a correlation analysis between protein age and tissue expression breadth for our data set. A significant positive association between the variables ($_{\text{Age of proteins}}\rho^{\text{Expression breadth}} = 0.308$, $P = 1.00 \times 10^{-6}$, $n = 1409$) strengthens the fact that lower tissue expression breadth of SET proteins are relatively younger than the NSET proteins. A comparison of protein age between SET and NSET proteins indicates a similar mode of age difference only within noncomplex forming group (table 1). All these results imply that the time of origin may have some influence to explain the evolutionary rate heterogeneity between SET and NSET proteins.

In an attempt to find the causes of target drug side effects, a recent article has highlighted that many of the drug side-effects stem from disruption of important biological processes/pathways (Liu and Altman 2015). Therefore, one plausible reason of detecting variation in *d*N or *d*N/*d*S between SET and NSET proteins could be their differences in functionalities which strongly influence protein age as well as the rate of molecular evolution (Kimura and Ohta 1974; Wolf et al. 2009b; Ohta 2011). Concentrating on biological functions (see "Materials and Methods" section) of proteins, we perceived that SET proteins are more multifunctional than NSET proteins (average $\text{pleiotropic index}_{\text{SET}\ [n=387]} = 20.052$, average $\text{pleiotropic index}_{\text{NSET}\ [n=1480]} = 15.724$, $P_{\text{MWT}} = 3.57 \times 10^{-7}$) where protein multifunctionality boosts its sequence conservation rates ($_{\text{pleiotropic index}}\rho^{dN} = -0.157$, $_{\text{pleiotropic index}}\rho^{dN/dS} = -0.155$, $P = 1.00 \times 10^{-6}$, $n = 1867$). Here pleiotropic index has been used to measure the multifunctionality of a protein. Based on our result, we infer that multifunctionality could be a promising determinant of evolutionary rates of SET/NSET proteins. When we performed correlation analyses between protein biological functions, their age and killer druggability, we found that killer druggability increases with protein multifunctionality ($_{\text{pleiotropic index}}\rho^{\text{killer druggability/target}} = 0.168$, $P = 1.00 \times 10^{-6}$, $n = 1867$) and reduces with their age (table 2). Our result affirms that more multifunctional younger drug targets are candidates for acquiring killing side effects (tables 2 and 3). Furthermore, to interpret

Table 1
Age Distribution of SET/NSET Proteins within Complex Forming and Noncomplex Forming Groups

Resource	Class	Proteins Those Form Complex	*P*-value	Noncomplex Forming proteins	*P* value
ProteinHistorian server	SET	≈885 Ma	1.82×10^{-1}	≈965 Ma	$1.10 \times 10^{-2**}$
	NSET	≈1183 Ma		≈1169 Ma	
Supplementary Material	Old	SET: 14.69%	5.50×10^{-1}	SET: 57.47%	$1.10 \times 10^{-3**}$
		NSET: 15.92%		NSET: 63.31%	
	New	SET: 4.64%	6.38×10^{-1}	SET: 21.13%	$3.49 \times 10^{-2**}$
		NSET: 4.10%		NSET: 13.71%	

NOTE.—Significant difference **($P < 0.05$) between pairs are highlighted in bold. Mann–Whitney U test was used to demonstrate the difference in protein numerical age data taken from ProteinHistorian.

Table 2
Killer Druggability Per Target Proteins Using Age Data from Two Different Resources

Gene Class	Killer Druggability/Target (Considering age data of Domazet-Lošo and Tautz)	Killer Druggability/Target (Considering Protein Historian server)
New	0.715	0.421
Old	0.210	0.248
P_{MWT} value	$1.30 \times 10^{-4**}$	$3.60 \times 10^{-2**}$

NOTE.—Significant difference **($P < 0.05$) between groups are highlighted in bold. Mann–Whitney U test was used to exhibit the differences between groups. In case of age data from ProteinHistorian, we considered proteins with age ≤ 500 Ma as "new" and proteins with age > 500 Ma as "old."

the impact of killer druggability on related biological parameters, we removed all SET proteins with killer druggability > 0 [55.41%; 215/388] from our entire data set. We, thereby, noticed that the differences in protein age (considering numerical age data: SET = 1080 Mya, NSET = 1172 Mya, $P_{MWT} = 3.33 \times 10^{-1}$; considering categorical age data: SET *vs.* NSET old: 76.88% vs.79.23%, SET vs. NSET new: 22.54% vs.17.81%, $P > 5.00 \times 10^{-2}$ in both cases) and protein multifunctionality ($P_{MWT} = 3.33 \times 10^{-1}$) no longer hold between SET and NSET groups. A consistent trend was obtained for noncomplex forming proteins where the *d*N difference exists for SET vs. NSET proteins (data not shown). From these observations, we conclude that higher killer druggability and multifunctionality have the potential to explain the lower evolutionary rates of recently emerged SET proteins over the NSET proteins. However, after removal of SET proteins with killer druggability > 0, the average *d*N of SET proteins elevates from 0.062 to 0.066, although the between group (SET vs. NSET) *d*N difference ($P_{MWT} = 1.90 \times 10^{-2}$) still persists in rest of the data set. These findings corroborate that target killer druggability strongly influences protein age and functionality in the evolutionary landscape and partially explains the conservation of SET proteins over NSET group.

Impact of Protein Age and Sub-Cellular Localization for the Evolutionary Rate Difference of SET vs. NSET Proteins

A significant evolutionary rate conservation of SET ($n = 173$) proteins with respect to NSET ($n = 1488$) group in rest of the data set (i.e. excluding drug targets with killing side effect > 0) indicates that some other factor(s) also influence the evolutionary rates of SET (with no killing side effect) and NSET proteins. A significant difference (Wilcoxon rank sum test: $P < 2.20 \times 10^{-16}$) in *d*N between randomized (10,000 simulated replicates) SET and NSET groups reassures that our result is insensitive to the nonuniform sample sizes of SET and NSET proteins. To determine the impact of gene age on *d*N in the data set of proteins exhibiting no killing side effects, we performed a correlation analysis between the two factors. A better correlation than whole data set (considering numerical age data: $\rho^{dN}_{\text{Age of proteins}} = -0.390$, $P = 1.00 \times 10^{-6}$, $n = 1626$) was observed by this way. When we compared the evolutionary rates of SET and NSET proteins separately within "old" and "new" age groups; the evolutionary rate variation was found only within new age group (considering both the age data resources; fig. 3). So as the case for noncomplex forming young proteins (Data set of numerical age data; categorical age data respectively: dN_{SET} vs. dN_{NSET} $P_{MWT} = 1.00 \times 10^{-3}$; dN_{SET} vs. dN_{NSET} $P_{MWT} = 6.00 \times 10^{-3}$). Since because, the sample size distributions of SET and NSET proteins within new age class are statistically insignificant ($Z = 1.524$; $P = 1.27 \times 10^{-1}$); our observation confirms that protein age is a significant feature that drives SET/NSET protein evolutionary rates in new age group irrespective of sampling distribution.

One possible reason for obtaining lower *d*N of SET proteins compared to the NSET proteins within the new age group could be their subcellular localization that largely imposes selective constraints on protein sequences in an order of intracellular *d*N/*d*S < membrane *d*N/*d*S < extracellular *d*N/*d*S (Julenius and Pedersen 2006; Wang et al. 2013b). In this regard, it is already published that membrane-embedded proteins are of principal therapeutic interest due to the ease of drug binding to membrane proteins (Wang et al. 2013b). Proceeding further, the proportions of membrane targets are found to be considerably higher in new SET proteins compared to the NSET proteins (considering ProteinHistorian (Capra et al. 2012) age data set: SET vs. NSET 18.02% vs. 6.95%, *Z* score = 5.030, $P < 1.00 \times 10^{-4}$; considering Domazet-Lošo and Tautz (2008) age data set: SET vs. NSET 10.46% vs. 4.22%, *Z* score = 3.588, $P = 3.00 \times 10^{-4}$).

Table 3

Comparison of Features between Drug Targets with Killing Side Effect(s) and with Nonkilling/No Side Effect(s)

Serial no.	Groups	Evolutionary Rates (dN)	Pleiotropic Index	Tissue Expression Breadth
1	Drug targeted proteins with killer druggability > 0	0.058	24.195	7
2	Drug targeted proteins with killer druggability = 0	00.081	15.636	12
P_{MWT} value		$3.48 \times 10^{-4**}$	$5.94 \times 10^{-13**}$	$2.09 \times 10^{-8**}$

NOTE.—Mann–Whitney U test was used to demonstrate the significant **($P_{MWT} < 0.05$) differences between groups. Bold data represent significant difference.

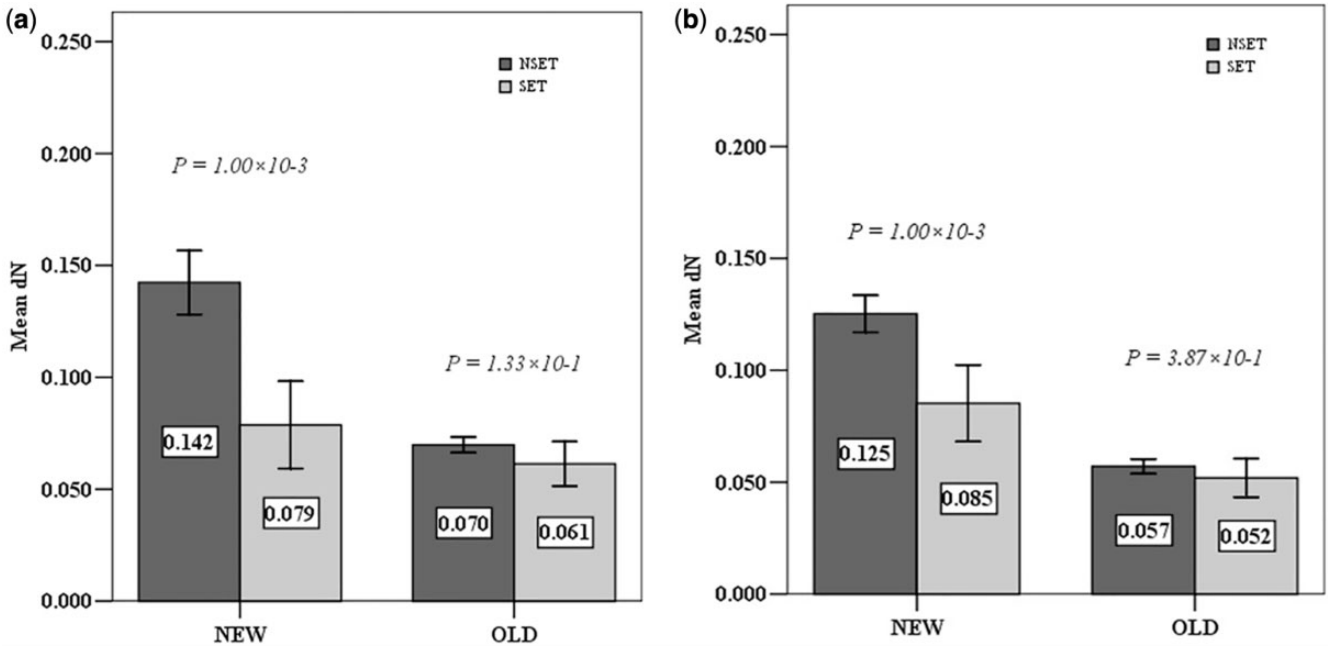

FIG. 3.—Comparisons of evolutionary rates (*d*N) of SET vs. NSET proteins within different age groups. In the figure, (*a*) the categorical age data are provided by Domazet-Lošo and Tautz (2008); (*b*) the numerical protein age data were obtained from ProteinHistorian (Capra et al. 2012). For numerical data, we considered proteins with age ≤ 500 Ma as "new" and rest as "old" proteins. The plots showing the importance of young/new gene age in the disparity of evolutionary rates of SET and NSET proteins. Error bars represent 95% confidence interval.

However, no such result was obtained for intracellular or extracellular target proteins (data not shown). When we estimated evolutionary rates of SET and NSET proteins within the new age class, *d*N of membrane localized SET proteins are found to be substantially lower than NSET proteins (average $d\mathrm{N}_{\mathrm{SET\ \textit{vs.}\ NSET}} = 0.0.086$ vs. 0.162, average $d\mathrm{N}_{\mathrm{SET\ \textit{vs.}\ NSET}} = 0.093$ vs. 0.182; considering numerical and categorical age data respectively; $P_{\mathrm{MWT\ [SET\ \textit{vs.}\ NSET]}} < 7.00 \times 10^{-3}$ in both cases). Our results do not alter considering noncomplex new gene class (table 4). To understand why membrane located SET proteins evolve slower than NSET proteins, we considered transmembrane proteins (proteins containing transmembrane helices) because such proteins often demonstrate tissue restricted expression and evolve slowly due to their elevated buried residue composition (Oberai et al. 2009; Ramskold et al. 2009; Begum and Ghosh 2014). Thus, we noticed that only within new membrane protein class, transmembrane helices share a significant negative correlation to protein evolutionary rates (considering ProteinHistorian (Capra et al. 2012) data: $\rho^{d\mathrm{N}}_{\mathrm{Transmembrane\ helices/protein}} = -0.367$, $P = 1.54 \times 10^{-5}$, $n = 132$; considering Domazet-Lošo and Tautz (2008) age data: $\rho^{d\mathrm{N}}_{\mathrm{Transmembrane\ helices/protein}} = -0.517$, $P = 1.04 \times 10^{-6}$, $n = 79$). However, such correlation does not persist considering the whole data set ($\rho^{d\mathrm{N}}_{\mathrm{Transmembrane\ helices/protein}} = 0.032$, $P = 1.62 \times 10^{-1}$, $n = 1866$) or considering data set after eliminating SET proteins with killing side effect(s) ($\rho^{d\mathrm{N}}_{\mathrm{Transmembrane\ helices/protein}} = 0.0.075$, $P = 2.00 \times 10^{-3}$, $n = 1652$). It indicates that relatively higher number of transmembrane helices may be present in SET proteins within new membrane class in support of their lower evolutionary rates and this is the case in reality [considering ProteinHistorian (Capra et al. 2012): average $\mathrm{transmembrane\ helices}_{\mathrm{SET}\ [n=31]} = 4.742$, average $\mathrm{transmembrane\ helices}_{\mathrm{NSET}\ [n=101]} = 3.139$, $P_{\mathrm{MWT}} = 3.00 \times 10^{-3}$; considering Domazet-Lošo and Tautz (2008) age data: average $\mathrm{transmembrane\ helices}_{\mathrm{SET}\ [n=18]} = 5.444$, average $\mathrm{transmembrane\ helices}_{\mathrm{NSET}\ [n=61]} = 3.361$, $P_{\mathrm{MWT}} = 1.50 \times 10^{-3}$]. These results support the fact that simultaneous effect of protein age

Table 4

Distribution of SET/NSET Membrane Proteins within Complex Forming and Noncomplex Forming New Age Class Considering Targets with Killer Druggability = 0

Resource	Proteins Those Form Complex	*Z* score, *P* value	Noncomplex Forming Proteins	*Z* score, *P* value
Age data provided by Domazet-Lošo and Tautz	SET: 1.16%	Z = 0.679,	SET: 9.30%	***Z* = 3.580,**
	NSET: 0.69%	$P = 4.97 \times 10^{-1}$	NSET: 3.54%	$P = 3.00 \times 10^{-4}$**
From age data of ProteinHistorian server	SET: 1.74%	*Z* = 0.743,	SET: 16.28%	***Z* = 5.088**
	NSET: 1.10%	$P = 4.57 \times 10^{-1}$	NSET: 5.84%	$P < 1.00 \times 10^{-4}$**

NOTE.—Significant difference **($P < 0.05$) between pairs and their corresponding *Z* scores are highlighted in bold.

(young) and membrane localization of target are the driving force for distinct evolutionary rates of SET and NSET proteins.

Discussion

In the present communication, we analyzed and compared the evolutionary rates of human drug side effect associated targets (SET) with respect to the targets having no available side effect data in the public repositories (NSET) to understand the biological features of targets contributing to drug side effect(s) in the evolutionary landscape. We observed that SET proteins evolve substantially slower than NSET proteins. To clarify the underlying reason of such fluctuations in evolutionary rates between SET and NSET proteins, we considered factors (gene expression levels, tissue expression breadth, target essentiality, total druggability, killer druggability, essential druggability of a target, complex association number, protein complex size, pleiotropic index, age of a protein, and protein subcellular localization) (Zhang and Li 2004; Alba and Castresana 2005; Drummond et al. 2006; He and Zhang 2006; Liao and Zhang 2006; Pal et al. 2006; Park and Choi 2010; Panda et al. 2012; Chakraborty and Ghosh 2013; Wang et al. 2013a, 2013b; Begum and Ghosh 2014; Zhang and Yang 2015) those either have known correlations or we conjectured to have relation with protein evolutionary rates. Thereby, we identified that noncomplex forming nature, time of origin of proteins, killer druggability of targets, multifunctionality, and subcellular localization play key roles to dictate the divergence in evolutionary rates between SET and NSET proteins among all other features.

To estimate protein evolutionary rates, we considered most common human-mouse orthologous pair due to the functional conservation between the two species in normal and pathological conditions (Liao and Zhang 2006; Cai et al. 2009; Gharib and Robinson-Rechavi 2011). However, it is now established that a shorter generation time in the murid lineage may possess problems to estimate *d*N/*d*S (Hoffman and Birney 2007). Moreover, it is well known that in human and mouse, *d*S is not constant and depends upon GC content, gene recombination rates, codon usage bias and several other parameters (Castresana 2002; Williams and Hurst 2002; Hoffman and Birney 2007). Therefore, to avoid any confusion, a possible alternative is to consider another mammal to estimate *d*N, *d*S, and *d*N/*d*S. To serve the purpose, we used human-cow orthologous pair to compute evolutionary rates of SET ($n = 375$) vs. NSET ($n = 1424$) proteins. When we replicated the whole analyses between SET vs. NSET proteins using human-cow orthologous pair, we noticed that the results remain absolutely similar to what we found using human-mouse orthologs (supplementary tables S1 and S2, Supplementary Material online). Even, in the new data set, the *d*S value between two groups ($P_{MWT} = 2.15 \times 10^{-1}$) does not vary. This result suggests that the difference in *d*N/*d*S in SET vs. NSET proteins strongly depend on the nonsynonymous evolutionary distance, *d*N (supplementary table S1, Supplementary Material online) and our result is free from any bias due to orthology selection.

Previous study has affirmed the fact that the number of essential proteins that bind to a drug is a crucial factor of drug side effects rather than the total number of drugs that bind to a target (Wang et al. 2013a). Regarding total druggability of a target, we draw a similar conclusion i.e. the evolutionary rates vary between SET and NSET groups independent of the total number of drug(s) against a target. Contemplating on essential targets, we noticed that the rates of molecular evolution do not vary between SET and NSET proteins, even though essential proteins are significantly overrepresented in SET group. Since, essential proteins are multifunctional in nature (Chakraborty and Ghosh 2013; Acharya et al. 2015); one plausible reason for not obtaining such evolutionary rate difference could be due to the equal number of biological process involvement ($P_{MWT} = 1.11 \times 10^{-1}$) of SET and NSET proteins in the essential group. Introducing essential druggability of a target, we detected slow evolutionary rates for proteins with higher essential druggability compared to proteins with lower essential druggability. Again, proteins are pooled into three bins according to their essential target druggability (bin 1 [number: 1–2]: Low, bin 2 [number: 3–4]: Medium, bin 3 [number ≥ 5]: High). No difference in *d*N even where variation in essential druggability exists (in bin 1 and bin 3; see "Results" section) suggests that the evolutionary rate heterogeneity between SET and NSET proteins does not solely

rely on target essential druggability. In this context, it is well known that disruption/alteration of a biological function often lead to the progression of human diseases which are overrepresented in targets with elevated drug side effects (Janjic and Przulj 2012; Begum and Ghosh 2014). When we estimated protein functionality in three bins (showing no difference in *d*N), we did not find any difference in protein functionality between SET and NSET groups within those bins (bin 1: $P_{MWT} = 7.64 \times 10^{-1}$, bin 2: $P_{MWT} = 8.23 \times 10^{-1}$, bin 3: $P_{MWT} = 5.70 \times 10^{-1}$). This study indicates that protein functionality, rather than protein essentiality or essential target druggability, is one of the key evolutionary rate determinants of SET and NSET proteins in the evolutionary landscape.

In our study, we found that within noncomplex forming group, SET and NSET proteins are under distinct selective constraints. However, within complex forming group, such evolutionary rate difference does not persist. One plausible reason might be essential proteins are largely involved in protein complexes to execute important biological functions (Chakraborty et al. 2010; Ruepp et al. 2010). Since, we did not notice any difference in the evolutionary rates and functionalities of SET and NSET proteins within the essential group; it could be a reason for not obtaining complex number variation between SET and NSET proteins. In this context, we know that disturbed essential genes may impact all the genes in an organism by causing complete cell death and therefore, majority of human disease genes are nonessential in nature (Goh et al. 2007; Han et al. 2013). We have previously seen that human disease genes are largely associated with drug side effects (Begum and Ghosh 2014). Therefore, in agreement to Goh et al. (2007), we here confirm that nonessential noncomplex forming proteins are the major candidates of drug side effects and influence the evolutionary trajectory of SET and NSET proteins.

Disruption of cell division and expression regulation can also lead to drug side effects in many cases (Liu and Altman 2015). Thus, indispensable factors of protein evolutionary rates are expression level of a gene as well as tissue expression breadth (Drummond et al. 2006; Park and Choi 2010). However, in contrast to our expectation, in both cases we noticed that conserved SET proteins are lowly expressed as well as tissue restricted in nature compared to NSET proteins. It may be due to the fact that restricted expression of SET proteins (table 3) may impede the spread of drug side effects (especially killing side effects) along the links of the interactome network to maintain the activity of their interaction partners and undergo reasonable purifying selection to keep the side effects restricted in a limited number of tissues. However, from evolutionary perspective, our results suggest that gene expression levels or tissue expression breadth alone is inadequate to explain the evolutionary rate difference between SET and NSET proteins.

Previous communications have claimed that phylogenetic age of a protein strongly influences the rates of protein evolution (Alba and Castresana 2005; Wolf et al. 2009b). Even though, SET proteins are highly conserved in nature, such proteins are found to be substantially younger than NSET proteins. Contrary to the previous findings i.e. ancient/old proteins are more multifunctional (Wolf et al. 2009b); we obtained a higher multifunctionality of druggable young proteins than ancient proteins. It is also found that such multifunctional young SET proteins are candidates those acquire undesirable killing side effects. Thus, higher killer druggability and multifunctionality supports the lower evolutionary rates of SET proteins over the NSET proteins. However, killer druggability cannot be a sole attribute to explain the discrepancy in evolutionary rates between SET and NSET proteins, since it is a characteristic of SET proteins. For better understanding, SET proteins with killing side effects were eliminated from our data set. We, thereby, found that membrane-embedded young transmembrane proteins significantly explain the evolutionary rate diversity between SET and NSET proteins.

Potential Caveat

The purpose of our study is to detect genomic and structural evolutionary features those can help to capture the general properties of side effect associated targets and in turn, may help to predict SET or NSET proteins based on the characteristic features employed in our analyses. Experimental validation of our outcome is time-consuming and therefore, the scope is quite limited. One potential caveat of our study is the incomplete set of experimental side effect data of SIDER v.2 (Kuhn et al. 2013) database which plausibly may introduce some bias in our results. To address the problem, we thought of revalidating our analyses with some experimental/manually curated side-effect databases of human. However, due to lack of experimental side-effect database apart from SIDER (Kuhn et al. 2013), we focused to use advanced version of SIDER (v.4.1) database to revalidate our analyses. In contrast to SIDER v.2, we identified 48 proteins whose annotations differ from the previous version (i.e. SET in the current data set which used to be NSET in SIDER2 and vice versa) and the SET protein number increased from 388 to 414 using the newer version (v.4.1) SIDER (supplementary table S2, Supplementary Material Online). Replicating the whole analyses with the newer data set of SET ($n = 414$) and NSET ($n = 1462$) proteins reflected similar trends of results (data not shown) as previously obtained, although the magnitudes differed. Such observation reconfirms the fact that our evolutionary rate parameters can be used further to predict unknown SET/NSET proteins in human.

SVM Training and Prediction of SET/NSET Proteins

To understand the influences of all the 13 evolutionary rate attributes (*d*N, *d*N/*d*S, gene expression level, tissue expression breadth, protein age, complex/noncomplex forming ability, functionality, subcellular localization, number of

Table 5
Efficiency Evaluation of Our Optimized SVM Model

Test Set	MCC	Sensitivity	Specificity	PPV	NPV	*F*-measure
Our data set using categorical gene age data[a]	0.4810	29.26%	99.74%	94.94%	85.30%	44.57%
Our data set using numerical gene age data[a]	0.5056	31.47%	99.75%	97.02%	85.11%	47.39%

[a]Performance measures were averaged over 10 randomized test sets to obtain a single value.

transmembrane helices, target total druggability, target essentiality, target essential druggability, and target killer druggability) those showing significant difference between SET and NSET proteins, we trained two mathematical SVM models (using gene age data from two different resources—model 1 incorporates categorical gene age data of Domazet-Lošo and Tautz (2008) and model 2 uses numerical gene age data of ProteinHistorian server (Capra et al. 2012)) with these parameters and optimized the kernel parameters C and γ to maximize the prediction accuracy as well as the MCC value of our model. Since, SVM uses numerical values of predictors; we assigned +1 and −1 value to the categorical predictors before scaling and training the model. A predictor which is unable to predict one class (small TP or TN) adequately is undoubtedly biased and less informative, and in such cases MCC can critically assess the bias of a result by acquiring a negative or nearly zero value (Dey et al. 2012; Kisslov et al. 2014). We found that the optimal values of C and γ for our SVM models are 32 and 0.0078125, respectively. When we trained our models using these kernel parameters and applied on test sets to predict SET/NSET proteins, we observed that the success rate (~86%) and all other performance measures for both the models (table 5) are quite impressive. When we excluded *d*N/*d*S attribute from our analyses, the optimal values of C and γ become 1 and 0.0078125, respectively. In that case, the success rates decrease by ~5% (accuracy of model 1 and model 2 are 80.92% and 81.29%, respectively) for both of our SVM models. Furthermore, this study reveals that using *d*N and *d*N/*d*S altogether have the potential to increase the prediction accuracy of our models.

In addition, we optimized the kernel parameters C and γ of our SVM models trained with all 13 evolutionary rate parameters using 5-fold cross-validation. In such a case, our data set was broken into five subsets: (i) four subsets are used to train the classifier and (ii) rest of the subset is used as a test set to evaluate prediction accuracy using the trained model (Dey et al. 2012). The process is replicated five times so that all subsets could be used for training and testing (Dey et al. 2012). The prediction accuracy achieved by this process remains almost similar to what we found before (cross validation accuracy of model 1 is 86.03% and the accuracy of model 2 is 85.98%).

DR. PRODIS, a recent open access web server, has been designed to identify side effect associated proteins in the human genome with a reported precision of ~57% and a recall of ~24% (Zhou et al. 2015). We mapped the annotations (SET/NSET) of DR. PRODIS to our data set of 1876 proteins to compare the performances of our SVM models with respect to DR. PRODIS web server. Drug side effect associated targets provided by the SIDER v.2 database (Kuhn et al. 2013) was used to benchmark the approach. We, thereby, noticed that our optimized SVM models outperform DR. PRODIS server in terms of MCC, accuracy, specificity, PPV, NPV, and F-measure, respectively, except for sensitivity measure. In contrast to the performance of our models as summarized in table 5, the MCC, accuracy, specificity, PPV, NPV, and F-measure of DR. PRODIS for the 1876 human proteins are −0.1056, 49.09%, 53.02%, 15.88%, 75.50%, and 21.66%, respectively. However, the sensitivity for our models are in the range of ~29% to 31.5% which is almost comparable to the sensitivity of DR. PRODIS (i.e. 34.02%) for our data set. Based on the MCC value, we infer that our models predict better than the previous existing prediction server. Even when we used human-cow orthologs, we achieved a prediction success rate of ~80% (for model 1 and model 2 prediction accuracy are 79.58% and 80.74%, respectively) using all 13 evolutionary rate attributes. Comparison of our prediction results in terms of MCC (model 1 and model 2 MCC are 0.2203 and 0.2259 respectively) to the results of DR. PRODIS further strengthens our conclusion that our prediction models work better than the existing prediction model irrespective of our orthology selection.

Concluding Remarks

To conclude, our systematic study reports the key factors leading to drug side effects in the framework of protein evolution. It is clear from this communication that the evolutionary origin, involvement in biological processes, noncomplex forming ability and localization of targets are the special features for predicting candidates with drug (killing) side effects. Our study has described the relative influences of killer druggability vs. essential druggability on protein evolutionary rates considering nonside effect associated (NSET) proteins as control. Furthermore, our results help to identify unknown SET/NSET proteins in human. Using all 13 evolutionary rate attributes incorporated in this study, our SVM models perform predictions that are much better than the existing prediction server. With the increasing coverage of the drug side effect

associated data in human, more interesting analyses can be performed to further dissect the properties of drug targets and the associated side effects and the accuracy of our prediction models can be further improved.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

Authors thankfully acknowledge the Department of Biotechnology, Government of India, for their financial support to Bioinformatics Centre, Tripura University. The authors especially thank Sushanta Deb, Mustafa Raza, Krishna Bhowmik, Jyotirmoy Das, and Manish Prakash Victor for their help.

Literature Cited

Acharya D, Ghosh T. 2016. Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. BMC Genomics 17:1.

Acharya D, Mukherjee D, Podder S, Ghosh T. 2015. Investigating different duplication pattern of essential genes in mouse and human. PLoS One 1010:e0120784.

Alba MM, Castresana J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. Mol Biol Evol. 22:598–606.

Begum T, Ghosh T. 2014. Elucidating the genotype-phenotype relationships and network perturbations of human shared and specific disease genes from an evolutionary perspective. Genome Biol Evol. 6:2741–2753.

Begum T, Ghosh TC. 2010. Understanding the effect of secondary structures and aggregation on human protein folding class evolution. J Mol Evol. 71:60–69.

Bin Goh W, Guo T, Aebersold R, Wong L. 2015. Quantitative proteomics signature profiling based on network contextualization. Biol Direct. 10(1):19.

Cai J, Borenstein E, Chen R, Petrov D. 2009. Similarly strong purifying selection acts on human disease genes of all evolutionary ages. Genome Biol Evol. 1:131–144.

Capra J, Williams A, Pollard K. 2012. ProteinHistorian: tools for the comparative analysis of eukaryote protein origin. PLoS Comput Biol. 8:e1002567.

Castresana J. 2002. Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. Nucleic Acids Res. 30:1751–1756.

Chakraborty S, Ghosh T. 2013. Evolutionary rate heterogeneity of core and attachment proteins in yeast protein complexes. Genome Biol Evol. 5:1366–1375.

Chakraborty S, Kahali B, Ghosh T. 2010. Protein complex forming ability is favored over the features of interacting partners in determining the evolutionary rates of proteins in the yeast protein-protein interaction networks. BMC Syst Biol. 4:1.

Chang C, Lin C. 2011. LIBSVM. A Library for Support Vector Machines. ACM Trans Intell Syst Technol (TIST). 2:27.

Dey S, Pal A, Guharoy M, Sonavane S, Chakrabarti P. 2012. Characterization and prediction of the binding site in DNA-binding proteins: improvement of accuracy by combining residue composition, evolutionary conservation and structural parameters. Nucleic Acids Res. 40:7150–7161.

Domazet-Loso T, Tautz D. 2008. An ancient evolutionary origin of genes associated with human genetic diseases. Mol Biol Evol. 25:2699–2707.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 102:14338–14343.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol. 23:327–337.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol. 17:68–74.

Flicek P, et al. 2013. Ensembl 2013. Nucleic Acids Res. 41:D48–D55.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. Science 296:750–752.

Georgi B, Voight B, Bucan M. 2013. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. PLoS Genet. 9:e1003484.

Gharib W, Robinson-Rechavi M. 2011. When orthologs diverge between human and mouse. Brief Bioinform. 12:436–441.

Goh KI, et al. 2007. The human disease network. Proc Natl Acad Sci U S A. 104:8685–8690.

Han H, Ohn J, Moon J, Kim J. 2013. Yin and Yang of disease genes and death genes between reciprocally scale-free biological networks. Nucleic Acids Res. 41:9209–9217.

He X, Zhang J. 2006. Toward a molecular understanding of pleiotropy. Genetics 173:1885–1891.

Hesterberg T, et al. 2010. Bootstrap methods and permutation tests. Chapter 16. In. Introduction to the practice of statistics. New York: W. H. Freeman. p. 11:16–57.

Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. Nature 411:1046–1049.

Hoffman M, Birney E. 2007. Estimating the neutral rate of nucleotide substitution using introns. Mol Biol Evol. 24:522–531.

Huang F, et al. 2002. Proteomics of Synechocystis sp strain PCC 6803—Identification of plasma membrane proteins. Mol Cell Proteomics 1:956–966.

Huang Y, et al. 2013. Recent adaptive events in human brain revealed by meta-analysis of positively selected genes. PLoS One 8:e61280.

Janjic V, Przulj N. 2012. Biological function through network topology: a survey of the human diseasome. Brief Funct Genomics 11:522–532.

Juan-Blanco T, Duran-Frigola M, Aloy P. 2015. IntSide: a web server for the chemical and biological examination of drug side effects. Bioinformatics 31:612–613.

Julenius K, Pedersen A. 2006. Protein evolution is faster outside the cell. Mol Biol Evol. 23:2039–2048.

Kimura M, Ohta T. 1974. On some principles governing molecular evolution. Proc Natl Acad Sci U S A. 71:2848–2852.

Kisslov I, Naamati A, Shakarchy N, Pines O. 2014. Dual-targeted proteins tend to be more evolutionarily conserved. Mol Biol Evol. 31:2770–2779.

Kuhn M, et al. 2013. Systematic identification of proteins that elicit drug side effects. Mol Syst Biol. 9:1611–1624.

Law V, et al. 2014. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 42:D1091–D1097.

Liao B, Zhang J. 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol. 23:530–540.

Liu T, Altman R. 2015. Relating essential proteins to drug side-effects using canonical component analysis: a structure-based approach. J Chem Inf Model. 55:1483–1494.

Mintseris J, Weng Z. 2005. Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci U S A. 102:10930–10935.

Nagaraj SH, Ingham A, Reverter A. 2010. The interplay between evolution, regulation and tissue specificity in the Human Hereditary Diseasome. BMC Genomics 11:S23.
Necsulea A, Semon M, Duret L, Hurst LD. 2009. Monoallelic expression and tissue specificity are associated with high crossover rates. Trends Genet. 25:519–522.
Oberai A, Joh NH, Pettit FK, Bowie JU. 2009. Structural imperatives impose diverse evolutionary constraints on helical membrane proteins. Proc Natl Acad Sci U S A. 106:17747–17750.
Ohta T. 2011. Near-Neutrality, Robustness, and Epigenetics. Genome Biol Evol. 3:1034–1038.
Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7:337–348.
Panda A, Begum T, Ghosh T. 2012. Insights into the evolutionary features of human neurodegenerative diseases. PLoS One 7:e48336.
Park S, Choi S. 2010. Expression breadth and expression abundance behave differently in correlations with evolutionary rates. BMC Evol Biol. 10:241.
Perez-Lopez A, et al. 2015. Targets of drugs are generally, and targets of drugs having side effects are specifically good spreaders of human interactome perturbations. Sci Rep. 5:10182.
Ramskold D, Wang E, Burge C, Sandberg R. 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol. 5:e1000598.
Ruepp A, et al. 2010. CORUM: the comprehensive resource of mammalian protein complexes-2009. Nucleic Acids Res. 38:D497–D501.
Tang CSM, Epstein RJ. 2007. A structural split in the human genome. PLoS One 2:e603.
Teichmann S. 2002. The constraints protein-protein interactions place on sequence divergence. J Mol Biol. 324:399–407.
Vishnoi A, Kryazhimskiy S, Bazykin G, Hannenhalli S, Plotkin J. 2010. Young proteins experience more variable selection pressures than old proteins. Genome Res. 20:1574–1581.
Wang D, Liu F, Wang L, Huang S, Yu J. 2011. Nonsynonymous substitution rate (Ka) is a relatively consistent parameter for defining fast-evolving and slow-evolving protein-coding genes. Biol Direct. 6:1.
Wang ET, et al. 2008. Alternative isoform regulation in human tissue transcriptomes. Nature 456:470–476.
Wang H, et al. 2009. A complex-based reconstruction of the *Saccharomyces cerevisiae* interactome. Mol Cell Proteomics 8:1361–1381.
Wang X, Thijssen B, Yu H. 2013a. Target essentiality and centrality characterize drug side effects. PLoS Comput Biol. 9:e1003119.
Wang X, Wang R, Zhang Y, Zhang H. 2013b. Evolutionary survey of druggable protein targets with respect to their subcellular localizations. Genome Biol Evol. 5:1291–1297.
White J, et al. 2013. Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. Cell 154:452–464.
Williams E, Hurst L. 2002. Is the synonymous substitution rate in mammals gene-specific? Mol Biol Evol. 19:1395–1398.
Wolf JBW, Kuenstner A, Nam K, Jakobsson M, Ellegren H. 2009a. Nonlinear dynamics of nonsynonymous (d(N)) and synonymous (d(S)) substitution rates affects inference of selection. Genome Biol Evol. 1:308–319.
Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009b. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proc Natl Acad Sci U S A. 106:7273–7280.
Zhang J, Yang J. 2015. Determinants of the rate of protein sequence evolution. Nat Rev Genet. 16:409–420.
Zhang L, Li W. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol. 21:236–239.
Zhou H, Gao M, Skolnick J. 2015. Comprehensive prediction of drug-protein interactions and side effects for the human proteome. Sci Rep. 5:11090.

Associate editor: Maria Costantini