Chapter 5

Towards a resolution on the ongoing controversies on the evolutionary rates differences between human disease and non-disease genes

“Evolution advances, not by a priori design, but by the selection of what works best out of whatever choices offer. We are the products of editing, rather than of authorship.”

George Wald
Human disease genes are widely analyzed to understand the disease prevention and treatment. It has been reported that disease genes evolve slower compared to the non-disease genes, though they are mostly tissue specific. However, tissue-specific genes evolve faster than the housekeeping genes. Thus, it remains elusive whether the nature of selective constraints differs between human disease and non-disease genes in housekeeping as well as in tissue-specific groups. In this study, we integrated human disease genes from OMIM, GAD, and HGMD databases and categorized all human genes into tissue-specific and housekeeping group based on RNA-Seq data. We noticed that in tissue-specific genes, disease genes evolve slower than that of non-disease genes. However, in housekeeping group, disease and non-disease genes were found to evolve at similar rates. Moreover, we established that in tissue-specific group, human disease genes are evolutionarily conserved due to their association in a large number protein complex and elevated gene expression levels. Further, we found that human disease genes are associated with conserve core biological processes. Although, gene expression breadth was found to be the most important factor, our regression analysis also suggested that the degree of association of a protein in protein complex followed by protein multifunctionality could independently modulate the evolutionary rate of human disease genes.

5.1. Introduction

In medical research, the preliminary aim is to provide the genetic basis of human diseases to improve the remedies of disease prevention and their treatment. In the last decade, researchers analyzed the genetic basis of the human diseases by exploring different aspects of disease genes such as their evolutionary pattern, functional property, gene essentiality, gene duplication, protein disorder content, protein-protein interaction network, etc. Evolutionary constraints on human disease genes could provide important clue on their phenotypic connectivity and therefore, a large number of studies have compared the evolutionary rates of human disease genes with that of non-disease genes. Initially, Smith and Eyre-Walker showed that human disease...
Towards a resolution...

genes evolve at a faster rate as compared to non-disease genes\textsuperscript{87}. However, in a subsequent study Huang and Winter found no significant difference in protein evolutionary rates between human disease and non-disease genes\textsuperscript{89}. Even an opposite trend, \textit{i.e.}, disease genes are evolutionarily conserved relative to non-disease genes was observed by López-Bigas and Ouzounis in their independent study\textsuperscript{88}. These conflicting results draw further attention to characterize the evolutionary forces operating on human disease genes. Parallel studies delineating the effects of gene expression pattern on protein evolutionary rate emphasized that evolutionary constraints of human proteins vary widely depending upon their gene expression breadth (number of tissues in which a gene is expressed). Housekeeping genes being shown to be evolutionarily conserved, whereas tissue-specific genes were found to be evolutionarily faster\textsuperscript{183,184}. In recent studies, researchers analyzed the evolutionary rates of different disease gene classes like monogenic, polygenic and neurodegenerative disease genes to unveil the signatures of molecular evolution in human disease genes\textsuperscript{138,176}. Interestingly, majority of these studies concluded that the disease genes are tissue-specific by nature\textsuperscript{87,185}, \textit{i.e.}, these genes are expressed in a narrower ranges of tissues. Simultaneously, studies on expression patterns and protein evolution concluded that the tissue-specific genes evolve rapidly\textsuperscript{183,184}. Therefore, it could be expected that the disease genes evolve fast, which was observed by Smith \textit{et al.}\textsuperscript{87}. However, the majority of these studies also reported that human disease genes evolve at a slower rate in spite of their tissue-specific nature\textsuperscript{88,138,185,186}. Thus, it raised the question, how disease genes remained conserved in spite of being expressed in the fewer number of tissues?

Moreover, most of the previous analyzes of gene expression in disease and non-disease genes were based on microarray experiments. However, statistical methods for analyzing microarray data are less capable of differentiating low gene expression pattern from experimental noises\textsuperscript{187}. Thus, there is a high possibility of error included in the previous studies when lowly expressed genes were considered. Therefore, misclassification may occur in distinguishing the tissue-specific and housekeeping genes. Interestingly, using EST and microarray dataset Zhu \textit{et al.} concluded that the information of the total number housekeeping genes was less documented than it is actually present\textsuperscript{188}. Using the next generation RNA-sequencing data Emig \textit{et al.} also estimated
higher proportion of housekeeping genes in their study compared to the microarray-based experiments $^{187}$. Therefore, in this study we used RNA-sequencing data to classify human genes in housekeeping and tissue-specific categories. Moreover, studies that dealt with the evolutionary rate of human disease genes were mainly concentrated on specific disease types and analyzed approximately 500 to 2000 human disease genes $^{87,138,173}$. Therefore, the apparent conflict in their results may be due to the inconsistency in the datasets. Notably, majority of those studies used OMIM database to collect human disease genes. However, Wang et al. $^{189}$ reported that the disease genes such as cancer and type 2 diabetes were underrepresented in the OMIM database. Thus, majority of the complex disease genes were not included in this database. Recently, to get a global view of human disease gene network Goh et al. $^{190}$ collected more than 12000 human disease genes from two different databases, viz., OMIM, and GAD. To establish a global trend, in this study, we took this approach and collected human disease genes from three different databases viz., OMIM, GAD, and HGMD databases.

A number of parameters were shown to dictate the evolutionary rate of human proteins $^{69,78,150}$. Among these possible determinants, gene expression levels were considered to have the strongest influence in constraining the rate of sequence evolution $^{58,69}$. In our previous study, we established that protein complex forming ability has substantial contribution in determining their evolutionary rates $^{32,33}$. In particular, we noticed that proteins participating in the large number of protein complexes are evolutionarily conserved. Although, it was noted that mutations in these proteins are susceptible to diseases $^{72}$; however, how protein complex forming nature dictates the evolutionary rate of human disease proteins still remain elusive. Recent progress in the field of evolutionary biology also emphasized that the rate of protein evolution can be constrained by their functional requirements $^{78}$. Along with protein multifunctionality, the involvement in core or regulatory processes of the proteins also shown to have a profound influence on their evolutionary rates $^{138}$. Therefore, in this study we considered the interplay of all these factors to analyze their relative contribution to human disease protein evolution in housekeeping and tissue-specific groups.

Finally, our results revealed that the human disease and non-disease genes are shaped by different evolutionary constraints in tissue-specific and housekeeping groups.
Here, we observed that disease genes evolve slower than non-disease genes in the tissue-specific groups. However, in housekeeping gene both disease and non-disease genes were found to evolve at a similar pace. Interestingly, our study revealed that disease genes are evolutionarily constrained from their higher gene expression level, higher protein complex association, and their higher protein multifunctionality in tissue-specific group. Independent influence of all these factors on the evolutionary rate of human disease proteins were confirmed from regression analysis. Based on results presented in this communication, here we proposed to consider the expression profile of the disease and non-disease genes for better understanding of evolutionary constraints operating on these genes.

5.2. **Materials and methods**

5.2.1. **Calculation of evolutionary rates**

To measure the evolutionary rates we calculated the average dN/dS ratio, which is the ratio of the number of non-synonymous substitutions per non-synonymous site (dN) to the number of synonymous substitutions per synonymous site (dS). We calculate the dN/dS ratio by comparing human gene sequences with one to one orthologous sequences in *Macaca mulatta, Rattus norvegicus, Pongo abelii, Mus musculus, Pan troglodytes, Bos Taurus, Nomascus leucogenys*, since they are mammalian species. We obtained whole coding sequences of all mammalian species from the Ensembl BioMart (version 74)\(^{95}\). Then the pair-wise alignment was performed using the ClustalW (version 2.0)\(^{98}\) for each set of orthologs gene pair. dN/dS values were calculated by Yang and Nielsen method\(^{165}\) using the PAML package (version 4)\(^{101}\). We took the average values of dN/dS for each human gene where at least three orthologous pair was present (*Additional_File_3: Average evolutionary rates for each human gene where at least three orthologous pair was present*). We get a total of 16515 human genes with at least 3 orthologs.

5.2.2. **Collection of human disease genes**

We collected disease genes from the Online Mendelian Inheritance in Man (OMIM; [http://www.ncbi.nlm.nih.gov/omim]\(^{131}\), the Genetic Association Database
Towards a resolution...

(GAD; http://geneticassociationdb.nih.gov/)\textsuperscript{132}, and the Human Gene Mutation Database (HGMD: http://www.biobase-international.com/product/hgmd)\textsuperscript{133}. In general, the OMIM database is focused on monogenic Mendelian diseases. However, the GAD and the HGMD are the general repositories of monogenic as well as complex disease genes. Thus, we compiled these three databases to collect human disease genes. We found 2608, 10724, 8329 disease genes in the OMIM, GAD and HGMD respectively. Compiling these three databases we found 11031 unique disease genes, rest of 5484 genes (out of 16515 human genes considered in this study) were denoted as non-disease genes (Additional_File_4: List of human disease and non-disease genes.).

5.2.3. Gene expression and expression breadth

Human gene expression data was collected from the RNA-Seq Atlas\textsuperscript{123}. At present, RNA-Seq Atlas holds gene expression profile of eleven normal human tissues. This database is based on next-generation sequencing technology. Therefore, RNA-Seq Atlas provides a much more comprehensive and unbiased view of gene expression compared to microarray-based approaches\textsuperscript{123}. We mapped 15144 genes with their expression levels (RPKM values) in our dataset. We measured the expression breadth (EB) of each gene as the number of tissues in which a gene is significantly over-expressed. A gene was considered to be over-expressed in a tissue, if it is expressed at least one standard deviation higher level than the mean of its expression across all of the eleven tissues\textsuperscript{191}. Disease and non-disease genes were classified into housekeeping (HK) and tissue-specific (TS) groups according to their expression breadths. Genes over-expressed in all eleven tissues were considered to be HK genes; whereas, genes over-expressed in one or two tissues were defined as TS genes. Finally, we got 3086 HK and 4236 TS genes in our dataset.

5.2.4. Protein multifunctionality

For each human disease and non-disease genes, the corresponding child term of Gene Ontology Biological Processes (GO-BPs) were retrieved from Ensembl BioMart
Towards a resolution...

We calculated the number of such unique GO-BP terms annotated to each protein and denoted that as its multifunctionality\textsuperscript{33,77,78}. In our dataset, we mapped 13593 genes with their functional information. We classified all human genes into high multifunctional (HF) and less multifunctional (LF) groups on the basis of their multifunctionality. Finally, we got 6792 HF genes with multifunctionality is at least five and rest of the 6801 genes were denoted as LF.

We also classified proteins function into two classes, viz., conserved biological processes and other biological processes according to the classification of Podder \textit{et al.}\textsuperscript{138}. They considered catalytic activity, macromolecule metabolism, biosynthesis protein metabolism, primary metabolism, mRNA splicing, mRNA processing, ATP and RNA binding, transporter activity, carrier activity, ion transport activity, protein transport, intracellular transport, protein localization, structural constituent of ribosome, structural molecule activity, nucleosome assembly, organelle organization and biogenesis, chromatin assembly or disassembly, DNA packaging as conserved biological processes. We found 6169 genes involved in conserved biological processes and rest of 7424 genes were found to be involved in other biological processes.

\textbf{5.2.5. Protein complex information}

Information about human protein-complex assemblies were collected from two major databases, CORUM (http://mips.helmholtz-muenchen.de/genre/proj/corum/index.html)\textsuperscript{118} and CHPC2012 (http://www1.i2r.a-star.edu.sg/~xlli/CHPC2012/CHPC2012.htm) and analyzed separately\textsuperscript{119}. The CORUM database provides information about experimentally characterized protein-complexes\textsuperscript{118}. However, the CHPC2012 database stores non-redundant high quality predicted protein-complex data\textsuperscript{119}. For each protein we calculated the protein-complex association number, \textit{i.e.}, number of protein-complex assemblies in which a protein subunit belongs and denoted that as “protein-complex number”\textsuperscript{32,162}. We found 2304, 2706 protein complex data in our dataset from CORUM and CHPC2012 respectively.
5.2.6. Software

We used SPSS (version 20.0), R language and environment and XLSTAT for all statistical calculations. In all statistical analysis, we used 95% level of confidence as a measure of significance. In-house Perl programs were used to prepare datasets.

5.3. Results & Discussion

5.3.1. Tissue-specific disease genes are evolutionarily conserved

In this study, we compared the average evolutionary rates (dN/dS) of 11031 human disease genes with that of 5484 non-disease genes. Here, we noticed that the average dN/dS value of human disease genes is significantly lower than the non-disease genes (average dN/dS_{disease} = 0.2644 (±0.0023), average dN/dS_{non-disease} = 0.3212 (±0.0039); Mann–Whitney U test, \( P < 1.0 \times 10^{-6} \)) [Figure 5.3.1]. Previously, López-Bigas et al.\(^8\) observed a similar trend with a much smaller dataset (they studied only 1567 disease genes). However, Smith et al.\(^8\) found that disease genes evolve at a faster rate, and they are tissue specific (TS) by nature. In accordance with the observation of Smith et al.\(^8\), we also found that most of the disease genes are expressed in tissue-specific fashion as compared to non-disease genes (60.87% disease and 50.47% non-disease genes are TS, Fisher’s Exact Test, \( P < 1.0 \times 10^{-6} \)). However, it has already been reported that TS genes evolve at a faster rate than the housekeeping (HK) genes\(^\text{183}\). Indeed, here we also noticed that TS genes evolve at a higher rate than the HK genes (average dN/dS_{TS} = 0.3331 (±0.0047), average dN/dS_{HK} = 0.2095 (±0.0032); Mann–Whitney U test, \( P < 1.0 \times 10^{-6} \)). The negative relationship between evolutionary rates and expression breadth (EB) (Spearman’s \( \rho_{EB \text{ vs. } dN/dS} = -0.1427, P < 1.0 \times 10^{-6} \)) further confirmed that ubiquitously expressed genes evolve at a slower pace than the tissue-specific genes\(^\text{87,184}\). Considering the tissue-specific nature of human disease genes it could be expected that these genes would evolve at a higher rate than non-disease genes. However, their slower evolutionary rate raises an important question regarding the fundamental relationship between protein evolutionary rates and gene expression breadth. In order to resolve this controversy, we analyzed the evolutionary rates of disease and non-disease genes in TS and HK
Towards a resolution...

groupseparately. We observed that disease genes evolve significantly slower than non-disease genes in TS group (average $dN/dS_{\text{disease}} = 0.3094 \pm 0.0052$, average $dN/dS_{\text{non-disease}} = 0.4030 \pm 0.0104$; Mann–Whitney U test, $P < 1.0 \times 10^{-6}$). However, no significant differences of protein evolutionary rate between disease and non-disease genes was observed in HK group (average $dN/dS_{\text{disease}} = 0.2096 \pm 0.0039$, average $dN/dS_{\text{non-disease}} = 0.2092 \pm 0.0056$; Mann–Whitney U test, $P = 7.0 \times 10^{-1}$) [Figure 5.3.1].

High gene expression level imposes a strong selective constraint on protein evolutionary rate\textsuperscript{33,69}. Thus, highly expressed genes are often found to have a lower evolutionary rate than the genes that expressed at lower level\textsuperscript{32,58}. Accordingly, we also found a significant negative correlation between expression level (EL) and evolutionary rates ($dN/dS$) (Spearman’s $\rho_{\text{EL vs. } dN/dS} = -0.1399$, $P < 1.0 \times 10^{-6}$) in our study. Therefore, it is likely that expression abundance may modulate the evolutionary rate of disease and non-disease genes in TS and HK gene pool. Interestingly, when we calculated their average expression level, disease genes showed 6 to 7 fold increased expression level than non-disease genes (average $\text{EL}_{\text{disease}} = 46.7856 \pm 6.1355$, average $\text{EL}_{\text{non-disease}} = 7.5493 \pm 0.6795$; Mann–Whitney U test, $P < 1.0 \times 10^{-6}$) in TS group. However, in HK group we found negligible difference in the expression level between disease and non-disease genes (average $\text{EL}_{\text{disease}} = 6.5973 \pm 0.2723$, average $\text{EL}_{\text{non-disease}} = 6.5820 \pm 0.4810$; Mann–Whitney U test, $P < 1.0 \times 10^{-6}$). Thus, these results emphasize that disease genes evolve at slower rate than the non-disease

\textbf{Figure 5.3.1:} Average evolutionary rates ($dN/dS$) between disease and non-disease genes. The statistical comparison performed by two-tailed Mann–Whitney U test.
genes might be due to the constraints imposed by their higher gene expression level in TS group.

The aforementioned results indicate that both EB and EL imposes strong selection pressure on the rates of protein evolution. However, Park and Choi\textsuperscript{184} observed that, in the mammalian genome EB is more tightly linked with protein evolutionary rates than EL. Thus, to explore the effect of EB and EL on protein evolutionary rates, we performed a regression analysis taking protein evolutionary rates as the dependent variable and EB and EL as the independent variables. Interestingly, we observed that EB exert stronger constraints on protein evolutionary rate than EL ($\beta_{EB} = -0.1597$, $P < 1.0 \times 10^{-6}$; $\beta_{EL} = 0.0038$, $P = 6.4 \times 10^{-1}$).

### 5.3.2. Protein-complex number influences evolutionary rate differences between disease and non-disease genes

Protein-complexes perform their prime functions by a coordinated mechanism through their protein subunits\textsuperscript{40}. Therefore, mutation in a single gene can render the total module dysfunctional\textsuperscript{72}. Thus, complex forming proteins are thought to be more often associated with human diseases\textsuperscript{72}. Accordingly, we observed an elevated proportion of disease genes in protein-complexes dataset (>75\%; Fisher’s exact test, $P < 1.0 \times 10^{-6}$). Moreover, previous studies have reported that proteins involved in several protein-complexes evolve at a slower rate than the proteins.

![Figure 5.3.2: Average complex number between disease and non-disease genes in tissue-specific genes using CORUM and CHPC2012 dataset. The statistical comparison performed by two-tailed Mann–Whitney U test.](image-url)
Towards a resolution...

...present in one or two protein-complex assemblies$^{32,33,162}$. To observe whether disease and non-disease genes share different protein complex assemblies we calculated their protein-complex numbers (the number of protein-complex assemblies in which a protein subunit belongs$^{32,33,162}$). Here, we found that disease genes have higher protein-complex numbers than the non-disease genes in our full dataset (Mann–Whitney U test, $P < 1.0 \times 10^{-6}$). The same trend also observed in the TS gene pool [Figure 5.3.2]. However, we found no significant difference in protein-complex number between disease and non-disease genes in HK gene pool (Mann–Whitney U test, $P = 5.0 \times 10^{-1}$). Earlier, it has been reported that protein-complex number constraints the rate of protein evolution by increasing their gene expression level$^{33}$. Similar results echoed in this study that protein-complex number influences gene expression level positively; whereas, it guide the protein evolutionary rate negatively [Table 5.3.2]. To further evaluate the effect of protein-complex number on protein evolutionary rate we randomly sampled disease and non-disease genes (we took 100 disease and 100 non-disease genes in each bin) into 1000 bins so that, the average protein-complex number of disease and non-disease genes in each bin remain statistically similar to one another. Interestingly, in more than 90% of those bins we found no significant differences in protein evolutionary rate between disease and non-disease genes (at 95% level of confidence). These results indicate that the differences in evolutionary rate between human disease and non-disease genes in TS group might be due to the differences in their protein-complex numbers.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Expression level</th>
<th>Multifunctionality</th>
<th>Evolutionary rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein complex number</td>
<td>CORUM</td>
<td>0.1414 ($P &lt; 1.0 \times 10^{-6}$)</td>
<td>0.2459 ($P &lt; 1.0 \times 10^{-6}$)</td>
</tr>
<tr>
<td></td>
<td>CHPC2012</td>
<td>0.1726 ($P &lt; 1.0 \times 10^{-6}$)</td>
<td>0.2944 ($P &lt; 1.0 \times 10^{-6}$)</td>
</tr>
<tr>
<td>Expression level</td>
<td>----</td>
<td>0.1811 ($P &lt; 1.0 \times 10^{-6}$)</td>
<td>----</td>
</tr>
<tr>
<td>Multifunctionality</td>
<td>0.1811 ($P &lt; 1.0 \times 10^{-6}$)</td>
<td>----</td>
<td>-0.2516 ($P &lt; 1.0 \times 10^{-6}$)</td>
</tr>
</tbody>
</table>

**TABLE 5.3.2**: Spearman’s correlation among expression level, protein complex number and evolutionary rate.
5.3.3. **Protein functionality plays a key role in controlling the evolutionary rate differences between disease and non-disease genes**

Functional constraints on proteins can significantly modulate their evolutionary rates\textsuperscript{78,138}. Thus, proteins with strong structural and functional requirements were often found to be evolutionarily conserved than proteins with no or weak constraints\textsuperscript{78}. Previously, Gillis and Pavlidis\textsuperscript{192} showed that disease genes tend to be multifunctional by nature. We also observed that disease genes are involved in more biological processes (BP) than non-disease genes (average $BP_{\text{disease}} = 8.0686 \pm 0.1116$), average $BP_{\text{non-disease}} = 3.0100 \pm 0.0614$; Mann–Whitney U test, $P < 1.0 \times 10^{-6}$). Additionally, we found a significant positive correlation between protein-complex number and protein multifunctionality [Table 5.3.2]. This result indicates that complex forming proteins might be involved in more biological processes than non-complex forming proteins. Indeed, we observed that complex forming proteins are multifunctional than non-complex forming proteins (CORUM: average $BP_{\text{complex}} = 12.8285 \pm 0.3326$), average $BP_{\text{non-complex}} = 6.7567 \pm 0.0859$; Mann–Whitney U test, $P < 1.0 \times 10^{-6}$; CHPC2012: average $BP_{\text{complex}} = 12.8992 \pm 0.2990$), average $BP_{\text{non-complex}} = 6.5182 \pm 0.0850$; Mann–Whitney U test, $P < 1.0 \times 10^{-6}$). It was also reported that the protein functionality negatively influences protein evolution rate\textsuperscript{78,193}. In our study, we also found a significant negative correlation between protein multifunctionality and protein evolutionary rate [Table 5.3.2]. Further, to confirm this result, we grouped our genes into high functional (HF) and less functional (LF) categories (see, materials and methods section). Here, we found that HF genes evolve at a slower rate than LF genes (average $dN/dS_{\text{HF}} = 0.2199 \pm 0.0025$), average $dN/dS_{\text{LF}} = 0.2989 \pm 0.0032$; Mann–Whitney U test, $P < 1.0 \times 10^{-6}$). Moreover, we observed that the disease genes are HF by nature (56.82\% of disease genes and 32.28\% of non-disease genes are HF; Fisher’s exact test, $P < 1.0 \times 10^{-6}$). Therefore, it is probable that disease genes in TS group evolve at a slower rate due to their higher multifunctionality. As our expectation, here we found that the TS disease genes are HF as compared to TS non-disease genes (50.17\% of TS disease genes are HF and 17.61\% of TS non-disease genes are HF, Fisher’s exact test, $P < 1.0 \times 10^{-6}$).
Previously, Podder and Ghosh\textsuperscript{138} observed that proteins involved in conserved biological processes such as metabolic processes, cellular transport, and protein synthesis evolve slowly than the proteins involved in other biological processes. Indeed, we also found that proteins involved in these processes evolve at a slower rate than the rest of other proteins (average $dN/dS_{\text{conserved}} = 0.2232$ (±0.0025), average $dN/dS_{\text{other}} = 0.2895$ (±0.0030); Mann–Whitney U test, $P < 1.0 \times 10^{-6}$). Moreover, we also observed that 49.35% of HK and 45.96% of TS genes were associated with conserved biological processes (Fisher’s exact test, $P = 9.2 \times 10^{-3}$). Interestingly, the proportion of genes involved in conserved biological processes was found to be similar between disease and non-disease genes in HK gene pool (54.01% of HK disease genes and 54.99% of HK non-disease genes involved in conserved biological processes, Fisher’s exact test, $P = 6.8 \times 10^{-1}$). However, in TS gene pool majority of the disease genes were found to be involved in conserved biological processes than the non-disease genes (54.34% TS disease genes and 41.40% TS non-disease genes were associated with conserved biological processes, Fisher’s exact test, $P < 1.0 \times 10^{-6}$). These results indicate that, the total number of unique GO biological processes and the number of conserved biological processes of a protein influence the rate of protein evolution simultaneously. To unveil the effect of these functional parameters on the rate of protein evolution, we performed a regression analysis taking these two parameters as the predictor variable of protein evolutionary rate. Our results revealed that, both these two functional parameters can independently influence the protein evolution rate. However, conserved biological processes are slightly more efficient than the total number of GO biological processes ($\beta_{\text{conserved biological processes}} = -0.1087$, $P < 1.0 \times 10^{-6}$; $\beta_{\text{GO biological processes}} = -0.1059$, $P < 1.0 \times 10^{-6}$) in controlling the protein evolutionary rate.

5.3.4. \textit{Relative contribution of different factors in controlling protein evolutionary rate}

The results presented so far suggest that gene expression, protein complex number, and protein function could significantly modulate the evolutionary rate of human disease and non-disease genes. To assess their relative contribution to the evolutionary
rate difference between human disease and non-disease genes we performed a linear regression. For gene expression we considered EB, as EB was found to have stronger influence protein evolutionary rate than EL. Similarly for protein function we considered the number of conserved biological processes rather than all GO biological processes. Thus, we performed an analysis on three predictor variables EB, number of conserved biological processes and protein complex number. From the values of regression coefficient it could be inferred that among these three variables, EB has the highest contribution ($\beta_{EB} = -0.1751, P_{EB} < 1.0 \times 10^{-6}$) followed by protein complex number ($\beta_{comolex number} = -0.0888, P_{comolex number} < 2.0 \times 10^{-6}$). Number of conserved biological processes was not found have any significant influence ($\beta_{comolex number} = -0.0319, P_{comolex number} > 7.9 \times 10^{-2}$) on protein evolution rate.

5.4. Conclusions

Studies have been conducted with the aim to characterizing the evolutionary forces operating on human disease and non-disease genes. For instance, Smith and Eyre-Walker first observed that lethal mutations are less effective in disease genes, and thus disease genes evolve at a higher rate than the non-disease genes 87. Later, with a set of 1567 human disease genes, Lopez-Bigas and Ouzounis showed that human disease genes are evolutionary conserved due to their higher selection pressure 88. In contrast to these observations Huang et al., reported that human disease and non-disease genes evolve at a similar rate with marginal difference in their evolutionary rate 89. Subsequently studies that attempted to resolve this controversy have also yielded dissimilar results 138,176. Thus, how human disease genes evolve with respect to human non-disease genes still remains an open question. Recent availability of large-scale gene expression data made it possible to characterize genes according to their tissue distribution. Detailed investigations on human housekeeping and tissue-specific genes have shown that selective constraints on human genes vary depending upon their housekeeping and TS nature 78. Therefore, in this study we classified human disease and non-disease genes according to their gene expression pattern. Here, we observed that the evolutionary rate of human disease and non-disease genes has taken a different turn in housekeeping and tissue-specific groups. Thus, our results revealed that the human disease and non-disease genes are shaped by
different selective constraints in TS group but have experienced similar constraint in HK group.

To explicate the reason behind the different rate of protein evolution in housekeeping and tissue-specific groups we considered a number of attributes contributing to the evolutionary rate difference between disease and non-disease genes. Previously, Bortoluzzi et al., have reported that genes involved in Mendelian diseases are highly expressed as compared to non-disease genes. Since, protein evolutionary rate varies according to their expression level, gene expression abundance was considered to be an important determinant of evolutionary rate heterogeneity between human disease and non-disease genes. In contrast, our study revealed that gene expression breadth rather than gene expression abundance has a stronger influence on the evolutionary rate of human disease and non-disease genes. Thus, it suggests that the extent to which a gene is expressed has a significant impact on protein evolutionary rate. Our systematic classification of human genes into tissue-specific and housekeeping groups further revealed that protein complex number could significantly modulate the evolutionary rates of disease and non-disease genes. Earlier studies indicated that proteins those participate in a large number of protein complexes need high gene expression level for their proper assemblage. In our previous study, we showed that proteins participating in several protein complexes have enhanced rate of transcription and mRNA translation and as a consequence a high gene expression level. Thus, protein complex number was found to share a positive correlation with gene expression, and hence, a negative correlation with protein evolutionary rate. In this study, we also noticed similar results. Moreover, here we noticed that the human disease-causing genes have a tendency to participate in a large number of protein complexes. Since, complex forming proteins are essential for cellular integrity, it was proposed that mutations in these proteins are likely to cause several human diseases. Hence, complex forming nature of proteins have been implicated in a number of human diseases such as Alzheimer disease, type 2 diabetes and amyotrophic lateral sclerosis, etc. By compiling protein complexes from different databases, here we confirmed that the majority of these proteins are associated with human diseases. Consequently, human disease genes were found to have higher protein complex number than the non-disease genes. However, this trend was observed mainly in
tissue-specific group, where human disease genes appeared to be evolutionarily conserved than the non-disease genes. Further, we found no significant difference of protein evolutionary rate between human disease and non-disease genes in HK group, which may be considered as due to the lack of difference in their protein complex number. As with protein complex number, gene expression abundance was also found to follow similar distributions in TS and HK gene groups. Therefore, their inter correlation may suggest that protein complex number is exerting its affects on protein evolutionary rate through gene expression abundance. However, from the regression analysis it was evidenced that both of these two parameters could independently guide the evolutionary rate of human disease and non-disease genes. Significant difference of evolutionary rate between human disease and non-disease genes in TS group was found to disappear on controlling the effect of protein complex number. Thus, it further vindicates the importance of protein complex number. Therefore, it could be inferred that both gene expression abundance and protein complex forming nature had played an important role in controlling the evolutionary rate of human disease and non-disease genes in TS group. However, gene expression abundance is more important determinant than protein complex number. In this context, it would be interesting to discuss how protein’s complex forming nature could constrain the rate of sequence evolution. Earlier, it was noted that proteins with a higher propensity to form complexes are associated with essential cellular functions \(^{160,161}\), and here we observed that these proteins are multifunctional by nature. Teichmann in his comprehensive analysis with yeast proteins reported that proteins in stable complexes have interfaces constrained by their interaction partners \(^{34}\). In addition, these proteins were proposed to preserve the geometry of their interface regions for proper complex assembly. Therefore, the strong structural and functional compulsions in these proteins may be attributed as the cause of their consequence conservation.

Recently, Podder and Ghosh \(^{138}\) observed that human disease genes require more advanced mechanism for fine tuning of their biological processes. Thus, their involvement in core biological processes was shown to restrain their sequence evolution \(^{138}\). In this study, we also observed that the evolutionary rates of human disease and non-diseases genes are determined by their functional attributes. Our results that tissue...
specific disease genes are HF by nature and are associated with conserved biological processes could partly explain their sequence conservation. Due to the functional importance proteins those participate in many biological processes were considered to be evolutionarily conserved \(^78\). Thus, protein multifunctionality was regarded as the dominant determinant of protein evolutionary rate \(^78\). Hence, the involvement of tissue specific disease genes in several processes could constraint their rate of sequence evolution. Subsequently, we observed that the functional specificity of proteins (core or regulatory) is more important parameter in explaining evolutionary rate variation between human disease and non-disease genes in TS group. Since, in HK group disease and non-disease genes were found to be equally likely for core processes, it is reasonable that they evolved at a similar rate. Thus, our results imply that along with protein complex number and gene expression abundance, protein’s association with core biological processes is a significant factor in determining their evolutionary rate. Their independent effect on protein evolutionary rate was confirmed from regression analysis.