Designability, aggregation propensity and duplication of disease-associated proteins

Philip Wong1, Andreas Fritz2 and Dmitri Frishman1,3,4

1Institute for Bioinformatics, GSF – National Research Center for Environment and Health, Ingolstädter Landstrasse 1, D-85764 Neuherberg, 2Biomax Informatics AG, Lochhamer Strasse 9, D-82152 Martinsried and 3Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftzentrum Weihenstephan, D-85350 Freising, Germany

Over 2000 proteins in the Ensembl human genome database have been linked with disease information from OMIM. In comparison with all human proteins, we find that disease-associated proteins tend to have less designable folds in terms of their SCOP family counts, suggesting that they are intrinsically less robust to mutation and environmental stress. Disease proteins also tend to have isoelectric points closer to neutrality and more alternating hydrophilic-hydrophobic amino acid stretches compared with the average human protein. These results suggest that protein aggregation is a significant phenomenon associated with diseases. Another finding in this work is that many disease proteins are highly sequence similar to other disease proteins, suggesting that gene duplication has contributed to the expansion of disease-prone protein families.

Keywords: aggregation/designability/disease/duplication

Introduction

A mystery in medicine that has largely remained unsolved is the question of how heritable diseases have evolved. The sequencing of the human and related genomes and the emergence of databases such as the Online Database of Mendelian Inheritance in Man (OMIM) (Hamosh et al., 2005), which contain mappings of genes to disease phenotypes, provide novel opportunities to explore this question. Although the definition of what phenotypes may be considered a disease can be highly debatable (Scully, 2004), the diseases currently associated with genes in databases are harmful to enough individuals as to be clinically recognized. One may hypothesize that by comparing disease-associated genes in these databases, termed disease genes, against genes without such annotation, termed non-disease genes, one can derive properties which predispose genes for diseases currently clinically identifiable. Knowledge of such properties can be applied to help predict or verify disease–gene relations derived from linkage and association studies (Botstein and Risch, 2003; Carlson et al., 2004) and to select targets for drug screens. Such knowledge is also crucial for understanding the evolution of heritable diseases.

Previously, disease genes were found to be more expressed (Bortoluzzi et al., 2003), longer (Smith and Eyre-Walker, 2003), more tissue-specific (Smith and Eyre-Walker, 2003; Winter et al., 2004), have more synonymous nucleotide substitutions (Huang et al., 2004) and have less members amongst slowly evolving housekeeping genes (Winter et al., 2004). Comparisons conducted by López-Bigas and Ouzounis (2004) at the protein level revealed that disease proteins tend to be longer, more conserved, phylogenetically more extended and have less highly conserved paralogs than the average human protein. They have subsequently exploited these differences to create a decision tree-based predictor of disease proteins.

A number of developments have allowed us to add to these studies. First, an intriguing hypothesis that proteins with more designable structures (i.e. proteins which have more sequences that encode their structures) were more robust to mutation and thermal stresses had been proposed (Li et al., 1996). In line with this hypothesis is the finding that proteins of a random sample of thermophiles exhibited a higher contact trace, a measure which correlates well with the designability, than a sample of mesophiles (England et al., 2003). We hypothesized that since proteins could be functionally impaired by mutations or environmental stresses, disease proteins would be less designable than non-disease proteins. A second development involved the discovery of sequence properties associated with protein aggregation (Chiti et al., 2003; DuBay et al., 2004). Many diseases have been associated with protein aggregation (Dobson, 2004; Ross and Poirier, 2004), but the extent of this phenomenon had not been assessed. One could test whether disease proteins are more aggregation prone than non-disease proteins in terms of these properties.

In this work, we compared disease and non-disease proteins from the Ensembl human database (Binney et al., 2004) in terms of designability and aggregation propensity. In addition, we assessed the likelihood that proteins highly sequence similar to disease proteins would also be associated with disease based on the current level of annotation. We validated our findings using a differently annotated database of human proteins provided by Biomax Informatics, containing roughly twice the number of proteins annotated with disease.

Materials and methods

Proteomes and disease annotation

A total of 34 111 proteins predicted to be encoded in the human genome were obtained from the Ensembl human v23.34e.1 database (Binney et al., 2004). We term this dataset the Ensembl protein dataset. OMIM (Hamosh et al., 2005) based disease annotations for human genes were obtained using the Ensmart tool (Hammond and Birney, 2004) and mapped to 2113 in the Ensembl protein dataset. OMIM is a database focused on heritable genetic diseases, most of which are of high penetrance. A larger set of 39 801 human proteins were obtained from the Biomax Human Genome Database.
High-quality disease proteins

We consider all proteins with any disease-related annotation as disease proteins for our analysis. This annotation varies from suggestions of disease susceptibility effects upon mutation to disease–gene associations identified by positional cloning or by multiple methods. Disease-causing mutations in OMIM, however, are marked with a number in parentheses indicating whether the mutation was positioned by mapping the wild-type gene (1), by mapping the disease phenotype itself (2) or by both approaches (3) (see http://www.ncbi.nlm.nih.gov/Omim/omimfaq.html). To check if our results were different using higher quality data, a subset of 1470 proteins was obtained from the 2113 Ensembl protein dataset by excluding proteins that were not associated with a disease marked with a ‘(3)’ in the OMIM annotation. We term this the ‘high-quality’ disease protein set.

Proteins associated with disease caused by amino acid substitution

Proteins associated with disease caused by amino acid substitution (\(\text{DP}_{\text{AA}}\)) were first screened for by text scanning OMIM entries associated with proteins found in the Ensembl and Biomax datasets using a Perl script. Only OMIM entries associated with a single protein in each dataset were included in order to avoid potential errors in mapping amino acid substitutions to the wrong protein. Potential \(\text{DP}_{\text{AA}}\)s were identified by amino acid substitutions defined by the pattern ANA (where A is a letter or group of letters representing an amino acid and N is a residue number with no possibility of ANA representing nucleotide substitutions) in the OMIM text. For example, L234E would be considered an amino acid substitution while A23C would be ignored for this study, as A and C could be potentially represent nucleotides. The list of potential \(\text{DP}_{\text{AA}}\)s was then refined manually.

Protein properties

In all cases in which disease and non-disease proteins were compared, only the largest protein encoded by each gene was included, as done by López-Bigas and Ouzounis (2004). Protein length, \(p_f\) and SCOP (Andreeva et al., 2004) assignments were obtained from the PEDANT system (Riley et al., 2005). SCOP folds were assigned to proteins if the corresponding sequences were within a BlastP (Altschul et al., 1997) \(E\)-value of \(10^{-6}\). The residues A, C, F, G, I, L, M, P, V, W and Y were considered to be hydrophobic and H, Q, N, S, T, K, R, D and E were considered hydrophilic in this study.

Designability

Protein designability was measured by counting the number of families in each fold contained in a given protein and taking the minimum. For example, if protein A contains three domains with folds F1, F2 and F3 and these folds in turn contain eight, three and seven families, respectively, protein A’s minimum family count would be three. By recording the minimum family count of the folds in proteins, we assessed their designability by computing the designability of their least designable fold.

Results and discussion

The comparison of disease and non-disease proteins was carried out in three steps. First, the full sets of disease proteins and non-disease proteins were compared. Second, as the study here involves comparing the sequence properties of proteins in which the disease state is attributed to the presence of altered proteins, we reduced our disease protein datasets to include only such proteins. Because of computational ease, we selected those proteins annotated with disease-causing amino acid substitutions as our reduced disease protein set (see Materials and methods) and repeated our comparison. Third, we identified a high-quality disease protein set within the Ensembl dataset (see Materials and methods) and again repeated our comparison. The conclusions drawn from comparisons using these repeated experiments were not different from those found using the full disease datasets (Supplementary material S3–S7 available at PEDS Online). Counts of proteins used in this study are listed in Table I.

Disease proteins tend to have folds with less families

Mutation or environmental change can disrupt and/or create aberrant function in proteins. Disease proteins were hypothesized to contain more often structures which are susceptible to perturbation by mutation or external stresses. Structures which are more designable tend to be more robust against mutation and thermal fluctuations (Zhang, 1997; Wingreen et al., 2004). Because a direct relationship exists between the number of sequences and the number of families in protein folds (Zhang et al., 1997), one can measure a protein fold’s designability, by counting the number of families in that fold. To provide an estimate of a protein’s structural susceptibility to perturbing influences, we counted the number of families in each SCOP domain fold (Andreeva et al., 2004) contained in the protein and recorded the minimum count (see Materials and methods). In other words, we assessed the designability of proteins by computing the designability of their least designable fold. We reason that if a domain which occupies a portion of a protein is destabilized and becomes misfolded owing to stress or mutation (within or outside the domain), it is likely that the function of the entire protein would be affected. Therefore, it was intuitive that we assessed designability in this way.

Analysis of our human datasets revealed that disease proteins tend to have significantly smaller minimum family counts than the average human protein. Disease proteins have a

<table>
<thead>
<tr>
<th></th>
<th>Ensembl</th>
<th>Biomax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>1132</td>
<td>3049</td>
</tr>
<tr>
<td>Non-disease</td>
<td>21 159</td>
<td>33 695</td>
</tr>
<tr>
<td>High-quality disease (\text{DP}_{\text{AA}})</td>
<td>772</td>
<td>–</td>
</tr>
<tr>
<td>Disease (\text{DP}_{\text{AA}})</td>
<td>631</td>
<td>1326</td>
</tr>
</tbody>
</table>

\(\text{DP}_{\text{AA}}\) denotes the number of disease and non-disease proteins in the Ensembl and Biomax datasets shown in the first two rows. The number of high-quality disease proteins extracted from the Ensembl proteome is shown in the third row (see Materials and methods). Only the largest protein isoform from each gene has been included. The last row shows the number of disease proteins associated with a disease causing amino acid substitution (\(\text{DP}_{\text{AA}}\) – see Materials and methods). Disease proteins represent 5% \(= 1132/(21 159 + 1132)\) and 9% of proteins in the Ensembl and Biomax datasets, respectively.
Disease proteins are more likely to aggregate

A possible consequence of structural perturbation by mutation or environmental change is that of misfolding or unfolding of proteins leading to aggregate formation. Aggregates or their precursors are cytotoxic and can cause cell death (Bucciantini et al., 2002, 2004). In comparing disease and non-disease proteins, we find that the former tend to have isolectric points closer to neutrality and more stretches of alternating hydrophobic–hydrophilic residues (of length 5 or more) than the latter (Figure 2A and B). Such properties have been implicated to increase aggregation rates of unfolded proteins in \textit{in vitro} experiments (Chiti et al., 2003; DuBay et al., 2004). These results suggest that disease proteins tend to be more aggregation prone than non-disease proteins and complement work by Dima and Thirumalai (2004), which suggested that low sequence correlation entropies, mixed charged-hydrophobic and charged-polar runs in proteins may be indicative of disease association and tendency to aggregate.

Furthermore, the number of alternating hydrophobic–hydrophilic stretches in proteins correlates well with the length of proteins (Pearson $R^2 > 0.8$) (Figure 3A). Longer proteins also tend to have isoelectric points closer to neutrality, although this attribute does not correlate well with length (Figure 3B). In addition, longer proteins, especially those containing multiple domains, must undergo many more folding processes to achieve their native states than short single-domain proteins. These properties are likely to predispose longer proteins for misfolding and subsequent aggregation. Indeed, \textit{in vitro} folding of relatively large eukaryotic proteins frequently fails. The finding that disease proteins tend to be much longer (Lopez-Bigas and Ouzounis, 2004) (~1.5x and 1.7x longer on average in the Ensembl and Biomax datasets, respectively) and less designable than non-disease proteins suggests that disease proteins are more likely to misfold upon mutation or stress. With isolectric points closer to neutrality and greater number of hydrophobic–hydrophilic stretches, disease proteins would likely have a greater potential for aggregation than non-disease proteins.

Many disease proteins are duplicates of other disease proteins

There are reasons to believe that sequence similarity to known disease proteins may also be a significant factor contributing to disease propensity. Protein length, designability, isolectric point and sequence stretch patterns have so far been implicated to contribute to disease propensity. All of these properties are dependent on the sequences of the corresponding proteins. In addition, highly sequence similar proteins are likely to share interacting partners (Yu et al., 2004) which may serve to link disease proteins functionally to non-disease proteins. Such functional linkage may indicate that both disease and non-disease proteins share a function that when disrupted would cause disease. Alternatively, disrupting the function of a non-disease protein functionally linked to a disease protein may subsequently disrupt the function of the disease protein and cause disease. Moreover, if the DNA sequences which encode
proteins are sufficiently similar, non-allelic homologous recombination, a mechanism associated with disease (Bailey et al., 2002; Shaw and Lupsky, 2004), may also occur. Hence one may hypothesize that annotating non-disease proteins, highly sequence similar to disease proteins, as disease proteins would be valid for many proteins. To confirm this, the proportions of disease proteins in the Ensembl and Biomax human databases with duplicates annotated with disease were assessed (Table II). Almost 40% of the disease proteins in the Biomax database (and 50% in the Ensembl database) have duplicates (paralogs) associated with disease. Over one-fifth of disease proteins have all duplicates associated with disease in both databases. Disease proteins represent 5 and 9% of proteins in the Ensembl and Biomax datasets, respectively (Table I). In contrast, the chance that a duplicate of a disease protein is also a disease protein is significantly higher than expected ($\chi^2$ test: $P < 0.01$) at 20 and 29% in the Ensembl and Biomax datasets, respectively (Table II). These findings strongly suggest that assigning disease status to non-disease proteins based on high sequence similarity is valid for many proteins.

Notably, many disease proteins have been associated with a different disease annotation than one of their duplicates. For example, the protein ATP2A2 (NP_001672) is annotated with Darier disease (OMIM: 124200), an autosomal dominant skin disorder, whereas its duplicate ATP2A1 is annotated as being associated with Brody myopathy (OMIM: 108730), a disorder characterized by painless muscle cramping and exercise-induced impairment of muscle relaxation. The high sequence similarity between the two proteins (within BlastP $E$-value of $10^{-70}$) suggests that they share a common evolutionary origin. The functional compromise of the two duplicates in different tissues results in two different disease phenotypes. A similar phenomenon has been reported for PYPAF1 and NOD2 (Albrecht et al., 2003).

The large proportion of disease proteins with duplicates associated with disease suggests that gene duplication is a significant phenomenon contributing to the expansion of disease-prone protein families. These families may contain proteins associated with different diseases. A clustering of OMIM entries based on sequence similarity of their associated

---

Fig. 2. Distributions of properties associated with protein aggregation in Ensembl disease and non-disease proteins. (A) Distance of isoelectric point from neutrality. (B) Hydrophilic–hydrophobic stretch count. Disease proteins tend to have isoelectric points closer to neutrality and more hydrophilic-hydrophobic stretches than non-disease proteins and the combined set of disease + non-disease proteins dataset (Mann–Whitney test, $P < 0.01$; Kolmogorov–Smirnov test, $D > 5\%$, $P < 0.01$).
proteins is shown in Figure 4. Examining such clusters allows the identification of disease families, which may yield new insight into how diseases may be related.

**Conclusion**

By comparing disease and non-disease proteins, we have shown that disease proteins tend to have folds with less families than non-disease proteins, suggesting that they are intrinsically more structurally vulnerable to mutation or environmental stresses. Disease proteins also tend to be longer, have isoelectric points closer to neutrality and more aggregation-prone stretches than non-disease proteins, suggesting that the former are more likely to aggregate than the latter upon unfolding or misfolding. Many disease proteins are duplicates of other disease proteins, reinforcing the notion that sequence similarity to known disease proteins can contribute substantially to disease propensity. These results were apparent even when we defined our disease protein set to include only those proteins with known disease causing amino acid substitutions and when we chose to use a high-quality disease protein set (Supplementary material S3, S5–7).

**Table II.** A large proportion of disease proteins have duplicates associated with disease

<table>
<thead>
<tr>
<th></th>
<th>Ensembl</th>
<th>Biomax</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of disease proteins</td>
<td>1132</td>
<td>3049</td>
</tr>
<tr>
<td>No. of disease proteins with at least one duplicate associated with disease</td>
<td>567 (50%)</td>
<td>1196 (39%)</td>
</tr>
<tr>
<td>No. of disease proteins with at least one duplicate associated with a different disease annotation</td>
<td>242 (21%)</td>
<td>882 (29%)</td>
</tr>
<tr>
<td>No. of disease proteins with all duplicates associated with a disease</td>
<td>249 (22%)</td>
<td>456 (15%)</td>
</tr>
<tr>
<td>Given a particular disease protein, the probability that a duplicate is also annotated with disease</td>
<td>20%</td>
<td>29%</td>
</tr>
</tbody>
</table>

*The number of disease proteins (with isoforms removed) for the Ensembl and Biomax human databases is shown in the first row. Proteins within a BlastP (Altschul et al., 1997) E-value of $10^{-30}$ or within 80% identity and length are considered duplicates of each other. At this threshold, duplicates are likely to have similar interacting partners (Yu et al., 2004). Subsequent rows show the number and proportions of disease proteins with duplicates associated with disease.

*Some differences in annotation are attributed to different strings used to annotate the same disease.
The reader should be aware that our results are based on incomplete data. We have defined the domains in human proteins using SCOP, which is biased towards domains which are soluble and commendable to structural determination. However, disease proteins tend to be relatively conserved than non-disease proteins (López-Bigas and Ouzounis, 2004). Therefore, even without consideration of what domains are defined by SCOP, one finds that disease proteins are more sequence restricted, consistent with the hypothesis that they have less designable structures (if any) than non-disease proteins.

Our results are also dependent on the accuracy of the human gene models in the human genomes used, the proteins predicted to be expressed using these gene models and the annotation associating proteins with various diseases. Our results are robust to two different human databases with different levels of disease annotation. We have also verified them against a high-quality subset of disease–gene relations.

The trends do not necessarily apply to specific populations or global trends gleaned from data currently stored in databases. Our results are dependent on the accuracy of the human gene models in the human genomes used, the proteins predicted to be expressed using these gene models and the annotation associating proteins with various diseases. Our results are robust to two different human databases with different levels of disease annotation. We have also verified them against a high-quality subset of disease–gene relations.

The probability that a protein becomes associated with disease depends on multiple factors, including the mutation type, the protein involved, the rest of the organism and the environment to which the organism is exposed. Our results are global trends gleaned from data currently stored in databases. The trends do not necessarily apply to specific populations or individuals. It would be of great interest to integrate information from epidemiological studies with the trends derived here to shed light on this matter.

The finding that families of sequence similar disease proteins exist suggests common origins and mechanisms to many of our modern diseases. Drugs targeting one member of a particular family may affect others in that family. Understanding the properties which predispose proteins for disease and how they may have evolved will perhaps aid the identification of novel gene-to-disease relations and the treatment of the associated diseases.

Acknowledgements

We thank Philipp Pagel for extremely helpful advice, Louise Riley and Martin Münsterkötter for creating the PEDANT database tables used in this investigation and Hans-Werner Mewes and Pawel Smialowski for useful comments. This work was funded by a grant from the German Federal Ministry of Education and Research (BMBF) within the BFAM framework (031U112C).

References


Received February 15, 2005; revised May 25, 2005; accepted August 3, 2005

Edited by Luis Serrano