Bioinformatical assay of human gene morbidity

Fyodor A. Kondrashov, Aleksey Y. Ogurtsov and Alexey S. Kondrashov*

National Center for Biotechnology Information, National Institutes of Health, 38a Center Drive, 6S602, Bethesda, MD 20892, USA

Received December 1, 2003; Revised February 3, 2004; Accepted February 17, 2004

ABSTRACT

Only a fraction of eukaryotic genes affect the phenotype drastically. We compared 18 parameters in 1273 human morbid genes, known to cause diseases, and in the remaining 16 580 unambiguous human genes. Morbid genes evolve more slowly, have wider phylogenetic distributions, are more similar to essential genes of Drosophila melanogaster, code for longer proteins containing more alanine and glycine and less histidine, lysine and methionine, possess larger numbers of longer introns with more accurate splicing signals and have higher and broader expressions. These differences make it possible to classify as non-morbid 34% of human genes with unknown morbidity, when only 5% of known morbid genes are incorrectly classified as non-morbid. This classification can help to identify disease-causing genes among multiple candidates.

INTRODUCTION

Every gene must be useful, since useless genes degenerate into pseudogenes. Still, only a minority of genes in a eukaryotic genome are essential. Some mutations of an essential gene, either dominant or recessive, cause drastically altered phenotypes, which are almost without exception lethal or deleterious. In contrast, mutations at non-essential genes affect phenotype and fitness only quantitatively (1). The proportion of essential genes was assayed for genomes of several prokaryotes and eukaryotes. Approximately 40% of genes in Saccharomyces cerevisiae (2,3), 30% in Caenorhabditis elegans (4–6), 30% in Drosophila melanogaster (7) and 25–35% in Danio rerio (8,9) and Mus musculus (10,11) are essential.

Human genetics is particularly concerned with morbid genes, which can harbor alleles that adversely affect health. Gene essentiality and morbidity are tightly related and essential genes form a subset of morbid genes. By definition, an essential gene is morbid, and a morbid gene responsible for a Mendelian disease is essential. Even those morbid genes which commonly act only as risk factors for complex diseases are often essential. Although heterozygous disease-predisposing alleles of such genes are incompletely penetrant, rare homozygotes may always be affected severely, as is the case with, for example, LPL (12) and BRCA2 (13). So far, over 1000 human morbid genes, affecting viability, fertility and/or longevity, have been described (14), but their total number may be closer to 10 000, if in humans, as in other vertebrates and eukaryotes, ~30% of all genes are morbid. Thus, >80% of human morbid genes probably still await discovery (15).

Discovering morbid genes responsible for a particular disease is currently among the main tasks of medical genetics. Linkage and/or association studies often point only to a relatively long region of the chromosome which contains many genes and identifying one or a few responsible genes among them may be difficult (see for example 16–20). Even if performed at very high resolution, such studies may fail to pinpoint the responsible gene, if the mutations are in distant (up to 1000 kb away) control elements (21). Thus, identification of the disease-causing gene would be greatly facilitated if morbidity (or essentiality, as a proxy) of every human gene were known a priori: a majority of genes probably cannot be responsible for any disease and, thus, should be immediately rejected. Thus, assaying morbidity (or essentiality) of a gene on the basis of its easily ascertainable, bioinformatical parameters would be highly desirable.

In a pioneering paper (22), Wilson and co-authors proposed that the rate of evolution of proteins encoded by essential genes, characterized by the per site number of non-synonymous substitutions $K_n$ between orthologous genes, is lower than in non-essential genes. Since then, whether nonessential genes evolve faster than essential genes has been a controversial issue' (23). Several authors claim that when two genomes are compared, $K_n$ in essential genes is, indeed, substantially lower than in non-essential genes (24–28), while others find only a weak correlation between protein essentiality and evolution rate (23). Also, there is a disagreement on why essential genes evolve more slowly. While some authors favor the original explanation of Wilson et al. (22), that essential genes are under stronger purifying selection (24–26), others claim that a lower $K_n$ in essential genes is due to positive selection in non-essential genes (27) or to the correlation of $K_n$ with another key parameter. The proposed key parameters are expression rate (28,29) and the numbers of paralogs (23), so that, in addition to a lower $K_n$, a higher expression rate and a higher number of paralogs, as well as a lower propensity for gene loss (30), may be useful in identifying essential genes.

Recently, a similar analysis of human morbid genes has been performed (31). Surprisingly, human–rodent $K_n$ in the studied human morbid genes was higher than in presumed non-morbid genes. Also, morbid genes were found to

*To whom correspondence should be addressed. Tel: +1 301 435 8944; Fax: +1 301 480 2290; Email: kondrashov@ncbi.nlm.nih.gov

Nucleic Acids Research, Vol. 32 No. 5. Published by Oxford University Press

Published online March 12, 2004
have higher synonymous divergence \( K_s \) in human–rodent comparisons, to encode longer proteins and to have narrower ranges of tissue expression (31) (Table 1). Using a linear combination of the four parameters, the authors were able to discriminate between their sets of morbid and non-morbid genes.

Here we report the results of an in silico analysis of morbidity of human genes based on the distributions of 18 parameters in genes which are known to be morbid and in the rest of the human genes. Many of these parameters are highly informative and together they produce a useful prediction of morbidity for every unambiguously known human gene.

### MATERIALS AND METHODS

We obtained the sequence and annotation of the longest isoform of every human gene from build 33 of the human genome (32). To eliminate poorly annotated genes we retained only those genes that produced at least one BLAST (33) hit with 95% identity that covered >80% of the their length when compared with the database of human mRNAs from GenBank. Human mRNA sequences were obtained from the Entrez retrieval system (34) by typing ‘complete mRNA OR cDNA’ as key words and excluding EST sequences in the limits option of Entrez. We also retained genes that showed at least one BLAST hit with an expected cut-off value at 10\(^{-20}\). A list of essential genes in \( D.melanogaster \) was obtained by typing ‘lethal’ into the phenotype search engine in FlyBase (38) and relating the FlyBase id from this list to those in the \( D.melanogaster \) NCBI genome annotation. Expression levels were estimated as the number of EST hits in the collection of human ESTs at NCBI, as described (39). Expression breadths were estimated as the number of tissue types for which EST hits are present in the collection, divided by the number of these hits. Numbers of matches of 5’ and 3’ intron edge sequences to splicing consensuses (40) were averaged for all introns of a gene.

Classification of genes was performed by a multilayer neural network of perceptron type (41) using our own software. The multilayer perceptron processes inputs, applies the sigmoid function to their linear combination and uses the result as input for the next neuron layer, which makes it possible to approximate any multidimensional surface (42). We used two hidden layers with four and two neurons, respectively. The output layer contains one neuron that...
generates output that characterizes similarity of a gene to the morbid and generic classes. All the genes were divided into three subsets: the training set (50%), the test set (25%) and the validation set (25%). The neural network was trained by minimizing the error function (the sum of squares of differences between the real and predicted classifications for the training set). The initial values for weights were assigned randomly and the iterative training steps were made while the error function for the test set decreased (43). The validation set was then used to characterize the neural network performance.

RESULTS

We consider the set of 1273 known human morbid genes and the set of all remaining 16 580 human genes. This generic set probably comprises ~50% of all human protein coding genes and must contain many non-morbid genes. We calculated distributions of a variety of bioinformatical parameters within the morbid and generic sets of genes. Figure 1 presents data on 18 parameters whose distributions within the two sets were substantially different.

Table 1 presents data on correlations among these 18 parameters within the generic and the morbid sets of genes. Obviously, correlations within the two sets are similar (the combined set also has very similar correlations; data not reported). We can see that most pairwise correlations are rather weak, with only few exceptions. In particular, $K_n$ is only rather weakly correlated with the number of paralogs or the expression level or breadth. Only $K_n$ and $K_s$, the similarities of a human protein to its closest homologs in D.melanogaster, C.elegans and A.thaliana, and the number of introns and the length of the encoded protein are strongly correlated. Thus, none of our 18 parameters is redundant, due to its very strong correlations with other parameters.

Eleven of these 18 parameters are particularly informative. Some values of these 11 parameters are rare among morbid genes and thus can be regarded as markers of non-mororbidity. Only seldom does a morbid gene have a human–mouse $K_n > 0.30$, belong to a gene family with >100 human paralogs, encode a protein of <800 amino acids, possess <4 introns of average length <400, carry splicing signals of quality <14.0, is expressed at a level <3 or contain <2% Ala, <2% Gly, >10% Lys or >5% Met (proportions of other amino acids were uninformative; data not reported). Two or more of these markers of non-mororbidity are possessed by 33% of generic genes, but by only 9% of morbid genes.

To further improve recognition of non-morbid genes, we used a neural network, which allowed us to utilize all 18 parameters. A neural network produces a value of the classification variable $X$ for each gene, morbed or generic. Figure 2 displays the process of network training. Figure 3 shows the distributions of $X$ produced by the trained neural network. If the threshold value of $X$ is chosen in such a way that 1, 5 or 10% of known morbid genes are incorrectly classified as non-morbid, the proportions of generic genes classified as non-morbid are 8, 34 and 44%, respectively (Fig. 3).

This classification makes it possible to offer an in silico prediction of morbidity for all 17 853 currently known unambiguous human genes. Each gene is placed into one of the following five categories: morbidity proven (1273 known morbid genes), morbidity (including early lethality) plausible (X > 0.60, 38% of generic genes), morbidity possible (0.53 < X < 0.60, 18%), morbidity unlikely (0.46 < X < 0.53, 10%) and morbidity very unlikely (X < 0.46, 34%).

The values of the 18 informative parameters of every analyzed gene are listed in a file ‘input.data’ at ftp://ftp.ncbi.nih.gov/pub/kondrashov/morbidity and the predicted morbidity of each of these genes, identified by the GI, LocusID and the name of the encoded protein is presented in a file ‘morbidity.prediction’ at the same address.

DISCUSSION

Morbid and generic sets of human genes are characterized by substantially different distributions of 18 bioinformatical parameters (Fig. 1). For a majority of these parameters, the differences between their distributions within morbid and generic genes can be approximately described by a shift: morbid genes possess high values of a particular parameter more (or less) often than generic genes. However, for the five parameters related to evolutionary conservatism (Fig. 1A–E), as well as for the number of paralogs (Fig. 1G), morbid genes are rarer among those possessing both high and low parameter values. Such differences can be used for the purpose of discrimination between morbid and non-morbid genes only by a non-linear method of classification, such as neural networks.

Many of the observed patterns are not surprising. In particular, stronger selection on morbid genes is expected to make them more evolutionarily conserved (22), leading to their higher similarity to murine orthologs and to homologs outside mammals. An elevated proportion of extremely conserved ‘super-morbid’ genes within the generic set, with human–mouse $K_n < 0.01$, human–mouse $K_s < 0.2$ and/or >90, >80 or >60% similarity to a gene from the D.melanogaster, C.elegans or A.thaliana genomes, respectively (Fig. 1A–E), may be due to genes whose loss-of-function alleles are early acting lethals. Indeed, early acting genes of Caenorhabditis briggsae have low values of $K_n$ in C.briggsae–C.elegans comparisons (44). Since morbidity of such genes is difficult to discover in humans, they mostly reside within the generic set.

It is also natural that, in comparison to generic genes, morbid genes are more similar to essential genes of D.melanogaster (Fig. 1F), encode longer proteins (Fig. 1H), possess more precise splicing signals (Fig. 1K) and are more highly and broadly expressed (Fig. 1L and M). In contrast, the reasons for intermediate numbers of paralogs (Fig. 1G) [a different pattern was reported for S.cerevisiae (3)], larger numbers of longer introns (especially, rarity of intronless genes, Fig. 1I and J) and shifted proportions of five amino acids (Fig. 1N–R) in morbid genes remain obscure. Still, a parameter which distinguishes morbid genes can be used to identify them, regardless of whether the observed pattern can be easily explained.

Our data indicate that morbid genes have a much lower average $K_n$ than generic genes, which is in accord with many observations on essential versus non-essential genes (24–30), assuming that morbidity and essentiality are tightly related. We are not sure why the opposite pattern was observed in Smith and Eyre-Walker (31). One possible explanation is that our samples (1273 morbid and 16 580 generic human genes) are more representative than the samples used in Smith and
Eyre-Walker (31) (387 morbid and 2024 provisionally non-morbid genes). Perhaps, these 2024 provisionally non-morbid genes have a disproportionaly high fraction of very conser-
vative early acting ‘super-morbid’ genes (which, in fact, are mostly essential and morbid).

Since we do not possess a large set of human genes which are certainly non-morbid, comparison of the known morbid genes with the remaining (generic) genes was the only feasible procedure. In such a comparison it made sense to concentrate on predicting non-morbidity (i.e. the lack of substantial

![Figure 1. Distributions of 18 parameters within 1273 known human morbid genes (black bars) and within all other 16 580 unambiguous human genes (gray bars). (A and B) \( K_s \) and \( K_n \) between a human gene and its murine ortholog. (C–E) The fraction of identical amino acids within the alignment of the protein coded by a human gene and the most similar protein of \textit{D.melanogaster} (C), \textit{C.elegans} (D) and \textit{A.thaliana} (E). (F) The proportions of genes for which the most similar gene in the \textit{D.melanogaster} genome is essential or non-essential. (G) The number of paralogs of a gene within the human genome. (H) The length of the protein encoded by a gene. (I) The number of introns within a gene. (J) The average length of introns within a gene. (K) The average quality of 5’ and 3’ splicing signals within introns of a gene. (L and M) The expression level or breadth of a gene. (N–R) The proportions of alanine (N), glycine (O), histidine (P), lysine (Q) and methionine (R) within the protein encoded by a gene.]
pathological consequences of all possible mutations) of some genes with unknown morbidity. Indeed, we can estimate the rate of false negative predictions of non-morbidity by measuring the proportion of known morbid genes which are incorrectly predicted to be non-morbid. In contrast, with the available data it would be impossible to determine how often a gene which we predict to be morbid is, in fact, non-morbid.

Since almost all correlations between our 18 parameters are low (Table 1), none of them is redundant and removing any one from consideration reduces our ability to discriminate between morbid and non-morbid genes (data not reported). Even with all 18 parameters used, our data did not allow us to achieve very low rates of false negative prediction of non-morbidity. If we insist that no more than 1% of known morbid genes are incorrectly predicted to be non-morbid, only 8% of generic genes are identified as non-morbid (Fig. 3), which is not very useful. However, if we accept a 5% false negative rate for non-morbidity predictions, morbidity of a third of generic genes can be rejected.

If the generic set is representative of the whole human genome, it contains ~30% morbid (including early lethal) genes (8–11). This fraction is probably even higher, perhaps as high as ~50%. Indeed, the generic set includes only unambiguous genes, and morbid genes are probably better studied and annotated, since their orthologs are easier to identify. Depending on the proportion of morbid genes within the generic set, classifying 34% of generic genes as non-morbid means that from ~40 to ~70% of non-morbid genes within this set are correctly identified as such, with only 5% of morbid genes misclassified.

A priori information on morbidity of human genes can help to identify genes responsible for a particular disease among multiple candidates. Obviously, genes that are more likely to be morbid should be considered first. Our analysis demonstrates that mutations at one third of all currently known unambiguous human genes are rather unlikely to pose a substantial health risk (Fig. 3) and ignoring such genes leads to only a low risk of overlooking the sought disease-causing gene. Using additional parameters can probably lead to even better discrimination between morbid and non-morbid genes. The relative simplicity of such bioinformatical analysis, in comparison to collecting experimental evidence (15), makes it worthwhile.

REFERENCES
