Published in final edited form as:

Haemophilia. 2017 January ; 23(1): e8–e17. doi:10.1111/hae.13107.

The importance of mRNA structure in determining the pathogenicity of synonymous and non-synonymous mutations in haemophilia

Nobuko Hamasaki-Katagiri^{1,†}, Brian C. Lin^{1,†}, Jonathan Simon^{1,†}, Ryan C. Hunt¹, Tal Schiller¹, Estelle Russek-Cohen², Anton A. Komar³, Haim Bar^{4,*}, and Chava Kimchi-Sarfaty^{1,*}

¹Laboratory of Hemostasis, Division of Hematology Research and Review, Center for Biologics Evaluation and Research, United States Food and Drug Administration, Silver Spring, Maryland, United States of America

²Division of Biostatistics, Center for Biologics Evaluation & Research, United States Food and Drug Administration, Silver Spring, Maryland, United States of America

³Center for Gene Regulation in Health and Disease, Department of Biological, Geological & Environmental Sciences, Cleveland State University, Cleveland, Ohio, United States of America

⁴Department of Statistics, College of Liberal Arts and Sciences, University of Connecticut, Storrs, Connecticut, United States of America

Abstract

Introduction—Mutational analysis is commonly used to support the diagnosis and management of haemophilia. This has allowed for the generation of large mutation databases which provide unparalleled insight into genotype-phenotype relationships. Haemophilia is associated with inversions, deletions, insertions, nonsense and missense mutations. Both synonymous and non-synonymous mutations influence the base pairing of messenger RNA (mRNA), which can alter mRNA structure, cellular half-life and ribosome processivity/elongation. However, the role of mRNA structure in determining the pathogenicity of point mutations in haemophilia has not been evaluated.

Aim—To evaluate mRNA thermodynamic stability and associated RNA prediction software as a means to distinguish between neutral and disease-associated mutations in haemophilia.

*Correspondence: Chava Kimchi-Sarfaty, Ph.D., Division of Hematology Research and Review, FDA | CBER | OBRR, 10903 New Hampshire Avenue, Silver Spring, MD 20993, Chava.Kimchi-Sarfaty@fda.hhs.gov. Haim Bar, Ph.D., Department of Statistics, College of Liberal Arts and Sciences, University of Connecticut, Storrs, Connecticut 06268, USA, Haim.Bar@uconn.edu.

[†]Equal contributor

Authors' contributions

NH-K, BL, JS, HB and CK-S designed the study. NH-K, BL, JS and HB collected the data and performed the analysis. NH-K, BL, JS, HB, RCH, TS, AAK and CK-S contributed to interpreting and presenting the results, drafting the manuscript and editing the manuscript.

Competing interests

The authors stated that they had no interests which might be perceived as posing a conflict or bias.

Methods—Five mRNA structure prediction software programs were used to assess the thermodynamic stability of mRNA fragments carrying neutral vs. disease-associated and synonymous vs. non-synonymous point mutations in *F8*, *F9* and a third X-linked gene, *DMD* (dystrophin).

Results—In *F8* and *DMD*, disease-associated mutations tend to occur in more structurally stable mRNA regions, represented by lower MFE (minimum free energy) levels. In comparing multiple software packages for mRNA structure prediction, a 101–151 nucleotide fragment length appears to be a feasible range for structuring future studies.

Conclusion—mRNA thermodynamic stability is one predictive characteristic, which when combined with other RNA and protein features, may offer significant insight when screening sequencing data for novel disease-associated mutations. Our results also suggest potential utility in evaluating the mRNA thermodynamic stability profile of a gene when determining the viability of interchanging codons for biological and therapeutic applications.

Keywords

RNA prediction software; haemophilia; synonymous mutations; *F8*; MFE; mRNA thermodynamic stability

Introduction

With the molecular underpinnings of haemophilia defined for over three decades, mutational analysis is now commonly performed to support diagnosis and management. In the majority of cases a causative genetic insult is identified, however, the assessment of novel and uncharacterized sequence variants remains challenging. The influence of mRNA structure in determining the pathogenicity of synonymous and non-synonymous point mutations remains incompletely evaluated. The curated collections of disease-associated and neutral point mutations in haemophilia provide the requisite information to reliably evaluate such genotype-phenotype questions. In our previous study [1, 2], mRNA structure/stability was identified as the second most significant variable, among more than 40 assessed RNA and protein characteristics, for distinguishing disease-causing and neutral point mutations in *F8* and *F9*. We previously showed that other variables such as primary sequence/structural conservation scores, as well as changes to charge, polarity, phosphorylation potential and GC content are informative in differentiating between neutral and pathogenic mutations.

Even slight alterations to mRNA folding can alter biological activity [3–5]. Altered stability of mRNA can lead to protein overproduction or underproduction, which has strong implications for the manifestation of human disease [6, 7]. There is now evidence that altering the native structure of mRNA via synonymous mutations may also impact the kinetics of translation and co-translational protein folding. The most common mutation underlying cystic fibrosis, F508 in the *CFTR* gene, includes not only the deletion itself but also a synonymous nucleotide substitution in the adjacent isoleucine codon. This synonymous mutation alters the *CFTR* mRNA structure and contributes significantly to the misfolding of the protein [8]. More globally, others have found that mRNA structure has a prominent influence on ribosome elongation rate. Experimental evidence indicates that

mRNA structure can impact the folding of a protein, specifically that mRNA secondary structures can serve as elongation brakes to control the speed and hence the fidelity of protein translation [9].

Because RNA structure and stability are onerous to assess using *in vitro* methods, *in silico* tools that render possible RNA conformations based solely on the nucleotide sequence provide a convenient medium for studying RNA structure. Computer algorithms, such as those used in mfold [10], KineFold [11] and ViennaRNA [12] have become adept at making large quantities of thermodynamic computations to predict plausible structures. Each of these software programs computes an mRNA thermodynamic stability value in the form of minimum free energy (MFE), a thermodynamic energy measurement that is based on intramolecular stacking and hydrogen bond interactions. MFE determination involves variables such as the system's temperature, entropy, enthalpy and ionic conditions and is representative of a variety of contributing sub-structural elements, such as loops, hairpins and a variety of turns [13]. A lower MFE value is representative of a more stable structure [14, 15]. For a given fragment length, the nucleotide sequence solely determines MFE computation. Longer fragments tend to generate more stacking and hydrogen bond interactions and therefore more stable structures. However, large coding mRNA molecules are often analyzed in arbitrarily determined segments, and these fragment lengths are not standardized among existing software programs.

In this study, we constructed databases of disease-associated and neutral point mutations in the coding region of three genes: the factor VIII gene (*F8*), the factor IX gene (*F9*) and a third X-linked gene, dystrophin (*DMD*), associated with Duchenne/Becker muscular dystrophy. This study uses these clinically relevant datasets with five highly utilized RNA prediction software (KineFold, mfold v3.6, NUPACK v3.0.4, remuRNA and ViennaRNA v2.1.7) to evaluate RNA thermodynamic stability in regions with synonymous and non-synonymous mutations. We observed significant differences between mRNA thermodynamic stability in regions with disease-associated vs. neutral mutations and synonymous mutations vs. non-synonymous mutations. The biological significance of these observations and possible application to recombinant protein design are discussed. Lastly, we evaluated ten unique mRNA fragments sizes (25, 51, 75, 101, 125, 151, 175, 201, 225 and 251 nucleotides (nt)) and demonstrated that each prediction software program has a unique regression pattern. A range of preferable fragment sizes for future *in silico* evaluation of mRNA structure is suggested.

Methods

Database Construction

Disease-associated single nucleotide mutations for coagulation factor VIII and factor IX, coded by *F8* and *F9* genes, respectively, were obtained from updated versions of databases previously described (Hamasaki-Katagiri et al., 2013), originally extracted from the Center for Disease Control and Prevention's (CDC) Haemophilia A and B Mutation Project (CHAMP/CHBMP) databases (<http://www.cdc.gov/ncbddd/hemophilia/champs.html>). Single nucleotide mutations for dystrophin, coded by the *DMD* gene (Dp427m isoform transcript, full length muscular type) associated with Duchenne and Becker muscular

dystrophy, were extracted from NCBI dbSNP (NM_004006.2). Neutral single nucleotide mutations in the *F8* and *F9* genes were extracted from NCBI dbSNP (NM_000132.3 and NM_000133.3).

Only single unique nucleotide mutations in the open reading frame (ORF) were included in the database; multiple SNPs in one allele, insertions/deletions, frameshift, nonsense mutations, and mutations which lack comprehensive genetic or clinical data were excluded. Data from mutations in the 5' and 3' UTR region and non-coding regions were also excluded from the analysis because regions such as the UTR play different roles in translation, and as a result, it is difficult to establish a direct correlation between the point mutation and RNA thermodynamic stability. Other mutation databases were not considered because most data sets do not explicitly categorize mutations. Primary mutation information, such as the position of mutation and nucleotide substitution, mutation type, and functional or clinical information, were obtained through NCBI Genbank and UniProt with entry numbers of P00451, P00740, and P11532 for *F8*, *F9* and *DMD*, respectively. Mutations in the *F8* and *F9* genes of all three severities (mild, moderate and severe) were collectively considered “disease-associated” mutations.

Summary of Point Mutation Databases

F8 (7,056 nt ORF): 1032 disease-associated (10 synonymous), 102 neutral (40 synonymous)
F9 (1,386 nt ORF): 134 disease-associated (3 synonymous), 25 neutral (13 synonymous)
DMD (11,058 nt ORF): 120 disease-associated (all non-synonymous), 28 neutral (13 synonymous)

RNA Folding Software

Five commonly used software programs were used to generate predictions of RNA conformation: KineFold, mfold v3.6, NUPACK v3.0.4, remuRNA and ViennaRNA v2.1.7. For all programs, the default parameters were used when computing the MFE values. For most programs, the default settings consisted of folding with 1 M NaCl and no divalent ions, such as Mg^{2+} and at 37 °C. The only exception was KineFold, for which, to ensure consistency and expediency, the random seed was set to 999, and the folding time to 1000 ms. KineFold's default setting utilizes co-transcriptional folding, but it also provides the option of performing a renaturation fold with 1 M NaCl and cooling from 99 °C to 37 °C (only default settings were used in this study). Summary of these five software programs are listed in Table 1. The three algorithms in mfold, NUPACK, and ViennaRNA all utilize a similar dynamic programming paradigm to enumerate a family of possible RNA structures and then select the most likely structure based on minimizing the predicted structure's free energy. However, this paradigm is forced to ignore all pseudoknotted RNA structures because their inclusion likely results in an exponential increase in both the runtime and storage space. KineFold allows for the inclusion of certain classes of pseudoknots by performing a stochastic simulation of the kinetic folding process. ViennaRNA has an option to consider pseudoknots (RNAPKplex component of the ViennaRNA package). However, this option was not used in this study because the program is generally used as an extension tool for evaluating the accessibility of regions for forming pseudoknotted structures. remuRNA operates differently from the other algorithms as it was created as an ensemble-

based program to compare between structural ensembles which include unique point mutations. In addition to MFE estimation, remuRNA also outputs a corresponding enthalpy value.

MFE Prediction

For each mutation, MFE values were computed for the associated RNA conformations at 10 different sequence lengths: 25, 51, 75, 101, 125, 151, 175, 201, 225, and 251 nt. For each mutation site and sequence length, a subsequence of that length, centered at that mutation site, was extracted. If the mutation site was too close to the head or tail of the mRNA sequence for the subsequence to be properly centered, it was excluded. Each RNA conformation tool was run for both the wild type and mutant variants of the subsequence (these variants differed by only a single nucleotide), and MFE was then calculated as $MFE_{\text{mutant}} - MFE_{\text{wild type}}$. The raw MFE values, outputted by the software programs, were used for all subsequent analysis. Average MFE values for each assessed nucleotide fragment lengths were computed as an average of the wild type MFE throughout the coding region (excluding sequences at the start of the gene sequence due to length restrictions). Average MFE values for each fragment length in the ORF of each gene were plotted for comparison as horizontal green lines in Fig. 1a, 1b, 2a, 2b, Fig S1a and S1b, Fig. S2a and S2b. Although MFE prediction values were generated for all software programs, remuRNA was arbitrarily chosen for analysis on the effect of mutation type and disease association on RNA thermodynamic stability.

Data Processing, Visualization, and Analysis

All data processing was performed using Excel and Python, and all data analysis and data visualization was performed using R and the ggplot2 package. A Python program was written to read in a table of gene mutations alongside a corresponding mRNA sequence, segment the sequence accordingly, pass the segments to the folding software and recombine the outputs into a data table. Box-plot visualizations were produced using the ggplot2 package in R: The horizontal central bar indicates the median; the boxes indicate the interquartile range (IQR); the whiskers indicate within 1.5 IQR of the upper and lower quartile; the dots indicate outliers. Statistical significance of MFE in the region encompassing each mutation type or effect was compared using a 2-sample t-test. In *F8*, this statistical test was performed for synonymous vs. non-synonymous and disease-associated vs. neutral mutations, both with and without the middle domain (amino acid #760-1667). The mutation distribution in *F8* was generated with both disease-associated and neutral mutations at a fragment length of 151 nt. A smoothed baseline MFE in Fig. 1c was generated by averaging the WT MFE of a 56 nt window along the open reading frame.

Fitted Linear Regression

For each protein, we performed a multivariate linear regression of the form:

$$MFE_i = \beta_0 + \beta_1 Length_i + (\beta_2 + \beta_3 Length_i) \cdot I(Mutant_i) + (\beta_4 + \beta_5 Length_i) \cdot I(mfold_i) \\ + (\beta_6 + \beta_7 Length_i) \cdot I(ViennaRNA_i) + (\beta_8 + \beta_9 Length_i) \cdot I(remuRNA_i) \\ + (\beta_{10} + \beta_{11} Length_i) \cdot I(NUPACK_i)$$

where I is the indicator function and β_0 is the MFE intercept. All other β variables within the linear regression equation correspond to the specific intercept value for each software. The model was fitted on points where the length >25 because for short sequence lengths, the trend in the predicted MFEs becomes nonlinear. For further analysis, we calculated the slope and intercept for each mutation sequence in each dataset with fragment length as the x variable and MFE as the y variable. The software algorithms were treated as sources of variation similar to treatments, and the blocks were the mutations. A randomized block ANOVA was used to compare the average slope among the five prediction methods, and then separately, the same analysis was performed for the intercept.

Approximating Inter-software Conversion Factors

In approximating the inter-software conversion factors, the distance between the software's MFE values at a particular fragment length was computed as:

$$\begin{aligned} \text{Software}_1: MFE_1 &= a_1 \times \text{Length} + b_1 \\ \text{Software}_2: MFE_2 &= a_2 \times \text{Length} + b_2 \\ MFE_{\text{difference}} &= (a_2 - a_1) \text{Length} + (b_2 - b_1) \end{aligned}$$

where a indicates the slope, and b represents the intercept value.

This MFE difference corresponds to the conversion factor that must be added to convert from Software 1 to Software 2. The estimated error of this factor was determined by:

$$\text{Error}_{\text{est}} = \frac{|MFE_{\text{Final}} - MFE_{\text{est Final}}|}{|MFE_{\text{Final}}|}$$

Correlation Plots (by Software)

Pearson correlation coefficients were calculated for the average predicted MFE values for each of the 10 sequence lengths for both the wild type and mutant sequences. A correlation score was predicted based on comparing two vectors, each with MFE values predicted by one of the five different software programs. These vectors were then used to generate the Pearson correlation coefficients. All five software programs and 10 different fragment lengths were assessed and plotted in a grid chart using R.

Results

Comparison of Disease-Associated vs. Neutral Mutations

The correlation between disease-associated mutations and MFE levels of mRNA fragments centered on mutation sites were analyzed using datasets of point mutation in the *F8*, *F9* and *DMD* genes. The MFE values for wild type *F8* mRNA in the region encompassing the mutation site, predicted by remuRNA, is shown in Fig. 1a. The MFE values of wild type mRNA fragment regions longer than 25 nt showed a statistically significant difference between disease-associated and neutral mutation sites ($p < 10^{-6} \sim 10^{-11}$). The significance is greater when the fragment size is increased. For all assessed fragment sizes, disease-associated mutations were observed in mRNA regions with a lower MFE compared to

mRNA regions containing neutral mutations. The correlation between mutation effect and the MFE value is apparent from the regression profiles for *F8* and *DMD* genes (dashed lines in Fig. 1 and Fig. S1a). We observed a statistically significant difference between the regression profiles of disease-associated vs. neutral mutations for *F8* ($p < 2 \times 10^{-16}$) and *DMD* ($p = 0.0297$). For *F9*, the MFE values are virtually identical between disease-associated and neutral mutations, suggesting a possible gene-specific relationship between MFE and mutation pathogenicity (Fig. S1b). As one way to evaluate this possibility, we performed an additional analysis of neutral and disease-associated *F9* mutations using KineFold, which employs co-transcriptional folding of mRNA into its algorithm. We observed no significant correlation between MFE and pathogenicity.

Factor VIII is known to have a unique large central region of the protein, encompassing the B-domain, which is cleaved out upon its activation. This region is considered to be dispensable for its function, and a variety of B-domain deleted factor VIII recombinant proteins exist as therapeutic drugs. Most likely, this contributes to a decreased number of entries of disease-associated mutations and a higher proportion of neutral mutations in this region as evident in Fig. 1c. In comparing the MFE of disease-associated mutations vs. neutral mutations with the local baseline subtracted, disease-associated mutations are distributed at lower MFE values than neutral mutations ($p = 0.0143$). We also performed the same analyses but with all mutations in the central region excluded, to account for the potential selection bias in *F8* sequencing data. This similarly showed that disease-associated mutations tend to occur in mRNA regions with lower MFE values relative to the baseline MFE, albeit with a lower defined value of statistical significance ($p = 0.00279$) (Fig. 1b). It should be noted that for both disease-associated and neutral mutations, the MFE between the wild type and mutant sequence is close to 0; thus, the change in MFE resulting from the presence of the mutation does not predict pathogenicity but rather the MFE inherent to the location of the mutation (Fig. S3c and S4c). We also assessed whether the mRNA region encoding for more highly conserved/functionally significant amino acid residues tend to have low MFE values relative to mRNA encoding lowly conserved residues, which could potentially underlie the direct relationship between MFE and pathogenicity. However, the correlation coefficients between amino acid/structural conservation score and MFE (as calculated by remuRNA) were low (0.14, 0.13 and 0.15 for 101, 125 and 151nt, respectively), suggesting other mechanisms for the relationship between MFE and pathogenicity in *F8* (Fig. S7).

Comparison of Synonymous vs. Non-synonymous Mutations

The MFE levels of mRNA fragments surrounding mutation sites were next compared between synonymous and non-synonymous mutations found in the *F8*, *F9* and *DMD* genes datasets. MFE values of *F8* mRNA in the region encompassing the mutation site, predicted by remuRNA are shown in Fig. 2a. The MFE values of wild type mRNA fragment regions longer than 25 nt show a statistically significant difference between synonymous mutations and non-synonymous mutations ($p < 0.09 \times 10^{-4}$). The regression profile for the *F8* gene shows a significant correlation with a p-value of 1.68×10^{-10} . For *F8* and *DMD*, across all fragment sizes, synonymous mutations were observed in mRNA regions with a higher MFE compared to regions where non-synonymous mutations are observed. The average MFE

difference between wild type mRNA fragments and mutant containing fragments is close to 0, further indicating that the mutation, regardless of classification, does not influence the mRNA thermodynamic stability in a consistent manner. Unique regression profiles are still seen between synonymous and non-synonymous mutations when excluding mutations in the middle region of FVIII, albeit with lower statistical significance than the full dataset ($p=0.0223$) (Fig. 2b).

It should be noted that a majority of documented disease-associated mutations are non-synonymous mutations. As a result, the observed differences in local mRNA MFE between synonymous and non-synonymous mutations may be heavily influenced by their status as disease-associated or neutral mutations. To assess mRNA thermodynamic stability in regions with synonymous vs. non-synonymous mutations in a less biased manner, we performed a separate analysis using only neutral mutations in *F8*. In this analysis, 40 synonymous neutral mutations were compared against 62 neutral non-synonymous mutations. Despite the smaller size of this dataset, when comparing the MFE around the neutral synonymous mutations vs. neutral non-synonymous mutations, synonymous mutations show a tendency to be located in mRNA regions with lower MFE values than non-synonymous mutations only when using fragment sizes of 101, 125 or 151 ($p=0.00976$, 0.00633 , and 0.0291 , respectively); however there are no significant differences when using shorter or longer fragment lengths. Moreover, when compared to the baseline MFE values, there are no significant differences between neutral synonymous and non-synonymous mutations ($p=0.0863$). This implies that there is no propensity for synonymous and non-synonymous mutations to occupy stable or unstable regions of mRNA; instead, these tendencies are driven by their pathogenic status. A similar evaluation of *F9* and *DMD* could not be performed due to the small number of synonymous mutations within these datasets. The small number of disease-associated synonymous mutations precludes the ability to make a statistically significant analysis of MFE values between neutral and disease-causing synonymous mutations.

MFE Patterns Across Fragment Lengths

The five prediction tools used here are KineFold, mfold v3.6, NUPACK v3.0.4, remuRNA and the RNAfold program in ViennaRNA v2.1.7, the characteristics of which are defined in Table 1 of the Methods section. Fig. 3 shows the average MFE values of mRNA at all mutation sites within our databases using ten different mRNA fragment sizes, ranging from 25 to 251 nt, predicted by five different prediction tools. The mRNA sequences for the *F8*, *F9* and *DMD* genes, including both disease-associated ($n=1032$, 134 and 120 for *F8*, *F9* and *DMD*, respectively) and neutral mutations ($n=102$, 25 and 28 for *F8*, *F9* and *DMD*, respectively), are constructed so that the mutation site is centered within the fragment length. The five prediction tools show a distinct and consistent trend in their patterns across the various RNA fragment sizes. KineFold and mfold have the lowest MFE values while remuRNA has the highest MFE values (Fig. 3a, 3b). As the length of the mRNA fragments increases, the MFE values decrease at a linear rate. As observed in the previous sections on the association of MFE with disease or synonymous mutations, the average MFE of the mRNA fragments of the wild type (Fig. 3a) and the mutant (Fig. 3b) around all the mutation sites are nearly identical, leading to average ΔMFE ($\text{MFE}_{\text{mutant}} - \text{MFE}_{\text{wild type}}$) values of

zero across the various mRNA fragment sizes (Fig. 3c). Similar trends are observed in the analysis using single nucleotide mutations of the *F9* and *DMD* genes (Fig. S5 and S6).

Establishing Correlations Among Prediction Tools

Plotting the average MFE values against mRNA fragment size shows clear linearity, where the p-values for all plots are less than 10^{-19} , and all R^2 values are greater than 0.84 (Fig. 3d). However, each software program exhibits unique slope and y-intercept values (Table 2).

In all three genes, the average slope across the software programs increases from the most negative slope to the least negative slope in the following prediction tool order: mfold, KineFold, ViennaRNA, NUPACK and remuRNA. Despite the specific differences in the slope and intercept for each prediction tool, these software programs correlate strongly to each other for *F8*, *F9* and *DMD*. Fig. 4 shows that mfold and ViennaRNA, which uses an algorithm modified from that of mfold, are highly correlated; i.e., the Pearson correlation factor is >0.8 . NUPACK and remuRNA also show a similarly high correlation. These results demonstrate a strong inter-software correlation in which the MFE levels from different methods deviate by only a small factor. In both the *F8* and *DMD* genes, the Pearson correlation is greater than 0.8 among all software programs. *F9* shows slightly lower correlations among the software programs, particularly between KineFold and the other four programs. However, the correlation is nonetheless considerably high, with a Pearson correlation greater than 0.7 across all assessed wild type and mutant fragment lengths.

For analyzing the statistical relationships in datasets generated by different software tools, the calculated linear regression statistics for each software program and gene can be effectively used to establish conversion factors between the software programs. For a given fragment length, the conversion from the predicted MFE value of software 1 to software 2 requires adding the MFE difference between the two software programs. This conversion factor must incorporate both the difference between slopes and the intercepts of the two software programs, as shown by the equations in the Methods section.

Discussion

The ability for point mutations to impact protein function outside the context of amino acid substitutions is not commonly considered in disease-association studies [16, 17]. However, any mutation can alter characteristics such as splice sites, miRNA binding, codon usage and translational/folding kinetics of a protein. Both synonymous and non-synonymous mutations may also impact mRNA structure and stability [18, 19].

The primary purpose of this study is to facilitate an understanding of mRNA thermodynamic stability in regions encompassing disease-associated vs. neutral and synonymous vs. non-synonymous mutations relevant to haemophilia. Our analysis of single nucleotide mutations found in the *F8* gene and in the *DMD* gene revealed that disease-associated mutations tend to occur in regions where the mRNA MFE is lower than regions accommodating neutral mutations. This observation can be made despite the unequal distribution of reported disease-associated and neutral mutations, and was further confirmed after accounting for data collection bias. Such a trend was observed in our previous analysis of *F8* mutations in

which a more limited interrogation of mRNA characteristics was performed [1]. Further analysis using only neutral mutations also indicated that non-synonymous mutations are highly associated with disease when they occur in regions where the mRNA MFE is lower than average of neighboring wild type MFE values (baseline). This indicates that mutations occurring in regions where the mRNA structure is more thermodynamically stable tend to be more deleterious, possibly contributing to altered expression, secretion and/or functional levels of the protein. It is to be emphasized that mutations could be deleterious even if the MFE value is not itself changed, depending on local MFE environment within the native mRNA molecule. Other factors such as the speed of elongation or binding of regulatory elements can also promote disease incidence. It cannot be excluded that such mRNA fragments may harbor important regulatory elements that, for example, may serve to recruit various regulatory proteins [20, 21]. In regard to the effects of the mutation on protein function, it is notable that both *F8* and *DMD* code for non-catalytic proteins, which may indicate that these single nucleotide changes affect interaction with other cellular components.

Factor replacement is the mainstay of haemophilia management. Today's recombinant versions of Factor VIII and Factor IX are of increasingly sophisticated design, which can involve the selective incorporation of synonymous and non-synonymous mutations into the coding sequence with the aim of prolonged half-life or increased protein expression yield. In the development of recombinant protein therapeutics, there is question as to whether such mutations and interchanged synonymous codons within the coding sequence could meaningfully alter protein function and structure. All biologic therapeutics undergo significant protein characterization and safety assessment prior to admission into clinical studies. However, product development is occasionally terminated prior to or after the initiation of clinical studies due to unexpected complications, which may include unforeseen effects of nucleotide substitutions initially assumed to be inconsequential.

This study also serves as a platform for assessing results of RNA structural and stability studies. Here, a systematic analysis and comparison of mRNA MFE prediction tools was completed using three genes, five prediction tools and ten RNA fragment lengths, revealing that these tools exhibit a similar linear regression profile across the assessed mRNA fragment sizes. In spite of these tools being developed independently and carrying slight parameter variances, the software used in this study show a high level of agreement in terms of their final MFE determination. Based on utilization of these tools in the assessment of mutation datasets, we have found no qualitative superiority among these five software programs. In considering the feasibility of calculations, which favors smaller fragment lengths, and achieving the highest possible biological correlation (the effect and type of the mutation), we conclude that 101–151 nucleotides is the most realistic mRNA fragment size range for MFE prediction and thus may serve as a plausible starting point for future assessment of point mutations. It is imperative that studies using such tools specify the prediction tool, nucleotide segment length and other relevant settings to allow for subsequent meta-analysis.

Conclusions

In this study, we find that disease-associated mutations tend to occur in more thermodynamically stable regions of mRNA in *F8* and *DMD*, which points to an unappreciated role for mRNA structure in determining pathogenicity of point mutations. This may suggest that interchanging codons or introducing mutations in less stable mRNA regions could be less likely to result in deleterious effects on protein expression or function. Undoubtedly, thermodynamic stability of mRNA is not the sole factor that determines the harmfulness of a given mutation. But our observations suggest that additional information may lie beyond the amino acid level. With the expanding role of bioengineered recombinant protein therapeutics, there may be merit to evaluating mutations included in the expression sequence at both the protein and RNA level. More broadly, this study encourages the continued use of RNA prediction tools to assess uncharacterized synonymous and non-synonymous mutations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This work was supported by funds from the laboratory of Hemostasis, Center for Biologics Evaluation and Research (CK-S), The American Heart Association [13GRNT17070025 to AAK] and The National Institutes of Health [1R15HL121779-01A1 to AAK]. Our contributions are an informal communication and represent our own best judgment. These comments do not bind or obligate FDA.

We are grateful to Dr. Laurence D. Hurst for the insightful discussions. We would also like to thank Mr. John Athey for his assistance in refining our data presentations.

List of Abbreviations

MFE	Minimum Free Energy
F8	Coagulation factor VIII
F9	Coagulation factor IX
DMD	dystrophin (Duchenne Muscular Dystrophy)
ORF	open reading frame
IQR	interquartile range

References

1. Hamasaki-Katagiri N, Salari R, Wu A, et al. A Gene-Specific Method for Predicting Hemophilia-Causing Point Mutations. *Journal of Molecular Biology*. 2013; 21:4023–4033.
2. Salari, R.; Kimchi-Sarfaty, C.; Gottesman, M.; Przytycka, T. Detecting SNP-Induced Structural Changes in RNA: Application to Disease Studies. In: Chor, B., editor. *Research in Computational Molecular Biology*. 7262. Springer; Berlin Heidelberg: 2012. p. 241-243.

3. Pleij CWA, Rietveld K, Bosch L. A new principle of RNA folding based on pseudoknotting. *Nucleic Acids Research*. 1985; 13:1717–1731. [PubMed: 4000943]
4. Russell R. RNA misfolding and the action of chaperones. *Front Biosci*. 2008; 13:1–20. [PubMed: 17981525]
5. Wang X, Lu Z, Gomez A, et al. N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*. 2014; 505:117–120. [PubMed: 24284625]
6. Parmley JL, Hurst LD. How do synonymous mutations affect fitness? *BioEssays*. 2007; 29:515–519. [PubMed: 17508390]
7. Wang D, Johnson AD, Papp AC, et al. Multidrug resistance polypeptide 1 (MDR1, ABCB1) variant 3435C>T affects mRNA stability. *Pharmacogenetics and Genomics*. 2005;15. [PubMed: 15864121]
8. Bartoszewski RA, Jablonsky M, Bartoszewska S, et al. A Synonymous Single Nucleotide Polymorphism in F508 CFTR Alters the Secondary Structure of the mRNA and the Expression of the Mutant Protein. *Journal of Biological Chemistry*. 2010; 285:28741–28748. [PubMed: 20628052]
9. Yang J-R, Chen X, Zhang J. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol*. 2014; 12:e1001910. [PubMed: 25051069]
10. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*. 2003; 31:3406–3415. [PubMed: 12824337]
11. Xayaphoummine A, Bucher T, Isambert H. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Research*. 2005; 33(suppl 2):W605–W610. [PubMed: 15980546]
12. Lorenz R, Bernhart S, Honer zu Siederdissen C, et al. ViennaRNA Package 2.0. *Algorithms Mol Biol Algorithms for Molecular Biology*: 2011. BioMed Central. 2011:1–14.
13. Mathews DH, Sabina J, Zuker M, et al. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of molecular biology*. 1999; 288:911–940. [PubMed: 10329189]
14. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*. 1981; 9:133–148. [PubMed: 6163133]
15. Zuker, M. *Computer Analysis of Sequence Data*. 25. Springer; New York: 1994. Prediction of RNA Secondary Structure by Energy Minimization; p. 267-294.
16. Hunt, R.; Sauna, Z.; Ambudkar, S., et al. Silent (Synonymous) SNPs: Should We Care About Them?. In: Komar, AA., editor. *Single Nucleotide Polymorphisms*. 578. Humana Press; 2009. p. 23-39.
17. Hunt RC, Simhadri VL, Iandoli M, et al. Exposing synonymous mutations. *Trends in Genetics*. 2014; 30(7):308–321. [PubMed: 24954581]
18. Chamary JV, Parmley JL, Hurst LD. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet*. 2006; 7:98–108. [PubMed: 16418745]
19. Presnyak V, Alhusaini N, Chen YH, et al. Codon Optimality Is a Major Determinant of mRNA Stability. *Cell*. 2015; 160:1111–1124. [PubMed: 25768907]
20. Cooper TA, Wan L, Dreyfuss G. RNA and Disease. *Cell*. 2009; 136:777–793. [PubMed: 19239895]
21. Moore MJ. From Birth to Death: The Complex Lives of Eukaryotic mRNAs. *Science*. 2005; 309:1514–1518. [PubMed: 16141059]

Fig. 1. MFE comparison of RNA structure encompassing disease-associated and neutral *F8* mutations

The average MFE values for wild type *F8* mRNA centered around mutation sites predicted by remuRNA are plotted against mRNA fragment size, either (A) across the entire *F8* transcript or (B) excluding the middle domain (amino acid #760-1667). Mutations are plotted according to their effect. The average wild type MFE values for each mRNA fragment size are shown in each plot as horizontal green lines. Linear regression profiles are shown as dashed lines, demonstrating the relationship between the average MFE value of mRNA centered at mutation site and the mRNA fragment size used. Corresponding plots for *F9* and *DMD* are shown in Supplementary Fig. S1. The comparison of average mRNA MFE levels based on mutation type of the *F8* gene across all five software programs is shown in Fig. S3. (C) Mutation distribution profile for the *F8* gene in our database with wild type MFEs predicted at a fragment length of 151 nt and disease-associated mutations shown in red and neutral mutations in blue. The light blue line indicates the MFE baseline values, which were calculated by smoothing the wild type *F8* transcript by 56 nt segments.

Fig. 2. MFE comparison of RNA structure encompassing synonymous and non-synonymous *F8* mutations

The average MFE values wild type *F8* mRNA around mutation sites predicted by remuRNA are plotted against mRNA fragment size, either (A) across the entire *F8* transcript or (B) excluding the middle domain (amino acid #760-1667). Mutations are plotted according to their classification. The average baseline MFE values for each mRNA fragment length are shown for each plot as horizontal green lines. Linear regression profiles are shown as dashed lines, demonstrating the relationship between the average MFE value of mRNA centered at mutation site and the mRNA fragment size used. Corresponding plots for *F9* and *DMD* are shown in Fig. S2. The comparison of average mRNA MFE levels based on mutation type of *F8*. Corresponding plots for *F9* and *DMD* are shown in Fig. S2. The comparison of average mRNA MFE levels based on mutation type of the *F8* gene across all five software programs is shown in Fig. S4.

Fig. 3. Comparison of *F8* mRNA fragment lengths and MFE calculated by five prediction tools
The average of mRNA MFE values around mutation sites predicted by five computational tools were plotted against the mRNA fragment length used. The average MFE values for single nucleotide mutations in the *F8* gene for (A) wild type, (B) mutant and (C) MFE ($MFE_{mutant} - MFE_{wild\ type}$) are indicated. Corresponding linear regression plots for wild type (left panel) and mutant (right panel) are shown in (D).

Fig. 4. Inter-software MFE correlations

The average MFE values for single nucleotide mutations in the *F8*, *F9* and *DMD* genes were evaluated to determine the Pearson correlation between prediction tools. Each dot represents the average MFE values for either wild type or mutant across the ten assessed fragment lengths.

Table 1

MFE Prediction Tools used in this study

Prediction tools (Reference)	Basic principles	RNA fragment size accepted	Condition used in this study
KineFold (Xayaphoummine <i>et al.</i> , 2005)	Stochastic folding algorithm, which includes prediction of pseudoknot and 'entangled' helical structures	< 400 bases for an immediate job, no explicit base limits for batch	Default (including pseudoknot) – co-transcriptional folding
Mfold (Zuker, 2003)	Free-energy minimization for unpseudoknotted structures	< 800 bases for an immediate job, < 9000 bases for batch	Default (37°C)
NUPACK (Zadeh <i>et al.</i> , 2011)	Free-energy minimization for unpseudoknotted structures	< 6000 bases	Default (37°C, 1.0M Na ⁺)
remuRNA (Salari <i>et al.</i> , 2013)	Boltzmann distribution of RNA secondary structure for unpseudoknotted structures; consideration of the relative entropy between structural ensembles	150 base window size suggested but no explicit base limits	Default
ViennaRNA (Lorenz <i>et al.</i> , 2011)	Free-energy minimization, partition function and base-pairing probabilities, which includes the optional prediction of pseudoknotted structures	< 4000 bases	Default (excluding pseudoknot)

Table 2

Linear regression statistics for *F8* mRNA fragments.

<i>F8</i>				
Wild type			Mutant	
Software	Y-intercept	Slope	Y-intercept	Slope
KineFold	5.3287 (0.1431) ^{a, b}	−0.2710 (0.0009)	5.2784 (0.1433)	−0.2705 (0.0009)
mfold	6.7018 (0.1536)	−0.2821 (0.0009)	6.6734 (0.1553)	−0.2816 (0.0009)
NUPACK	3.8642 (0.1363)	−0.2370 (0.0008)	3.8308 (0.1370)	−0.2365 (0.0008)
remuRNA	4.1980 (0.1373)	−0.2257 (0.0008)	4.2183 (0.1389)	−0.2251 (0.0008)
ViennaRNA	6.2718 (0.1585)	−0.2682 (0.0010)	6.2010 (0.1596)	−0.2674 (0.0010)

^a Standard deviations are shown in parentheses.

^b All *p*-values are significant ($< 10^{-19}$), and all R^2 values are greater than 0.84.