UMD-predictor, a new prediction tool for nucleotide substitution pathogenicity – application to four genes: FBN1, FBN2, TGFBR1, and TGFBR2.

Mélissa Yana Frédéric, Marine Lalande, Catherine Boileau, Dalil Hamroun, Mireille Claustres, Christophe Béroud, Gwenaëlle Collod-Béroud

UMD-Predictor, a New Prediction Tool for Nucleotide Substitution Pathogenicity—Application to Four Genes: *FBN1*, *FBN2*, *TGFBR1*, and *TGFBR2*

Mélissa Yana Frédéric,[1,2] Marine Lalande,[1] Catherine Boileau,[3,4] Dalil Hamroun,[1,5] Mireille Claustres,[1,2,5] Christophe Béroud,[1,2,5] and Gwenaëlle Collod-Béroud[1,2*]

[1]INSERM, U827, Montpellier, F-34000 France; [2]Université Montpellier1, UFR Médecine, Montpellier, F-34000 France; [3]INSERM, U781, Paris, F-75015 France; [4]AP-HP, Hôpital Ambroise Paré, Laboratoire de Biochimie, d'Hormonologie et de Génétique moléculaire, Boulogne, F-92100 France; [5]CHU Montpellier, Hôpital Arnaud de Villeneuve, Laboratoire de Génétique Moléculaire, Montpellier, F-34000 France

ABSTRACT: Approximately half of gene lesions responsible for human inherited diseases are due to an amino acid substitution, showing that this mutational mechanism plays a large role in diseases. Distinguishing neutral sequence variations from those responsible for the phenotype is of major interest in human genetics. Because in vitro validation of mutations is not always possible in diagnostic settings, indirect arguments must be accumulated to define whether a missense variation is causative. To further differentiate neutral variants from pathogenic nucleotide substitutions, we developed a new tool, UMD-Predictor®. This tool provides a combinatorial approach that associates the following data: localization within the protein, conservation, biochemical properties of the mutant and wild-type residues, and the potential impact of the variation on mRNA. To evaluate this new tool, we compared it to the SIFT, PolyPhen, and SNAP software, the BLOSUM62 and Yu's Biochemical Matrices. All tools were evaluated using variations from well-validated datasets extracted from four UMD–*LSDB* databases (UMD–*FBN1*, UMD–*FBN2*, UMD–*TGFBR1*, and UMD–*TGFBR2*) that contain all published mutations of the corresponding genes, that is, 1,945 mutations, among which 796 different substitutions corresponding to missense mutations. Our results show that the UMD-Predictor® algorithm is the most efficient tool to predict pathogenic mutations in this context with a positive predictive value of 99.4%, a sensitivity of 95.4%, and a specificity of 92.2%. It can thus enhance the interpretation of variations in these genes, and could easily be applied to any other disease gene through the freely available UMD® generic software (http://www.umd.be).

KEY WORDS: bioinformatics; pathogenicity; missense mutation; prediction tool

Introduction

The most common form of genetic variation in the human genome occurs as a single nucleotide polymorphism (SNP). It is now recognized that at least 10 million SNPs with a minor allele frequency greater than 1% are present in the human genome [Kruglyak and Nickerson, 2001]. Most of these variations are located in intergenic regions, and therefore do not result in any yet known phenotypic variation. However, others are located in coding regions or may affect splicing of various genes, and thus have a direct phenotypic impact. In addition, thousands of genes involved in human genetic diseases have been characterized and hundreds are now available for genetic testing. Thus, thousands of mutations are identified yearly in diagnostic laboratories worldwide. The analysis of these mutations reveals that the most frequent event is a substitution (missense or nonsense mutations as well as mutations affecting splice sites), which account for 67% (35,608 out of 53,200) of cases according to the HGMD database (http://www.hgmd.cf.ac.uk/ac/). New technologies have thus been developed to identify disease-causing mutations and the complete sequencing of a gene is now routinely performed. This approach results in the identification of many variations including SNPs as well as pathogenic mutations. The recognition of these two classes of variations is thus a major challenge of human genetics and diagnostic laboratories. So far, the access to an in vitro functional test to validate pathogenic mutations is restricted to a few genes, and is usually not available in diagnostic settings. In the absence of a functional test, the segregation of the mutation in affected family members (note that this approach does not differentiate a pathogenic mutation from an SNP in linkage disequilibrium with it; in addition, it is only possible if DNA from family members is available), the absence of this variation in a panel of at least 200 independent control chromosomes, the biochemical nature of the substitution, the protein region where the variation is located, and the degree of conservation among species are some of the arguments in favor of a pathogenic mutation. The collection of these data is often both time-consuming and costly. The availability of Locus Specific DataBases (LSDB) now provides valuable information to help in the decision process. Nevertheless, although much effort is dedicated to the collection of mutations in these LSDBs, many families harbor private mutations for which no data are yet available. Various attempts have been made to develop prediction tools that can evaluate the pathogenic potential of a given variation. Predictions regarding missense mutations can be supported by comparative evolutionary analysis to establish whether mutations are situated in conserved genomic regions. Several tools

The first two and last two authors contributed equally to this work.

*Correspondence to: Gwenaëlle Collod-Béroud, INSERM U827, Institut Universitaire de Recherche Clinique, 641 av du doyen Gaston Giraud, 34093 Montpellier Cedex 05, France. E-mail: gwenaelle.collod-beroud@inserm.fr

have been developed to perform this type of analysis. One matrix: BLOSUM62 [Henikoff and Henikoff, 1992] and three programs: SIFT [Ng and Henikoff, 2001, 2003], PolyPhen [Ramensky et al., 2002], and SNAP [Bromberg and Rost, 2007] are known for their accuracy to provide arguments in favor or against causality of nucleotide variations. BLOSUM62 (BLOcks SUbstitution Matrix) is an amino acid substitution matrix based on local multiple alignments of an unselected protein set of related sequences [Henikoff and Henikoff, 1992]. Therefore, position-specific information is lost in the BLOSUM62 matrix. The SIFT program (Sorting Intolerant From Tolerant) [Ng and Henikoff, 2003] uses sequence homology to predict whether an amino acid substitution will affect protein function, and hence, potentially confer a phenotype (http://blocks.fhcrc.org:sift:SIFT.html). SIFT considers the position at which the change occurred and the type of amino acid change. The PolyPhen program (Polymorphism Phenotyping) (http://www.bork.embl-heidelberg.de/PolyPhen/) predicts the possible impact of an amino acid substitution on the structure and function of a human protein using straight forward physical and comparative considerations. Finally, the SNAP program (Screening for Non-Acceptable Polymorphisms) utilizes various biophysical characteristics of the substitution (http://www.rostlab.org/services/SNAP), as well as evolutionary information and structural features to predict whether or not a variation is likely to alter protein function (in either direction: gain or loss). Theoretically determining whether an amino acid substitution is neutral or not is also possible by using physicochemical properties. Yu (2001) has reported an amino acid substitution matrix, "Biochemical matrix," depending on 48 qualitative physicochemical properties describing side-chain structure and functional groups, optical properties, hydrophobicity/charge/acid base properties and size (volume and side-chain length) (http://cmgm.stanford.edu./biochem218/Projects%202001/Yu.pdf).

These tools were successfully used with, in some cases, a high level of confidence. Nevertheless, it remains too low for clinical purposes prohibiting their use for clinical diagnosis [Tchernitchko et al., 2004]. Since the early 1990s, we have been involved in the design of LSDBs and generic tools to build such databases [Beroud et al., 2000, 2005]. We are thus daily confronted with the challenge of predicting the pathogenic impact of nucleotide substitutions. We therefore developed a new tool called UMD-Predictor® to address this challenge. This tool not only takes into account the impact of a nucleotide substitution at the protein level but also at the transcript level. Therefore, it is also able to predict the impact on splicing signals such as acceptor and donor splice sites as well as auxiliary splicing sequences such as Exonic Splicing Enhancer and Silencer (ESE and ESS). In this work, we evaluated the efficiency of this new tool in comparison to other reference tools: SIFT, Polyphen, SNAP, as well as BLOSUM62 and Biochemical matrices. To test these tools, we searched for a well-characterized human set of pathogenic mutations and polymorphisms. We chose international reference LSDBs with a stringent validation process and selected a set of 796 different missense mutations. It includes mutations from the international UMD–*FBN1* LSDB that includes 697 different missense mutations [Collod-Beroud et al., 2003; Faivre et al., 2007], the UMD–*TGFBR2* LSDB (63 mutations) [Frederic et al., 2008a], the UMD–*FBN2* LSDB (16 mutations) [Frederic et al., 2008b], and the UMD–*TGFBR1* LSDB (20 mutations) (data not published). For each gene, the full name of the locus, MIM numbers, and the diseases involved are indicated in Table 1. All these missense mutations are usually private and their pathogenicity has been evaluated in reported publications by indirect arguments such as family segregation and their absence in control populations in the absence of an in vitro test. Thus, the availability of a prediction tool would be of major importance and help in this context.

Materials and Methods

Genes and Mutations

The reference sequences used to describe mutations are: *FBN1* (GenBank NM_000138.3, original sequence L13923), *FBN2* (GenBank NM_001999), *TGFBR1* (GenBank NM_004612.2, encoding the longer isoform), and *TGFBR2* (GenBank NM_001024847.2, original sequence BC040499). All mutations are described using the cDNA numbering system with +1 corresponding to the A of the ATG translation initiation codon in the reference sequence, according to journal guidelines (www.hgvs.org/mutnomen). The initiation codon is codon 1.

Table 1. Genes Studied

LSDB	Gene	Protein	Diseases
UMD-*FBN1*	*FBN1*	Fibrillin-1, MIM# 134797	• Marfan syndrome (MIM# 154700) • Mitral valve prolapse, aortic dilation, skin and skeletal manifestations syndrome (MASS; MIM# 604308) • Mitral valve prolapse syndrome (MVP; MIM# 157700) • Isolated ectopia lentis (EL; MIM# 129600) with relatively mild skeletal features • Weil-Marchesani syndrome (WM; MIM# 266700)
UMD-*FBN2*	*FBN2*	Fibrillin-2, MIM# 134797	• Congenital contractural crachnodactyly (CCA; MIM# 121050)
UMD-*TGFBR1*	*TGFBR1*	Transforming Growth Factor-Beta Receptor type I, MIM# 190181	• Loeys-Dietz syndrome (MIM# 608967) • Furlong syndrome [Ades et al., 2006] • Marfan syndrome [Matyas et al., 2006] • Familial Thoracic Aortic Aneurysms and Dissections [Matyas et al., 2006] • Shprintzen-Goldberg syndrome [Stheneur et al., 2008]
UMD-*TGFBR2*	*TGFBR2*	Transforming Growth Factor-Beta Receptor type II, MIM# 190182	• Marfan syndrome type II [Mizuguchi, 2004] • Loeys-Dietz syndrome (MIM# 610380) • Familial Thoracic Aortic Aneurysm and Dissection (TAAD2) [Pannu et al., 2005] • Human nonpolyposis colorectal cancer (HNPCC6; MIM# 120435)

For each studied gene, the full name of the locus, MIM numbers and associated diseases are indicated.

Table 2. Organisms and Protein Sequences Used for SIFT Analysis

FBN1 Gene	*FBN2* Gene	*TGFBR1* Gene	*TGFBR2* Gene
Homo sapiens (ENSG00000166147)	*Homo sapiens* (ENST00000388849)	*Homo sapiens* (ENSG00000106799)	*Homo sapiens* (ENSG00000163513)
Felis catus (ENSFCAT00000015361)	*Felis catus* (ENSFCAG00000009685)	*Danio rerio* (ENSDARG00000017494)	*Felis catus* (ENSFCAG00000004384)
Tupaia belangeri (ENSTBEG00000003021)	*Danio rerio* (ENSDARG00000051896)	*Oryzias latipes* (ENSORLG00000018380)	*Gallus gallus* (ENSGALG00000011442)
Gallus gallus (ENSGALT00000007955)	*Oryzias latipes* (ENSORLG00000005344)	*Takifugu rubripes* (SINFRUG00000164470)	*Xenopus tropicalis* (ENSXETG00000014480)
Xenopus tropicalis (ENSXETP00000019281)	*Dasypus novemcinctus* (ENSDNOG00000015778)	*Felis catus* (ENSFCAG00000012235)	*Tetraodon nigroviridis* (GSTENG00012314001)
Oryzias latipes (ENSORLG00000002614)	*Myotis lucifugus* (ENSMLUG00000008374)	*Dasypus novemcinctus* (ENSDNOG00000011143)	*Oryzias latipes* (ENSORLG00000014158)
Tetraodon nigroviridis (GSTENT00018079001)	*Gallus gallus* (ENSGALG00000014686)	*Myotis lucifugus* (ENSMLUG00000015003)	*Danio rerio* (ENSDARG00000034541)
Takifugu rubripes (SINFRUG00000160719)			*Takifugu rubripes* (SINFRUG00000120388)
Danio rerio (ENSDARG00000040013)			

For each gene the organism name and the protein reference sequence extracted from Ensembl' are given.

BLOSUM62 Matrix

Each possible amino acid change is assigned a score. Positive scores are associated with conservative changes and negative scores with less conservative changes. As there is no position-specific information, the same matrix was used for all genes and is implemented in each UMD–LSDB.

Biochemical Matrix

Each possible amino acid substitution is assigned a score. A value superior or equal to 0.05 defines a valid or neutral substitution based on physicochemical data. The results are represented as a unique normalized matrix of substitution scores. As for BLOSUM62, the same matrix was used for all genes and is implemented in each UMD–LSDB.

SIFT Matrices

The efficiency of the SIFT software is highly sensitive to the set of sequences used for alignment. Thus, for each gene, analyses have been performed with the alignment of sequences with rational progression (mammals, chicken, frog, and fish) and no more than 90% homology. Chosen sequences are described in Table 2. Protein sequences have been aligned in Clustal prior to analysis by SIFT. Results are reported as "deleterious or not" according to scores (substitutions with less than 0.05 being deleterious). Each gene-specific matrix has been added to the corresponding UMD–LSDB.

PolyPhen Analysis

PolyPhen requests are only available one by one via the Internet. Results are reported as "benign" (when PSIC score difference is ≤ 0.05) and "possibly damaging" or "probably damaging" (when PSIC score difference is > 0.05). An "unknown" status is given when PolyPhen is not able to make a prediction. Contacts with the Webmaster have been necessary to prepare cache files to test large proteins like *FBN1* and *FBN2*.

Snap Analysis

Multiple substitution requests are possible at once. For each tested variation, a reliability index (confidence in prediction) and an expected accuracy are given. Variations are listed as "neutral" or "nonneutral." Contacts with the Webmaster were also needed to prepare cache files to test large proteins like *FBN1* and *FBN2*.

UMD-Predictor®

The predicted secondary structures for each of the 4 genes (*FBN1*, *FBN2*, *TGFBR1*, and *TGFBR2*) were annotated according to module organization from SWISS-PROT (accession numbers P35555, P35556, P36897, and P37179, respectively). In addition, we annotated the sequences for Highly Conserved Domains (HCD). These HCDs are residues for which a strong conservation is associated with a functional and/or a structural key role. The functional/structural arguments are: (1) cysteines involved in disulfide bonds and the correct folding of the protein; (2) amino acids potentially involved in calcium-binding [Dietz and Pyeritz, 1995], calcium affinity controlling stability and rigidity of the microfibrillar structure; (3) glycine implicated in domain–domain packing [Downing et al., 1996]; (4) posttranslational modifications such as N-linked glycosylation, as glycosylation can affect folding, conformation, secretion, stability, and biological activity or even β-hydroxylation; (5) Furine/Pace sites involved in proteolytic processing [Lonnqvist et al., 1998; Reinhardt et al., 1996]; (6) metalloprotease sites (MMP) as matrix-degrading proteinases are crucially important in the remodeling of connective tissues [Hindson et al., 1999]; (7) phosphorylated amino acids [Luo and Lodish, 1997]; (8) amino acids involved in ATP binding; (9) proton acceptor; (10) RGD motif recognized by integrin receptors [Ritty et al., 2003], and (11) highly conserved amino acids of unknown function.

The UMD-Predictor® tool provides a combinatorial approach of several arguments for a given variation. It takes into account its location at the protein level, that is, in which domain and whether the amino acid is involved in a structural or biological function (data from HCD). It checks for the degree of conservation (data from SIFT) and estimates the differences in biochemical properties between the WT and the substituted amino acid (data from BLOSUM62 and Yu's Biochemical matrix). Finally, it is well known that missense mutations can have an effect on mRNA splicing. The pre-mRNA splicing machinery recognizes exons and joins them together to form mRNAs with intact translational reading frames [Shapiro and Senapathy, 1987]. Regulatory

Table 3. Predicted Pathogenicity Evaluation for Mutations Localized in Each of the Four Genes: *FBN1, FBN2, TGFBR1, and TGFBR2*

Gene names	*FBN1*		*FBN2*		*TGFBR1*		*TGFBR2*	
Number of published mutations	1,756		32		33		124	
Number of independent event giving rise to missense mutation[a]	697		16		20		63	
Prediction	No. of correctly predicted mutation	%	No. of correctly predicted mutation	%	No. of correctly predicted mutation	%	No. of correctly predicted mutation	%
SIFT	607	87	9	56	20	100	56	89
PolyPhen	564	81	10	63	19	95	58	92
SNAP	607	87	14	88	19	95	55	87
BLOSUM62	589	85	16	100	11	55	43	68
Biochemical values	523	75	14	88	10	50	36	57
UMD predictor	663	95	16	100	20	100	60	95

[a]Number of missense mutations. Recurrent mutations are counted only once. The number of correctly predicted mutations means that reported pathogenic mutations in the various LSDBs are predicted as pathogenic mutations by the various tools.

elements involved in the splicing machinery are localized in introns or exons such as exonic splicing enhancers (ESEs) or silencers (ESSs). Previously established sequences to score probable ESE motifs of four human SR proteins (SF2/ASF, SC35, SRp40, and SRp55) [Cartegni et al., 2002] were used to analyze the potential loss or gain of ESEs at the site of a missense variation. Another category of exonic mutations that alter splicing is provided by single base pair changes that could either introduce novel splice sites that substitute for the wild-type sites or remove the wild-type splice sites. Consensus values (CVs) of potential donor and acceptor splice sites in the vicinity of missense mutations have been calculated according to the CVs for each nucleotide at each splice site's position using an algorithm derived from Senapathy et al. [1990] and Cartegni et al. [2002] to determine the impact of the variation on splice sites [Beroud et al., 2005].

The UMD-Predictor® tool computes all elements (structure, biochemistry, splicing, conservation) providing for each a specific strength based on their relative impact. For each of the seven elements, values are given based on the predicted impact of the variation (see below). To normalize the predictions on a scale from 0 to 100, the formula also includes an "a" constant value. The pathogenicity of a given variation is thus given by the formula:

$$P = a + \sum_{i=1}^{7} X_{(i,j)}$$

$X_{(i,j)}$ refers to a matrix table with "i" corresponding to the element (1 to 7) and "j" to the UMD value associated with the original element's value. For example: BLOSUM62 original values range from "−4" to "11." The $X_{(i,j)}$ for original value "−4" is "−30," whereas the $X_{(i,j)}$ value for "11" is "+30." The value range for each element has been arbitrarily determined to avoid any bias induced by a trial and error approach using a limited set of data. We used the following rules: key elements such as global conservation (BLOSUM62) and biochemical data (Biochemistry matrices) have the strongest impact (−30 to +30) on prediction as well as HCD (+30), whereas values for secondary elements have a reduced impact. SIFT, which partially overlaps biochemistry and conservation but adds a new level of information (conservation data for a specific protein from various species), has a reduced range of −20 to +20. The effect on auxiliary splicing sequences has a limited impact with a range of 0 to +10. Finally, when a wild-type splice site is abolished, the value was set to +80 to significantly impact the prediction.

From the normalized scale range, we also empirically defined four segments to favor positive prediction: a value of less than 50 is associated with the prediction of a nonpathogenic mutation annotated as "polymorphism," a value of 50 to 64 is associated with the prediction of a "probable polymorphism," a value of 65 to 74 is associated with the prediction of a "probably pathogenic" mutation, while a value above 74 is associated with the prediction of "pathogenic" mutation.

Results and Discussion

Efficiency of the Different Tools and Matrices

To date, more than 1,700 *FBN1* mutations have been published or reported in the UMD–*FBN1* database [Collod et al., 1996; Collod-Beroud et al., 1997, 1998, 2003] among which 1,031 are missense mutations (59.2%) resulting from 697 different mutational events. In our analysis, each recurrent mutation has been evaluated only once. In the same way, 20/33 missense *TGFBR1* mutations and 63/124 *TGFBR2* missense mutations have been analyzed. No recurrent mutation has been reported for the 16/32 *FBN2* missense mutations. To evaluate the efficiency of the UMD-Predictor® tool to detect pathogenic substitutions, we compared it to various approaches including matrices (BLOSUM62 and Yu's Biochemical matrix) or tools (SIFT, PolyPhen, and SNAP).

The first approach involves matrices defined for global studies of the genome. They are based on biochemical properties only (Yu's Biochemical matrix) or with additional conservation (BLOSUM62, SIFT). Only SIFT is influenced by the specific gene sequence. Results from the various tools are presented in Table 3. The second approach involves tools that take into account conservation between species, SWISS-PROT annotation, and 3D structural parameters (PolyPhen and SNAP).

Like the tools from the first category, the UMD-Predictor® tool takes into account biochemical properties and conservation. It also partially involves the 3D structure with the inclusion of the HCD data, and is the only tool that includes the impact of a missense mutation on splicing. Overall, it correctly predicts 95% of *FBN1* and *TGFBR*2 missense mutations and 100% of *FBN2* and *TGFBR1* mutations (Table 3).

Specificity of Prediction for Mutations

The 23 mutations of the *FBN1* gene (described in Table 4) that potentially inactivate the donor splice sites of the corresponding exons were correctly predicted by the UMD-Predictor® tool. Surprisingly, and whereas the other tools do not check for splice site modifications, 17 are correctly predicted by PolyPhen, 18 by SNAP, 19 by HCD, and 17 by SIFT, whereas only 10 and 7 were

Table 4. FBN1 Missense Mutations Potentially Affecting mRNA

Nomenclature c.	Nomenclature p.	Number of records	References	Structure	HCD	Conservation	SIFT probability	BLOSUM62	Biochemical values	SNAP	PolyPhen	ESE modification	Splice size
c.I64G>C	p.GIy55Ala	2	Stheneur [2009]	LTBP-like	conserved AA in LTBP-like	0.75	**0.00**	0.00	0.75	Neutral	**>0.05**		**DSSI [72.73→58.91]**
c.1147G>A	p.Glu383Lys	1	Loeys [2001]	TGFBP#01		1	0.00	1.00	0.54	**Non-neutral**	≤0.05	−SF2/ASF [2.56] −SF2/ASF [2.13]	**DSSI [86.55→73.27]**
c.1468G>T	p.Asp490Tyr	1	Hayward [1997a]	cbEGF-like#03	Ca2+ binding	1	**0.00**	**−3.00**	**0.29**	**Non-neutral**	**>0.05**		**DSSI [79.82→66.18]**
c.1960G>A	p.Asp654Asn	2	Halliday [2002], Katzke [2002]	TGFBP#02	conserved AA in TGFBP	1	**0.00**	1.00	0.75	Neutral	**>0.05**		**DSSI [86.55→73.27]**
c.2113G>A	p.Ala705Thr	2	Ades [1995], Stheneur [2009]	TGFBP#02		1	0.50	0.00	**0.29**	Neutral	≤0.05		**D55I [88→74.73]**
c.2167G>T	p.Asp723Tyr	1	(PC)	cbEGF-like#07	Ca2+ binding	1	**0.00**	**−3.00**	**0.29**	**Non-neutral**	**>0.05**	−SF2/ASF [2.93]	**DSSI [87.45→73.82]**
c.2728G>A	p.Asp910Asn	1	Collod-Beroud [2003]	cbEGF-like#10	Ca2+ binding	1	**0.00**	1.00	0.75	Neutral	**>0.05**		**DSSI [73.27→60]**
c.2728G>T	p.Asp910Tyr	1	Stheneur [2009]	cbEGF-like#10	Ca2+ binding	1	**0.00**	**−3.00**	**0.29**	**Non-neutral**	**>0.05**		**DSSI [73.27→59.64]**
c.2854G>C	p.Asp952His	1	Stheneur [2009]	TCFBP#03	conserved AA in TGFBP	1	**0.00**	**−1.00**	0.58	**Non-neutral**	**>0.05**		**DSSI [77.27→63.45]**
c.3589G>C	p.Asp1197His	1	Stheneur [2009]	cbEGF-like#15	Ca2+ binding	1	**0.00**	**−1.00**	0.58	**Non-neutral**	**>0.05**		**DSSI [91.82→78]**
c.3712G>A	p.Asp1238Asn	1	Yuan 1999	cbEGF-like#16	Ca2+ binding	1	**0.00**	1.00	0.75	**Non-neutral**	**>0.05**		**DSSI [89.64→76.36]**
c.3964G>A	p.Asp1322Asn	1	Stheneur [2009]	cbEGF-like#18	Ca2+ binding	1	**0.00**	1.00	0.75	**Non-neutral**	≤0.05	−SF2/ASF [2.68]	**DSSI [85.82→72.55]**
c.4210G>T	p.Asp1404Tyr	1	Hayward [1997b]	cbEGF-like#20	Ca2+ binding	1	**0.00**	**−3.00**	**0.29**	**Non-neutral**	**>0.05**	−SF2/ASF [2.68] −SRD40 [3.52]	**DSSI [88→74.36]**
c.4582G>T	p.Asp1528Tyr	1	(PC)	TGFBP#04	conserved AA in TGFBP	1	**0.00**	**−3.00**	**0.29**	**Non-neutral**	**>0.05**		**DSSI [86.55→72.91]**
c.5788G>A	p.Asp1930Asn	1	Liu [1997/98]	cbEGF-like#29	Ca2+ binding	1	**0.00**	1.00	0.75	**Non-neutral**	**>0.05**		**DSSI [85.82→72.55]**
c.5788G>C	p.Asp1930His	1	(PC)	cbECF-like#29	Ca2+ binding	1	**0.00**	**−1.00**	0.58	**Non-neutral**	≤0.05		**DSSI [85.82→72]**
c.6313G>A	p.Glu2105Lys	1	(PC)	TGFBP#06		0.75	0.30	1.00	0.54	**Non-neutral**	≤0.05	−SF2/ASF [1.97]	**DSSI [79.82→66.55]**
c.6379G>T	p.Asp2127Tyr	1	Matsukawa [2001]	cbEGF-like#32	Ca2+ binding	1	**0.00**	**−3.00**	**0.29**	**Non-neutral**	**>0.05**		**D5SI [82.36→68.73]**
c.6871G>C	p.Asp2291His	1	Stheneur [2009]	cbEGF-like#36	Ca2+ binding	1	**0.00**	**−1.00**	0.58	**Non-neutral**	≤0.05		**DSSI [94.73→80.91]**
c.7819G>A	p.Asp2607Asn	1	(PC)	cbEGF-like#42	Ca2+ binding	0.75	**0.50**	1.00	0.75	**Non-neutral**	**>0.05**		**DSSI [73.64→60.36]**
c.2677G>A	p.Asp893Asn	1	Loeys [2001]	Hybrid motif#02		1	**0.60**	1.00	0.75	Neutral	**>0.05**		**PDSS [77.27]** **DSSI [86.55→73.27]**
c.3463G>A	p.Asp1155Asn	3	Milewicz [1996], Biggin [2004], (PC)	cbEGF-like#14	Ca2+ binding	1	0.00	1.00	0.75	**Non-neutral**	**>0.05**		**PASS [75.66]** **DSSI [87.45→74.18]**
c.5296G>A	p.Asp1766Asn	1	(PC)	cbEGF-like#25	Ca2+ binding	1	**0.00**	1.00	0.75	**Non-neutral**	**>0.05**		**PASS [74.76]** **DSSI [89.64→76.36]**

PC: personal communication; DSSI: Donor splice site inactivated; PDSS: Potential donor splice site. For each mutation, the nucleotide and protein names are given according to nomenclature guidelines (http://www.hgvs.org/). **CONSERVATION:** The score ranges from 0 for unconserved to 1 for fully conserved. **SIFT** scores that are ≤0.05 are considered to be deleterious (bold text). In **BLOSUM62,** positive scores are associated with conservative changes and negative scores (bold text) with less conservative changes. In **BIOCHEMICAL VALUE**, a value below 0.05 defines an invalid substitution concerning physicochemical properties (bold text). In **SNAP**, variations are listed as "neutral" or "non-neutral". In **PolyPhen**, PSIC score differences above 0.05 define invalid substitutions ("possibly damaging" or "probably damaging"). ESE and splice site modifications are evaluated by the UMD tool. Deleterious modifications are shown in bold text.

Table 5. Prediction for the Five Nonsynonymous Substitutions of the *FBN1* Gene Initially Described as Polymorphisms in Databases But Predicted as Pathogenic by at Least One of the Tools

Mutation name	c.399C>G	c.986T>C	c.1087G>A	C.2656G>A	c.3442C>G
Protein name	p.His133Gln	p.Ile329Thr	p.Gly363Ser	p.Ala986Thr	p.Pro1148Ala
Biochemical matix		+		+	+
SNAP	ND	ND	ND	ND	ND
SIFT		+	+		+
BLOSUM62		+			+
PolyPhen	+		+	+	
UMD predictor		+	+	+	+

Positive symbol = pathogenic mutation; empty boxes are predicted polymorphism; ND = not determined despite many attempts.

Table 6. Sensitivity and Specificity of the Various Prediction Tools and Matrices

	Mutations	Sensitivity	Mutations[a]	Sensitivity[a]	Polymorphisms	Specificity
Biochemical Data	*583/796*	73.2%	*575/767*	75.0%	*NA*	*NA*
PolyPhen	*651/796*	81.8%	*633/767*	82.5%	*NA*	*NA*
BLOSUM62	*659/796*	82.8%	*646/767*	84.2%	*NA*	*NA*
SIFT	*692/796*	86.9%	*669/767*	87.2%	*NA*	*NA*
SNAP	*695/796*	87.3%	*673/767*	87.7%	*NA*	*NA*
UMD predictor	*759/796*	95.4%	*730/767*	95.2%	*4/51*	92.2%

Comparison of sensitivity and specificity of tools and matrices. NA = nonapplicable because of the limited number of available polymorphisms.
[a]Only mutations not affecting splice sites are considered.

predicted by the BLOSUM62 and Biochemistry matrices, respectively. These results underline the limits of predictions based only on tools examining the effect of the amino acid variation at the protein level. In fact, 15 of these amino acids are involved in calcium binding by cb EGF-like modules, and four are highly conserved between species. The corresponding theoretical mutant proteins were therefore predicted to be pathogenic by PolyPhen, HCD, and SIFT, whereas these proteins are probably not produced because of nonsense mediated mRNA decay.

Specificity of Prediction for Variations

To evaluate the specificity of predictions, we analyzed the 50 polymorphisms reported in the *FBN1* database. Among them, 41 are nucleotide substitutions leading to a synonymous change. Only the UMD-Predictor® tool can evaluate these variations and their possible effect at the mRNA level. The nine remaining substitutions result in a nonsynonymous change, and have been predicted as pathogenic or as polymorphic. Overall, PolyPhen, SIFT, Biochemical matrix, and BLOSUM62 predicted as pathogenic mutations 33.3% (3/9), 44.4% (4/9), 11.1% (1/9), and 22.2% (2/9) of reported polymorphisms, respectively (Table 5). Because of its unique design, the UMD-Predictor® is also able to predict the pathogenic impact of synonymous variations and was tested on all polymorphisms. It predicted that 8% (4/50) could, in fact, be pathogenic mutations. Because only the UMD-Predictor® tool is able to perform a prediction for these 50 polymorphisms, the specificity was only assessed for this tool (Table 6).

Because all tools predicted as pathogenic at least one of the nine substitutions previously reported as polymorphisms, we reviewed each variation in order to reassess its status. Among the four *FBN1* polymorphisms reported and predicted as pathogenic mutations by the UMD-Predictor® tool and other tools (Table 5), c.986T>C (p.Ile329Thr) is reported in the dbSNP database as rs363850. This variation has been found neither in the 184 chromosomes from CEPH nor in the 280 chromosomes from the HapMap set of European and Asian populations. It has been found in individual CH9(F) from the OEFNER set and in individual GM19127, a Nigerian female from the Yoruba population used in the HapMap Sub-Saharan African set. This variation was not transmitted to the daughter (GM19129) of this woman (Coriell pedigree Y077). No phenotypic information is available from these samples. This variation does not involve a conserved residue of the *FBN1* protein. The c.1087G>A (p.Gly363Ser) variation is reported in the dbSNP database as rs363855. This variation has not been found in the 418 chromosomes from the HapMap project or in the 120 chromosomes from CEPH. It has been reported in individual PG1137(M) from the OEFNER set for whom no phenotypic information is available. This variation involves a conserved residue from TGFBP module #1. The c.3442C>G (p.Pro1148Ala) variation is also reported in the dbSNP database as rs140598. This variation has been reported with a frequency for the minor G allele of 0.214 from the HapMap–HCB Asian population, of 0.393 from the HapMap–JPT Asian population, of 0.009 from the HapMap–YRI Sub-Saharan African population, of 0.178 from the Autosome population and 0.075 from the MITOGPOP6 population. It has not been reported in the HapMap–CEU European population. This variation does not involve a conserved residue and could therefore be considered as a nondisease variation. Furthermore, it has a value of 65 using the UMD-Predictor® tool, which is the lower limit for variations considered as probable pathogenic mutations. The c.2956G>A (p.Ala986Thr) variation involves a conserved residue from TGFBP module #3. It has been reported by Mine Arslan-Kirchner's team [Rommel et al., 2002]. This sequence variation has been identified in three unrelated cases. The first case was a male subject with suspected Marfan syndrome (mitral valve prolapse, joint hypermobility, but also mental retardation). No other change was identified in the *FBN1* gene by SSCP. The parents were tested and the asymptomatic mother was also shown to carry this variation. The second case was a female subject with suspected Marfan syndrome (dilated aortic root, scoliosis, myopia, long slender fingers). No other change was identified in the *FBN1* gene by SSCP. The sister and the mother were tested and the healthy sister was shown to carry this variation. The third case was a female subject with Marfan

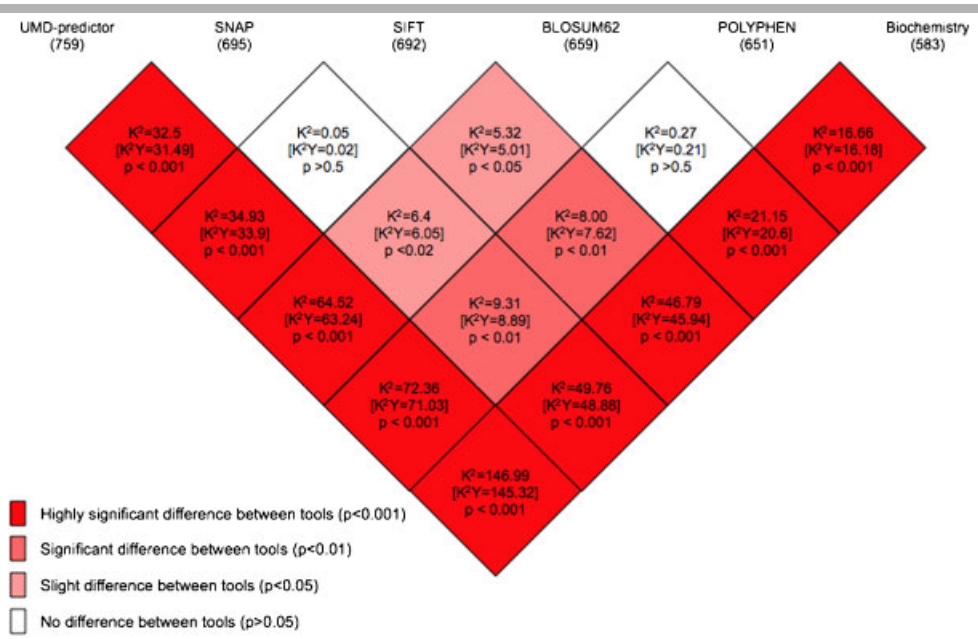

Figure 1. Statistical comparison of the six tools or matrices for the *FBN1*, *FBN2, TGFBR1,* and *TGFBR2* genes. Tools are compared two by two. The Pearson chi-square (K^2) uncorrected for continuity and the Yates chi-square corrected for continuity (K^2Y) as well as *p*-values are given for each comparison. Significance is always given for the left tool in comparison to the right tool. For example, the UMD-Predictor® tool is significantly more efficient than the SNAP (chi value = 32.5 and *p*-value <0.001) or SIFT (chi value = 34.93 and *p*-value <0.001).

syndrome (pectus carinatum, pes planus, joint hypermobility, dilatation of the ascending aorta, spontaneous pneumothorax). One of her relative meets MFS criteria independently. No other change was identified in the *FBN1* gene by SSCP. No relatives were tested. Testing was also performed on 100 controls and the variation was identified in two subjects (data kindly provided by Katrin Rommel and Mine Arslan-Kirchner). All these data are in favor of a nonpathogenic variation.

Conclusion

The prediction of the pathogenic impact of a given missense mutation is one of the biggest challenges for human genetics and molecular diagnostic laboratories. We developed the UMD-Predictor® tool to predict pathogenic substitutions. Our tool uses a combinatorial approach that includes the following data: effect of the amino acid variation at the protein level (i.e., in which protein domain the mutation is located and whether it involves a key residue), conservation, biochemical properties of the mutant and the wild-type residues and the potential impact of the variation on mRNA (search for creation/suppression of potential splice sites or auxiliary splicing sequences). We evaluated its effectiveness at correctly distinguishing pathogenic mutations from polymorphisms in the set of substitutions reported for four genes: *FBN1*, *FBN2*, *TGFBR1,* and *TGFBR2*. The comparison with previously reported tools (SIFT, PolyPhen, and SNAP) as well as frequently used matrices (BLOSUM62 and Biochemistry) revealed that the UMD-Predictor® tool was the most efficient tool to predict the pathogenic impact of missense mutations and substitutions in the context of the four studied genes (*FBN1, FBN2, TGFBR1*, and *TGFBR2)*. It gave statistically better results than all other tools ($p<0.001$) closely followed by SNAP and SIFT that gave similar results, BLOSUM62 and Polyphen, and finally Biochemistry matrix, the less efficient prediction tool (Fig. 1). To avoid any bias based on the specificity of the UMD-Predictor® tool, we removed data from variants affecting splice sites (Table 6). Results were not significantly different, underlining the strength of a combinatorial approach for prediction even without the inclusion of information concerning the impact of the mutation on mRNA.

Overall, UMD-Predictor® has a positive sensitivity of 95.4% and a specificity of 92.2%. The positive predictive value is 99.5% and the negative predictive value 56%. This low negative value is mainly related to the small number of polymorphisms used for this analysis and secondarily to the parameters chosen to define the four segments. The specificity and the positive predictive values are good enough to allow the use of UMD-Predictor® tool results in diagnostic settings for the four genes presented in this study. Thus, its use can enhance the interpretation of missense variation not only within these genes but also within virtually any gene. Additional validations should be performed for other genes coding for structural proteins but also for other proteins. Already data for additional genes such as the DYSF gene involved in various myopathies support results presented here [Sarkozy et al., 2008]. Indeed, in the study of 40 missense *DYSF* gene mutations (for which the pathogenic or nonpathogenic status has been evaluated) 36 out of 40 variations were correctly predicted by the UMD-Predictor® tool with a sensitivity of 85.7%, a specificity of 100%, a positive predictive value of 100%, and a negative predictive value of 75% [Krahn et al., in press]. These data support findings reported in our study. Similar results were also found for germline mutations of the *CDKN2A* gene involved in melanoma-prone families with 13 out of 14 variations validated by various in vitro experiments and correctly predicted by the UMD-Predictor® tool (C. Kannengiesser, personal communication).

These results demonstrate that the combinatorial approach used by the UMD-Predictor® tool gives better results than tools or matrices using only one or a few characteristics. The addition of other annotation criterion such as the identification of critical residues (based on 3D structure or protein function) is currently under evaluation to improve this tool. The full set of predictions for missense mutations of the *FBN1*, *FBN2*, *TGFBR1*, *and TGFBR2*

genes are available through our Web site. Note that the UMD-Predictor® tool is integrated into the UMD® software [Beroud et al., 2005] that is freely available at http://www.umd.be, and could therefore be used to predict substitutions' pathogenicity for virtually any human gene.

Acknowledgments

M.Y.F. and M.L. are supported by a grant from the European Community. This work was supported by grants from AFM (Association Française contre les Myopathies). The research leading to these results has also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 200754—the GEN2PHEN project.

References

Ades LC, Sullivan K, Biggin A, Haan EA, Brett M, Holman KJ, Dixon J, Robertson S, Holmes AD, Rogers J, Bennetts B. 2006. FBN1, TGFBR1, and the Marfan-craniosynostosis/mental retardation disorders revisited. Am J Med Genet A 140A:1047–1058.

Beroud C, Collod-Beroud G, Boileau C, Soussi T, Junien C. 2000. UMD (Universal mutation database): a generic software to build and analyze locus-specific databases. Hum Mutat 15:86–94.

Beroud C, Hamroun D, Collod-Beroud G, Boileau C, Soussi T, Claustres M. 2005. UMD (Universal Mutation Database): 2005 update. Hum Mutat 26:184–191.

Bromberg Y, Rost B. 2007. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35:3823–3835.

Cartegni L, Chew S, Krainer A. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285–298.

Collod G, Beroud C, Soussi T, Junien C, Boileau C. 1996. Software and database for the analysis of mutations in the human FBN1 gene. Nucleic Acids Res 24:137–140.

Collod-Beroud G, Beroud C, Ades L, Black C, Boxer M, Brock DJ, Godfrey M, Hayward C, Karttunen L, Milewicz D, Peltonen L, Richards RI, Wang M, Junien C, Boileau C. 1997. Marfan Database (second edition): software and database for the analysis of mutations in the human FBN1 gene. Nucleic Acids Res 25:147–150.

Collod-Beroud G, Beroud C, Ades L, Black C, Boxer M, Brock DJ, Holman KJ, de Paepe A, Francke U, Grau U, Hayward C, Klein HG, Liu W, Nuytinck L, Peltonen L, Alvarez Perez AB, Rantamäki T, Junien C, Boileau C. 1998. Marfan Database (third edition): new mutations and new routines for the software. Nucleic Acids Res 26:223–229.

Collod-Beroud G, Le Bourdelles S, Ades L, Ala-Kokko L, Booms P, Boxer M, Child A, Comeglio P, De Paepe A, Hyland JC, Holman K, Kaitila I, Loeys B, Matyas G, Niuytinck L, Peltonen L, Rananmaki T, Robinson P, Steinmann B, Junien C, Béroud C, Boileau C. 2003. Update of the UMD-FBN1 mutation database and creation of an FBN1 polymorphism database. Hum Mutat 22:199–208.

Dietz H, Pyeritz R. 1995. Mutations in the human gene for fibrillin-1 (FBN1) in the Marfan syndrome and related disorders. Hum Mol Genet 4(Spec No): 1799–1809.

Downing A, Knott V, Werner J, Cardy C, Campbell I, Handford P. 1996. Solution structure of a pair of calcium-binding epidermal growth factor-like domains: implications for the Marfan syndrome and other genetic disorders. Cell 85:597–605.

Faivre L, Collod-Beroud G, Loeys BL, Child A, Binquet C, Gautier E, Callewaert B, Arbustini E, Mayer K, Arslan-Kirchner M, Kiotsekoglou A, Comeglio P, Marziliano N, Dietz HC, Halliday D, Deroud C, Bonithon-Kopp C, Claustres M, Muti C, Plauchu H, Robinson PN, Adès LC, Biggin A, Benetts B, Brett M, Holman KJ, DeBacker J, Coucke P, Francke U, De Paepe A, Jondeau G, Boileau C. 2007. Effect of mutation type and location on clinical outcome in 1,013 probands with Marfan syndrome or related phenotypes and FBN1 mutations: an international study. Am J Hum Genet 81:454–466.

Frederic MY, Hamroun D, Faivre L, Boileau C, Jondeau G, Claustres M, Beroud C, Collod-Beroud G. 2008a. A new locus-specific database (LSDB) for mutations in the TGFBR2 gene: UMD–TGFBR2. Hum Mutat 29:33–38.

Frederic MY, Monino C, Marschall C, Hamroun D, Faivre L, Jondeau G, Klein H-G, Neumann L, Gautier E, Binquet C, Maslen C, Godfrey M, Gupta P, Milewitz D, Boileau C, Claustres M, Béroud C, Collod-Béroud G. 2008b. FBN2 gene: new mutations, locus specific database (UMD–FBN2) and genotype–phenotype correlations. Human Mutation 30:181–190.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919.

Hindson V, Ashworth J, Rock M, Cunliffe S, Shuttleworth C, Kielty C. 1999. Fibrillin degradation by matrix metalloproteinases: identification of amino- and carboxy-terminal cleavage sites. FEBS Lett 452:195–198.

Krahn M, Béroud C, Labelle V, Nguyen K, Bernard R, Bassez G, Figarella-Branger D, Fernandez C, Bouvenot J, Richard I, Ollagnon-Roman E, Bevilacqua JA, Salvo E, Attarian S, Chapon F, Pellissier JF, Pouget J, Hammouda el H, Laforêt P, Urtizberea JA, Eymard B, Leturcq F, Lévy N. 2009. Hum Mut 30:E345–E375.

Kruglyak L, Nickerson DA. 2001. Variation is the spice of life. Nat Genet 27: 234–236.

Lonnqvist L, Reinhardt D, Sakai L, Peltonen L. 1998. Evidence for furin-type activity-mediated C-terminal processing of profibrillin-1 and interference in the processing by certain mutations. Hum Mol Genet 7:2039–2044.

Luo K, Lodish H. 1997. Positive and negative regulation of type II TGF-beta receptor signal transduction by autophosphorylation on multiple serine residues. EMBO J 16:1970–1981.

Matyas G, Arnold E, Carrel T, Baumgartner D, Boileau C, Berger W, Steinmann B. 2006. Identification and in silico analyses of novel TGFBR1 and TGFBR2 mutations in Marfan syndrome-related disorders. Hum Mutat 27:760–769.

Mizuguchi T, Collod-Beroud G, Akiyama T, Abifadel M, Harada N, Morisaki T, Allard D, Varret M, Claustres M, Morisaki H, Ihara M, Kinoshita A, Yoshiura K, Junien C, Kajii T, Ohta T, Kishino T, Furukawa Y, Nakamura Y, Niikawa N, Boileau C, Matsumoto N. 2004. Heterozygous TGFBR2 mutations in Marfan syndrome. Nat Genet 36:855–860.

Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. Genome Res 11:863–874.

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814.

Pannu H, Fadulu V, Chang J, Lafont A, Hasham S, Sparks E, Giampietro P, Zaleski C, Estrera A, Safi H, Shete S, Willing MC, Raman CS, Milewicz DM. 2005. Mutations in transforming growth factor-beta receptor type II cause familial thoracic aortic aneurysms and dissections. Circulation 112:513–520.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894–3900.

Reinhardt DP, Sasaki T, Dzamba BJ, Keene DR, Chu ML, Gohring W, Timpl R, Sakai LY. 1996. Fibrillin-1 and fibulin-2 interact and are colocalized in some tissues. J Biol Chem 271:19489–19496.

Ritty TM, Broekelmann TJ, Werneck CC, Mecham RP. 2003. Fibrillin-1 and -2 contain heparin-binding sites important for matrix deposition and that support cell attachment. Biochem J 375(Pt 2):425–432.

Rommel K, Karck M, Haverich A, Schmidtke J, Arslan-Kirchner M. 2002. Mutation screening of the fibrillin-1 (FBN1) gene in 76 unrelated patients with Marfan syndrome or Marfanoid features leads to the identification of 11 novel and three previously reported mutations. Hum Mutat 20:406–407.

Sarkozy A, Bushby K, Beroud C, Lochmuller H. 2008. 157th ENMC International Workshop: patient registries for rare, inherited muscular disorders 25–27 January 2008 Naarden, The Netherlands. Neuromuscul Disord 18:997–1001.

Senapathy P, Shapiro MB, Harris NL. 1990. Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. Methods Enzymol 183:252–278.

Shapiro MB, Senapathy P. 1987. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res 15:7155–7174.

Stheneur C, Collod-Beroud G, Faivre L, Gouya L, Sultan G, Le Parc JM, Moura B, Attias D, Muti C, Sznajder M, Claustres M, Junien C, Baumann C, Cormier-Daire V, Rio M, Lyonnet S, Plauchu H, Lacombe D, Chevallier B, Jondeau G, Boileau C. 2008. Identification of 23 TGFBR2 and 6 TGFBR1 gene mutations and genotype–phenotype investigations in 457 patients with Marfan syndrome type I and II, Loeys-Dietz syndrome and related disorders. Hum Mutat 29:E284–E295.

Tchernitchko D, Goossens M, Wajcman H. 2004. In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics. Clin Chem 50:1974–1978.

Yu K. 2001. Theoretical determination of amino acid substitution Groups based on qualitative physicochemical properties. http://cmgm.stanford.edu/biochem218/Projects%202001/Yu.pdf.