Biochemistry & Physiology: Open Access

van den Bergh, Biochem Physiol 2016, 5:1
DOI: 10.4172/2168-9652.1000197

Research Article Open Access

Common Pitfalls and Novel Opportunities for Predicting Variant Pathogenicity

Tom van den Bergh[1,2]*, Bas Vroling[2], Remko KP Kuipers[1,3], Henk-Jan Joosten[1,2] and Gert Vriend[3]

[1]*Laboratory of Systems and Synthetic Biology, Wageningen University, Wageningen, The Netherlands*
[2]*Bio-Prodict, Nijmegen, The Netherlands*
[3]*CMBI, Radboud University Medical Centre, Nijmegen, The Netherlands*

Abstract

The prediction of missense variant pathogenicity is normally performed using analyses of multiple sequence alignments optionally augmented with analyses of the (predicted) protein structure. The most straightforward way, though, is to search the literature to see whether this variant has already been described. Variant data from homologous proteins are also valuable because mutations in a homologous protein often have similar effects as mutations at the equivalent residues of the protein of interest. Transferring variant data seems trivial but is seriously hampered by the fact that homologous residue positions have different numbers in different species. This problem is even bigger when to proteins have such low sequence identities that they can no longer be aligned based on their sequences only and their structures need to be compared to align them accurately. The protein superfamily analysis software suite 3DM solves these problems, because 3DM is a system that combines high quality structure based multiple sequence alignments in which aligned residues have the same number, with all published mutant and variant data for human and all other species. We have used 3DM to analyze nine human proteins for which many disease-related variants are known. This study reveals that mutation data can be transferred even between very distant homologous proteins. Thus, protein superfamily information systems, such as 3DM, offer a wealth of unused information that can be used in the analysis of human variants.

Keywords: DNA diagnostics; Fabry; Long QT syndrome; Protein superfamily; 3DM

Introduction

Rapidly evolving gene sequencing technologies have revealed the relation between mutations in genes and the onset of symptoms that can be assigned to corresponding genetic disorders. More than a hundred thousand unique pathogenic variants are available from the Human Gene Mutation Database (HGMD) [1] that are known to cause over ten thousand different monogenic disorders [2] and almost four thousand genes already have been described that are involved in polygenic disorders [3]. Next generation gene sequencing efforts, on the other hand, have revealed that harmless single nucleotide polymorphisms (SNPs) are even more frequent in the human genome. More than ten million DNA variations have been uncovered in the human genome, of which about 4% are located in gene coding regions and about half of those (2%) are non-synonymous SNPs (nsSNPs) that thus result in an amino acid change in the corresponding proteins [4]. In fact, it was estimated by Crawford et al. [4] that on average each gene contains five nsSNPs that are present in more than 5% of the human population. Only a small fraction of these nsSNPs have an effect on the function of the corresponding protein and can be classified as pathogenic. It is evident that gene sequencing is a promising method for diagnosis of genetic disorders, but the frequent occurrence of benign variants drastically hampers routine diagnosis of genetic disorders.

Mutation databases, such as HGMD, OMIM [5], and protein specific databases, such as for example the P53 mutation databases [6,7], can be used as reference for previously identified pathogenic variants. In the daily practice of DNA diagnostics one obviously encounters many variants for which these databases have no information available yet. In such cases one normally resorts to literature searches or, if that fails, to any of a series of tools, such as Polyphen-2 [8], SIFT [9], and HOPE [10] that have been developed for the *in silico* evaluation or prediction of the effects of variants on gene and protein function. However, in many cases variant effects are known for homologous proteins. The concept of evolution is that all homologs share a common ancestor and thus it makes sense that variants on equivalent positions would cause a similar effect (e.g. cause a disease).

The first step in transferring mutability data between proteins is to identify the equivalent positions. However, to confidently determine which positions are equivalent is difficult unless the proteins in the alignment are so similar that aligning them becomes trivial. Jordan et al. [11] showed that the publicly available mutant severity prediction methods sometimes produce very poor alignments, and thus questionable predictions. Therefore, protein superfamily information systems that use structural alignments are needed to analyze these proteins and enable data transfer between proteins that are more distantly related but still share a common fold.

We have previously described 3DM [12], a system that can generate superfamily systems that fully integrate sequence data with literature data, mutation information, and three-dimensional structures. The 3DM multiple sequence alignments are derived from structure superpositions. This makes them more correct than commonly available alignments, and it allows for larger numbers of sequences to be reliably included in the alignments. 3DM systems contain all available sequence variants (protein- and DNA sequences of all splice variants, with and

***Corresponding author:** Tom van den Bergh, Laboratory of Systems and Synthetic Biology, Wageningen University, Wageningen, The Netherlands and Bio-Prodict, Nijmegen, The Netherlands, Tel: 0031 24 845 7988; E-mail: tvandenbergh@bio-prodict.nl

Received January 05, 2016; **Accepted** January 27, 2016; **Published** February 03, 2016

Citation: van den Bergh T, Vroling B, Kuipers RKP, Joosten HJ, Vriend G (2016) Common Pitfalls and Novel Opportunities for Predicting Variant Pathogenicity. Biochem Physiol 5: 197. doi: 10.4172/2168-9652.1000197

without leader sequences). We have generated 3DM systems for five proteins that are involved in long QT syndrome (gene names: KCNQ1, KCNH2, SCN5A, SCN1A, KCNJ2) and for four members of the amylase superfamily that are involved in Fabry disease, Schindler- or Kanzaki disease, glycogen storage disease, and cystinuria (gene names: GLA, NAGA, GBE1, SLC3A1, respectively).

In this work we show that the mutability of residues can still be transferred even on a superfamily scale where proteins are sequentially very different. The sequence identity between the proteins in our manuscript is as low as 10% and thus these proteins can no longer be aligned by sequence alignment programs. Therefore, protein superfamily information systems that use structural alignments, such as 3DM, are needed to analyze these proteins and transfer data for pathogenicity predictions. Additionally, we show that automated literature mining software can outperform manually curated databases such as HMGD, both in terms of the number of unique mutations extracted, as well as the depth of information per mutation.

Method

3DM information systems

The 3DM software that generates superfamily information systems is extensively described elsewhere [12,13], and will here only be discussed briefly. A structure based multiple sequence alignment (MSA) forms the backbone of each information system. 3DM uses protein structure data to determine which regions are structurally conserved. These regions are termed core regions and 3DM normally uses only these superfamily core regions to generate the superfamily alignments. All sequences and structures are renumbered so that residues aligned in the MSA get the same number throughout the information system. This enables the transfer of data and knowledge between proteins, and facilitates literature searches for mutations in homologs.

Multiple sequence alignments

To predict the pathogenicity of non-synonymous variants the quality of alignments is of much greater importance than the completeness of the MSA. Structure based sequence alignment methods, which are used by default in the 3DM systems, tend to produce alignments of higher quality and deeper coverage than classical MSA methods. However, for the long QT related genes, structure information is available for only small parts of the proteins. To enable high-quality predictions, we implemented a method that aligns only those parts of the sequences that can be aligned with great confidence, similar to the way the PROTOMAT algorithm produces what they call BLOCKS: ungapped regions of aligned proteins [14]. For the parts of the long QT related proteins for which structural information is available, structure-based alignments were produced and were subsequently merged with the sequence-based MSAs.

Mutation data

Mutations were extracted from the literature by the 3DM Mutator module [13]. PubMed was queried for papers containing mutations related to the protein members of the two protein superfamilies here investigated. For the alpha-amylase superfamily Mutator scanned 11,471 full-text papers, whereas mutation-related information for the potassium channel superfamily was extracted from 41,253 full-text articles. In total, this resulted in 5,219 and 65,891 mutations for the alpha-amylase and potassium channel superfamilies, respectively.

Results

We investigated the transferability of several types of information among members of protein superfamilies, and the power and limitations of automatically extracting mutation data from the literature.

Pathogenic variants tend to cluster at equivalent positions

The basic assumption that allows for the transfer of mutation data between protein family members is that mutations at equivalent positions in homologous proteins tend to result in similar structural and functional effects. Based on this assumption it is to be expected that structurally related proteins have equivalent locations where mutations are well tolerated and locations where mutations are prone to result in detrimental effects. We have tested this hypothesis by analyzing the pathogenic variants that were extracted from the literature by Mutator, in the two protein sets. None of the pathogenic variants are observed with a minor allele frequency of 1% or higher in the ExAC population database [15,16]. The first test set consists of 381 aligned pathogenic variants in four human proteins of the α-amylase superfamily (GLA, NAGA, GBE1, and SLC3A1) that, when mutated, can result in Fabry disease, Schindler- or Kanzaki disease, glycogen storage disease, or cystinuria, respectively. These proteins are sequentially very distantly related to each other. Sequence identities of these proteins range between 10% and 57% as shown in Table 1, which makes it (almost) impossible to align them correctly using sequence based alignment method that are normally used by the standard variant prediction tools (e.g. SIFT or PolyPhen [8,9]), but due to their similar protein structures they can still be aligned correctly (see Figure 1). The 381 pathogenic variants are observed at 158 structurally different positions. Figure 2 shows how often pathogenic variants are observed at corresponding positions in these four proteins.

For our second test set, the four potassium channels, structural information is present only for a very small fraction of these proteins, which hampers the transfer of mutation data between these proteins. SCN5A and SCN1A are sequentially closely related and these two can reliably be aligned over nearly the full lengths of their sequences. However, KCNQ1 and KCNH2 can only reliably be aligned at 156 positions that

Protein 1	Protein 2	Core identity	Aligned positions	Mutations protein 1	Mutations protein 2	Overlap	P-value
GLA	SLC3A1	0.15	210	150	37	31	0.046
GLA	NAGA	0.57	209	149	8	8	0.062
GLA	GBE1	0.10	194	140	12	8	0.787
GLA	SLC3A1, NAGA	-	210	150	45	39	0.006
GLA	SLC3A1, NAGA, GBE1	-	210	150	57	47	0.021
KCNQ1	KCNH2	0.17	156	109	126	96	0.001
SCN5A	SCN1A	0.73	1576	476	271	101	0.003

Table 1: The overlap of mutations between different diseases related proteins.

From left to right the columns show the two proteins of the comparison; the sequence identity of the aligned region; the number of aligned positions; the number of positions in both proteins at which pathogenic mutations have been observed, the number of those that overlap, and the p-value for the overlap to be random determined by a permutation analysis.

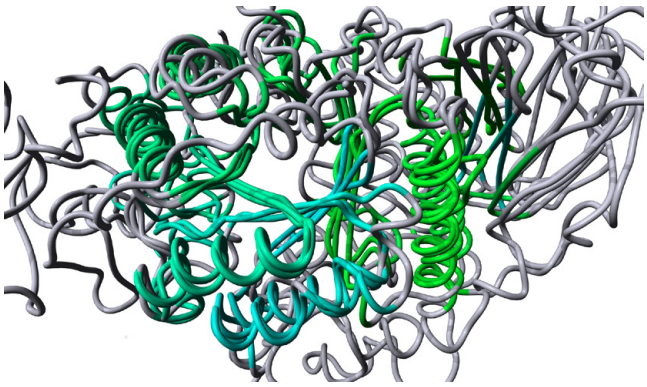

Figure 1: Structural alignment of subfamily representative structures for the four human proteins in the alpha amylase superfamily.

GLA and NAGA are both represented by 1R47 chain B, SLC3A1 is represented by structure 2DH3 chain A, and GBE1 is represented by structure 1M7X chain A. The blue to green regions are considered structurally conserved while the gray regions are structurally variable in this alignment. The blue to green gradient visualizes the order of the conserved regions from the N- to C- terminus respectively.

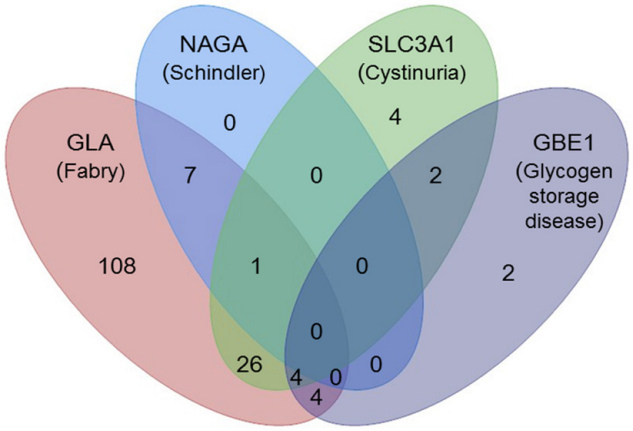

Figure 2: The relation of pathogenic mutations of different homologous proteins and their corresponding diseases.

Overlapping parts of the ovals represent equivalent protein positions and the numbers are the number of positions for which mutation data is detected. For instance, there are 31 positions for which pathogenic mutations have been detected in both GLA and SLC3A1 proteins.

are structurally conserved. This is the transmembrane region of these proteins. No structure data is available for SCN5A and SCN1A and these proteins can reliably be aligned to KCNQ1 and KCNH2 at only 29 of these 156 positions. Due to the absence of structural information and the limited number of mutation data for the four potassium channels, the significance of the overlap of mutation data could only be determined for SCN5A and SCN1A and for the 156 structural conserved positions of KCNQ1 and KCNH2. For these two datasets an even greater overlap is observed of positions that are disease related in both families. Table 1 provides the numerical details of these analyses.

Variation in close homologs is indicative of mutation-tolerant positions

It is commonly accepted that mutations at conserved positions are likely to be pathogenic, and many MSA-based software packages (e.g. SIFT; [9]) that aim at predicting the significance of mutations for a disease state implicitly use this concept. If an alignment consists only of highly similar sequences, then obviously most positions will be observed as conserved. If in a sequence alignment all sequences are more than 90% sequence identical to the human sequence then obviously the MSA contains only sequences from species that are closely related to homo sapiens, and consequently, any variability observed in this MSA is likely to also be acceptable in the human sequence. To test this hypothesis, we compared the ratio of pathogenic variants at conserved positions with the ratio of pathogenic variants at non-conserved positions. To ensure that this test was statistically meaningful we only used two of the nine human proteins (one from each super-family) for which a large number of different pathogenic variants (>250) are available. Table 2 shows that pathogenic variants are less frequently observed at variable positions in alignments composed of only highly similar sequences. For instance, for 179 of the 420 positions in the alignment of closely related GLAs pathogenic variants (stop codons and deletions excluded) have been reported in the HGMD database. The alignment composed of sequences that are at least 90% identical to GLA contains 387 conserved positions. For 176 (45%) of these positions pathogenic mutations have been described. For the variable positions this is five times less, because for only 3 of the remaining 33 variable positions (9%) pathogenic variants have been described. To determine the significance of the lower frequency of pathogenic mutations at variable positions a p-value was determined, which was <0.01 for all factor values from Table 2.

Mutation data extraction: automated mining outperforms manually curated approaches

HGMD [1] is the *de facto* standard source for mutation information. Like any manually curated database, the high quality of the HGMD comes at the cost of incompleteness. It was shown recently that Mutator is able to extend the HGMD [17]. Figure 3 show that Mutator extracts significantly more mutations from the literature than human experts. This does, of course, not invalidate systems like HGMD because the HGMD also provides details on the effect of a variant. Even though natural language parsing software is developing very rapidly the days that the successor of Mutator will also always correctly extract the effects of those mutations from the literature are not near. In contrast to the HGMD database that stores each unique mutation once, 3DM collects all publications that describe any particular mutation. This can, for example, be two publications that describe the same mutation detected in different patients with different onsets or different symptoms. We

	GLA		KCNQ1	
Identity	factor	p-value	factor	p-value
0.9	5.0	<0.001	2.24	0.0045
0.85	4.0	<0.001	2.25	<0.001
0.8	3.5	<0.001	2.07	<0.001
0.75	3.2	<0.001	1.98	<0.001
0.7	2.8	<0.001	2.07	<0.001
0.65	2.8	<0.001	2.00	<0.001
0.6	2.6	<0.001	1.91	<0.001

Table 2: The relation between pathogenicity of mutations at conserved positions versus variable positions.

The left column represent the sequence identity compared to the human sequence. The factor column indicates how much more often pathogenic mutations are found at 100% conserved positions than at variable positions. For instance, using an alignment composed of sequences that are 90% or more identical to GLA this factor is 5.0, which means that the percentage of conserved positions at which pathogenic mutations have been observed is 5.0 times higher than positions at which at least one of the aligned homologs has a different residue type than the human sequence. Clearly, human mutations are more easily tolerated at positions that are variable in highly related sequences.

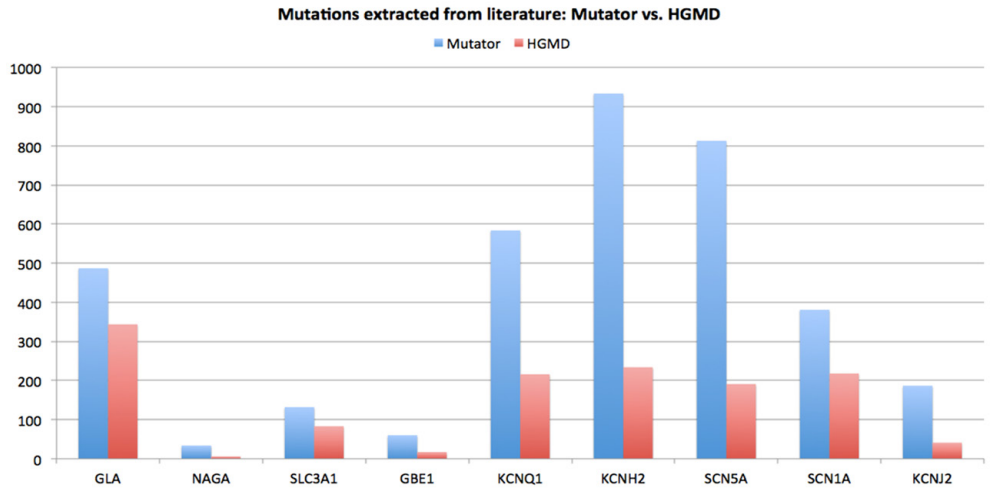

Figure 3: Comparison of the number of unique mutations extracted by Mutator and HGMD for the α-amylase and LQT protein family members.

have frequently observed that contradicting information is reported for the same mutation in different patients.

Discussion

We have made a number of interesting observations. First, we find that variants are more likely to be pathogenic if they occur in structurally conserved regions of a protein [13]. Second, we find that it is much less likely that variants are pathogenic if they occur at positions that are variable in alignments composed of only highly similar proteins. From this observation follows that if a close homolog of a human protein has a different residue, it is more likely that other residue types are allowed at the equivalent position of the human protein. Third, we show that there is a large overlap in alignment positions where pathogenic mutations occur even among distantly related human proteins of a superfamily. Therefore, when a missense mutation is pathogenic to its host organism, the chance that a mutation at the equivalent position in a homologous protein (to any residue) is also pathogenic is much higher. These observations show, as was hinted at previously [10], that pathogenicity is much more determined by the location of the mutation in the protein than by the type of amino acid that is introduced. We can conclude that the use of protein superfamily systems can extensively add previously unused data for the investigation of human disease related variants. These revelations can function as very useful predictive features for variant effect prediction models. The availability of a protein superfamily data integration system is valuable for such a model, since it can provide predictive features that otherwise would be missing, such as mutation data for very distant homologs. In fact, these models have been generated for the LQT related genes. We show that the use of superfamily data largely increases the accuracy of variant effect predictions (publication in progress).

References

1. Stenson PD, Ball EV, Howells K, Phillips AD, Mort M, et al. (2009) The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. Hum Genomics 4: 69-72.
2. World Health Organization (2011) Genes and human disease.
3. Cooper DN, Chen JM, Ball EV, Howells K, Mort M, et al. (2010) Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. Hum Mutat 31: 631-655.
4. Crawford DC, Akey DT, Nickerson DA (2005) The patterns of natural variation in human genes. Annu Rev Genomics Hum Genet 6: 287-312.
5. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA (2000) Online Mendelian Inheritance in Man (OMIM). Hum Mutat 15: 57-61.
6. Béroud C, Soussi T (2003) The UMD-p53 database: new mutations and analysis tools. Hum Mutat 21: 176-181.
7. Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, et al. (2002) The IARC TP53 database: new online mutation analysis and recommendations to users. Hum Mutat 19: 607-614.
8. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248-249.
9. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31: 3812-3814.
10. Venselaar H, Te Beek TA, Kuipers RK, Hekkelman ML, Vriend G (2010) Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics 11: 548.
11. Jordan DM, Kiezun A, Baxter SM, Agarwala V, Green RC, et al. (2011) Development and validation of a computational method for assessment of missense variants in hypertrophic cardiomyopathy. Am J Hum Genet 88: 183-192.
12. Kuipers RK, Joosten HJ, van Berkel WJ, Leferink NG, Rooijen E, et al. (2010) 3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities. Proteins 78: 2101-2113.
13. Kuipers R, van den Bergh T, Joosten HJ, Lekanne dit Deprez RH, Mannens MM, et al. (2010) Novel tools for extraction and validation of disease-related mutations applied to Fabry disease. Hum Mutat 31: 1026-1032.
14. Henikoff S, Henikoff JG (1991) Automated assembly of protein blocks for database searching. Nucleic Acids Res 19: 6565-6572.
15. Exome Aggregation Consortium, Lek M, Karczewski K, Minikel E, Samocha K, et al. (2015) Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv.
16. Song W, Gardner SA, Hovhannisyan H, Natalizio A, Weymouth KS, et al. (2015) Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification. Genetics in Medicine: Official Journal of the American College of Medical Genetics.
17. Stenson PD, Cooper DN (2010) Prospects for the automated extraction of mutation data from the scientific literature. Hum Genomics 5: 1-4.