Types and effects of protein variations.

Vihinen, Mauno

*Published in:* Human Genetics

**DOI:**
10.1007/s00439-015-1529-6

2015

**Link to publication**

**Citation for published version (APA):**
TYPES AND EFFECTS OF PROTEIN VARIATIONS

Mauno Vihinen

Department of Experimental Medical Science, Lund University, BMC D10, SE-22184 Lund, Sweden

Correspondence to Mauno Vihinen, Department of Experimental Medical Science, Lund University, BMC D10, SE-22184 Lund, Sweden, tel. +46 72 526 0022, fax +46 46 222 0296, e-mail mauno.vihinen@med.lu.se
ABSTRACT

Variations in proteins have very large number of diverse effects affecting sequence, structure, stability, interactions, activity, abundance and other properties. Although protein-coding exons cover just over 1% of the human genome they harbor an disproportionately large portion of disease-causing variants. Variation ontology (VariO) has been developed for annotation and description of variation effects, mechanisms and consequences. A holistic view for variations in proteins is made available along with examples of real cases. Protein variants can be of genetic origin or emerge at protein level. Systematic names are provided for all variation types, a more detailed description can be made by explaining changes to protein function, structure and properties. Examples are provided for the effects and mechanisms, usually in relation to human diseases. In addition, the examples are selected so that protein 3D structural changes, when relevant, are included and visualized. Here, systematics is described for protein variants based on VariO. It will benefit the unequivocal description of variations and their effects and further reuse and integration of data from different sources.

Keywords: Variation Ontology; VariO; variation annotation; mutation; protein variant; effects, consequences and mechanisms of variants
INTRODUCTION

Variations manifest their effects at different ways. Of the identified disease-causing cases, a large proportion appears in protein coding DNA and RNA sequences, although the protein-coding regions constitute only about 1.3% of human genome. Due to the large number of different functions in which proteins are involved, also the effects of variants are widely different. In addition to functions, protein variants can change structures, properties, interactions and other characteristics in several ways. List of the effects would be really long and include e.g. interactions, stability, electrostatic effects, protein packing, local and global structural changes and so on.

Many deleterious variants are straightforward to explain, such as large deletions, protein truncations, amphigoric amino acid insertions, deletions and indels, and many other types. Usually, the most difficult ones are minor alterations, most often amino acid substitutions. The effects on proteins are more difficult to study experimentally than those on DNA or RNA. Hybridization and sequence complementarity-based methods work (almost) equally well for all nucleotide sequences, whereas a dedicated analysis protocol is needed for every individual protein and often even for individual variants [see e.g. (Perniola and Musco 2014; Storz and Zera 2011; Yates and Sternberg 2013)].

Experimental methods are the first choice to study the detailed effects of variants; however, these kinds of data are often missing. In such cases, bioinformatics predictors may be useful (Kucukkal et al. 2014; Thusberg and Vihinen 2009; Zhang et al. 2012). Several prediction methods have been developed based on e.g. evolutionary information, protein physical and chemical features, and Gene Ontology annotations to detect function affecting or harmful variants (Calabrese et al. 2009; Kircher et al. 2014; Li et al. 2009; Niroula et al. 2015; Olatubosun et al. 2012; Schwarz et al. 2014) or those having more specific effects such as in aggregation (Conchillo-Sole et al. 2007; Maurer-Stroh et al. 2010; Trovato et al. 2007) or stability (Capirotti et al. 2005; Cheng et al. 2006; Yin et al. 2007) or affecting specific proteins such as mismatch repair system (Ali et al. 2012), protein kinases (Izarzugaza et al. 2013) or hemophilia-related proteins (Hamasaki-Katagiri et al. 2013). The majority of these kinds of methods are dedicated for amino acid substitutions. Assessments of performance of these tools indicated that tolerance (Thusberg et al. 2011), protein stability (Khan and Vihinen 2010), protein disorder (Ali et al. 2014) and localization (Laurila and Vihinen 2011) predictors, and likely also other kinds of methods, which have not yet been assessed, have widely varying performance.

Three-dimensional structures of proteins are very helpful for interpretation of the effects and mechanisms of variations. If experimental structures obtained with X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy are not available, it may be possible to model the protein structure. Retrospective analysis of computer models indicated that biological and medical inferences can be quite reliable even for difficult modeling cases with low sequence identities (Khan and Vihinen 2010).

Statistical studies of variation types and protein structures have provided useful insight. Disease-causing variants have a clearly distinct amino acid distribution compared to benign variants (de Beer et al. 2013; Schaafsma and Vihinen 2014; Steward et al. 2003; Vitkup et al. 2003), also in protein secondary structural elements (Khan and Vihinen 2007). Protein stability is frequently altered due to variations (Ferrer-Costa et al. 2002; Wang and Moult 2001; Yue et al. 2005).
The majority of the harmful variants are loss-of-function variants; however, gain-of-function variants are identified at increasing rate. Usually the latter ones have constitutive activity and have harmful effects due to not being regulated.

All the experimental and computational approaches mentioned above can be combined to explain and annotate variations and their consequences. In this article, effects of protein variations are discussed mainly in relation to diseases and in the framework of the Variation Ontology (VariO) (Vihinen 2014a), which allows systematic description of variation effects, consequences and mechanisms whether of experimental or of predicted origin. Examples are included to highlight the different features of variants. Some previous studies have reviewed protein variation effects, mainly in relation to protein three-dimensional (3D) structures (Ferrer-Costa et al. 2002; Furnham et al. 2012; Stefl et al. 2013; Steward et al. 2003; Wang and Moult 2001; Yue et al. 2005), but without systematic coverage and systematics.

**DATABASES FOR PROTEIN VARIATIONS**

Several freely available and commercial databases contain information about proteins and their variants. Swiss-Var (http://www.swissvar.expasy.org) at UniProtKB/Swiss-Prot (http://www.uniprot.org/) contains notions of disease-causing variants. Variation information is available in locus-specific databases (LSDBs), which are collected for individual genes/diseases or groups of them, or from central databases such as the Human Gene Mutation Database (HGMD, http://www.hgmd.org/ or the commercial version with up-to-date information at http://www.biobaseinternational.com/), dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) and ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/). LSDBs can be searched either from the Human Genome Variation Society web site (http://www.hgvs.org/dblist/glsdb.html), the LOVD listing (http://grenada.lumc.nl/LSDB_list/), or from the GEN2PHEN project (http://www.gen2phen.org/data/lsdbs). In addition, WAVE lists databases and also (partial) lists of variants in them (http://www.bioinformatics.ua.pt/WAVE/). Protein 3D structures are collected to the Protein Data Bank (PDB, http://www.rcsb.org/pdb/). An example of an effect-specific database is ProTherm (http://www.abren.net/protherm/) for protein thermodynamic changes upon variation often for other organisms than human and in many cases including engineered variations.

**VARIATION ONTOLOGY**

To allow efficient use, reuse, search and integration of information it is essential to describe it in a systematic way. This applies also to variation information. Recently, VariO (http://www.variationontology.org/) was introduced for the systematic description of variation effects, consequences and mechanisms (Vihinen 2014a). The ontology is used to annotate information in databases at the three molecular levels: DNA, RNA and protein. Each of these levels contains further terms for variation type, function, structure and various properties. The discussion of variation effects will be made based on VariO. VariO annotations are always made in relation to a reference state, e.g. reference sequence or wild-type enzyme activity.

VariO annotations consist of two parts including the VariO prefix and a number such as VariO:0012 and name, in this case “protein variation type”. The number with the prefix is mandatory for annotations. This article is organized according to the VariO and is thus divided into the four major sublevel variation types, protein function, structure and properties. VariO terms are used in the subtitles and otherwise written in quotation marks. For the sake of
clarity, in this article, the prefix and the number have been omitted. Detailed guidelines are available for VariO annotations (Vihinen 2014b). Annotations for databases can be made with VariOtator annotation tool. For the examples in the text, full VariO annotations together with Evidence Code (ECO, http://www.evidenceontology.org/) annotations are available at http://www.variationontology.org/annotation/protein_examples. The number of databases utilizing VariO annotations is increasing, including UniProtKB/Swiss-Prot (Famiglietti et al. 2014).

**PROTEIN VARIATION TYPE**

Variation type in VariO provides a description for a variation in English (see Fig. 1). Human Genome Variation Society (HGVS) nomenclature provides systematic description of variations (den Dunnen and Antonarakis 2001); however, these names may be difficult to interpret. For example p.(K28_K29delinsW) means predicted deletion of two lysines (K28 and K29) and their replacement by a single tryptophan without experimental evidence at protein level. Variation type terms of VariO provide a brief description with commonly used terms and are not intended to replace HGVS names, instead to provide easily understandable description for human readers and computer applications.

There are two types of “protein variation origin”, namely “protein variation of genetic origin” and “variation emerging at protein level”. Variants of genetic origin have appeared on DNA (or RNA) level and therefore directly affect the protein. An example is R525Q substitution in Bruton tyrosine kinase (BTK) causing X-linked agammaglobulinemia (XLA) (Vihinen et al. 1994), which is caused by g to a transition in the coding gene. The variations emerging at protein level are one of four types. They are either “artificial protein variation”, “epigenetic protein variation”, “mistranslated protein” or “post-translational modification”. Phosphorylation of BTK tyrosine kinase domain at Y551 is a “post-translational modification” (Mahajan et al. 1995). Post-translational modifications are frequent; however, only a few of the 600 known modifications in RESID database (http://www.pir.georgetown.edu/resid) are frequent and not all of them appear in human.

“Mistranslated proteins” due to editing-defective tRNA synthetase cause neurodegeneration because of misfolding protein (Lee et al. 2006). The epigenetic variations include “protein structural inheritance” and “proteinaceous infection” of prions. Structural inheritance with unknown mechanism has been noted in centrosomes, organelles which are the main microtubule-organizing center and regulator of cell cycle progression (Wilson 2008). Asymmetric cell division or centrosome inheritance may have effect on cancer (Izumi and Kaneko 2012). Prions are discussed in detail later in the text.

Protein engineering has been widely used in an attempt to modify properties of proteins, especially those having commercial utility e.g. to alter their stability or to increase activity or specificity. A prime example of an engineered protein is subtilisin, a protease that digests proteins, properties of which have been extensively modified with random and targeted mutagenesis [see (Bryan 2000)]. Engineering of the NEMO protein at N-terminus affects thermal stability and ligand binding (Guo et al. 2014) is an example of “artificial protein variation”.

“Protein variation classification” in VariO divides the variants into “amino acid insertion”, “amino acid deletion”, “amino acid indel”, “amino acid substitution” or “missing protein”. All the following variants in BTK lead to XLA: R525Q “amino acid substitution” (Vihinen et al. 1994), deletion of 28 amino acids from SH3 domain and a linker region cause “amino acid deletion” (Zhu et al. 1994), and “amino acid insertion” at S604 (Holinski-Feder et al. 1998). Insertions and indels can be of two types, either the sequence is retained after the
insertion/indel site or it has changed, in which case they are called as amphigoric variations. “Protein truncation” is a special case of deletion where deletion occurs either in N- or C-terminus. Indels originate due to both insertion and deletion.

The different types of protein variations in a short protein sequence for alanine, alanine, serine... in one letter code AASEQWENCE are depicted in Fig. 2. There are altogether 8 categories of variation types. “Sequence retaining amino acid deletion” in Bruton tyrosine kinase (BTK) deletes, due to skipping of exon 8, altogether 21 amino acids from the C-terminus of Src homology 3 (SH3) domain and linker connecting to SH2 domain (Zhu et al. 1994). The protein is produced but non-functional and is a cause for XLA, a rare primary immunodeficiency, due to block in B cell maturation. The truncated domain alone does not fold correctly (Chen et al. 1996). “Amphigoric amino acid insertion” because of a single-base insertion to the BTK gene region encoding PH domain changes the sequence after the insertion site at E59 and creates new premature stop codon truncating the modified C-terminus of the protein (Holinski-Feder et al. 1998). The truncated protein has no activity and cannot be detected.

ClC5 variations lead to Dent’s disease, X-linked disorder with low molecular weight proteinuria, hypercalciuria, nephrocalcinosis, nephrolithiasis, and often also renal failure. One of the variants, R718X, deletes 28 C-terminal amino acids (D’Antonio et al. 2013). Although the truncated protein is produced, it has enhanced protease susceptibility suggesting an at least partly misfolded and destabilized protein. This variant is a “protein truncation”.

Eight amino acid “sequence-retaining amino acid indel” in SERPINC1 is a complex variant that consists of a two amino acid deletion, a seven residue inserted repeat and a three amino acid insertion (Martínez-Martínez et al. 2012). It causes type II antithrombin deficiency due to missing serine protease inhibitory activity. Patients suffer from impaired blood clotting when having vascular damage. The indel appears in the highly conserved β-strand forming part of a β-sheet (Martínez-Martínez et al. 2012). The hyperstable form of the variant has a new β-strand that displaces the reaction center loop (Fig. 3a) thereby inactivating the protein. A “sequence-retaining amino acid insertion” of two residues in the BTK kinase domain at position 604 (Holinski-Feder et al. 1998) affects α-helix G and causes XLA (Mao et al. 2001).

“Amino acid substitutions” describe changes of individual amino acids. These are typically caused by nucleotide substitutions at DNA, and called for missense variants at RNA level. “Missing protein” describes a situation where no protein is produced. This can be due to e.g. nonsense mediated decay (NMD) destroying RNA molecules containing premature stop codons, or variants to start codons preventing translation. The E31K substitution and R41X truncation in Wiskott–Aldrich syndrome (WAS) protein (WASP) are examples of “missing protein” (Jin et al. 2004). WAS is a rare recessive X-linked primary immunodeficiency. WAS patients suffer from thrombocytopenia, eczema and diarrhea in addition to immunodeficiency.

VARIATION AFFECTING PROTEIN FUNCTION

Protein functions are presented in Fig. 4. Changes to the function are divided into seven categories. “Effect on protein catalytic function” describes changes to enzyme activity. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology EC numbers (http://www.chem.qmul.ac.uk/iubmb/enzyme/) can be used for more detailed description of affected enzyme activity. The R525Q substitution at the catalytic site inactivates BTK kinase catalytic function (Vihinen et al. 1994). “Effect on protein information transfer” can be used to describe e.g. hormone effects. Insulin is a peptide
hormone in which variants lead to for example neonatal diabetes due to impaired regulation of carbohydrate and fat metabolism (Støy et al. 2007).

“Effect on protein movement” is used to describe changes to movement of contractile proteins such as actin and myosin. Several myopathies with muscle weakness and muscle structural changes emerge due to some 200 known variants in α-actin (Nowak et al. 2013). These diseases are related to muscle weakness and skeletal muscle structural lesions. “Effect on protein recognition” term describes selective and non-covalent interactions of proteins e.g. for immunoglobulins and other immunological recognition molecules. Insertion to immunoglobulin gene for common γ2 chain (Cγ2) causes IgG2 deficiency (Tashita et al. 1998). The patients are susceptible for sinopulmonary infections.

“Effect on protein storage” describes changes to proteins acting as amino acid or metal cation stores, such as ferritin and casein. Ferritin light chain variants lead to autosomal dominant neuroferritinopathy due to iron deposition to brains (Lehn et al. 2012). The accumulation of iron to brain causes adulthood disease with slow progression. “Effect on transport function of protein” can be used to annotate changes that affect active transport functions of proteins such as membrane transporters. G87R substitution in zinc transporter ZnT-2 (SLC30A2) impairs zinc transport because the protein is mislocalized and has decreased stability (Lasry et al. 2012). The variant has ER retention. The disease affects many cellular processes as Zn2+ is required for activity of several enzymes. Infants are especially vulnerable for Zn deficiency, as the prolonged metal deprivation causes growth retardation and neurophysiological changes.

“Effect on structural protein” defines alterations in proteins providing support for cells and tissues including keratin, collagen and others. Variations to type II collagen affect the structural protein of connective tissue and result in conditions with varying severity ranging from lethal to relatively mild (Kannu et al. 2012).

The terms for protein function are for fundamental protein activities in cells. One protein may have several functions. The term describing the function(s) affected by the variation should be used for annotation.

**VARIATION AFFECTING PROTEIN STRUCTURE**

Folding is a prerequisite for activity of most proteins. Structural changes due to variations have a wide range from minor local alterations to drastic global changes. Structure terms are first divided into two categories, those having “effect on protein 3D structure” and those of “epigenetic protein modification” (Fig. 5). The epigenetic modifications include “epigenetic protein complex structure” and “prion formation”.

The effect on protein 3D structure is further divided into terms “complex 3D structural change”, “effect on protein dynamics”, “effect on protein quaternary structure” and “effect on protein tertiary structure”. Terms for protein dynamics cover those affecting allosteric site, induced fit, structural disorder, and structural flexibility. The largest number of terms relates to tertiary structure including e.g. amino acid size, interaction site, secondary structural element and protein folding. Structure and property terms in VariO can be modified by attributes to provide more detailed descriptions. ECO terms are used to indicate the evidence for the annotations. VariO website includes examples for each of the terms along with ECO annotations.

*Effect on protein dynamics*

Proteins are dynamic molecules and in constant motion but the basic scaffolding usually remains the same during the fluctuations. Variations can affect normal dynamics in several
ways. M257Y is a constitutively active variant of G-protein-coupled receptor (GPCR) that affects the dynamics of structural alteration between active and inactive conformation (Tsukamoto and Farrens 2013). The transition is essential for the protein function. Protein with the gain-of-function variant is in the active conformation even in the absence of the activating external signal.

P275L substitution in phenylalanine hydroxylase (PAH) has an “effect on allosteric site” regulation by cofactor (Gersting et al. 2008) causing phenylketonuria (PKU), a metabolic disorder, which, if untreated, leads to intellectual disability, seizures and other problems. PAH is needed to degrade amino acid phenylalanine to tyrosine. The variant has decreased affinity for tetrahydrobiopterin, the cofactor, thereby disturbing allostery. The substrate affinity, however, is even increased. PKU is the most common inborn error in amino acid metabolism in European-descent populations.

Apurinic/apyrimidinic endonuclease 1 (APE1) is required for DNA damage correction by base excision and nucleotide incision repair mechanisms. The enzyme–substrate complex undergoes induced fit upon binding to lesion. Substitution K98A in APE1 has an “effect on induced fit” preventing it and decreasing the stability of enzyme–substrate complex (Timofeyeva et al. 2011).

Protein structures typically fold to ordered 3D structures. Exceptions are intrinsically disordered proteins and regions. Disordered regions can adopt diverse temporary conformations and have large numbers of binding partners (Oldfield et al. 2008). An example is tumor suppressor p53, which contains large disordered regions that interact with several different partners. Variations at p53 are hallmarks in numerous cancers. “Effect on structural disorder” status is related to certain diseases. Misfolded proteins have four possible fates in cells. They are either degraded, refolded, they form aggregates or they are sequestered. Single amino acid replacement D54G in creatine kinase leads to muscle creatine kinase deficiency (Feng et al. 2007). The variant has partially misfolded structure, and although the protein is produced, the activity is impaired. Creatine kinase catalyzes the production of ATP by transferring phosphate group from phosphoserine. Muscles, especially heart, consume lots of high-energy ATP molecules.

Structural flexibility is essential for many protein functions (Vihinen 1987). Laforin regulates cellular autophagy. Harmful variations are responsible for progressive form of myoclonus epilepsy called Lafora disease. K87A substitution in laforin has reduced stability due to “effect on structural flexibility” (Srikumar and Rohini 2013). Laforin has two activities, protein dual specificity phosphatase and glycogen binding. K87A eliminates the glycogen-binding activity without effect on phosphatase activity. This is because the activities are in different domains.

Effect on quaternary structure

Protein quaternary structures are formed of interacting subunits. Variations with “effect on quaternary structure forming interaction” can be harmful by preventing or weakening complex formation. Congenital cataract-causing A2V substitution in the eye lens βB2-crystallin folds normally and forms βB2/βA3-crystallin dimer; however, it prevents tetramer formation, the protein complex in eyes (Xu et al. 2012). The tetramer structure is shown in Fig. 3b. The patients have from birth cloudy of the lens of the eye.

Effect on tertiary structure

This is the largest category for protein structural variations depicting the numerous ways protein 3D structures can be altered.
Effect on amino acid size

Cores of globular proteins are typically tightly packed and therefore amino acid substitutions by larger or smaller side chains can be harmful. “Effect on protein packing” can be due to “protein overpacking” or “protein cavity formation”. An example of the latter is I2A substitution of invariant residue in chain A of insulin, which generates a cavity into the hydrophobic protein core and leads to reduced thermal stability (Xu et al. 2002). The structure of insulin is shown in Fig. 3c both for the normal and the variant form. Overpacking appears when the variant is larger than the original residue which leads to sterical clashes and at least local structural change. Figure 3d shows a situation where the larger side chain of G584 to W variant (Futatani et al. 2001) does not fit into the core of the BTK kinase domain structure without structural alterations due to clashes with a large number of amino acids.

Effect on protein folding rate

Protein folding is a complex process, the rate of which can be altered by variations. Even minor amino acid substitutions including Y25F and V68A have major effects to the folding rate of muscle acylphosphatase (AcP) (Chiti et al. 1999). AcP is a widely studied model because it folds slowly and is thus easy to investigate.

Effect on protein interaction site

Proteins function via interactions with other molecules. Variations affecting interactions can be described very detailed using VariO attributes.

Type I interferons (IFNs) are cytokines having immunomodulatory, antiviral and antiproliferative activities. Substitution R120A and variations at surrounding amino acids in the α-helix D in human IFNα2 have “effect on protein-binding site” by decreasing interaction with receptor subdomain IFNAR-1 while the affinity for IFNAR-2 subunit remains unchanged (Pan et al. 2008). The sites in IFN-α2 structure are shown in Fig. 3e. An example of variation having an “effect on protein catalytic site” is BTK in which variants R525Q and D430E at catalytic amino acids cause XLA (Fig. 3f) (Vihinen et al. 1994).

Effect on post-translational modification

Numerous proteins are modified post-translationally after synthesis. These modifications often have important regulatory functions. “Deletion of post-translational modification site” can occur due to a variation. BTK kinase domain is activated by Y551 phosphorylation at activation loop. Variation Y551H inactivates the protein by preventing activation although the protein is expressed (Aghamohammadi et al. 2006).

T168N substitution leads to “generation of novel posttranslational modification site”, a new N-glycosylation site, to interferon receptor IFNγR2 causing severe mycobacterial disease (Vogt et al. 2005). Because the variant protein contains extra carbohydrate moiety it has increased molecular weight.

Proteolytic processing is a common mechanism in protein activation and thus variations can have an “effect on protein processing”. “Effect on protein splicing” does not occur in human, but in some other organisms. Indel I143NT in the Ca²⁺-binding loop of guanylate cyclase-activating protein 1 (GCAP1) confers the protein susceptible for proteolysis (Nishiguchi et al. 2004) and is thus annotated as “variation generating a novel protein-processing site”. The variant leads to autosomal-dominant cone degenerate in retinal diseases. Variation R628P in integrin α3β1, which acts as receptor for laminins, is a “variation preventing polypeptide processing” of seven-bladed β-propeller structure of the protein (Yamada and Sekiguchi 2013). Upon maturation, the propeller structure is cleaved to produce heavy and light chains,
which remain connected by a disulfide bond. The variant protein transport to cell surface is
deficit from endoplasmic reticulum (ER) to Golgi apparatus. The variant is associated with
disorders in the lung, kidney, and skin due to the protein’s wide expression during the
development in epithelial organs.

Effect on protein secondary structural element

Protein 3D folds contain secondary structural elements including helices, strands, coils and
various turn motifs. Disease-related variations emerge often in secondary structural elements
and have typical characteristics (Khan and Vihinen 2010). Secondary structural elements are
usually less tolerant for substitutions than coils or irregular loops. Examples of secondary
structural element-affecting variants are H362Q and L358F in BTK Src homology 2 (SH2)
domain, which impair the structure of the domain based on functional and structural studies
(Mattsson et al. 2000). The structure of SH2 domain is shown in Fig. 3g. These variants have
description “effect on protein helix” and more specifically “effect on right-handed protein
helix” of type “effect on alpha helix”. Examples are not provided for all the possible
secondary structural elements as there are so many of them (see Fig. 5) and they all follow the
same principles of annotation.

Variations with “effect on transmembrane polypeptide” have either “effect on membrane
protein extramembrane region” or “effect on membrane protein intramembrane region”.
Proteolipid protein (PLP1) variants are causative of hereditary spastic paraplegia (HSP) type
2. The disease is characterized by progressive leukodystrophy and dysmyelination leading to
axonal degeneration. PLP1 has four transmembrane helices connected by extramembrane
loops. A30P alteration in the transmembrane helix 1 leads to the retaining of the protein in ER
instead of transport to cellular membrane (Noetzli et al. 2014). The second extracellular
domain of PLP1 contains two disulfide bridges. C184R substitution breaks the bridge and
leads to Pelizaeus–Merzbacher disease, another dysmyelination disorder (Fukumura et al.
2011). This is an “effect on membrane protein extramembrane region”.

Effect on protein fold

Protein folds are in constant motion and dynamics is essential also for functions. The folds
tolerate changes at some positions but are sensitive at other parts. The type of the variation is
important for the obtained effect.

“Changed domain orientation” in adenylate kinase 4 (AK4) caused by L171P variation leads
to a movement of LID domain (Liu et al. 2009). AK4 has two conformations, open and
closed, dependent on ligand binding (Fig. 3h). The L171P variant reorients the LID domain
and has effect on ligand binding. AK4 catalyzes the reversible ATP (GTP) production from
ADP (GDP).

Variation C260Y in hereditary hemochromatosis (HFE) protein is one of the most common
variants causing hereditary hemochromatosis where patients have excessive intestinal
absorption of dietary iron. C260Y alteration prevents disulfide bridge formation with
structural consequences and thus has “effect on protein disulfide formation” (Bennett et al.
2000) (Fig. 3i).

“Protein conformational change” can be either minor (local) or global. An example of “local
conformation change” is caused by ΔF508 deletion in the cystic fibrosis transmembrane
conductance regulator (CFTR). The deletion causes a localized conformational change at
residues 509–511 (Lewis et al. 2010) (Fig. 3j). Since the region is thought to be in contact
with transmembrane domain of the protein it is harmful and causes cystic fibrosis,
autosomal recessive disease, because of abnormal transport of chloride and sodium across an epithelium affecting lungs and digestive system.

Protein elongation

Elongation of the protein chain can appear at both ends. Variation at the terminal codon for Bβ chain for fibrinogen in a patient with fibrinogen Kyoto VI causes dysfibrinogenemia (defective fibrinogen clot formation). The variation leads to addition of 12 extra C-terminal amino acids (Okumura et al. 2012) and is “C-terminal protein elongation”.

c.-14C>T transition leads to generation of upstream start codon in the IFITM5 gene for interferon-induced transmembrane protein 5 (Semler et al. 2012). The new start site is in frame and it leads to “N-terminal protein elongation” of the produced protein by five additional residues. The elongation is harmful and causes osteogenesis imperfecta with a bone fragility and susceptibility to fractures even after minor trauma.

“Protein fusion” is quite common variant in cancers but appears also in some other diseases. Philadelphia chromosome originates from reciprocal translocation of chromosomes 9 and 22 fusing genes Abl1 for Abelson murine leukemia viral oncogene homolog 1 and BCR coding for breakpoint cluster region protein (Advani and Pendergast 2002). The fusion is characteristic for chronic myelogenous leukemia (CML), 95 % of patients carrying the variant. The fusion protein varies between patients depending on the actual fusion location. The variant protein is constitutively active protein tyrosine kinase, which acts unregulated and contributes to the disease.

Epigenetic protein modification

Epigenetic changes are common at DNA and RNA levels and affect also proteins. “Epigenetic protein complex formation” may appear even in man. “Prion formation” occurs in transmissible spongiform encephalopathies in man in Creutzfeldt–Jakob disease (CJD), kuru, Gerstmann–Sträussler–Scheinker (GSS) syndrome and familial insomnia (FFI). There are other forms in cattle and in sheep and goats. Common to these diseases is progressive degeneration of the central nervous system resulting in dementia, motor dysfunction and death.

Amyloid fibrils accumulate in brains in the prion diseases caused by irreversible structural conformational change of prion protein. The formation of pathogenic conformation that is insoluble and resistant to proteolytic degradation is a signature for prion diseases. The pathogenic form stimulates the conversion of normal proteins, thus prions are called infectious proteins. The prion protein has normal functions and the protein is abundant in brain (Roucou et al. 2004).

Lots of studies have been devoted for the prion conformational conversion. More than 40 variants, including G131V, S132I and A133V, in human prion are associated with the disease (Chen et al. 2010). The details of the conversion are not yet fully revealed. The native prion protein structure is shown in Fig. 3k. The disease form cannot be studied with experimental structure determination methods.

VARIATION AFFECTING PROTEIN PROPERTY

In addition to variation type, functional and structural alterations, variations affect properties of proteins. The properties range from pathogenicity to protein physical characteristics. Many of the properties are further defined in VariO with attributes, especially with quality attributes.
Association of protein variation to pathogenicity

The association of a variation to pathogenicity varies depending on the type of variation, its location etc. An example of “disease-causing” variation is R525Q in BTK (Vihinen et al. 1994). Along with the other variants listed in BTKbase (Piirilä et al. 2006), it is associated with the disease, XLA. The other attributes of this property are “disease associated” and “not related to clinical phenotype”.

Conservation of protein variation site

Sequence conservation at the variation site can be defined by conservation attributes such as “conserved region”, “covariant position” and “invariant region”. An example of the latter is W124C variation in BTK pleckstrin homology (PH) domain (Fiorini et al. 2004). PH domains share the same overall fold but have very low sequence identity of 17 % (Shen and Vihinen 2004). W124 is the only nearly invariant residue in the entire protein domain family. Alterations are deleterious in BTK in this αA helix position.

An example of “conserved region” is BTK SH2 domain R288. It is conserved in a large protein family. Alterations are disease-causing also in this hotspot site which accumulates altogether 8 R288Q cases, 29 patients in 21 families have R288W alteration, and in addition there are two patients with deletions starting at this site [see BTKbase and SH2base (Lappalainen et al. 2008; Piirilä et al. 2006)]. The invariant and conserved regions are depicted in the BTK SH2 domain in Fig. 3l.

Covariant amino acids are coevolving, meaning that viable variation in one site requires also the other site to change. Covarying amino acids are typically interacting with each other, but may have also other reason for conservation as seen in BTK PH domain-binding site (Shen and Vihinen 2004). Analysis of neutrophil elastase (ELANE) homologs revealed a network of covarying amino acids (see Fig. 3m) several of which are involved in congenital or cyclic neutropenia (lack of neutrophils) (Thusberg and Vihinen 2006).

Effect on protein abundance

Alterations to produced protein amounts are frequent in many diseases. Some variations in Wiskott–Aldrich syndrome protein (WASp) lead either to decreased protein abundance (e.g. L39P) in X-linked thrombocytopenia (XLT) or missing protein (e.g. R13X) in Wiskott–Aldrich Syndrome (WAS) (Jin et al. 2004). XLT is a milder disease than WAS.

Effect on protein accessibility

The localization of an amino acid on the surface or in protein core is one of the essential functional and physical properties of residues within proteins. Variants can affect protein accessibility, such as Y279C in SH2-domain-containing protein-tyrosine phosphatase (SHP2) that causes LEOPARD syndrome due to altered accessibility of the substrate recognition surface (Yu et al. 2013), see Fig. 3n. Although the phosphatase activity is reduced there is gain-of-function phenotype because the variants are hypersensitive for upstream activators and they have prolonged substrate turnover.

R234W is a “variation burying exposed region” in cellular retinaldehyde-binding protein (CRALBP) causing a domino-like conformational rearrangement of a number of side chains in the protein structure (He et al. 2009). The variant reduces the rate of pigment regeneration in eye and leads to Bothnia dystrophy, an autosomal recessive disease with high prevalence in northern Sweden. The structural changes are shown in Fig. 3o.
Δ508 deletion in CFTR is an example of a “variation exposing buried region”, see Fig. 3j. Due to the deletion, the conformation of the loop is changed so that the side chain of V510 is exposed compared to the normal protein (Lewis et al. 2010) (Fig. 6a).

Effect on protein activity

Numerous variations and variation mechanisms affect protein activity. It is obvious that activity alterations are largely harmful. However, the degree of activity required normally is dependent on the protein. Adenosine deaminase deficiency is a defect in purine nucleotide metabolism and most harmful for T cells. The defect causes severe combined immunodeficiency (SCID), but only when the enzyme activity is below 0.1 % of that in normal enzyme (Hershfield 2003). Thus, it is important to study which level of change is harmful, not just to look at the extent of change. This is important to keep in mind also in regard to other protein properties.

“Effect on protein affinity” has an example of immunoglobulin E (IgE) binding to CD23 receptor on B cells. Ca\textsuperscript{2+} binding to CD23, lectin-like head domain, triggers a conformational change in a loop via P250 trans to cis conformational change (Fig. 6b) (Yuan et al. 2013). The conformational change increases IgE affinity by 30-fold to the lectin-like head domain of CD23 and variations in Ca-binding residues (E249A and D270A) reduce the binding. Note that although $K_m$ for reactions following Michaelis–Menten (M–M) kinetics is often thought to be associated with affinity of enzyme to substrate, it is not always the case. Therefore, changes of $K_m$ and related features should be annotated with “effect on enzyme reaction kinetics”.

“Effect on enzyme reaction kinetics” altering in case of M–M kinetics $k_{\text{cat}}$, $K_m$, $V_{\text{max}}$, $k_{\text{cat}}/K_m$ etc., is often harmful. Examples are N158D and E232K in multifunctional enzyme (MFE) type 2 involved in β-oxidation and causing D-bifunctional protein (D-BP) deficiency when altered. N158D variant decreases $V_{\text{max}}$ and $k_{\text{cat}}/K_m$ ratio and increases both $K_m$ and $k_{\text{cat}}$ values (Mehtälä et al. 2013). E232K alteration has quite normal $K_m$ value; however, the other kinetic characteristics are changed. N158 is involved in substrate binding while K232 is participating in protein dimerization (Fig. 6c).

Variations in the DEF site interaction pocket in p38α MAP kinase decrease phosphorylation of some substrates but not others (Tzarum et al. 2013). The Y258A variant has such “effect on protein specificity” (see Fig. 6d). Variants at the DEF pocket significantly reduce the phosphorylation of ATF-II, Elk-1 and MBP without the effect on MK2 kinase phosphorylation.

Effect on protein charge

Charge and polarity of protein surface are important characteristics for molecular interactions. Change of M712T in UDP-N-acetylglucosamine 2-epimerase/N-acetyl-mannosamine kinase leads to hereditary inclusion body myopathy because of “effect on protein isoelectric point” lowering the IEP of the protein (Weidemann et al. 2011). The variant generates a novel phosphorylation site which contributes to the surface charge.

“Effect on protein electrostatics” has been shown to be the major effect for several XLA-causing variants in the BTK PH domain (Okoh and Vihinen 1999). Figure 6e shows all the currently known PH domain variation positions (Piirilä et al. 2006) and effects to electrostatics. A large number of the variants change the electrostatics of the domain and thereby affects its interactions with cellular membrane and protein partners.

Effect on protein degradation
Degradation is a normal fate of proteins once they are misfolded. Variant E130D in formylglycine-generating enzyme (FGE) is rapidly degraded although it is correctly localized and relatively active, leading to metabolic disorder multiple sulfatase deficiency (MSD) (Schlotawa et al. 2013). Posttranslational activation of sulfatases is reduced because the variant protein has low stability and it is therefore degraded.

Effect on protein interaction

Variations frequently affect protein interactions, including those in 2Cl/H\(^+\) exchanger ClC-5 responsible for X-linked recessive Dent’s disease. A large number of 39 studied variants modify the protein dimer interface, including G513R, I524K and G512R (Lourdel et al. 2012).

Extensive analysis of variation effects in human neutrophil elastase (HNE)-causing variants either in congenital or cyclic neutropenia include several which have an “effect on residue contact energy”, the strongest ones including ELANE variants A57T, L121P, C151S and C151Y (Thusberg and Vihinen 2006) (Fig. 6f).

More detailed description of interaction effects can be made using the interaction attributes in VariO.

Effect on protein solubility

Variations reducing protein solubility have either “effect on protein aggregation” or “effect on protein inclusion body formation”. Aggregation is a common mechanism in neurodegenerative diseases. L55P variation in transthyretin (TTR, thyrosine-binding prealbumin) causes early-onset familial amyloidotic polyneuropathy (McCutchen et al. 1993). Comparison of crystal structures for the normal and the variant-containing forms indicates disruption of hydrogen bonds between β-strands leading to altered surface that can explain the aggregation of the protein (Sebastião et al. 1998) (Fig. 6g).

Inclusion body formation is the mechanism of P56S variant in vesicle-associated membrane protein (VAPB) major sperm protein domain (MSP) leading to amyotrophic lateral sclerosis (ALS) (Shi et al. 2010). P56 has a cis conformation i.e. the side chain is on the same side of protein backbone as the preceding and following amino acids. This conformation is critical for the loop structure. The variant protein is unstructured and insoluble and forms inclusion bodies (Shi et al. 2010).

Effect on protein stability

Stability is a fundamental protein property. Variations can either increase of decrease stability. The most commonly studied form of stability is thermal stability; however, there are also other forms of stability e.g. toward denaturants, proteases or organic solvents. Study of 20 expressed variant proteins of mitochondrial DNA (mtDNA) helicase, implicated in a number of mitochondrial diseases, revealed a spectrum of thermal stability affecting variants (Longley et al. 2010). R354P increases the stability, while A359T and F485L have the most reduced stability.

Effect on protein subcellular localization

For the proper function of a protein it is essential that it is correctly localized within the cell (or outside of it). Proteins of cellular compartments have targeting signals which can be altered by variations or the localization can be otherwise impaired. Mislocalized proteins can be harmful due to being active in wrong compartment and due to lacking activity in the normal environment.
Amino acid substitutions, including R529Q, in the SH3 domain of tetratricopeptide repeat-containing protein 2 (SH3TC2) affect the localization to cell membrane and cause Charcot–Marie Tooth (CMT) disease type 4, a demyelinating neuropathy (Lupo et al. 2009).

Protein function change

Changes to protein function, both “protein gain of function” and “protein loss of function”, are frequent consequences of variations. M257Y substitution is constitutively active “protein gain of function” form of rhodopsin, a GPCR family member. The variant alters the dynamics of conformational change between the active and inactive forms (Tsukamoto and Farrens 2013). “Antimorphic protein variation” has antagonizing activity, i.e. opposite activity compared to normal state. The dominant negative effect appears for example in 637E and K673R variants in signal transducer and activator of transcription 1 (STAT1) SH2 domain resulting in impaired STAT1-mediated responses to IFN-γ and IL27 leading to susceptibility to mycobacterial disease (Tsumura et al. 2012).

“Neomorphic protein variation” has novel activities. D444V substitution of mitochondrial diisopamide dehydrogenase (DLD) activates cryptic protease activity (Babady et al. 2007). The protein forms normally a homodimer and variants into the dimer interface enhance proteolytic activity and partial or complete inactivation of DLD activity. D444V variant causing severe metabolic disorder has reduced DLD activity and the novel protease activity. The variant misses hydrogen bonds between the monomers and exposes the catalytic site for the protease (Fig. 6h).

There are numerous examples above about “protein loss of function” variants such as BTK R525Q in BTK and E31K in WAS (Jin et al. 2004; Vihinen et al. 1994). Usually it is not necessary to annotate loss of function as it is evident already by functional annotations.

CONCLUSIONS

A comprehensive overview of variation effects on proteins ranging from variation types to functional and structural effects is presented. The effects, mechanisms and consequences are discussed based on VariO annotations and examples often include protein 3D structures for such cases. VariO provides a kind of periodic system for these descriptions. Systematic description of variants allows unequivocal description, easy reuse, integration, and further studies of observed variants and effects.

REFERENCES


Martinez-Martinez I, Johnson DJ, Yamasaki M, Navarro-Fernández J, Ordóñez A, Vicente V, Huntington JA, Corral J (2012) Type II antithrombin deficiency caused by a large in-


19


Vihinen M (2014a) Variation Ontology for annotation of variation effects and mechanisms. Genome Res 24: 356-64. doi: 10.1101/gr.157495.113


FIGURE LEGENDS

Fig. 1. Protein variation types provide information for variation classification and origin.

Fig. 2 Amino acid variation types. The possible variation types with examples in a short protein sequence. Amino acid deletions, indels and insertions are of two types, sequence retaining or amphigoric. Protein truncations can appear either at N- or C-terminus. Original sequence left in the variants is underlined.

Fig. 3 Examples of variation effects at protein structures. a The normal (left PDB entry 1E04) and hyperstable (right 4EB1) forms of SERPINC1. The β-sheet A is in yellow except for the reactive center loop (RCL) in red. In the hyperstable form, the RCL is in blue, the insertion in red and the surface loop with altered conformation in orange. b Eye lens βB2-crystallin (2BB2) tetramer structure for bovine protein. The monomers are coded with different colors. The variant affects tetramer formation. c Normal insulin structure, left (4INS), and cavity forming variant, right (1K3M). For NMR structures the best representative structure, in this case chain number 1, is used, otherwise the first one was taken. The two chains after proteolytic cleavage are color coded as yellow and cyan in the left. The structural alterations generating a cavity are shown with magenta to the right. d Substitution G584W (red) causes a number of sterical clashes (indicated by green lines for red colored atoms) when introduced into the core of BTK kinase domain (1K2P). The variant is harmful because it causes structural alterations. e IFN-α2 (1ITF) showing residue R120 (red) in which variations affect binding and residues N65, L80, Y85 and Y89 (yellow) essential for IFNAR-1 binding. f BTK kinase domain structure (1K2P). The catalytic residues K430 (top) and R525 (below) are in red and causative of XLA when substituted. ATP binding amino acids are in cyan. g BTK SH2 domain (2GE9). The secondary structural elements are color coded, α-helices in yellow and β-strands in cyan. Amino acid substitutions at L358 (top) and H362 (below) affect α-helix. h The open (left 2AR7) and closed forms (right 3NDP) of adenylate kinase 4. Variations at L171 (red) alter the conformation of LID domain (yellow) in the structure to the right. i Variation at C260 (red) prevents formation of disulfide bridge with C203 (yellow) in haemochromatosis HFE protein (1A6Z). j Local conformational change at CFTR protein. Residues 509–511 in yellow have altered surface accessibility when the normal form (left 2BBO) is compared to ΔF508-containing form (right 2BBBT). k Native structure of prion protein (1QLX). The α-helices are in cyan and β-strands in yellow. Irreversible structural conformational change turns the protein infectious and capable of converting structures of native prion proteins. l Sequence conservation in the 58 seed sequences for SH2 domain in Pfam (PF00017). The most conserved sites are blue and the most variable ones red. m Coevolving ELANE positions (3Q76). Those amino acids involved in disease are in red others are in yellow. These sites appear in crucial positions at protein core. n Surface of the normal (left 2SHP) and Y279C variant (right 4DGX) forms of SHP2. The variant has major impact on the accessibility of residues. The central P-loop, pTyr-loop, Q-loop, D′ E-loop and WPD-loop are in yellow and residue 279 in red. o Variation R234W (right 3HX3) buries exposed region of normal CRALBP (left 3HY3). Residue 234 is in red (color figure online).

Fig. 4. Protein functional variations, clockwise from top: protein catalytic function, protein information transfer, protein movement, protein recognition, protein storage, structural protein, and transport function of protein.

Fig. 5. Protein structural variation. Organization of the descriptive VariO terms, which facilitate very detailed annotation of observed effects.
Fig. 6. Examples of variation effects at protein structures. a Variation ΔF508 in CFTR exposes the side chain of V510 (right 2BBT) when compared to the normal protein (left 2BBO). b Conformational transition of P250 (red) in CD23 (4G9A and 4G96) lectin-like head domain has an effect on protein affinity. Surrounding Ca$^{2+}$ binding residues E249 and T251 are in yellow. The Ca$^{2+}$ ion is shown in magenta. c Multifunctional enzyme type 2 (1ZBQ). The two chains of the dimer are in cyan and blue. Variation at E232 (magenta) has an effect on enzyme kinetics. Catalytic site residues are in yellow, residues at the lip of the cavity leading to active site in pink, and K246 forming salt bridge together with E232 in green. This salt bridge stabilizes the dimer as it appears between the monomers. d Variation at Y258 (red) has an effect on protein specificity by altering the conformation of the DEF site interaction pocket (yellow) in p38 MAP kinase (4GEO). e Several variants in BTK PH domain (1BTK) have an effect on protein electrostatics. The surface of the normal (left) and variant version of the domain (right) are colored according to the electrostatics. To the structure right, all PH domain changes affecting polarity/charge are combined. f Variations for A57T, L121P and C151S or Y (red) in ELANE (3Q76) have effects on residue contact energies. g Substitution of L55 in red (left 1TTA) by proline (right 5TTR) alters the surface and structure of transthyretin and leads to protein aggregation. h Substitution at residue 444 (red) affects the dimer formation of mitochondrial dilipoamide dehydrogenase (1ZMC). The conformational change alters the catalytic site (yellow) formed of residues from both the subunits (cyan and gray) and exposes catalytic site for neomorphic protease activity (magenta) (color figure online)
<table>
<thead>
<tr>
<th>Protein variation type</th>
<th>Protein variation of genetic origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>protein variation classification</td>
<td>protein variation emerging at protein level</td>
</tr>
<tr>
<td>amino acid deletion</td>
<td>artificial protein variation</td>
</tr>
<tr>
<td>protein truncation</td>
<td>epigenetic protein variation</td>
</tr>
<tr>
<td>sequence retaining amino acid deletion</td>
<td>protein structural inheritance</td>
</tr>
<tr>
<td>amino acid indel</td>
<td>proteinaceous infection</td>
</tr>
<tr>
<td>amphigoric amino acid indel</td>
<td>mistranslated protein</td>
</tr>
<tr>
<td>sequence retaining amino acid indel</td>
<td>post translationally modified</td>
</tr>
<tr>
<td>amino acid insertion</td>
<td>protein</td>
</tr>
<tr>
<td>amphigoric amino acid insertion</td>
<td></td>
</tr>
<tr>
<td>sequence retaining amino acid insertion</td>
<td></td>
</tr>
<tr>
<td>amino acid substitution</td>
<td></td>
</tr>
<tr>
<td>missing protein</td>
<td></td>
</tr>
</tbody>
</table>

Fig 1.
Fig. 2
Fig 3.
Fig. 4.
Fig. 5.
Fig. 6.