Introduction

UniProt provides human disease information with extensive cross-references to disease relevant databases such as: Medical Subject Headings (MeSH)\(^1\) and Online Mendelian Inheritance in Man (OMIM)\(^2\), Figure 1. In order to further enhance the functional annotations relevance to biomedical research; UniProt has recently developed a pipeline for importing protein altering variants from globally recognised genetic variant repositories with the aim to extend the manually curated set of natural protein altering variants provided by UniProt. By combining these resources UniProt has become a more relevant resource for biomedical research and drug target identification. Here we describe how users of UniProt can develop methodologies to utilise the described cross-references, protein structure and functional annotations to explore how structural, functional and chemical ligand annotations can be utilised to identify relationships between a protein and disease causing variants.

Methodology

Utilising the extensive cross-referencing within UniProt (Figure 2), InterPro domains for protein targets with disease annotation were probabilistically matched to ChEMBL\(^3\) ligands based upon ChEMBL’s activity score. Imported variants from the Ensembl variation\(^4\) and COSMIC\(^5\) databases were mapped to the InterPro\(^6\) domains. Variants within binding pockets was determined using PDBe\(^7\) binding pocket definitions when a high resolution structure with ligands bounds was available for the protein target. Finally any variant within a defined binding pocket was also mapped to ProSite patterns to determine if the variant was within the pattern.

Mapping diseases to InterPro Domains, Variants and ChEMBL Compounds

- 4,246 diseases mapped to 2,337 InterPro domains.
- 316 InterPro domains from 510 protein entries matched to 3,601 ChEMBL ligands.
- Somatic variants have been found within the binding pockets of proteins associated to specific cancer types.

Mitogen-activated Protein Kinase 4 Example

- A somatic variant Serine-233 to Alanine-233 was identified using the mapping methodology.
- This COSMIC variant was found to be within the binding pocket of Mitogen Activated Protein Kinase 4 (MAPK4).
- As illustrated in Figure 3 residue 233 is in close proximity of the phosphate group of the inhibitor Phosphoaminophosphonic acid-adenylate ester. Serine-233 can form hydrogen bonds to the inhibitor whilst this is lost upon mutation to alanine.
- Residue 233 is a non-conserved residue within the ProSite pattern for MAPK4.

Conclusions

These results illustrate that UniProtKB is a useful resource for biomedical research and therefore from these findings UniProt is planning, in collaboration with its partner databases, to develop new services to aid biomedical research and drug discovery pipelines.

References