BITS :: Call for Abstracts 2021 - Oral communication

Type	Oral communication
Session	Protein structure and function
Title	A large-scale analysis of human protein missense variations collected from HUMSAVAR and ClinVar
All Authors	Castrense Savojardo(1,*), Giulia Babbi(1,*), Matteo Manfredi(1), Pier Luigi Martelli(1), Rita Casadio(1,2)

Affiliation

(1) Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Italy
(2) Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies (IBIOM), Italian National Research Council (CNR), Bari, Italy
* Authors equally contributed to the work

Motivation

The advent of modern sequencing technologies allows the collection of an unprecedented amount of data about missense single-residue variations (SRVs) in protein sequences, that could be related to genetic diseases. These data are stored in public databases such as HUMSAVAR, the UniProt dataset of human missense variants annotated in Swiss-Prot, and ClinVar, the NCBI resource of relationships among human variations and disease phenotypes.

We precedently characterised disease-related variations occurring in proteins with a resolved 3D structure, and we studied proteins having missense variants mapped in PFAM domains. Here, we present a curated dataset derived from the union of the SRVs extracted from HUMSAVAR and ClinVar, and we show the physico-chemical and functional characterization of the pathogenic and benign SRVs, highlighting their different features that are preserved between the resource databases.

Methods

We collected SRVs from HUMSAVAR (release: 8/04/2021) and ClinVar (release: 29/03/2021), which both annotate SRVs into different classes, according to their clinical significance: Pathogenic or Likely Pathogenic (P/LP), Benign or Likely Benign (B/LB), and Uncertain Significance (US). We retained P/LP SRVs occurring on canonical UniProt sequences and unambiguously associated with diseases reported in OMIM or MONDO. All B/LB SRVs occurring on the same set of proteins were also collected. We finally filtered out all SRVs labelled as somatic or US. SRVs derived from HUMSAVAR and ClinVar were separately analysed and then merged in a comprehensive dataset (Union).

We characterised residue solvent accessibility of all sequences using an in-house predictor based on deep learning and already adopted for a large-scale analysis of HUMSAVAR SRVs. Protein flexibility was predicted from sequence using the recently released Medusa predictor. The five levels of flexibility assigned by Medusa were reduced into two classes: “Rigid”, merging prediction “0” and “1”, and “Flexible”, merging prediction “2”, “3”, and “4”. Intrinsically disordered regions were retrieved from the MobiDB database by means of the API provided by authors. Pfam domains were localised on protein sequences using the Pfam release 33.1 annotation file for Human proteins downloaded from the Pfam FTP server.

Results

The combination of SRVs reported in HUMSAVAR and ClinVar leads to 75,927 curated SRVs coming from 3,627 proteins. Some 59% and 41% of all SRVs are annotated as P/LP and B/LB, respectively. It is important to notice that HUMSAVAR and ClinVar provide different and complementary SRV sets, in particular for those labelled as B/LB. Indeed, the two datasets share only about 5%, 30%, and 72% of the B/LB SRVs, P/LP SRVs, and proteins collected in the Union dataset, respectively.

In the subsequent analyses we investigated to which extent the SRV datasets share physico-chemical and structural features, despite their small intersection. We took advantage of computational methods for predicting protein solvent accessibility, protein flexibility and disorder for positions carrying P/LP and B/LB SRVs, and we compared results obtained on the ClinVar, HUMSAVAR, and Union datasets.

Our analysis shows that P/LP and B/LB SRVs are endowed with distinctive features and these differences are similar across the analysed datasets. In particular, we observed that P/LP SRVs are significantly more abundant in buried regions (about 75%), while B/LB SRVs are often located in solvent-exposed regions (about 60%). The significance of these results is further supported by the fact that the number of buried and exposed residues predicted on the complete protein sequences are roughly equal.

Analysis of SRV flexibility shows similar figures, with P/LP SRVs more often occurring in rigid positions and B/LB SRVs more frequent in flexible regions, to be contrasted with a 50/50% prediction on the complete sequences. The analysis of intrinsically disordered regions is clearly biased by the uneven background distribution: only 14% of disordered residues are overall predicted in the whole dataset of sequences. In spite of this, we observe a slight tendency of P/LP SRVs to be more abundant in structured, not disordered regions, than B/LB. Interestingly, all these conclusions are rather independent of the considered dataset (i.e. HUMSAVAR, ClinVar, and Union).

Info

-

Figure

-

Availability	-

Corresponding Author

Name, Surname	Giulia, Babbi
Email	giulia.babbi3@unibo.it
Submitted on	07.05.2021