OXFORD

Structural bioinformatics

3DBIONOTES v3.0: crossing molecular and structural biology data with genomic variations

Joan Segura*, Ruben Sanchez-Garcia , C. O. S. Sorzano and J. M. Carazo

Spanish National Institute for Bioinformatics (INB ELIXIR-ES) and Biocomputing Unit, National Centre of Biotechnology (CSIC)/Instruct Image Processing Centre, 28049 Madrid, Spain

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Abstract

Motivation: Many diseases are associated to single nucleotide polymorphisms that affect critical regions of proteins as binding sites or post translational modifications. Therefore, analysing genomic variants with structural and molecular biology data is a powerful framework in order to elucidate the potential causes of such diseases.

Results: A new version of our web framework 3DBIONOTES is presented. This version offers new tools to analyse and visualize protein annotations and genomic variants, including a contingency analysis of variants and amino acid features by means of a Fisher exact test, the integration of a gene annotation viewer to highlight protein features on gene sequences and a protein–protein interaction viewer to display protein annotations at network level.

Availability and implementation: The web server is available at https://3dbionotes.cnb.csic.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: Spanish National Institute for Bioinformatics (INB ELIXIR-ES) and Biocomputing Unit, National Centre of Biotechnology (CSIC)/Instruct Image Processing Centre, C/ Darwin n° 3, Campus of Cantoblanco, 28049 Madrid, Spain. jsegura@cnb.csic.es,jsegurmor@gmail.com

1 Introduction

Next-generation sequencing has flooded many databases with biomedical data where single-nucleotide variations are associated with phenotypes or diseases (Zerbino *et al.*, 2018). This information comprises collections of variant–disease pairs that can be used to infer which genomic variations might be involved in a particular disease. However, changes on the biochemical or structural features of the affected amino acids (if applicable) can be more informative in order to understand the causes of diseases. For that reason, some of the existent resources compiling variant–disease knowledge also annotates protein residues with biochemical features displaying what properties could be affected (Dingerdissen *et al.*, 2018).

In this work, we present a new version of 3DBIONOTES (Segura *et al.*, 2017; Tabas-Madrid *et al.*, 2016) where different analysis tools and viewers have been integrated to find how genomic variants may affect the different protein residues. 3DBIONOTES is a web framework that integrates biological annotations and structural information of proteins from multiple sources (see Supplementary Section S1). In this version the application computes Fisher's exact test in order to find what biochemical or structural features are statistically affected by the variants associated to a particular disease. Moreover, a new panel displays those annotated regions where the co-occurrence between protein features and variants are statistically enriched. In addition, a gene annotation viewer has been fully integrated to display protein features at gene level. Also, a protein–protein interaction (PPI) viewer has been included in such a way that the different annotations can be displayed at network level. Finally, and as an additional tool, a new type of query, request by set of proteins, has been implemented to explore and analyse PPI networks. Moreover, custom annotations, including variants, can be submitted and analysed with the biological features integrated in the application.

2 New features

2.1 Gene annotation viewer

In this version of 3DBIONOTES, a gene annotation viewer has been fully integrated (see Supplementary Section S2). This panel displays gene information from ENSEMBL database (Zerbino *et al.*, 2018); the collected information includes: introns, exons, codifying regions and genomic variants. Moreover, ENSEMBL gene sequences are aligned with UniProt (UniProt Consortium, 2018) and PDB (Burley *et al.*, 2018) amino acids in such a way that protein annotated regions can be highlighted on gene sequences and vice versa (see Supplementary Section S2).

2.2 Genomic variants contingency analysis

In order to find which protein regions are statistically affected by the genomic variants associated to a particular disease, Fisher's exact test between the different annotations and the variants associated to the different diseases is computed. The main objective is to find co-occurrence of protein residues between the different structural or biochemical annotations and the variants associated to diseases. For example, most cancer related variants of the KRAS human protein map on its nucleotide binding region (see Section 3.1 and Supplementary Section S3).

2.3 Exploring PPI networks

A new type of query to request information for a set of proteins is now available. Moreover, a panel to visualize PPI networks using a graph-based representation has been integrated. This panel displays the physical binding between proteins when the information for a multimeric entry is requested or the PPIs that have been experimentally observed when a given set of proteins is submitted. In the first case, the contacts are computed using a distance threshold of 6 í between heavy atoms. For the second case, PPI data is collected from Interactome3D (Mosca *et al.*, 2013). Moreover, the network panel can display annotations at network level using a similar approach as dSysMap (Mosca *et al.*, 2015) (see Supplementary Section S4).

2.4 Submitting custom annotations

The application supports the submission of custom data in such a way that users can analyse their own genomic variants or other annotations and compare them with 3DBIONOTES integrated data. The submitted information is fully integrated and the different visualization and analysis tools can be used to display and process the external data.

3 Use cases

3.1 Analysis on the human KRAS genomic variants

In this example we have analysed the genomic variants associated to KRAS human protein (UniProt accession P01116). KRAS protein is a GTPase that acts as a signalling switch in many transduction pathways including cell proliferation. The active state of KRAS occurs when the protein is bound to GTP. In this state, the protein recruits and activates other growth factors and cell signalling receptors. Upon GTP hydrolysis and conversion to GDP, KRAS is inactivated. KRAS mutations are known to be involved in different diseases such as multiple cancer types or neurofibromatosis (Simanshu, et al., 2017). We used 3DBIONOTES to analyse the co-occurrence of KRAS variants associated to diseases with the different biochemical annotations. The main reason was to check whether those variants occur in particular regions of KRAS or randomly distributed. Supplementary Figure S4 and Supplementary Tables S1 and S2 display the analysis panel of

3DBIONOTES and clearly show that many of those variants occur in the 'Nucleotide Binding Site' annotated regions. KRAS acts as on/off switch in many processes and its active or inactive form depends on the interaction with GTP or GDP, respectively. Then, mutations affecting KRAS GTP/GDP-binding sites may affect its activation and therefore, many cell regulatory processes.

3.2 GNB1 neurodevelopmental disability

This example illustrates how 3DBIONOTES can be used to analyse external variants. We have collected the variants of the G protein subunit beta (GNB1) associated to neurodevelopmental delay, hypotonia and seizures available in the work of Petrovski *et al.* (2016) (see Supplementary Table S3). GNB1 protein modulates transmembrane signalling pathways controlled by G protein-coupled receptors. We have requested the PPI network information for the GNB1 protein (UniProt accession P62873) and attached the collected variants to 3DBIONOTES. When the variants are mapped to the PPI network, many of them appear affecting the binding sites between GNB1 and other G proteins. Moreover, the contingency analysis identified that the co-occurrence between many of the GNB1-biding sites and the submitted variants was statistically significant (see Supplementary Fig. S8). Consequently, mutations of the GNB1-binding sites may affect the interaction with other G proteins and, thus, some of the cell signalling pathways involving G proteins.

Funding

References

Burley,S.K. *et al.* (2018) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res*, 47, D464–D474.

Dingerdissen,H.M. *et al.* (2018) BioMuta and BioXpress: mutation and expression knowledgebases for cancer biomarker discovery. *Nucleic Acids Res.*, 46, D1128–D1136.

Mosca,R. *et al.* (2013) Interactome3D: adding structural details to protein networks. *Nat. Methods*, 10, 47–53.

Mosca,R. *et al.* (2015) dSysMap: exploring the edgetic role of disease mutations. *Nat. Methods*, 12, 167–168.

Petrovski,S. *et al.* (2016) Germline De Novo Mutations in GNB1 Cause Severe Neurodevelopmental Disability, Hypotonia, and Seizures. *Am. J. Hum. Genet.*, 98, 1001–1010.

Segura,J. *et al.* (2017) 3DBIONOTES v2.0: a web server for the automatic annotation of macromolecular structures. *Bioinformatics*, 33, 3655–3657.

Simanshu,D.K. *et al.* (2017) RAS proteins and their regulators in human disease. *Cell*, 170, 17–33.

Tabas-Madrid,D. *et al.* (2016) 3DBIONOTES: a unified, enriched and interactive view of macromolecular information. *J. Struct. Biol.*, 194, 231–234.

UniProt Consortium. (2018) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, 47, D506–D515.

Zerbino,D.R. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, 46, D754–D761.