Journal of Structural Biology 194 (2016) 231–234

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Technical Note

3DBIONOTES: A unified, enriched and interactive view of macromolecular information

D. Tabas-Madrid [1], J. Segura [*,1], R. Sanchez-Garcia, J. Cuenca-Alba, C.O.S. Sorzano, J.M. Carazo

GN7 of the Spanish National Institute for Bioinformatics (INB) and Biocomputing Unit, National Center of Biotechnology (CSIC)/Instruct Image Processing Center, C/Darwin n° 3, Campus of Cantoblanco, 28049 Madrid, Spain

ARTICLE INFO

Article history:
Received 9 December 2015
Received in revised form 2 February 2016
Accepted 5 February 2016
Available online 10 February 2016

Keywords:
Protein structure
Protein sequence
Protein annotations
Database integration

ABSTRACT

With the advent of high throughput techniques like Next Generation Sequencing, the amount of biological information for genes and proteins is growing faster than ever. Structural information is also rapidly growing, especially in the cryo Electron Microscopy area. However, in many cases, the proteomic and genomic data are spread in multiple databases and with no simple connection to structural information.

In this work we present a new web platform that integrates EMDB/PDB structures and UniProt sequences with different sources of protein annotations. The application provides an interactive interface linking sequence and structure, including EM maps, presenting the different sources of information at sequence and structural level. The web application is available at http://3dbionotes.cnb.csic.es.

1. Introduction

Currently, high-throughput techniques are producing massive amounts of genomic and proteomic information, feeding most relevant biological databases such as UniProt (UniProt Consortium, 2015) and ENSEMBL (Cunningham et al., 2015), extending the amount of available annotations for genes and proteins. Indeed, these annotations are essential contributions to the study of protein and gene functions. However, structural information is a key element required for a deeper understanding of the molecular properties that allow proteins to perform specific tasks. Therefore, depicting genomic and proteomic information over structural data would offer a more complete picture in order to understand how proteins and genes behave in the different cellular processes.

In this work we present a web platform –3DBIONOTES– that integrates proteomic and functional annotations with structural data, providing a unified and interactive view of the different sources of information. The main interface comprises three panels: the 3D viewer, the protein sequence viewer and the annotations panel. The three views are interactively connected and the different annotations can be displayed both at sequence level, highlighting the amino acids of a selected annotation, and at structural level, mapping the corresponding residues into the protein structure.

2. 3DBIONOTES framework

The 3DBIONOTES project aims to integrate the different levels of molecular biology information into an intuitive environment where protein sequences and structures are represented in a single and interactive graphical interface. In its current version, 3DBIONOTES offers a unified view of three of the most relevant protein databases: UniProt (UniProt Consortium, 2015), PDB (Berman et al., 2014) and EMDB (Lawson et al., 2011), onto which other sources of biological annotations are also provided, such as PhosphoSitePlus (Hornbeck et al., 2015), Immune Epitope DB (Vita et al., 2015), BioMuta (Wu et al., 2014) and dSysMap (Mosca et al., 2015).

2.1. 3DBIONOTES server

The web server was implemented using the Ruby on Rails application framework. The server performs three major tasks: first, it maps the residue identifiers of PDB structures with the amino acid indexes of UniProt sequences; second, it provides the relation between EMDB maps, PDB structures and UniProt sequences; and third, it supplies the protein annotations. The first task is carried out using the SIFTS (Structure Integration with Function, Taxonomy and Sequence) resource provided by the EBI (Velankar et al., 2013), this resource is a collection of XML files that offer a residue-level mapping between UniProt and PDB entries. For the second task, providing the relation between EMDB, PDB and UniProt entries, the server uses the EBI REST web services

* Corresponding author.
E-mail address: jsegura@cnb.csic.es (J. Segura).
[1] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

http://dx.doi.org/10.1016/j.jsb.2016.02.007
1047-8477/

(Meldal et al., 2015). In particular, three different services are used: the 'fitted' EMDB-service is used to obtain the PDB structures fitted within an EMDB map, the 'uniprot' SIFTS-service that relates the PDB chain identifiers with their corresponding UniProt accession and, finally, the 'best_structure' SIFTS-service, that allows to retrieve a list of PDB structures and chains related with a specific UniProt entry. Consequently, the web server can be accessed using any class of identifier, such as EMDB codes, PDB identifiers or UniProt accessions. Finally, the server also supplies the protein annotations collected from UniProt, PhosphoSitePlus, Immune Epitope DB, BioMuta and dSysMap. The UniProt and dSysMap annotations are collected using the web services provided for this purpose. In turn, PhosphoSitePlus and BioMuta provide their data in tab format files that were parsed and stored in a MySQL database. Finally, Immune Epitope DB information is available as a MySQL exportable file. Once the raw annotations are collected, the server builds an array of features where each array element describes a particular annotation, identifying its database source, type, subtype, start, end and description. This design is very easy to extend, since introducing a new source of protein annotations only involve the development of a data collector system and a new parser.

Regarding maintenance of this application, PDB structures, UniProt features and dSysMap annotations are retrieved on the fly and thus, they are updated automatically. However, EMDB maps, PhosphoSitePlus data and BioMuta annotations are stored locally and a bash script is executed weekly to keep this information updated.

2.2. 3DBIONOTES client

The web client is responsible for the data representation and also for providing an interactive environment connecting protein sequences, structures and annotations. The client comprises three major panels (Fig. 1): the structural panel (Fig. 1A), the sequence panel (Fig. 1B) and the annotation panel (Fig. 1C). The structural panel uses JSmol viewer (Jmol, 2013) in order to display PDB structures and EMDB maps. The sequence panel shows the alignment between the sequences of PDB structures and their corresponding UniProt sequences, to that end we used a bespoken version of the BioJS (Gomez et al., 2013) 'Sequence' package (Gomez and Jimenez, 2014). The annotation panel consists of a modified version of the EBI-UniProt protein annotations viewer (UniProt – EBI, 2014). All panels are interconnected, leading to a graphic interactivity between them; thus, when a protein annotation from the annotation panel is clicked, the protein sequence region related to the annotation and the residues in the corresponding chain of the structural panel are highlighted. Additionally, when a segment of sequence in the sequence panel is selected, the annotation panel displays the selected region so that the particular annotations falling inside are marked and the corresponding residues of the structural panel are highlighted. Querying 3DBIONOTES is performed through a web form and the application accepts any identifier from EMDB, PDB or UniProt.

3. Use cases

3.1. Human APC/C-Cdh1-Emi1 ternary complex

The Anaphase-Promoting complex is a cell-cycle regulatory macromolecule that triggers the transition from metaphase to anaphase. The activation of the complex depends on its association with one of the coactivator proteins. In this example we have explored the interaction between the APC/C complex and the Cdh1 coactivator subunit (EMDB entry EMD-2924). This interaction is negatively regulated by phosphorylation and Cdh1 is inactivated and prevented to interact with APC/C when residues S40,

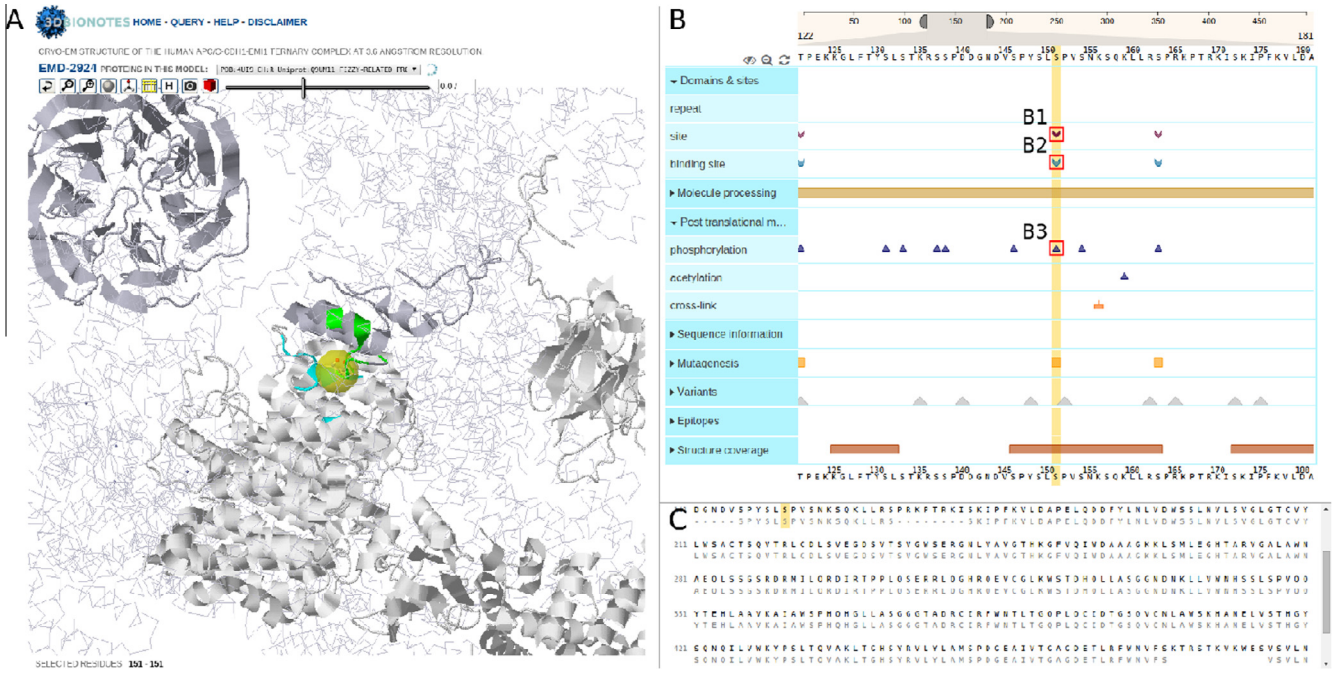

Fig. 1. Human APC/C-Cdh1-Emi1 ternary complex. Screenshot of the 3DBIONOTES interface presenting the case when the server is queried with the EMDB entry EMD-2924. (A) The structural panel, on the left hand side, highlights the Cdh1 coactivator subunit in dark gray color and cartoon style together with the Anaphase-Promoting complex subunit 1 (APC1) in light gray color and also cartoon style. The S151 amino acid of Cdh1 is displayed within a yellow sphere and those residues that are closer than 10 Å are highlighted in green if the residues belong to the Cdh1 protein and in cyan color if are contained in the APC1 subunit. (B) The annotation panel, on the top right hand side, is centered on the S151 amino acid of Cdh1 (UniProt ID Q9UM11) and the annotations related with the phosphorylation and regulation properties of S151 are remarked within red squares. (B1) This annotation indicates that the s151 amino acid is a regulatory site for the interaction between Cdh1 and APC1. (B2) Binding site annotation for a kinase–substrate interaction, in this case the Cell division protein kinase 2 protein binds to the S151 residue. (B3) Annotation indicating that the S151 residue is a phosphorylation site. (C) The sequence panel, on the bottom right hand side, shows the amino acid sequence of the UniProt entry Q9UM11 (Cdh1 subunit), marking in orange color the S151 residue.

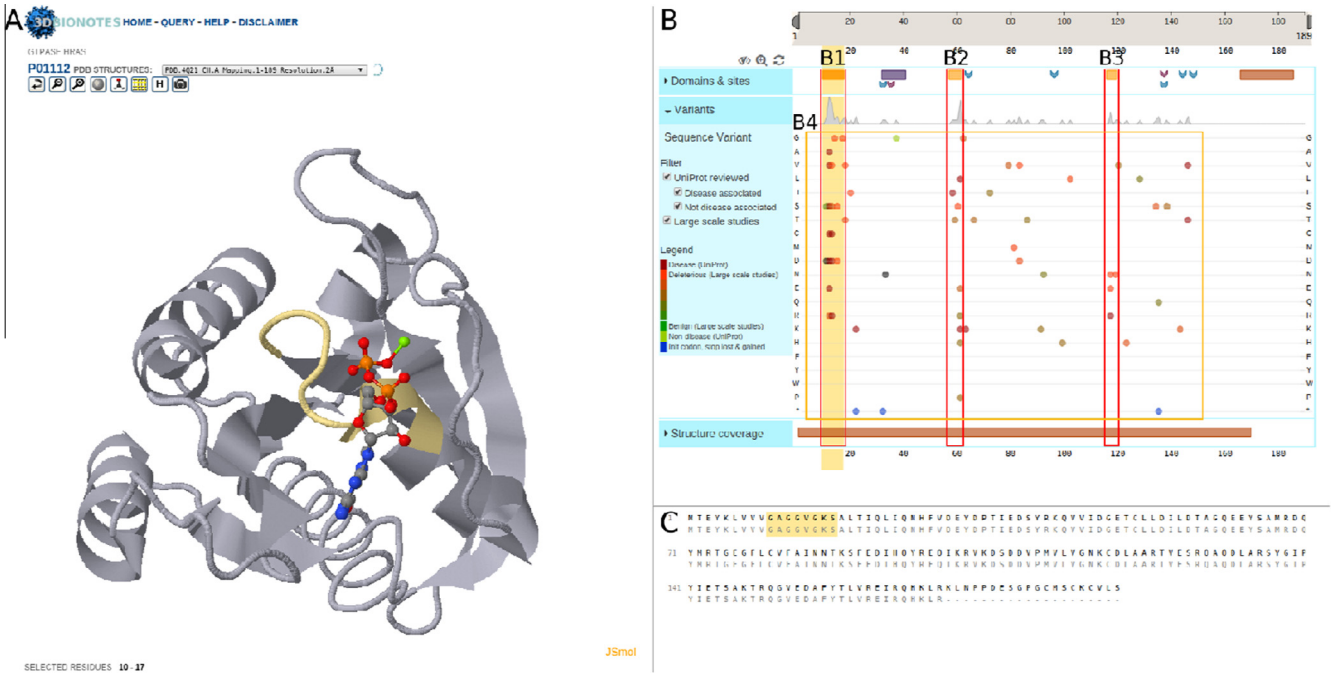

Fig. 2. Human GTPase HRas protein. Screenshot of the 3DBIONOTES interface presenting the case when the server is queried with the UniProt ID P01112. (A) The structural panel, on the left hand side, shows the structure of the HRas protein represented by the PDB code 4Q21. The GTP/GDP binding region comprised between residues 10 to 17 is highlighted in orange, at the same time that a GDP molecule is shown in a ball and sticks atomic representation. (B) The annotation panel, on the upper right hand side, shows the annotations associated to HRas (UniProt ID P01112), marking in orange the GTP/GDP binding region comprised between amino acids 10–17. The red rectangles (B1, B2 and B3) contain those annotations that fall within the three different GTP/GDP binding regions, while the orange rectangle (B4) contains the genetic variant annotations. (C) The sequence panel, on the lower right hand side, displays the UniProt sequence of P01112, where the sequence segment corresponding to the GTP/GDP binding region comprised between amino acids 10–17 is highlighted in orange.

S151 and S163 of Cdh1 are phosphorylated (Chang et al., 2015). Fig. 1 shows a screenshot of 3DBIONOTES for the EMDB entry EMD-2924. We have focused on the interaction between the APC/C complex and the Cdh1 subunit and, in particular, on the S151 residue of Cdh1 (Fig. 1A). The structural panel displays, in a cartoon schema, the structure of the Cdh1 subunit (dark gray); the S151 residue is shown within a yellow sphere, and in a light gray color and cartoon style the structure of the Anaphase-Promoting complex subunit 1 (APC1), which is the unique subunit of the APC/C complex that interacts with the S151 residue of the Cdh1 protein. The application highlights in different colors those amino acids that are closer than 10 Å to the S151 residue of Cdh1, green color for those residues belonging to the same Cdh1 chain and cyan color for those belonging to other subunits (in this case APC1). Among the different features of the annotation panel (Fig. 1B), we find the information related with phosphorylation of residue S151, in this case collected from PhosphoSitePlus database (Fig. 1B3). Also, this residue is annotated as a substrate–kinase binding site that interacts with the Cell division protein kinase 2 (Fig. 1B2). Finally, the annotation panel also shows a feature indicating that this residue regulates the interaction with the APC1 protein (Fig. 1B1). Similar information was also found for the other regulatory residues S40 and S163 of the Cdh1 subunit.

3.2. GTPase HRas

GTPase HRas (HRas) is a cell division regulatory protein that acts as a molecular on/off switch for the propagation of signals from cellular membrane receptors. In its active form, HRas binds to GTP and recruits the specific proteins necessary for signal propagation; finally, upon conversion of GTP to GDP, HRas is turned off. Due to its importance in cellular division, HRas is involved in multiple cancer processes and has been shown to be a proto-oncogene (Kiaris et al., 1995). In this example we have used 3DBIONOTES to explore the structure and annotations available for the HRas protein (UniProt ID P01112). Fig. 2 displays the annotations and structure of HRas when the PDB code 4Q21 is selected (Fig. 2A) for its structure representation. Among the different features in the annotation panel (Fig. 2B), we have highlighted the three regions of the HRas protein that interact with GTP/GDP (Fig. 2B1, B2 and B3). The application is highlighting the GTP/GDP binding region comprised between amino acid indexes 10–17 (Fig. 2B1 and C), at the same time that the structure residues corresponding to this region are emphasized in orange color (Fig. 2A). The genetic variants section (Fig. 2B4) shows how most of the known disease associated variants (more than 60%) fall within one of these regions. This example illustrates the importance of the GTP/GDP biding sites of the HRas protein and their relation with different disease processes.

4. Conclusions

In this work we have presented a new platform that pursues the integration of multiple sources of molecular biology data in the context of structural information. Currently, the application integrates protein sequence and structure in a unique environment, unifying the accession to three of the most relevant protein resources: EMDB, PDB and UniProt. The design of the application allows for the easy integration of additional data bases, and in its first version 3DBIONOTES already incorporates protein annotations from PhosphoSitePlus, Immune Epitope DB, BioMuta and dSysMap, paving the way to the incorporation of other proteomics and genomics data sources in future releases.

Funding

This work was supported by the Instituto de Salud Carlos III, project number PT13/0001/0009 funding the Spanish National

Institute of Bioinformatics, the Spanish Ministry of Economy and Competitiveness through grants AIC-A-2011-0638 and BIO2013-44647-R and the European Union (EU) and Horizon 2020 through grant CORBEL (INFRADEV-1-2014-1 – Proposal: 654248), EGI-Engage project under Grant number 654142 and ELIXIR-EXCELERATE (INFRADEV-1-2015-1 – Proposal: 676559). C.O.S. Sorzano is recipient of a Ramón y Cajal fellowship. J. Segura is recipient of a Juan de la Cierva fellowship.

References

Berman, H.M., Kleywegt, G.J., Nakamura, H., Markley, J.L., 2014. The Protein Data Bank archive as an open data resource. J. Comput. Aided Mol. Des. 28, 1009–1014.

Chang, L., Zhang, Z., Yang, J., McLaughlin, S.H., Barford, D., 2015. Atomic structure of the APC/C and its mechanism of protein ubiquitination. Nature 522, 450–454.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Giron, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Kahari, A.K., Keenan, S., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Aken, B.L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S.M., Spudich, G., Trevanion, S.J., Yates, A., Zerbino, D.R., Flicek, P., 2015. Ensembl 2015. Nucleic Acids Res. 43, D662–D669.

Gomez, J., Jimenez, R., 2014. Sequence, a BioJS component for visualising sequences. F1000Res. 3, 52.

Gomez, J., Garcia, L.J., Salazar, G.A., Villaveces, J., Gore, S., Garcia, A., Martin, M.J., Launay, G., Alcantara, R., Del-Toro, N., Dumousseau, M., Orchard, S., Velankar, S., Hermjakob, H., Zong, C., Ping, P., Corpas, M., Jimenez, R.C., 2013. BioJS: an open source JavaScript framework for biological data visualization. Bioinformatics 29, 1103–1104.

Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., Skrzypek, E., 2015. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. 43, D512–D520.

Jmol. 2013. Jmol: an open-source Java viewer for chemical structures in 3D.

Kiaris, H., Spandidos, D.A., Jones, A.S., Vaughan, E.D., Field, J.K., 1995. Mutations, expression and genomic instability of the H-ras proto-oncogene in squamous cell carcinomas of the head and neck. Br. J. Cancer 72, 123–128.

Lawson, C.L., Baker, M.L., Best, C., Bi, C., Dougherty, M., Feng, P., van Ginkel, G., Devkota, B., Lagerstedt, I., Ludtke, S.J., Newman, R.H., Oldfield, T.J., Rees, I., Sahni, G., Sala, R., Velankar, S., Warren, J., Westbrook, J.D., Henrick, K., Kleywegt, G.J., Berman, H.M., Chiu, W., 2011. EMDataBank.org: unified data resource for CryoEM. Nucleic Acids Res. 39, D456–D464.

Meldal, B.H., Forner-Martinez, O., Costanzo, M.C., Dana, J., Demeter, J., Dumousseau, M., Dwight, S.S., Gaulton, A., Licata, L., Melidoni, A.N., Ricard-Blum, S., Roechert, B., Skyzypek, M.S., Tiwari, M., Velankar, S., Wong, E.D., Hermjakob, H., Orchard, S., 2015. The complex portal – an encyclopaedia of macromolecular complexes. Nucleic Acids Res. 43, D479–D484.

Mosca, R., Tenorio-Laranga, J., Olivella, R., Alcalde, V., Ceol, A., Soler-Lopez, M., Aloy, P., 2015. DSysMap: exploring the edgetic role of disease mutations. Nat. Methods 12, 167–168.

UniProt – EBI, 2014. A visualization tool for protein features and variations.

UniProt Consortium, 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43, D204–D212.

Velankar, S., Dana, J.M., Jacobsen, J., van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., O'Donovan, C., Martin, M.J., Kleywegt, G.J., 2013. SIFTS: structure integration with function, taxonomy and sequences resource. Nucleic Acids Res. 41, D483–D489.

Vita, R., Overton, J.A., Greenbaum, J.A., Ponomarenko, J., Clark, J.D., Cantrell, J.R., Wheeler, D.K., Gabbard, J.L., Hix, D., Sette, A., Peters, B., 2015. The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 43, D405–D412.

Wu, T.J., Shamsaddini, A., Pan, Y., Smith, K., Crichton, D.J., Simonyan, V., Mazumder, R., 2014. A framework for organizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual Environment (HIVE). Database (Oxford) 2014, bau022.