Andrew Nightingale[1], Tunca Dogan[1], Diego Poggioli[1], Maria Martin[1] and the UniProt Consortium[1,2,3]

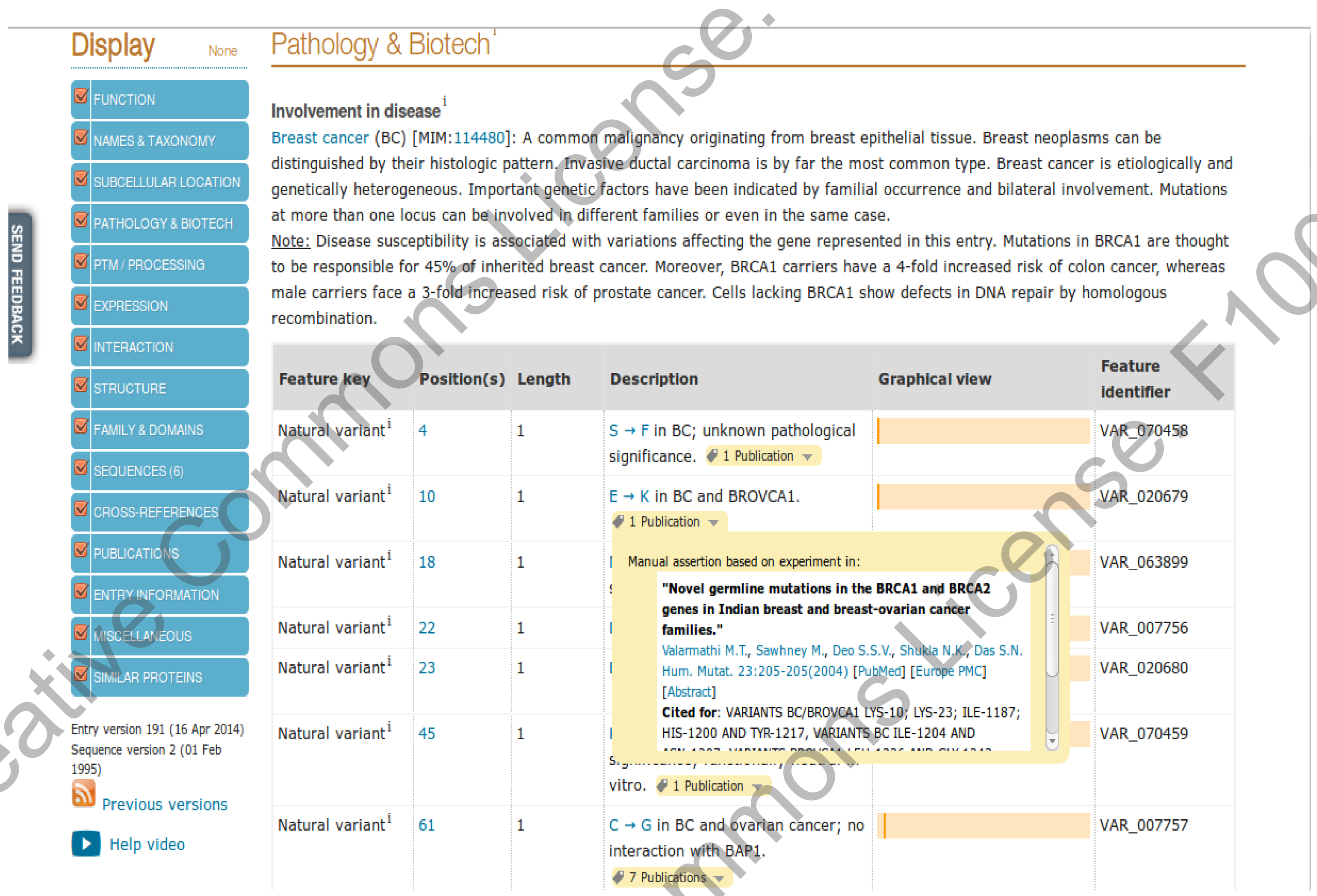
1 EMBL-European Bioinformatics Institute, Cambridge, UK
2 SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
3 Protein Information Resource, Georgetown University, Washington DC & University od Delaware, USA

The Role of UniProt's Protein Sequence Databases in Biomedical Research

EMBL-EBI

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.


Display None
FUNCTION
NAMES & TAXONOMY
SUBCELLULAR LOCATION
PATHOLOGY & BIOTECH
PTM / PROCESSING
EXPRESSION
INTERACTION
STRUCTURE
FAMILY & DOMAINS
SEQUENCES (6)
CROSS-REFERENCES
PUBLICATIONS
ENTRY INFORMATION
MISCELLANEOUS
SIMILAR PROTEINS
Entry version 191 (16 Apr 2014)
Sequence version 2 (01 Feb 1995)
Previous versions
Help video
SEND FEEDBACK

Pathology & Biotech

Involvement in disease

Breast cancer (BC) [MIM:114480]: A common malignancy originating from breast epithelial tissue. Breast neoplasms can be distinguished by their histologic pattern. Invasive ductal carcinoma is by far the most common type. Breast cancer is etiologically and genetically heterogeneous. Important genetic factors have been indicated by familial occurrence and bilateral involvement. Mutations at more than one locus can be involved in different families or even in the same case.

Note: Disease susceptibility is associated with variations affecting the gene represented in this entry. Mutations in BRCA1 are thought to be responsible for 45% of inherited breast cancer. Moreover, BRCA1 carriers have a 4-fold increased risk of colon cancer, whereas male carriers face a 3-fold increased risk of prostate cancer. Cells lacking BRCA1 show defects in DNA repair by homologous recombination.

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Natural variant	4	1	S → F in BC; unknown pathological significance. 1 Publication		VAR_070458
Natural variant	10	1	E → K in BC and BROVCA1. 1 Publication		VAR_020679
Natural variant	18	1			VAR_063899
Natural variant	22	1			VAR_007756
Natural variant	23	1			VAR_020680
Natural variant	45	1	vitro. 1 Publication		VAR_070459
Natural variant	61	1	C → G in BC and ovarian cancer; no interaction with BAP1. 7 Publications		VAR_007757

Manual assertion based on experiment in:
"Novel germline mutations in the BRCA1 and BRCA2 genes in Indian breast and breast-ovarian cancer families."
Valarmathi M.T., Sawhney M., Deo S.S.V., Shukla N.K., Das S.N.
Hum. Mutat. 23:205-205(2004) [PubMed] [Europe PMC] [Abstract]
Cited for: VARIANTS BC/BROVCA1 LYS-10; LYS-23; ILE-1187; HIS-1200 AND TYR-1217, VARIANTS BC ILE-1204 AND

Figure 1: Disease and natural variant annotation for UniProtKB/SwissProt entry for Human BRCA1

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.

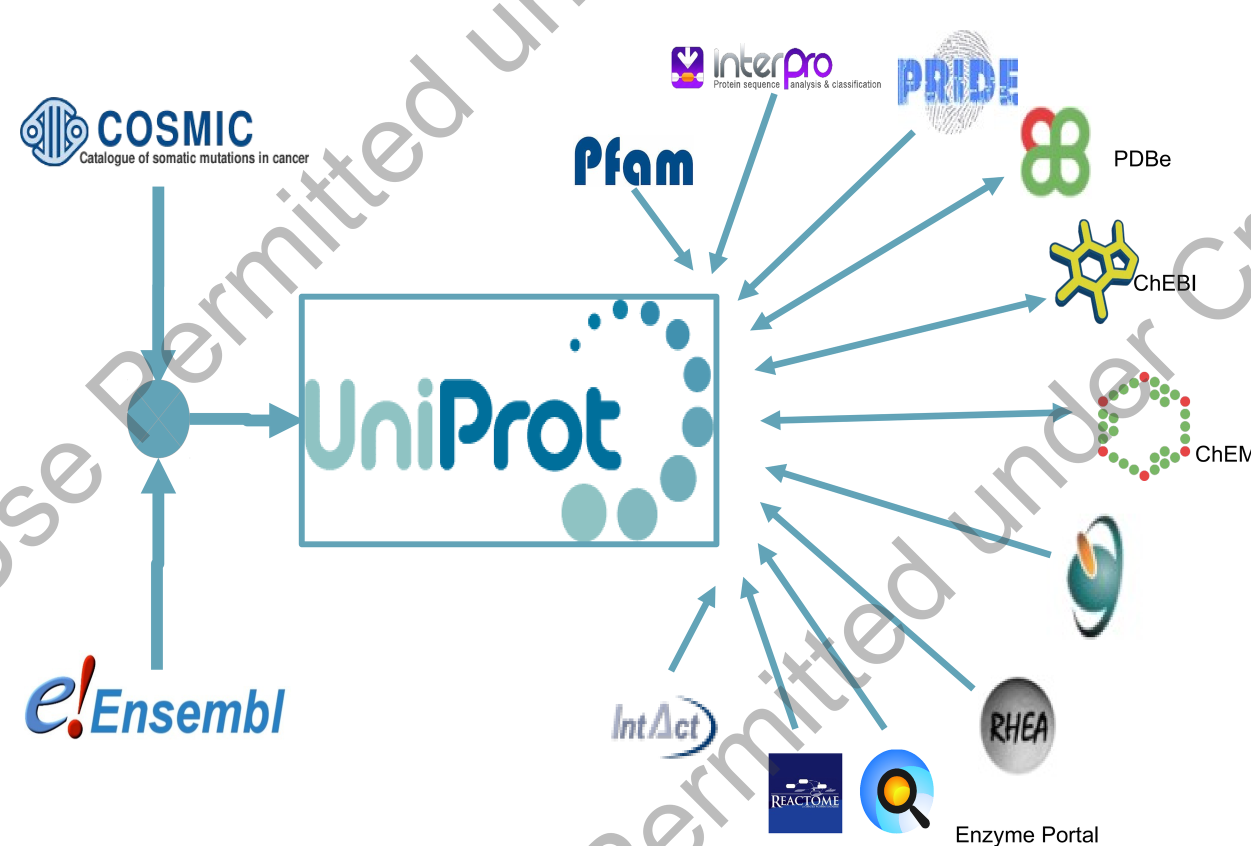


Figure 2: Extensive cross-referencing within UniProt

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

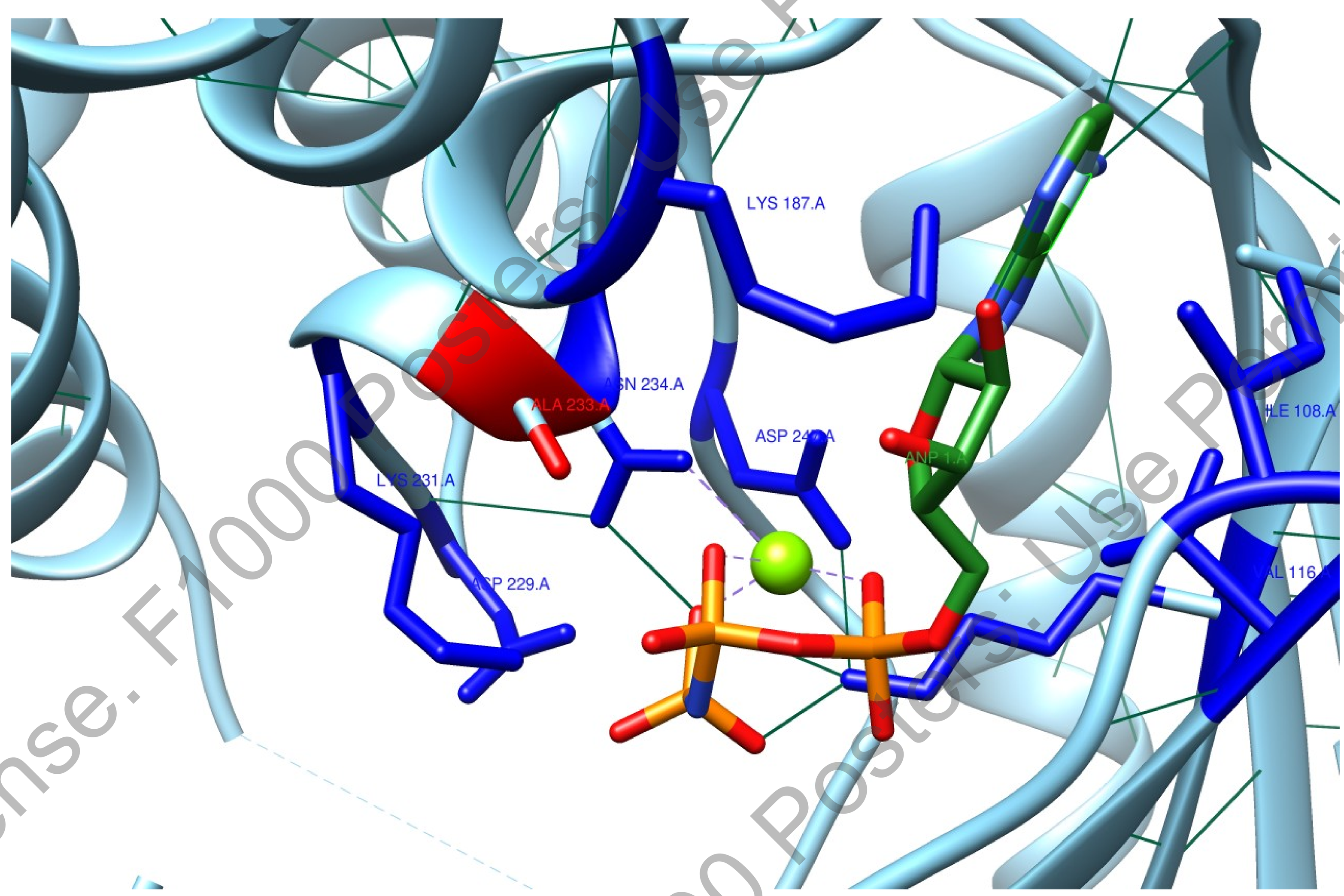


Figure 3: MAPK4 binding pocket with analogue inhibitor bound and p.Ser233Ala variant.

- 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 (MAPK)4.
- 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

1. **Medical subject headings.** *Bull Med Libr Assoc* **51**:114–6 (1963).
2. **Online Mendelian Inheritance in Man, OMIM®**. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) (2014). www.omim.org
3. **ChEMBL: a large-scale bioactivity database for drug discovery.** *Nucleic Acids Research* **40**(D1):D1100-D1107 (2011)
4. **Ensembl 2014**. *Nucleic Acids Research* **42**(1):D749-55 (2014)
5. **COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.** *Nucleic Acids Research* **39**(suppl 1): D945-D950 (2011)
6. **InterPro in 2011: new developments in the family and domain prediction database.** *Nucleic Acids Research* ***40**(D1): D306-D312* (2012)
7. **PDBe: Protein Data Bank in Europe.** *Nucleic Acids Research*, **42**(D1), D285-D291 (2014),

UniProt is mainly supported by NIH (grant 4U41HG006104-04).
With additional funding support from EMBL-EBI.

EMBL-EBI
Wellcome Trust Genome Campus
Hinxton, Cambridgeshire, CB10 1SD, UK
Tel. +44 (0) 1223 494 444
comms@ebi.ac.uk
www.ebi.ac.uk

F1000 Posters: Use Permitted under Creative Commons License.