6-26-2020

Protein Annotation of Breast-cancer-related Proteins with Machine-learning Tools

Arli Aditya Parikesit

Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences, Jakarta 13210, Indonesia, arli.parikesit@i3l.ac.id

David Agustriawan

Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences, Jakarta 13210, Indonesia

Rizky Nurdiansyah

Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences, Jakarta 13210, Indonesia

Follow this and additional works at: <https://scholarhub.ui.ac.id/science>

Recommended Citation

Parikesit, Arli Aditya; Agustriawan, David; and Nurdiansyah, Rizky (2020) "Protein Annotation of Breast-cancer-related Proteins with Machine-learning Tools," *Makara Journal of Science*: Vol. 24 : Iss. 2 , Article 6.

DOI: 10.7454/mss.v24i1.12106

Available at: <https://scholarhub.ui.ac.id/science/vol24/iss2/6>

This Article is brought to you for free and open access by the Universitas Indonesia at UI Scholars Hub. It has been accepted for inclusion in Makara Journal of Science by an authorized editor of UI Scholars Hub.

Protein Annotation of Breast-cancer-related Proteins with Machine-learning Tools

Cover Page Footnote

The authors would like to thank the Institute for Research and Community Services of the Indonesia International Institute for Life Sciences (i3I) for their heartfelt support. Thanks also goes to Direktorat Riset dan Pengabdian Masyarakat, Direktorat Jenderal Penguatan Riset dan Pengembangan Kementerian Riset, Teknologi dan Pendidikan Tinggi Republik Indonesia for providing Hibah Penelitian Dasar DIKTI/LLDIKTI III 2019 No. 1/AKM/PNT/2019. Lastly, thanks to Dr. Arun Prasad Pandurangan from the MRC Laboratory of Molecular Biology, Cambridge, UK, for providing troubleshooting assistance with the SUPERFAMILY database

Protein Annotation of Breast-cancer-related Proteins with Machine-learning Tools

Arli Aditya Parikesit*, David Agustriawan, and Rizky Nurdiansyah

Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences,
Jakarta 13210, Indonesia

*E-mail: arli.parikesit@i3l.ac.id

Received December 31, 2019 | Accepted April 20, 2020

Abstract

One of the primary contributors to the mortality of women is breast cancer. Several approaches are used to cure it, but recurrence occurs in 79% of the cases because the underlying mechanism of the protein molecules is not carefully examined. The goal of this research was to use machine-learning tools to elucidate conserved regions and to obtain functional annotations of breast-cancer-related proteins. The sequences of five breast-cancer-related proteins (BRCA2, BCAR1, BCAR3, BCAR4, and BRMS1) and their annotations were retrieved from the UniProt and TCGA databases, respectively. Conserved regions were extracted using CLUSTALX. We constructed a phylogenetic tree using the MEGA 7.0. SUPERFAMILY database to obtain fine-grained domain annotation. The tree revealed that the BRCA2 and BCAR4 protein sequences are located in a clade, which indicates that they have overlapping functions. Several protein domains were identified, including the SH2 and Ras GEF domains in BCAR3, the SH3 domain in BCAR1, and the BRCA2 helical domain, the nucleic-acid-binding protein, and tower domain. We found that no protein domains could be annotated for BCAR4 or BRMS1, which may indicate the presence of a disordered protein state. We suggest that each protein has distinct functionalities that are complementary in regulating the progression of breast cancer, although further study is necessary for confirmation. This protein-domain annotation project could be leveraged by the complete integration of mapping with respect to gene and disease ontology. This type of leverage is vital for obtaining biochemical insights regarding breast cancer.

Keywords: breast cancer, breast-cancer-related protein, machine learning, protein annotation, sequence analysis

Introduction

As reported by WHO, breast cancer is considered to be one of the most dangerous diseases for women [1]. In the United States, breast cancer is reported to comprise 15.2% of all the cancer cases and almost 7% of the mortality rate [2]. A survey by the Ministry of Health of Indonesia found that symptom of breast cancer are detected in 2.6 of every 1000 breast lesion samples [3]. This disease has long been treated by conventional means, i.e., surgery, radiation, and chemotherapy, which are undertaken only after the cancer has entered a late stage and has metastasized [4].

However, because the biomolecular mechanism of breast cancer has been determined, the development of more modern diagnostic approaches and therapeutics are now feasible. Moreover, the field of molecular medicine has employed the machine-learning approach to diagnose the lesion images, metabolomics, genomics, medical informatics predictors, and structural bioinformatics of breast cancer, which significantly facilitate

the development of highly accurate diagnostics [5]. Moreover, machine-learning models have been employed to annotate medical informatics and pharmacogenomics data to diagnose and provide expert judgments regarding the progression of breast cancer [6,7]. Based on recent research, several mutations have been found in particular genes that could play some part in the progression of breast cancer. The most well-known genes are BRCA1 and BRCA2, as the mutations of those genes have been identified in breast-cancer cell-line samples [5,6]. Moreover, a transcriptomics biomarker is being developed to detect the progression of triple-negative breast cancer (TNBC) by leveraging its non-coding (nc)RNA signatures [9].

The development of transcriptomics-based diagnostics and therapy is recognized as challenging, especially with respect to clinical trials. In this regard, studying the proteomics expression of breast cancer is expected to provide significant relevant information. Several proteomics-based breast-cancer drugs have been

approved by the FDA, including Anastrozole®, which targets hormone-receptor-positive (HR+) breast cancer, Trastuzumab® for HER2-positive (HER2+) breast cancer, and Fulvestrant (Faslodex®) for HR+ and HER2-negative breast cancers with no hormone therapy [8–10]. An exhaustive list is linked with the FDA database that is available online [13]. As noted above, combining molecular medicine with machine-learning methods provides an interesting approach for obtaining information about the–omics repertoire of breast cancer. In this regard, the shortcoming of the machine-learning method is the possibility of determining any introduction of bias in the data annotations, which could introduce redundancy in reports [14]. However, the advantages of the machine-learning method for making gene or protein annotations regarding breast cancer far outweigh its shortcomings, i.e., mainly its ability to predict a fine-grained molecular mechanism, its extensibility to broad-range–omics studies, and the provision of data annotations for breast cancer survivors [15–17].

With respect to molecular proteomics, the protein domain is widely recognized as an independent molecular evolutionary unit that plays an important role in cancer progression. An approach to structural bioinformatics research has been devised that simulates the molecular mutations in the TP53 and ER proteins directly related to breast cancer [14,15]. The role of disordered proteins and their phylogeny are also computed. Moreover, to obtain complete population-size samples of cancer patients, a dedicated database has been developed, namely The Cancer Genome Atlas (TCGA) [20]. The availability of this dedicated database and the growing interest in proteomics research have enabled the growth of machine learning-based tools for annotating the occurrence of protein domains that have a role in the progression of breast cancer [19–20]. This method employs all the latest protein annotation or proteomics tools. Notable proteomics databases such as SUPERFAMILY, SCOP, and PFAM incorporate the hidden Markov model method for predicting previously unidentified protein domains and folds [21–23]. Moreover, the STRING database employs a graph algorithm with a statistical probability model for devising a nodes and edges repertoire [26]. In this respect, the SUPERFAMILY protein domain database was developed to provide annotations regarding the progression of disease [27]. Based on the SCOP classification that incorporates the machine-learning approach of the SUPERFAMILY database, protein domains are classified into families, superfamilies, and folds. Protein families comprise domains with similar sequences but different features, superfamilies comprise domains with common ancestors, and folds comprise domains with similar structures [26,27]. The SCOP classification is the industry standard for protein domain annotations.

Several software packages and pipelines have been developed for making extensive domain annotations, and these have attracted the interest of the scientific community [30,31]. This effort has been strengthened by the more fine-grained annotation of protein–protein interaction networks of cancer-related proteins [30–32]. To this end, these efforts should be solidified as a basis for developing a blueprint for constraining the menace of breast cancer and developing drug designs for breast cancer [34,35]. The development of proteomics information pipelines will facilitate the examination of the protein domain repertoire in the progression of breast cancer. The objective of this research is to use machine-learning tools to annotate the protein domains responsible for the progression of breast cancer. These annotations are expected to provide fine-grained information for use in constructing a blueprint for the development of drugs based on a complete analysis of proteomics data regarding functional and conserved regions, with annotations for TCGA-database-based proteins that represent a significant risk of breast cancer as the drug target. The reason protein sequences were examined rather than genes is due to the availability of 3D structural data that could shed a light on functional annotations.

Materials and Methods

Our choice of research methodology was inspired by existing pipelines that have undergone significant modifications with respect to existing indicators and parameters [37,38]. This research was conducted using a standard MacBook Pro Laptop with MacOSX version 10.13.6, 512 GB of HDD, and 16 GB of RAM. The employment of a Mac-based laptop was crucial for leveraging a graphics subsystem with proven scientific computation performance [39–42]. To navigate and search for associations between genes and breast cancer, the phrase ‘breast cancer’ was used to search for the appropriate genes in the TCGA database (<https://portal.gdc.cancer.gov/>). After identifying the names of the genes, the associated protein sequences were downloaded from the UNIPROT database (<https://www.uniprot.org/>). All available sequence retrieval procedures were utilized to obtain a consensus regarding the sequences. The evolutionary history was inferred using the maximum likelihood (ML) method based on the Jones–Taylor–Thornton (JTT) matrix-based model in the MEGA7 package, along with its default parameters. The ML method was chosen for its ability to iterate many different evolutionary models to improve its reliability as a general statistics model. Moreover, the ML method has been proven to be the most accurate phylogenetic method for estimating branch length and other parameters [64,65]. The tree with the highest log likelihood (–1709.1737) was also identified. The initial tree(s) for the heuristic search were obtained automatically by applying the neighbor-join and BioNJ

algorithms to a matrix of pairwise distances estimated using the JTT model and then selecting the topology with the highest log-likelihood value. The tree is drawn to scale, with branch lengths measured based on the number of substitutions per site. The analysis involved five amino-acid sequences. Positions with less than 95% site coverage were eliminated. After the construction of the tree, SUPERFAMILY and STRING database searches were performed to annotate the protein domains and their respective interactions. For the SUPERFAMILY database search (<http://supfam.org/>), the HMMSCAN significant E-value hit was 0.03, whereas the reported E-value hit was 1, which were leveraged as default values. Then, as part of the SUPERFAMILY database project, annotations of disordered proteins were searched using the D2P2 database (<http://d2p2.pro/>), which only provides hits with 100% identity. Thus, the minimum required interaction score of the STRING database (<https://string-db.org/>) is a medium confidence rating (0.400, the default parameter) [41-49]. The downloaded Protein Data Bank (PDB) files linked from the PFAM database (<http://pfam.xfam.org/>) were visualized using Chimera version 1.13.1 [54].

Results

The results are divided into five subsections with their respective calculation times for each cycle in brackets, i.e., protein sequence retrieval (10 minutes), protein phylogeny (15 minutes), SUPERFAMILY domain annotations (5 minutes), PFAM 3D annotations (5 minutes), STRING annotations (5 minutes), and the disordered protein annotations (5 minutes), the last four of which were obtained using machine-learning tools.

Protein sequence retrieval. The search for entries regarding genes associated with breast cancer in the TCGA database yielded the identification of five genes with supporting references that are associated with a significant risk of breast cancer and serve as drug targets, namely BRMS1, BRCA2, BCAR1, BCAR3, and BCAR4 [55–58]. As the TCGA mainly annotates genes with complete population data, this does not mean that other genes have no association with breast cancer. They are simply not annotated because there is insufficient population data for establishing a strong association with the progression of breast cancer. To

retrieve the protein sequences, the gene names were simply queued into a search box in the UNIPROT database (Table 1). As all of the proteins are associated with the phrase “breast cancer” due to their importance in the progression of this disease and their extensive annotations in the TCGA database, it is interesting to examine whether or not these proteins have a common ancestor.

Protein phylogeny. Based on the UNIPROT database, the respective protein function can be obtained by accessing its annotations. The BRMS1 protein functions as a translational repressor that regulates the anti-apoptotic gene and inhibits the metastatic stage of cancer. The function of the BRCA2 protein is to regulate homologous recombination, avoid genomic instability, and maintain the integrity of DNA repair [59]. The function of the BCAR1 protein is to perform docking for cell adhesion and migration, as well as to mediate anti-estrogen resistance [60]. The function of the BCAR3 protein is to regulate the signaling pathway during the proliferation of breast cancer. Lastly, the BCAR4 protein functions as an oncoprotein that induces tamoxifen resistance to breast cancer. The aberration of these genes has a significant influence on the progression of breast cancer. To shed light on the clustering of these proteins, we generated a phylogenetic tree by performing a MEGA7 phylogenetic tree computation of the proteins, as shown in Figure 1.

Figure 1. Molecular Phylogenetic Analysis of BRCA2, BCAR1, BCAR3, BCAR4, and BRMS1 by the Maximum Likelihood Method via the MEGA7 Application

Table 1. Breast-cancer-related Proteins from the UNIPROT Database

No.	Protein ID	Description
1.	sp Q9HCU9 BRMS1_HUMAN	Breast cancer metastasis-suppressor 1 (BRMS1)
2.	sp P51587 BRCA2_HUMAN	Breast cancer type 2 susceptibility protein (BRCA2)
3.	sp P56945 BCAR1_HUMAN	Breast cancer anti-estrogen resistance protein 1 (BCAR1)
4.	sp O75815 BCAR3_HUMAN	Breast cancer anti-estrogen resistance protein 3 (BCAR3)
5.	tr D3DUG6 D3DUG6_HUMAN	Breast cancer antiestrogen resistance 4 (BCAR4)

The results of our phylogeny analysis indicated that the BRCA2 protein was in the same cluster as the BCAR4 protein, although it was evident that the BCAR1 protein is considered to be a distinct and unique cluster relatively unrelated to other proteins. The phylogeny analysis revealed an interesting domain cluster consisting BRCA2 and BCAR4, both of which are subjects of continuous drug development for breast cancer. It could be true that these proteins were clustered due to their extensive annotations in molecular research, although a common molecular evolutionary history may also play a part. It is also feasible that the development of diagnostics and therapeutic agents involving these two protein domains are aligned to some extent.

SUPERFAMILY domain annotation. To annotate the distribution of protein domains, we utilized the SUPERFAMILY database. Figure 2 shows the domain distribution of the five annotated breast-cancer-related proteins.

The BRCA2 protein was found to have two annotated domains (Figure 2a), with the colored boxes in the protein illustration representing different domains. The blue box is the BRCA2 helical domain (scop id of 81872), which has a SUPERFAMILY E-value of $1.44\text{e-}87$. The red box indicates nucleic acid-binding proteins (scop id of 50249), which have a SUPERFAMILY E-value of $1.71\text{e-}45$. The BCAR1 protein belongs to one annotated domain (Figure 2b), which is light blue in the figure, representing the SH3-domain (scop id of 50044) with a SUPERFAMILY E-value of $3.54\text{e-}19$. The BCAR3 protein has two annotated domains (Figure 2c), with the red box indicating the SH2 domain (scop id of 55550) with a SUPERFAMILY E-value of $4.58\text{e-}25$ and the light blue box the Ras GEF domain (scop id of

48366) with a SUPERFAMILY E-value of $1.57\text{e-}36$. The E-value threshold for hit significance is a value equal to or smaller than $10\text{e-}4$.

In the figure, we can see that there are no annotated domains for either the BCAR4 or BRMS1 protein, which are shown as straight lines in the protein model (Figures 2d and 2e) and have E-values close to zero (0). This means that the query results are very similar to the template in the database, which indicates that the designated protein domain has a highly probability of existing. The descriptions in the SCOP database regarding the function of these protein domains are limited. Therefore, these outputs were extrapolated to the PFAM database, which provides more detailed descriptions. Moreover, the BCAR4 and BRMS1 proteins merit more attention to address the lack of domain annotations. To provide structural and functional annotations for a protein domain, 3D structural data in the protein domain must be accessed from both the PFAM and RCSB databases.

PFAM 3D annotation. The PFAM database provides access to 3D-visualized annotations of the protein domain, whereas the SUPERFAMILY with the SCOP database lacks this feature. The PDB file-based visualization was obtained from the RCSB website, which is directly linked with the PFAM database.

Homologous recombination repair, an important process for cancer avoidance, is the main role of the BRCA2 protein in humans. Both the under- and overexpression of the BRCA2 protein are found in sporadic breast-cancer cases. The main feature of the BRCA2 helical domain, which is the backbone of the BRCA2 protein, is its helical and beta-hairpin structure (Figure 3).

Figure 2. Protein Domain Annotations of a) BRCA2, b) BCAR1, c) BCAR3, d) BCAR4/D3DUG6, and e) BRMS1 from SUPERFAMILY Database

Figure 3. Annotated domain in the BRCA2 protein; the BRCA2 helical, nucleic-acid binding, and tower domains (PDB ID: 1MIU) that share a common structure. Visualized using Chimera version 1.13.1. Different colors represent different secondary-structure folds (<https://pfam.xfam.org/family/PF09169>)

The BCAR3 protein comprises two domains, SH2 and RAS GEF. The SH2 domain mainly regulates the signal transduction and expression of oncoproteins, with assistance from tyrosine kinases (Figure 4a). The RAS GEF domain is a smart molecular switch that catalyzes the hydrolytic reaction from guanosine triphosphate (GTP) to guanosine diphosphate (GDP), which ensures the balance of these biochemical species in the cell for the activation of the GTPase enzyme (Figure 4b).

The interaction between adaptor proteins and tyrosine kinases is the primary feature of the SH3 domain, while also acting as a mediator in the assembly of protein complexes (Figure 5).

In the PDB, the BRMS1 protein does not have a complete structure, with only fragments of the N-terminal region provided (http://www.rcsb.org/pdb/results/results.do?tab_toshow=Current&qrid=6E9CAA53). The BCAR4 protein also lacks a complete structure and even any fragments in the PDB database. However, the UNIPROT database provides annotations of homologous information about these protein domains, with slightly more detail at the domain level. The next step was to identify protein–protein interactions to better understand the protein features.

STRING database annotations for protein–protein interaction. Protein–protein interactions (PPIs) are annotated in the STRING database. Figure 6 shows the PPIs for the BRCA2, BCAR1, BCAR3, and BRMS1 proteins. The interaction intensity is expressed as a

score with a maximum of 1, which indicates the most intense interaction. In Figure 6a, we can see that the protein that interacts with BRCA2 is PALB2, with an interaction score of 0.99. PALB2 is as an auxiliary protein for BRCA2 that assists with homologous repair [61]. In Figure 6b, we can see that the protein that interacts with BCAR1 is PXN, with an interaction score of 0.999. The function of PXN is to assist with the membrane attachment of the cytoskeleton protein. Figure 6c shows that intense interaction has occurred between the BCAR3 and BCAR1 proteins, with an interaction score of 0.983. Figure 6d shows interaction between BRMS1 and ARID4A, with an interaction score of 0.996. The function of ARID4A is to support interaction with the retinoblastoma protein. There are no annotations for the BCAR4 protein in the STRING database. As such, we could not validate the occurrence

Figure 4. BCAR3 Protein with a) SH2 (PDB ID: 2cr4) and b) RAS GEF Domains (PDB ID: 6d52). Visualized using Chimera Version 1.13.1. Different Colors Represent Different Secondary-structure Folds (Source: <https://pfam.xfam.org/family/sh2> and <https://pfam.xfam.org/family/PF00617>)

Figure 5. BCAR1 Protein with SH3 Domain Visualized using Chimera Version 1.13.1. Different Colors Represent Different Secondary-structure Folds (Source: PDB ID: 6NMW and <https://pfam.xfam.org/family/PF00018>)

of PPI between BCAR4 and BRCA2. However, interaction may occur as in the phylogeny study they were found to share the same protein cluster. Subtle annotations could be the result of a disordered protein feature, which is described in the next section. We note that the proteins that interact with BCAR4, BRCA2, BCAR1, and BRMS1 may have potential for being

leveraged for further study toward a rational drug design because the latest biomedical research indicates that they play important roles in the progression of breast cancer [57,62–67]. In future work, consideration should be given to whether genes are up- or down-regulated in breast cancers.

Figure 6. Protein–protein Interactions of: a) BRCA2, b) BCAR1, c) BCAR3, and d) BRMS1 (<https://string-db.org/>)

ENSPYP00000004685, ENSP00000353042, ENSGGOP00000028472

Figure 7. Disordered Protein Annotations for BRMS1. (Source: <http://d2p2.pro>)

Disordered protein database of SUPERFAMILY.

Next, we searched in the D2P2 database for the signatures of disordered proteins without annotated domains. Since the existence of such a disordered protein could be indicated by the absence of protein-domain annotations, this would explain the lack of result for the BCAR4 protein. However, there was a hit for the BRMS1 protein (Figure 7). In this regard, all the disorder predictors (PONDR VL-XT, PONDR VSL2b, PrDOS, PV2, Espritz (all variants), IUPred (all variants), and ANCHOR) showed significant hits in the protein as signatures of disorders.

An explanation of the predictors is provided in a published protocol [68]. Moreover, the top of Figure 7 shows three available isoforms for the BRMS1 protein. BRMS1 is considered to be a disordered protein with no annotated SUPERFAMILY domain, although it shows hits in the PFAM domain, as indicated by the red block in Figure 7. There is at least a 75% agreement among different predictors, as shown by the green block with the label 'Predicted Disorder Agreement'. There are also five post-translational modifications of the protein domain, as indicated by the 'PTM sites' label. These results confirm that the BRMS1 protein is indeed disordered. Although the structural disorders of the BCAR4 and BRMS1 proteins make their domain information unavailable, we note that the 3-D structures of the proteins are intact and can still be used to facilitate drug design. However, the missing domain

information could hamper our understanding of the chemical bonds and our ability to obtain a more-detailed resolution of the 3-D structure to identify what detracts from the overall stability of the protein [69]. One possible solution to this problem would be to wait for the next update to the structural disorders in the SUPERFAMILY database.

Discussion

Due to the incompleteness of the domain annotations, some possible strategies must be devised to address this problem. A gene prediction package could be utilized to leverage the existence of protein domains that provide the function and the structure of these genes. Homology modeling and ab-initio methods could be used to predict the structures of the BCAR4 and BRMS1 proteins, which are feasible within the boundaries of available computational power. In our computational study, we did not leverage any of the predicted data to determine their degree of alignment with information generated by wet laboratory research groups. As the indicators of protein domain annotations, structural annotation, PPI, and disordered proteins can shed light on both BCAR4 and BRMS1. More research is needed in this area. Moreover, using structural bioinformatics tools, the feasibility of using the PPI partners of the annotated breast-cancer-related proteins as drug candidate targets could be considered. Further research should be conducted on proteins with unannotated domains as

possible signatures of a disordered protein. For proteins without a clearly defined tertiary structure, the domain repertoire should be annotated using a different approach. This effort will be much more effective when the SUPERFAMILY database has successfully integrated its D2P2 disordered protein database into one integrated platform. On this platform, strategies could be developed to establish a disease ontology of the protein domain. To this end, the relation between protein domains, genes, and disease annotation could be elucidated in finer detail.

Phylogenetic tree and PPI studies have shown that the current molecular simulation protocol is insufficient for effective drug design as proteome data has been used without consideration of the networking context of the interacting proteins. In the current approach, structural bioinformatics studies are conducted with very limited knowledge of PPIs or gene networks. To some extent, solid knowledge about the protein domain repertoire is subtly handled in structural annotations. Only indicators that provide fine-grained annotations can be secured, mainly from sequence retrieval and phylogenetic tree studies. Annotations from other indicators should be strengthened. To do so, advanced machine-learning-based tools or even deep-learning tools could be used to facilitate completion of these annotations. Future structural bioinformatics studies should incorporate a more holistic approach to increase the success rate of drug designs. The complexity of proteomics studies remains daunting as protein domain rearrangement or co-occurrence must be incorporated as the basis of biomedical informatics studies and no method has yet been developed to incorporate these data into the structural bioinformatics approach.

Conclusion

Based on the results of this phylogenetic study, the BCAR4 and BRCA2 proteins can be inferred to be in the same cluster. Furthermore, the BCAR4 protein was found to have no annotated domain in the SUPERFAMILY, STRING, PDB, or D2P2 databases. As such, to develop a blueprint for drug design, determination of the correlation and interaction between the BCAR4 and BRCA2 proteins will require further in-depth annotations. Moreover, the absence of SUPERFAMILY annotations for the BRMS1 protein indicates that it may be in a disordered state. However, STRING studies offer hope regarding the availability of other proteins that could serve as targets for drug design. This drug design effort would be made more effective by incorporating the results of protein-domain rearrangement and co-occurrence studies using more advanced machine-learning-based annotation tools such as the gene prediction pipeline.

Acknowledgments

The authors would like to thank the Institute for Research and Community Services of the Indonesia International Institute for Life Sciences (i3l) for their heartfelt support. Thanks also goes to Direktorat Riset dan Pengabdian Masyarakat, Direktorat Jenderal Penguatan Riset dan Pengembangan Kementerian Riset, Teknologi dan Pendidikan Tinggi Republik Indonesia for providing Hibah Penelitian Dasar DIKTI/LLDIKTI III 2019 No. 1/AKM/PNT/2019. Lastly, thanks to Dr. Arun Prasad Pandurangan from the MRC Laboratory of Molecular Biology, Cambridge, UK, for providing troubleshooting assistance with the SUPERFAMILY database.

References

- [1] WHO. 2015. WHO | Breast cancer: prevention and control. WHO.
- [2] Surveillance Epidemiology and End Results Program. 2019. Cancer stat facts: Female breast cancer. Natl. Cancer Inst. 1–10.
- [3] Kemenkes-RI. 2013. Profil kesehatan Indonesia, Jakarta.
- [4] National Cancer Institute. 2019. Definition of conventional treatment—NCI Dictionary of Cancer Terms. National Cancer Institute.
- [5] Parikesit, A.A., Ramanto, K.N. 2019. The usage of deep learning algorithm in medical diagnostic of breast cancer. *Malaysian J. Fundam. Appl. Sci.* 15: 274–281, <http://dx.doi.org/10.11113/mjfas.v15n2.1231>.
- [6] Stark, G.F., Hart, G.R., Nartowt, B.J., Deng, J. 2019. Predicting breast cancer risk using personal health data and machine learning models. *PLoS One*. 14, <http://dx.doi.org/10.1371/journal.pone.0226765>.
- [7] Ming, C., Viassolo, V., Probst-Hensch, N., Chappuis, P.O., Dinov, I.D., et al. 2019. Machine learning techniques for personalized breast cancer risk prediction: Comparison with the BCRAT and BOADICEA models. *Breast Cancer Res.* 21, <http://dx.doi.org/10.1186/s13058-019-1158-4>.
- [8] James, C.R., Quinn, J.E., Mullan, P.B., Johnston, P.G., Harkin, D.P. 2007. BRCA1, a potential predictive biomarker in the treatment of breast cancer. *Oncologist*. 12: 142–50. <http://dx.doi.org/10.1634/theoncologist.12-2-142>.
- [9] Eades, G., Wolfson, B., Zhang, Y., Li, Q., Yao, Y., et al. lincRNA-RoR and miR-145 regulate invasion in triple-negative breast cancer via targeting ARF6. *Mol. Cancer Res.* 13: 330–8, <http://dx.doi.org/10.1158/1541-7786.MCR-14-0251>.
- [10] Sestak, I., Smith, S.G., Howell, A., Forbes, J.F., Cuzick, J. 2018. Early participant-reported symptoms as predictors of adherence to anastrozole in the International Breast Cancer Intervention Studies

- II. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. 29: 504–509, <http://dx.doi.org/10.1093/annonc/mdx713>.
- [11] Piccart-Gebhart, M.J., Procter, M., Leyland-Jones, B., Goldhirsch, A., Untch, M., *et al.* 2005. Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer. *N. Engl. J. Med.* 353: 1659–1672, <http://dx.doi.org/10.1056/NEJMoa052306>.
- [12] Lee, C.I., Goodwin, A., Wilcken, N. 2017. Fulvestrant for hormone-sensitive metastatic breast cancer. *Cochrane Database Syst. Rev.* 2017: CD011093, <http://dx.doi.org/10.1002/14651858.CD011093.pub2>.
- [13] NCI. 2019. Drugs approved for breast cancer—National cancer institute, NIH-NCI.
- [14] Parikesit, A.A., Stadler, P.F., Prohaska, S.J. 2011. Evolution and quantitative comparison of genome-wide protein domain distributions. *Genes (Basel)*. 2: 912–24. <http://dx.doi.org/10.3390/genes2040912>.
- [15] Zubor, P., Kubatka, P., Kajo, K., Dankova, Z., Polacek, H., *et al.* 2019. Why the gold standard approach by mammography demands extension by multiomics? Application of liquid biopsy mirna profiles to breast cancer disease management. *Int. J. Mol. Sci.* 20, <http://dx.doi.org/10.3390/ijms20122878>.
- [16] Chang, C.C., Chen, S.H. 2019. Developing a novel machine learning-based classification scheme for predicting SPCs in breast cancer survivors. *Front. Genet.* 10, <http://dx.doi.org/10.3389/fgene.2019.00848>.
- [17] Denkiewicz, M., Saha, I., Rakshit, S., Sarkar, J.P., Plewczynski, D. 2019. Identification of breast cancer subtype specific micrnas using survival analysis to find their role in transcriptomic regulation. *Front. Genet.* 10, <http://dx.doi.org/10.3389/fgene.2019.01047>.
- [18] Toy, W., Shen, Y., Won, H., Green, B., Sakr, R.A. *et al.* 2013. ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. *Nat. Genet.* 45: 1439–1445. <http://dx.doi.org/10.1038/ng.2822>.
- [19] Chitralla, K.N., Yeguvapalli, S. 2014. Computational screening and molecular dynamic simulation of breast cancer associated deleterious non-synonymous single nucleotide polymorphisms in TP53 gene. *PLoS One*. 9: e104242. <http://dx.doi.org/10.1371/journal.pone.0104242>.
- [20] Parikesit, A.A. 2018. Kontribusi aplikasi medis dari ilmu bioinformatika berdasarkan perkembangan pembelajaran mesin (machine learning) terbaru. *Cermin Dunia Kedokt.* 45: 700–703.
- [21] Jensen, L.J., Bateman, A. 2011. The rise and fall of supervised machine learning techniques. *Bioinformatics*. 27: 3331–3332. <http://dx.doi.org/10.1093/bioinformatics/btr585>.
- [22] Bartoli, L. 2011. Bioinformatics: a machine-learning approach to the prediction of protein, structure, function and interactions. *Altran Ital. Technol. Rev.* 5–20.
- [23] Larranaga, P., Zhang, Y.Q., Rajapakse, J.C., Larranaga, P., Zhang, Y.Q., *et al.* 2006. Machine learning in bioinformatics. *Brief. Bioinform.* 7: 86–112. <http://dx.doi.org/10.1093/bib/bbk007>.
- [24] Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G. 1998. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511790492>.
- [25] Eddy, S.R. 2004. What is a hidden Markov model? *Nat. Biotechnol.* 22: 1315–6, <http://dx.doi.org/10.1038/nbt1004-1315>.
- [26] Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., *et al.* 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43: D447–52. <http://dx.doi.org/10.1093/nar/gku1003>.
- [27] Fang, H., Gough, J. 2013. DeGO: Database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic Acids Res.* 41: D536–44, <http://dx.doi.org/10.1093/nar/gks1080>.
- [28] Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., Murzin, A.G. 2014. SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.* 42: D310, <http://dx.doi.org/10.1093/NAR/GK11242>.
- [29] Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S.E., Hubbard, T.J.P., *et al.* 2008. Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Res.* 36: D419–25, <http://dx.doi.org/10.1093/nar/gkm993>.
- [30] Kemena, C., Bornberg-Bauer, E. 2019. A roadmap to domain based proteomics. *Methods Mol. Biol.* pp. 287–300, http://dx.doi.org/10.1007/978-1-4939-8736-8_16.
- [31] Moore, A.D., Held, A., Terrapon, N., Weiner, J., Bornberg-Bauer, E. 2013. DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins. *Bioinformatics*. 30: btt640, <http://dx.doi.org/10.1093/bioinformatics/btt640>.
- [32] Wu, X., Al-Hasan, M., Chen, J.Y. 2014. Pathway and network analysis in proteomics. *J. Theor. Biol.* 362: 44–52, <http://dx.doi.org/10.1016/j.jtbi.2014.05.031>.
- [33] Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S.K., Teichmann, S.A., Weiner, J. 2005. The evolution of domain arrangements in proteins and interaction networks. *Cell. Mol. Life Sci.* 62: 435–45, <http://dx.doi.org/10.1007/s00018-004-4416-1>.
- [34] Kuang, X., Han, J.G., Zhao, N., Pang, B., Shyu, C.-R. *et al.* 2012. DOMMINO: a database of macromolecular interactions. *Nucleic Acids Res.* 40: D501–6, <http://dx.doi.org/10.1093/nar/gkr1128>.
- [35] Parikesit, A.A. 2018. The construction of two and three dimensional molecular models for the miR-31 and its silencer as the triple negative breast cancer biomarkers. *Online J. Biol. Sci.* 18: 424–431, <http://dx.doi.org/10.3844/ojbsci.2018.424.431>.

- [36] Parikesit, A.A., Anurogo, D. 2018. 3D prediction of breast cancer biomarker from the expression pathway of Lincrna-Ror/Mir-145/Arf6. *J. Sains Teknol.* 2: 10–19. <https://ojs.uph.edu/index.php/FaSTJST/article/view/1012>.
- [37] Parikesit, A.A., Utomo, D.H., Karimah, N. 2018. Protein domain annotation of Plasmodium spp. Circumsporozoite Protein (CSP) using hidden markov model-based tools. *J. Biol. Indones.* 14: 185–190, <http://dx.doi.org/10.14203/jbi.v14i2.3737>.
- [38] Parikesit, A.A., Nurdiansyah, R. 2019. The challenge of protein domain annotation with supervised learning approach: A systematic review. *J. Mat. Sains.* 24: 1–9.
- [39] Cense, J.-M. 1989. MolDraw: Molecular graphics for the Macintosh. *Tetrahedron Comput. Methodol.* 2: 65–71, [http://dx.doi.org/10.1016/0898-5529\(89\)90030-4](http://dx.doi.org/10.1016/0898-5529(89)90030-4).
- [40] Smith, T.J. 1995. MOLView: A program for analyzing and displaying atomic structures on the Macintosh personal computer. *J. Mol. Graph.* 13: 122–125, [http://dx.doi.org/10.1016/0263-7855\(94\)00019-O](http://dx.doi.org/10.1016/0263-7855(94)00019-O).
- [41] Baro, J.A., Hughes, H.C. 1991. The display and animation of full-color images in real time on the Macintosh computer. *Behav. Res. Methods Instrum. Comput.* 23: 537–545, <http://dx.doi.org/10.3758/BF03209994>.
- [42] Oliver, A., del Molino, J., Cañellas, M., Clar, A., Bibiloni, A. 2019. VR Macintosh Museum: Case Study of a WebVR Application, in: Springer, Cham. pp. 275–284, http://dx.doi.org/10.1007/978-3-030-16184-2_27.
- [43] Yu, Y., Nakhleh, L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics.* 16: S10, <http://dx.doi.org/10.1186/1471-2164-16-S10-S10>.
- [44] Yu, Y., Dong, J., Liu, K.J., Nakhleh, L. 2014. Maximum likelihood inference of reticulate evolutionary histories. *Proc. Natl. Acad. Sci.* 111: 16448–16453. <http://dx.doi.org/10.1073/PNAS.1407950111>.
- [45] Tomczak, K., Czerwińska, P., Wiznerowicz, M. 2015. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* 19: A68–77. <http://dx.doi.org/10.5114/wo.2014.47136>.
- [46] Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Alpi, E., et al. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* <http://dx.doi.org/10.1093/nar/gkw1099>.
- [47] Sievers, F., Higgins, D.G. 2014. Clustal omega. *Curr. Protoc. Bioinforma.* <http://dx.doi.org/10.1002/0471250953.bi0313s48>.
- [48] Hall, B.G. 2013. Building phylogenetic trees from molecular data with MEGA. *Mol. Biol. Evol.* 30: 1229–1235, <http://dx.doi.org/10.1093/molbev/mst012>.
- [49] Jeanquartier, F., Jean-Quartier, C., Holzinger, A. 2015. Integrated web visualizations for protein-protein interaction databases. *BMC Bioinformatics.* 16: 195, <http://dx.doi.org/10.1186/s12859-015-0615-z>.
- [50] Morais, D.A.D.L., Fang, H., Rackham, O.J.L., Wilson, D., Pethica, R. et al. 2011. SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res.* 39: D427–34, <http://dx.doi.org/10.1093/nar/gkq1130>.
- [51] Finn, R.D., Clements, J., Arndt, W., Miller, B.L., Wheeler, T.J. et al. 2015. HMMER web server: 2015 update. *Nucleic Acids Res.* 43: W30–8, <http://dx.doi.org/10.1093/nar/gkv397>.
- [52] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A. et al. 2011. The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* 39: D561–D568. <http://dx.doi.org/10.1093/nar/gkq973>.
- [53] Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S. et al. 2017. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45: D362–D368, <http://dx.doi.org/10.1093/nar/gkw937>.
- [54] Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., et al. 2004. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* 25: 1605–1612, <http://dx.doi.org/10.1002/jcc.20084>.
- [55] Lin, L.-Z., Cai, M.-G., Dai, Y.-C., Zheng, Z.-B., Jiang, F.-F. et al. 2017. BRMS1 gene expression may be associated with clinico-pathological features of breast cancer. *Biosci. Rep.* 37, <http://dx.doi.org/10.1042/BSR20170672>.
- [56] Shimelis, H., Mesman, R.L.S., Nicolai, C.V., Ehlen, A., Guidugli, L., et al. 2017. BRCA2 hypomorphic missense variants confer moderate risks of breast cancer. *Cancer Res.* 77: 2789–2799, <http://dx.doi.org/10.1158/0008-5472.CAN-16-2568>.
- [57] Wallez, Y., Riedl, S.J., Pasquale, E.B. 2014. Association of the breast cancer antiestrogen resistance protein 1 (BCAR1) and BCAR3 scaffolding proteins in cell signaling and antiestrogen resistance. *J. Biol. Chem.* 289: 10431–44, <http://dx.doi.org/10.1074/jbc.M113.541839>.
- [58] Godinho, M.F.E.E., Wulfschlegel, J.D., Look, M.P., Sieuwerts, A.M., Sleijfer, S. 2012. BCAR4 induces antioestrogen resistance but sensitises breast cancer to lapatinib. *Br. J. Cancer.* 107: 947–955, <http://dx.doi.org/10.1038/bjc.2012.351>.
- [59] UniprotKb. 2019. UniProtKB–P51587 (BRCA2_HUMAN), UniprotKb.
- [60] UniprotKb. 2019. UniProtKB–P56945 (BCAR1_HUMAN), UniprotKb.

- [61] Nepomuceno, T.C., De Gregoriis, G., De Oliveira, F.M.B., Suarez-Kurtz, G., Monteiro, A.N. 2017. The role of PALB2 in the DNA damage response and cancer predisposition. *Int. J. Mol. Sci.* 18, <http://dx.doi.org/10.3390/ijms18091886>.
- [62] Brinkman, A., Flier, S.V.D., Kok, E.M., Dorssers, L.C.J. 2000. BCAR1, a human homologue of the adapter protein p130Cas, and antiestrogen resistance in breast cancer cells. *J. Natl. Cancer Inst.* 92: 112–120, <http://dx.doi.org/10.1093/jnci/92.2.112>.
- [63] Godet, I., Gilkes, D.M. 2017. BRCA1 and BRCA2 mutations and treatment strategies for breast cancer. *Integr. Cancer Sci. Ther.* 4, <http://dx.doi.org/10.15761/icst.1000228>.
- [64] Godinho, M., Meijer, D., Setyono-Han, B., Dorssers, L.C.J., Van Agthoven, T. 2011. Characterization of BCAR4, a novel oncogene causing endocrine resistance in human breast cancer cells. *J. Cell. Physiol.* 226: 1741–1749. <http://dx.doi.org/10.1002/jcp.22503>.
- [65] Van Agthoven, T., Dorssers, L.C.J., Lehmann, U., Kreipe, H., Looijenga, L.H.J. 2015. Breast cancer anti-estrogen resistance 4 (BCAR4) drives proliferation of IPH-926 lobular carcinoma cells. *PLoS One.* 10: e0136845, <http://dx.doi.org/10.1371/journal.pone.0136845>.
- [66] Liu, X., Ehmed, E., Li, B., Dou, J., Qiao, X. *et al.* 2016. Breast cancer metastasis suppressor 1 modulates SIRT1-dependent p53 deacetylation through interacting with DBC. *Am. J. Cancer Res.* 6: 1441–9.
- [67] Li, Q., Xiao, L., Harihar, S., Welch, D.R., Vargis, E. Zhou, A. 2015. In vitro biophysical, microspectroscopic and cytotoxic evaluation of metastatic and non-metastatic cancer cells in responses to anti-cancer drug. *Anal. Methods.* 7: 10162–10169, <http://dx.doi.org/10.1039/c5ay01810b>.
- [68] Oates, M.E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M.J., *et al.* 2013. D2P2: Database of disordered protein predictions. *Nucleic Acids Res.* 41: D508–16, <http://dx.doi.org/10.1093/nar/gks1226>.
- [69] Prakash, A., Bateman, A. 2015. Domain atrophy creates rare cases of functional partial protein domains. *Genome Biol.* 16, <http://dx.doi.org/10.1186/s13059-015-0655-8>.