REVIEW Genomics & Informatics Vol. 8(4) 165-169, December 2010

Bioinformatics Resources of the Korean Bioinformation Center (KOBIC)

Byungwook Lee[1], In-Sun Chu[1], Namshin Kim[1], Jinhyuk Lee[1], Seon-Yong Kim[1], Wan Kyu Kim[2] and Sanghyuk Lee[1,2]*

[1]Korean Bioinformation Center (KOBIC), KRIBB, Daejeon 305-806, Korea, [2]Ewha Research Center for Systems Biology (ERCSB), Ewha Womans University, Seoul 120-750, Korea

Abstract

The Korean Bioinformation Center (KOBIC) is a national bioinformatics research center in Korea. We developed many bioinformatics algorithms and applications to facilitate the biological interpretation of OMICS data. Here we present an introduction to major bioinformatics resources of databases and tools developed at KOBIC. These resources are classified into three main fields: genome, proteome, and literature. In the genomic resources, we constructed several pipelines for next generation sequencing (NGS) data processing and developed analysis algorithms and web-based database servers including miRGator, ESTpass, and CleanEST. We also built integrated databases and servers for microarray expression data such as MDCDP. As for the proteome data, VnD database, WDAC, Localizome, and CHARMM_HM web servers are available for various purposes. We constructed IntoPub server and Patome database in the literature field. We continue constructing and maintaining the bioinformatics infrastructure and developing algorithms.

Keywords: bioinformatics programs, databases, web server, KOBIC

Introduction

The Korean Bioinformation Center (KOBIC) is a national research center for bioinformatics, founded in 2001 previously called National Genome Information Center (NGIC). KOBIC plays a key role in various bioinformatics research areas including genomics, proteomics, systems biology, and personalized medicine. We are responsible for the integration and management of bioresource and biodiversity information from various research labs and institutions across the country. We also provide a centralized data access portal to promote data sharing and utilization among research groups.

In an effort to support bioinformatics and genomics research in Korea, we develop numerous bioinformatics algorithms and applications, with emphases on i) next generation sequencing (NGS), ii) systems bioinformatics, iii) biomedical informatics, and iv) structural informatics. These resulting pipelines and knowledgebases are provided to biologists and bioinformaticians to facilitate the biological interpretation of OMICS data. In addition, we actively participate in many international collaborative projects for research and education.

Here we present bioinformatics resources including biological databases and tools constructed in KOBIC, which are tremendously useful to understand and process various biological data. All resources discussed are available from the KOBIC home page at www.kobic.re.kr. Table 1 provides an at-a-glance summary of these resources.

Genome and Transcriptome

Pipelines for NGS data

NGS has led to previously unimaginable amounts of data. To cope with the deluge of NGS data, we constructed NGS analysis pipelines, which can be divided into two major parts based on the processing data types: genome and transcriptome. In the genome pipeline, NGS read sequences were mapped to reference genome and then from the mapped reads we can identify SNP and structural variations such as Indel, CNV, and translocations. The genome pipeline also has *de novo* assembly pipeline for analysis of NGS reads with no reference genome. For the transcriptome research, we constructed integrated analysis pipelines to process RNA-seq, epigenome, and small RNA data generated from NGS technology. The transcriptome pipeline provides researchers with efficient ways to deal with the exprimental transcriptome data, allowing them to get information such as how different alleles of a gene are expressed, detect post-transcriptional mutations or identify gene fusions.

*Corresponding author: E-mail sanghyuk@kribb.re.kr
Tel +82-42-879-8500, Fax +82-42-879-8519
Accepted 2 December 2010

Table 1. A summary of selected web-based databases and tools provided by the KOBIC

Resources	Description	URL	Reference
NEUMA	Accurate quantification tool of transcriptome from RNA-Seq data by effective length normalization.	http://neuma.kobic.re.kr	(Lee, *et al.*, 2010)
miRGator v2.0	Integrated system for functional investigation of microRNAs.	http://miRGator.kobic.re.kr	(Cho, *et al.*, 2011)
ESTpass	Web-based server for processing and annotating expressed sequence tag (EST) sequences.	http://estpass.kobic.re.kr	(Lee, *et al.*, 2007)
CleanEST	Database server that classified dbEST libraries and removes contaminants.	http://cleanest.kobic.re.kr/	(Lee and Shin, 2009)
MDCDP	Microarray database for cancer diagnosis and prognosis to provide information on potential prognostic or diagnostic markers.	http://integromics.kobic.re.kr/mdcdp	
VnD	Structure-centric database of disease-related SNPs and drugs.	http://vnd.kobic.re.kr:8080/VnD	(Yang, *et al.*, 2011)
WDAC	Web server for measuring domain architecture similarity to identify homolog of multidomain proteins.	http://wdac.kr	(Lee and Lee, 2009)
Localizome	Web server for identifying transmembrane topologies and TM helices of eukaryotic proteins utilizing domain information.	http://localodom.kobic.re.kr/LocaloDom	(Lee, *et al.*, 2006)
CHARMM-HM	CHARMM-based Homology Model Builder.	http://psb.kobic.re.kr/charmm-hm	
IntoPub	Database server to provide web server information for processing biological data.	http://into.kobic.re.kr	
Patome	Database server of containing biological sequence data disclosed in patents and published applications, as well as their analysis information.	http://genepatent.kr	(Lee, *et al.*, 2007)

NEUMA

NEUMA (Normalization by Expected Uniquely Mappable Area) (Lee, *et al.*, 2010) is a novel, efficient and intuitive approach of estimating mRNA abundances from the whole transcriptome shotgun sequencing (RNA-Seq) (Marguerat and Bahler, 2010) data. NEUMA is based on effective length normalization using uniquely mappable areas of gene and mRNA isoform models. Using the known transcriptome sequence model such as RefSeq, NEUMA pre-computes the numbers of all possible gene-wise and isoform-wise informative reads: the former being sequences mapped to all mRNA isoforms of a single gene exclusively and the latter uniquely mapped to a single mRNA isoform. The results are used to estimate the effective length of genes and transcripts, taking experimental distributions of fragment size into consideration. We propose that NEUMA could make a standard method in quantifying gene transcript levels from RNA-Seq data.

miRGator v2.0

miRGator v2.0 (Cho, *et al.*, 2011) is an integrated database of microRNA (miRNA)-associated gene expression, target prediction, disease association and genomic annotation, which aims to facilitate functional investigation of miRNAs. The database contains information about (i) human miRNA expression profiles under various experimental conditions, (ii) paired expression profiles of both mRNAs and miRNAs, (iii) gene expression profiles under miRNA-perturbation (e.g. miRNA knockout and overexpression), (iv) known/predicted miRNA targets and (v) miRNA-disease associations. In total, $>$8,000 miRNA expression profiles, $\sim$300 miRNA-perturbed gene expression profiles and $\sim$2000 mRNA expression profiles are compiled with manually curated annotations on disease, tissue type and perturbation. The miRGator serves as a reference database to investigate miRNA expression and function.

ESTpass

ESTpass (Lee, *et al.*, 2007) is a web-based server for processing and annotating sequence data from expressed sequence tag (EST) projects. ESTpass accepts a FASTA-formatted EST file and its quality file as inputs, and it then executes a back-end EST analysis pipeline consisting of three consecutive steps. The first is cleansing the input EST sequences. The second is clustering and assembling the cleansed EST sequences using d2_cluster (Burke, *et al.*, 1999) and CAP3 (Huang and Madan, 1999) programs and producing putative transcripts. From the CAP3 output, ESTpass detects chimeric EST sequences which are confirmed through comparison with the nr database. The last step is annotating the putative transcript sequences using RefSeq (Pruitt, *et al.*, 2007), InterPro (Hunter, *et al.*, 2009), GO (Barrell, *et al.*, 2009) and KEGG (Arakawa, *et al.*, 2005) gene databases according to user-specified options.

The major advantages of ESTpass are the integration of cleansing and annotating processes, exhaustive annotation, and email reporting to inform the user about the progress and to send the analysis results.

CleanEST

CleanEST (Lee and Shin, 2009) is a database server that classified dbEST (Radeva, *et al*., 2008), the EST division of GenBank, libraries and removed contaminants. In the CleanEST, all dbEST libraries were classified according to species and sequencing center. In addition, human EST libraries were further classified by anatomical and pathological systems according to eVOC (Kelso, *et al*., 2003) ontologies. For each dbEST library, we provide two different cleansed sequences: 'pre-cleansed' and 'user-cleansed'. To generate pre-cleansed sequences, we cleansed sequences in dbEST by alignment of EST sequences against well-known contamination sources: UniVec, Escherichia coli, mitochondria and chloroplast (for plant). To provide user-cleansed sequences, we built an automatic user-cleansing pipeline, in which sequences of a user-selected library are cleansed on-the-fly according to user-selected options.

MDCDP

MDCDP (Microarray Database for Cancer Diagnosis and Prognosis) is a microarray database to provide information on potential prognostic or diagnostic markers identified through the integrative analyses of cancer gene expression datasets. We have collected publicly available gene expression data sets encompassing about 5,000 samples in 13 different tissues with clinical information on patient's survival. Available information in MDCDP is as follows: prognostic information for each gene, results of differential gene expression analyses, results of gene set enrichment analyses, Kaplan-Meier survival analyses across cancer subtypes, and heatmaps of expression levels for each gene.

Proteome

VnD

VnD (the variations and drugs) is a consolidated database containing information on diseases, related genes and genetic variations, protein structures and drug information. VnD was built in three steps. First, we integrated various resources systematically to deduce catalogs of disease-related genes, single nucleotide polymorphisms (SNPs), protein mutations and relevant drugs. Next, we carried out structure modeling and docking simulation for wild-type and mutant proteins to examine the structural and functional consequences of non-synonymous SNPs in the drug-related genes. Finally, we investigated the structural and biochemical properties relevant to drug binding such as the distribution of SNPs in proximal protein pockets, thermo-chemical stability, interactions with drugs and physico-chemical properties. The VnD database would be a useful platform for researchers studying the underlying mechanism for association among genetic variations, diseases and drugs.

WDAC

WDAC (Weighed Domain Architecture Comparison) (Lee and Lee, 2009) is used to identify homolog of multi-domain proteins. The key ideas of WDAC are the use of weight scores for domain promiscuity and combining domain architecture comparison with sequence similarity method. WDAC assigned a weight score to Pfam domain extracted from RefSeq proteins based on its abundance and versatility to distinguish these promiscuous domains from conventional protein domains. To measure the similarity of two domain architectures, cosine similarity is used. WDAC combined sequence similarity with domain architecture comparisons to identify proteins belonging to the same domain architecture.

Localizome

The Localizome (Lee, *et al*., 2006) server predicts the transmembrane (TM) helix number and TM topology of a user-supplied eukaryotic protein and presents the result as an intuitive graphic representation. It utilizes hmmpfam to detect the presence of Pfam (Finn, *et al*., 2010) domains and a prediction algorithm, Phobius (Kall, *et al*., 2007), to predict the TM helices. The results are combined and checked against the TM topology rules stored in a protein domain database called LocaloDom. LocaloDom is a curated database that contains TM topologies and TM helix numbers of known protein domains. It was constructed from Pfam domains combined with Swiss-Prot annotations and Phobius predictions. The Localizome server is a highly accurate and comprehensive information source for subcellular localization for soluble proteins as well as membrane proteins.

CHARMM-HM

CHARMM-HM is a web-based service dedicated to automated homology modeling based on the CHARMM (Chemistry at HARvard Macromolecular Mechanics)

program. The core module of CHARMM-HM is a homology model builder, written in FORTRAN and implemented in CHARMM. It assists researchers in building homology models using existing CHARMM global optimization methods. A user provides the input, alignment or sequence, to the web server, and receives the output via electronic mail. The generated structures and their quality evaluation scores are kept on the web server for the further analysis.

Literature

IntoPub

IntoPub (Informatics tools in PubMed) is a web-based database server to provide web server information for processing biological data. To construct IntoPub, we obtained abstracts containing web resources information from the PubMed (Sequeira, *et al*., 2001) database and annotated the obtained abstracts automatically and manually. This database server provides information on comprehensive and up-to-dated tools and databases covering all the biological fields. New web resources introduced by PubMed have been updated automatically on IntoPub.

Patome

With the advent of automated and high-throughput techniques, the number of patent applications containing biological sequences has been increasing rapidly. Patome (Lee, *et al*., 2007) is a database server, which contains biological sequence data disclosed in patents and published applications, as well as their analysis information. The disclosed sequences were annotated with RefSeq database and then were linked to Entrez Gene, OMIM and GO databases. From these results, patents are mapped into genes. The gene?patent mapping table can be used to identify whether a particular gene or disease is related to patenting.

KOBIC Computer Facilities

A large-scale computer system is an essential component of bioinformatics research. Thus we manage high-performance servers, clusters, and storage for building and maintaining KOBIC's biological resources. Currently we have high performance clusters with more than 2,000 CPU cores, storage system of 1,500 TB for large scale data, more than 250 machines for biological analysis and web systems, and 10Gbps network switches for high speed data transmission.

Conclusion

The huge quantities of data that are now being generated by life-science researchers provide unforeseen challenges and opportunities. The biological research has been fundamentally changed by shifting towards asking questions on a genome-wide scale rather than one gene at a time. In addition, NGS methods are providing the technology to identity individual genomes, to quantify expression, to measure biodiversity and to study cancer differentiation. To deal with these situations, KOBIC continue building and maintaining the bioinformatics infrastructure and developing algorithms. Most of the resources described here include documentation, other explanatory material, and references to collaborators and data sources on the respective Web sites. A user support staff is available to answer question at help@kobic.kr

Acknowledgements

The work was supported by the KRIBB Research Initiative Program. We would like to thanks all KOBIC members for developing and maintaining KOBIC bioinformatics resources.

References

Arakawa, K., Kono, N., Yamada, Y., Mori, H., and Tomita, M. (2005). KEGG-based pathway visualization tool for complex omics data. *In Silico Biol.* 5, 419-423.

Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C., and Apweiler, R. (2009). The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucl. Acids Res.* 37, D396-403.

Burke, J., Davison, D., and Hide, W. (1999). d2_cluster: a validated method for clustering EST and full-length cDNAsequences. *Genome Res.* 9, 1135-1142.

Cho, S., Jun, Y., Lee, S., Choi, H.S., Jung, S., Jang, Y., Park, C., Kim, S., and Kim, W. (2011). miRGator v2.0 : an integrated system for functional investigation of microRNAs. *Nucl. Acids Res.* 39, D158-162.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R., and Bateman, A. (2010). The Pfam protein families database. *Nucl. Acids Res.* 38, D211-222.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Res.* 9, 868-877.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A.F., Selengut, J.D., Sigrist, C.J., Thimma, M., Thomas, P.D., Valentin, F., Wilson, D., Wu,

C.H., and Yeats, C. (2009). InterPro: the integrative protein signature database. *Nucl. Acids Res.* 37, D211-215.
Kall, L., Krogh, A., and Sonnhammer, E.L. (2007). Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucl. Acids Res.* 35, W429-432.
Kelso, J., Visagie, J., Theiler, G., Christoffels, A., Bardien, S., Smedley, D., Otgaar, D., Greyling, G., Jongeneel, C.V., McCarthy, M.I., Hide, T., and Hide, W. (2003). eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res.* 13, 1222-1230.
Lee, B., and Lee, D. (2009). Protein comparison at the domain architecture level. *BMC Bioinformatics* 10 Suppl 15, S5.
Lee, B., and Shin, G. (2009). CleanEST: a database of cleansed EST libraries. *Nucl. Acids Res.* 37, D686-689.
Lee, B., Hong, T., Byun, S.J., Woo, T., and Choi, Y.J. (2007). ESTpass: a web-based server for processing and annotating expressed sequence tag (EST) sequences. *Nucl. Acids Res.* 35, W159-162.
Lee, B., Kim, T., Kim, S.K., Lee, K.H., and Lee, D. (2007). Patome: a database server for biological sequence annotation and analysis in issued patents and published patent applications. *Nucl. Acids Res.* 35, D47-50.
Lee, S., Lee, B., Jang, I., Kim, S., and Bhak, J. (2006). Localizome: a server for identifying transmembrane topologies and TM helices of eukaryotic proteins utilizing domain information. *Nucl. Acids Res.* 34, W99-103.
Lee, S., Seo, C.H., Lim, B., Yang, J.O., Oh, J., Kim, M., Lee, B., and Kang, C. (2010). Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucl. Acids Res.* 38, 1-10.
Marguerat, S., and Bahler, J. (2010). RNA-seq: from technology to biology. *Cell Mol. Life Sci.* 67, 569-579.
Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucl. Acids Res.* 35, D61-65.
Radeva, M., Hofmann, T., Altenberg, B., Mothes, H., Richter, K.K., Pool-Zobel, B., and Greulich, K.O. (2008). The database dbEST correctly predicts gene expression in colon cancer patients. *Curr. Pharm. Biotechnol.* 9, 510-515.
Sequeira, E., McEntyre, J., and Lipman, D. (2001). PubMed Central decentralized. *Nature* 410, 740.
Yang, J.O., Oh, S., Ko, G., Park, S.J., Kim, W.Y., Lee, B., and Lee, S. (2011). VnD: a structure-centric database of disease-related SNPs and drugs. *Nucl. Acids Res.* 39, D939-944.