Bioinformatics and Biosystems 2006, Vol. 1, No. 2, pp. 95-98

Retrieving Protein Domain Encoding DNA Sequences Automatically Through Database Cross-referencing

Yoon-sup Choi[1], Jae-Seong Yang[1], Sung Ho Ryu[1,2], Sanguk Kim[1,2,3]

[1]School of Interdisciplinary Bioscience and Bioengineering,
[2]Department of Molecular and Life Science,
[3]Biological Research Information Center,
Pohang University of Science and Technology, Pohang, Kyungbuk, 790-784, Republic of Korea

Abstract

Recent proteomic studies of protein domains require high-throughput and systematic approaches. Since most experiments using protein domains, the modules of protein-protein interactions, require gene cloning, the first experimental step should be retrieving DNA sequences of domain encoding regions from databases. For a large scale proteomic research, however, it is a laborious task to extract a large number of domain sequences manually from several inter-linked databases. We present a new methodology to retrieve DNA sequences of domain encoding regions through automatic database cross-referencing. To extract protein domain encoding regions, it traverses several inter-connected database with validation process. And we applied this method to retrieve all the EGF domain encoding DNA sequences of homo sapiens. This new algorithm was implemented using Python library PAMIE, which enables to cross-reference across distinct databases automatically.

Introduction

Genome projects are generating vast amounts of data that provide the existence of thousands of new gene products, especially the list of proteins responsible for cellular regulation. However it does not immediately reveal what these proteins do, nor how they are assembled into the molecular machines and functional networks that control cellular behavior (Pawson et al., 2003). Cellular processes and overall molecular architectures of all organisms are largely mediated through elaborate scaffolds of protein-protein interactions. Thus, the high-throughput strategies to study protein-protein interactions, such as yeast two-hybrid screening, have been developed to describe the protein interaction networks and to construct the protein interaction maps in model organisms (Uetz et al., 2000, Li et al., 2004, Ghavidel et al, 2005). However, proteins interact with more than one partner at a time, it is difficult to interpret large scale protein-protein interactions (Santonico et al., 2005). Protein domains represent the modular nature of proteins, which fold independently and often perform specific tasks. While protein domains could interact with several binding partners, they are the single binding modules and interact with only one partner at a time (Santonico et al., 2005). Thus, the domain knowledge can help to obtain a clearer representation of the protein networks.

The experiments using protein domains need to extract the sequences of domain encoding regions from distinct databases for gene cloning and protein expression, although this process often performed manually (Yu et al., 2004). However, for the high-throughput proteomic experiments, the manual retrieval is daunting due to the following three reasons. First, it needs to collect the information of hundreds or thousands of protein domains for large scale experiments. Second, domain knowledge is not located in a single source so that one should cross-refer separately updating interconnected databases. Third, iterative extraction process can be erroneous since databases sometimes contain dubious entries and point to missing links. Thus, proper decision making policies are essential to eliminate the database entry errors and to validate the results. Therefore, there are needs to develop bioinformatics methodology for retrieving genetic information of domains encoding region to conduct

Corresponding Author: Sanguk Kim (Email:sukim@ postech. ac.kr)
This work was supported by the Korea Research Foundation Grant by the Korean Government (MOEHRD) (KRF-2005-070-C00095) and POSTECH BSRI research fund-2005.

large scale proteomic researches.

Here we developed a methodology to extract protein domain encoding DNA sequence automatically from three distinct databases: Pfam, UniProt and GeneBank (Finn et al., 2006, Wu et al., 2006, Benson et al., 2006) using Python library PAMIE. The algorithm also includes the validation process to verify the retrieved data. We applied this method to extract all the EGF domain encoding regions of homo sapiens for further large-scale proteomic experiments. The EGF (Epidermal Growth Factor) domain is a widely distributed, independently folding protein module that is thought to play a general role in extracelluar events such as adhesion, coagulation, and receptor-ligand interactions (Downing et al, 1996).

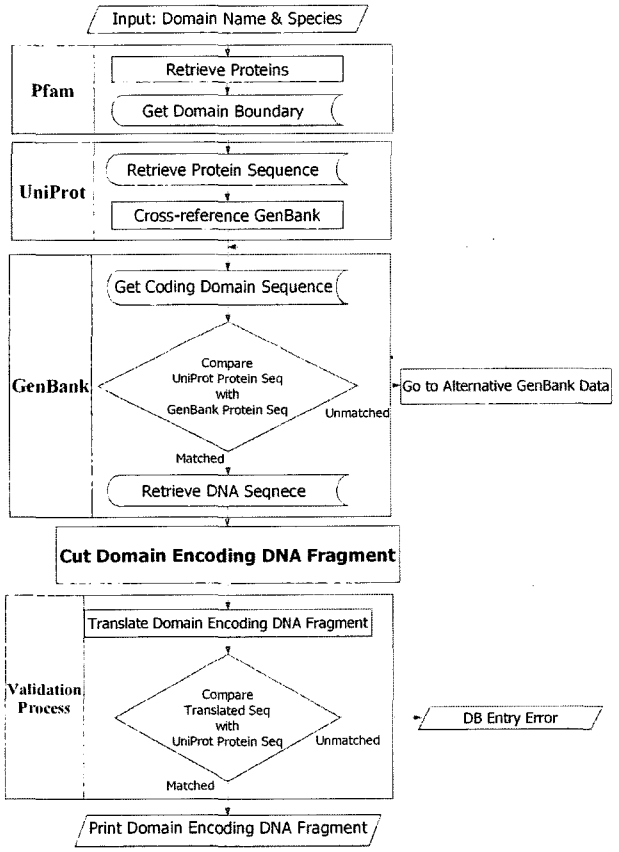

Figure 1. The Algorithm of retrieving domain encoding sequences through database cross-referencing

Algorithm

To retrieve EGF domain encoding DNA sequences, the process traverses three separately updating databases: Pfam, UniProt, GenBank (Figure 1). In Pfam, it retrieves the list of human proteins containing the EGF domain and the domain boundary of each protein sequence. Moving to UniProt through links from Pfam, the process collects protein sequences and cross-references to GenBank mRNA data. When a UniProt entry has more than one cross-reference mRNA from GenBank, this program finds the best match by comparing the protein sequences of UniProt with each GenBank entry. Then, it extracts corresponding EGF domain encoding DNA fragments from GenBank using domain boundary information retrieved from Pfam. After data acquisition, it goes through validation process. To verify the extracted DNA sequences, the translated amino acid sequences using codon table were compared to the protein sequences from UniProt. If errors are found including stop codons in the middle of extracted sequence or unusual nucleotides (other than A, T, G, C), or mismatches compared to UniProt entries, the process provide the proper error messages.

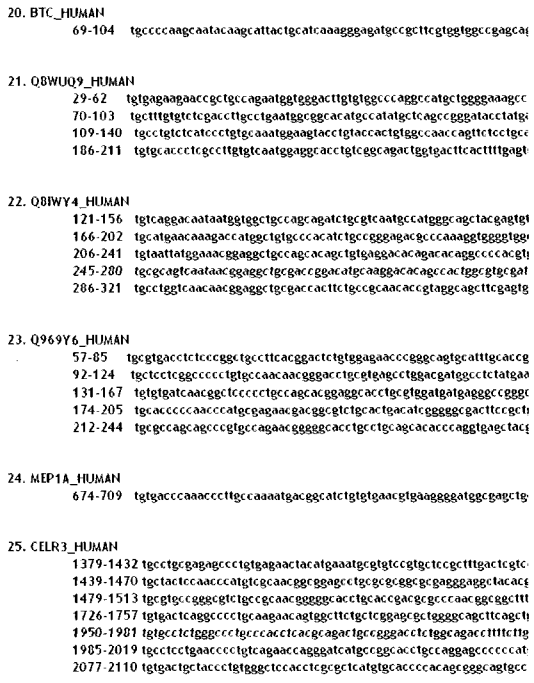

Figure 2. Part of the extracted human EGF domain encoding sequences: The result consists of protein names, domain boundaries, and domain encoding DNA sequences.

Table 1. EGF domain extracted from the database cross-referencing.

Domain encoding sequences & Erroneous entries	number
Proteins having EGF domains	2280
Human proteins having EGF domains	292
EGF domains from homo sapiens	985
mRNA sequence data not found	25
Erroneous entries detected from the validation process	49

Implementation & Results

We have implemented this method in python with PAMIE, Python Automated Module for Internet Explorer (http://pamie.sourceforge.net/). Using the PAMIE library functions, it is possible to click the links, automatically fill out the text box, push the radio button, or get the HTML source from the web site. So it enabled the application to traverse three distinct databases iteratively. Furthermore, the web interface for searching general protein domain and species search could be easily implemented. The source code is available upon request for academic purpose. We applied this method to the extraction of EGF domain encoding regions of homo sapiens. It took approximately 4 hours to complete the whole process. The results and statistics are shown in Table 1 and the part of extracted result is represented in Figure 2. We found 2280 proteins in Pfam have EGF domains. They include 292 human proteins, which have 985 EGF domains. The process could find the missing links and erroneous entries in cross-linked databases. We found 25 proteins miss-linked, so no mRNA sequence data found in GenBank that linked to Uniprot. Another 49 proteins have errors and detected during validation process, which have stop codons in the middle of sequence, unusual nucleotides (other than A, T, G, C) or mismatches compared to UniProt entries.

Limitations

Our method depends on the network-connected databases so that the process will not be available when the referred databases were down. And the process will not retrieve nucleotide sequences, if the corresponding mRNAs are not found in GenBank. In this case, user gets error massage, 'No mRNA DATA'.

Acknowledgement

We thank SBI members, Jiyoung Park and Chungoo Park for critical reading of the manuscript and helpful comments.

References

[1] Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2006) GenBank. Nucleic Acids Research, 34: 16-20.

[2] Downing,A.K., Knott,V., Werner,J.M., Cardy,C.M., Campbell,I.D. and Handford,P.A. (1996) Solution structure of a pair of calcium-binding epidermal growth factor-like domains: implications for the Marfan syndrome and other genetic disorders. Cell, 85: 597-605

[3] Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R., Eddy,S.R., Sonnhammer,E.L. and Bateman,A. (2006) Pfam: clans, web tools and services. Nucleic Acids Research, 34: 247-251.

[4] Ghavidel,A., Cagney,G. and Emili,A. (2005) A skeleton of the human protein interactome. Cell, 122: 830-832.

[5] Li,S., Armstrong,C.M., Bertin,N., Ge,H., Milstein,S., Boxem,M., Vidalain,P.O., Han,J.D., Chesneau,A., Hao,T., Goldberg,D.S., Li,N., Martinez,M., Rual,J.F., Lamesch,P., Xu,L., Tewari,M., Wong,S.L., Zhang,L.V., Berriz,G.F., Jacotot,L., Vaglio,P., Reboul,J., Hirozane-Kishikawa,T., Li ,Q., Gabel,H.W., Elewa,A., Baumgartner,B., Rose,D.J., Yu ,H., Bosak,S., Sequerra,R., Fraser,A., Mango,S.E., Saxton,W.M., Strome,S., van den Heuvel,S., Piano,F., Vandenhaute,J., Sardet,C., Gerstein,M., Doucette-Stamm,L., Gunsalus,K.C., Harper,J.W., Cusick,M.E., Roth,F.P., Hill,D.E. and Vidal,M. (2004) A map of the interactome network of the metazoan C. elegans. Science, 303: 540-543.

[6] Pawson,T. and Nash,P. (2003) Assembly of cell regulatory systems through protein interaction domains. Science, 300: 445-452.

[7] Santonico,E., Castagnoli,L. and Cesareni,G. (2005) Methods to reveal domain networks. Drug Discov. Today, 10: 1111-1117.

[8] Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P., Qureshi-Emili,A., Li,Y., Godwin,B., Conover,D., Kalbfleisch,T., Vijayadamodar,G., Yang,M.,

Johnston,M., Fields,S. and Rothberg,J.M. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature, 403: 623-627

[9] Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M., Martin,M.J., Mazumder,R., O'Donovan,C., Redaschi,N. and Suzek,B. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Research, 34: 187-191.

[10] Yu,J.W., Mendrola,J.M., Audhya,A., Singh,S., Keleti,D., DeWald,D.B., Murray,D., Emr,S.D. and Lemmon,M.A.. (2004) Genome-wide analysis of membrane targeting by S. cerevisiae pleckstrin homology domains. Molecular Cell, 13: 677-688.