모티프 자원에 대한 통합 접근 검색 메소드 구현

Nguyen Thu Trang, 이 범 주, 류 근 호
충북 대학교 데이터베이스 & 바이오인포메틱스 연구실
e-mail : {trangnt, bjlee, khryu}@dblab.chungbuk.ac.kr

Implementation of Integrated Search Method for Motif Resources

Nguyen Thu Trang, Bum Ju Lee, Keun Ho Ryu
Lab. of Database & Bioinformatics, Chungbuk National University

요 약

Motif which is a recurrent feature shared by a significant segments from the proteins belonging to different families can be used to predict functional or structural properties of the protein, or to describe features common to biologically related proteins. Although there are many existed motif databases, they are unstandardized and built up by different methods. So it is difficult to get enough motif information when researches use motif databases for classification, prediction, etc. Using logical integration of biological based on cross-references between these databases is the common solution for information tracking through www. However, the problem such as the motif resource tracking method for making a united motif information page emerges from cross-reference.
In this paper we propose the motif resource tracking method based on the existed motif databases for easy and comfortable interface better than in the case of using each motif database. The result which is given by this method is the review of necessary information about any motif in PROSITE database.

1. Introduction

Nowadays, the development of computer programs and database resources for bioinformatics applications makes the supply of biology information more easily and comfortably. This means that the field of biology is becoming increasingly dependent on the computer software, and at that time the attention of some researchers is the analysis and extracting information from large biological resources [3]. Motifs are the “conserved” segments between protein sequences that are extracted from the proteins belong to different families. It is the smallest unit executing the role of protein structure and function. So the motif detection is suggestive of function preserved by evolution. Existed motif databases such as InterPro [9, 10], PROSITE [6, 7, 8], BLOCKS [11, 12, 13], PRINTS [14] and Pfam [15, 16] created the use of each different method. Therefore, when extracting motif information, researches give the results of different formats because motif databases are unstandardized [2, 3]. This disunity makes researchers difficult to synthesize and compare motif information between different resources.

On the other hand, almost the existed motif database has the organic linkage mechanism among biologically various fragment information [1]. When researchers want to get enough any motif information, they have to link to many databases and the received result can be the redundant or coincident information. For solving these problems, we suggest the system with motif resource tracking method to get its information. Though this system, users can see the link between different databases obviously and get enough the necessary information about any motif [2].

This paper is organized as followed. In section 2, a detailed survey of the previous related work will be introduced and section 3 will describe a motif resource tracking method which is applied in our system. The implementation of this system and comparison among other systems will be presented in the fourth section. Finally, in section 5, some conclusions and the outline directions for future work will be presented.

2. Related Work

In this section, we talk about the existing motif databases and

the linkage between them through cross-references.
Each motif database such as PROSITE, BLOCKS, PRINTS and Pfam generates heterogeneous motif formats using individually different methods and has particular features. The focus of each database is different although they share a common interest in protein sequence classification. So the given information of each database is also different. Each database has its strength and weaknesses but in general, all these databases are abundant resources and useful tools for protein sequence analysis [1, 2, 3]. Nowadays, the common trend is integration different databases into a single, coherent protein-signature resource to take full advantage of each individual database. Such databases are SRS, InterPro, MetaFam, etc. The important motif databases and corresponding cross-reference to other databases are as follows.

SRS (Sequence Retrieve System)
The European Bioinformatics Institute (EBI) is the centre of research and services in bioinformatics. It manages and analyses databases of biological data including DNA, protein sequences and macromolecular structures [17]. SRS, one of the products of Lion BioScience AG, is the world's premier data integration, analysis and display tool for bioinformatics, genomic and related data. It can be used to browse the various biological sequence and literature databases the EBI has available. The integration method used by SRS allows all data and tools to be accessible through a single interface [18].

InterPro
InterPro is an integrated documentation resource for protein families, domains and functional sites, which integrated information of component databases such as PROSITE, PRINTS, Pfam and ProDom into a single comprehensive format [10]. Each entry is the group of PROSITE patterns and profiles, PRINTS fingerprints, ProDom domains and Pfam, SMART and TIGRFAMs HMMs that provide methods for identifying the same domain with in protein sequences. Moreover, a list of pre-computed matches against the whole of the SWISS-PROT and TrEMBL databases is given based on InterPro entry hits. Because of the use of various motif resources, InterPro provides quality control mechanisms for assessing individual methods and supports wide coverage. However, its disadvantage is not enough to include motif 3D structure and classification information [1, 2, 9].

PROSITE
PROSITE is a database of both patterns and profiles. Each patterns in PROSITE are built from alignments of related sequences which are taken from a variety of sources [6, 8]. This database uses SWISS-PROT entries to assess the matching of core pattern with protein families. Beside some advantages, patterns also have their limitations across whole sequences, which is why PROSITE also creates profiles to complement the pattern [2, 7]. PROSITE provides ScanProsite, MotifScan tools to scan a sequence or a pattern against other databases such as SWISS-PROT, PDB and Pfam. The relationship of PROSITE with the SWISS-PROT database allows the evaluation of the sensitivity and specificity of the PROSITE motifs and their periodic reviewing. The result page which is retrieved though the searching process in PROSITE contains PDB cross-references but it only gives the structure image without detailed information such as PDB sequence.

BLOCKS
One of the main protein signature databases is BLOCKS, a collect of multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. This database is based on InterPro entries with sequences from SWISS-PROT and TrEMBL and with cross-referances to PROSITE and/or PRINTS and/or SMART, and/or PFAM and/or ProDom entries. However, BLOCKS database with its searching tools such as BlockSearcher or BlockMaker stores and produces purely multi blocks and distances between them observed for sequences in the protein family but not gives specific information about them such as their properties or structure [11, 12, 13].

PRINTS
Built and maintained at the University of Manchester, PRINTS is a public domain database which uses "fingerprints" as diagnostic signatures in a variation on the pattern recognition methods [1, 2]. A fingerprint is a group of conserved motifs used to characterize a protein family; its diagnostic power is refined by iterative scanning of a SWISS-PROT/TrEMBL composite. The occurrence of these conserved areas across the whole is taken into account but not focused on small conserved areas. In a database search, creating many fingerprints would be greater chance of identifying proteins at the superfamily as well as the family and subfamily levels. About cross-references to others, PRINTS is a companion to the BLOCKS, PROSITE, Pfam and ProDom database [14]. But fingerprints do not support weight, secondary structure information and similarity data.

Pfam
Pfam database uses Hidden Markov Models (HMMs) as a way of creating diagnostic signatures for protein families, domains and conserved protein region [15]. Each family in Pfam is represented by two multiple sequence alignments and two profile-HMMs. Pfam is actually composed of two sets of families: Pfam-A families are based on curated multiple alignments whereas Pfam-B is an automatic (no manual curation) clustering of the rest of SWISS-PROT and TrEMBL derived from ProDom database [2, 16].

3. Motif Resources Access

Among existed motif database, PROSITE is the first one created and continuously evolved since. Our motif resource tracking method is built based on the database file PROSITE.DAT, a computer readable file that contains all the information necessary to programs that will scan sequence(s) with pattern and/or matrix. The table which is created by this file consists of name, accession number, description and pattern for each entry. Because of the relationship between PROSITE, SWISS-PROT and PDB, the cross-reference linkage to other databases can be tracked obviously. The common information about description, consensus pattern and references for any motif can be extracted though PROSITE access number (i.e. PS00001). Figure 1 shows the functions which are used in the tracking method.

Each sequence region described in a PROSITE entry can have or not corresponding structural data in PDB. These data are listed by the PDB code at the "3D" (3D-structure) lines in the database file of PROSITE. The "DR" (Database Reference) lines in this file are used as pointers to the SWISS-PROT entries that picked up (or missed) by the pattern being described in the PROSITE entry. However, our tracking method does not use this cross-reference to get corresponding sequences because the one-to-many relationship between PROSITE patterns and SWISS-PROT makes the motif tracking more complicated. To turn a difficulty, SWISS-PROT information will be extracted through PDB cross-references.

From the chosen PDB cross-reference, we can get the detailed sequence and coordination of atoms in this sequence. The PDB sequence has FASTA format which is accepted for many multiple sequence alignment programs. In FASTA format, the first word on the line is the name of sequence; the rest of the line is a description of sequence and the remaining lines contain the sequence itself. Moreover, based on the coordinate of atoms in the sequence, the three-dimension image is also given to provide a view of motif structure for users.

In the PDB flat file, the DBREF records which are in primary structure section provide cross-reference links between PDB sequences and the corresponding database entry or entries. Based on this record, cross-reference to SWISS-PROT are listed and chosen to continue in tracking to other motif databases.

```
public class GetProFrame extends JFrame {
        //Get summarized information of this Prosite number
        public void getSumInfo(String acc) {…}

        //Get PDB cross-reference for this Prosite number
  public static String getPDB(String acc) {…}

        //Get information from PDB database
        public void getPDBSequence() {…}
        public int getPDBCoordinate() {…}
        public void getImage() {…}

        //Get SWISS-PROT cross-reference
        public void addSwiss() {…}
        public void getSwissSequence() {…}

        //Check cross-reference to other databases
        public void checkOtherDB() {…}
        //Functions to get corresponding information
        //from the InterPro, BLOCKS, Pfam and PRINTS
        //based on the choice of users
}
```

Figure 1 Used functions in motif resource tracking method

As we knew in the related work section, SWISS-PROT database closely related other database by its cross-reference such as InterPro, BLOCKS, Pfam and PRINTS, etc. The "DR" lines (Database cross-Reference) which are in SWISS-PROT flat file are used as pointers to information related in all cross-referenced databases. From those cross-references, the tracking method continues to check whether the next corresponding database is extracted related information.

4. Implementation and Evaluation

For building this method, we used WindowXP as operating system and Java (version 1.5+) which supports for network and user interface programming as programming language. Figure 2 shows the user interface of this system with the motif information which is extracted through cross-referenced motif databases such as PROSITE, PDB, SWISS-PROT, etc.

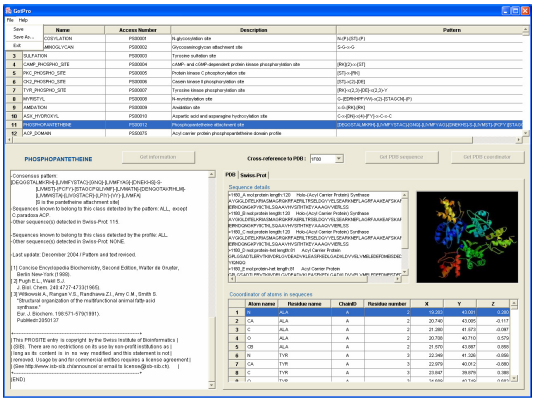

Figure 2 User interface of our system with the summary information (in the left text area), PDB detailed sequence in FASTA (in the top-right text area), 3-dimension image and the table of atom coordinate for chosen motif in the table of entries

Through this motif tracking process in existed databases, we gave the full view of motif information for users. With the PROSITE.DAT flat file as input, our system helps users in tracking motif information step by step. Users can get all necessary information of any motif in PROSITE database such as summary information, primary structure (sequence), coordinate of atom and its three-dimension structure image…After that all information can be saved in one page which is useful in motif information synthesis through the interface as in Figure 3.

Each existed database has its advantages and individual strengths of the different methods with the tremendous data. However, no system has enough function and information because it is built for the specific purpose. By motif resource tracking, our system somewhat carried out the data synthesis, supplement necessary information and concurrently reduce the problem of redundant or coincident information. This thing is also the improvement based on disadvantages of existed database which has cross-references to others.

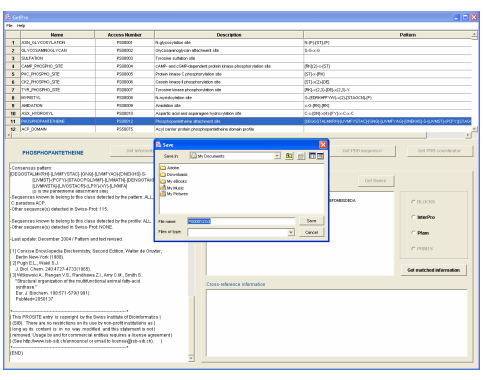

Figure 3 Users can choose the folder and file name to save motif information into one file

5. Conclusion

With the large influx of raw sequence data into protein databases, the problem is the synthesis biological knowledge for particular and classification protein sequences. Using correct tools and resources is to get enough information very important, especially when analysis the smallest units of protein sequence – motifs to find sequences in the same families.

Therefore, this method enables us to solve the problem of redundant or coincident information, standardization result between motifs resources. Also, our system provides the full view of motif information through the tracking of existed database such as PROSITE, PDB, SWISS-PROT, InterPro, BLOCKS, etc.

Future works will develop this system to the web-based integration and search system. At the time, it can support the prediction and classification protein structure based on the integrated information through existed protein databases.

Reference

[1] Bum Ju Lee, Heon Gyu Lee, Keun Ho Ryu. Integration of Motif Resource for Protein Analysis. Bioinfor2005
[2] Nicola J Mulder and Rolf Apweiler. Tools and resources for identify protein families, domains and motifs. Genome Biology, 3(I): reviews 2001.1-2001.8
[3] A.C.Siepel, A.N.Tolopko, A.D.Farmer, P.A.Steadman, F.D.Schilkey, B.D.Perry, W.D.Beavis. An integration platform for heterogeneous bioinformatics software components. IBM Systems Journal, Vol. 40, No 2, 2001
[4] A. Siepel, A.Farmer, A.Tolopko, M.Zhuang, P.Mendes, W.Beavis and B.Sobral. ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources. Bioinformatics, Vol. 17, No 1.2001: 83-94
[5] SWISS-PROT and TrEMBL
[http://www.expasy.org/sprot/sprot-top.html]
[6] Ingvar Eidhammer, Inge Jonassen, William R.Taylor. Protein Bioinformatics. Pages 142-146, 2004
[7] Philipp Bucher, Kevin Karplus, Nicolas Moeri and Kay Hofmann. A flexible motif search technique based on generalized profiles. Computers Chem. Vol. 20, No 1: 3-23, 1996
[8] PROSITE
[http://www.expasy.org/prosite/]
[9] R.Apweiler, T.K.Attwood, A.Bairoch, A.Bateman, E.Birney, M.Biswas, P.Butcher, L.Cerutti, F.Corpet, M.D.R.Croning, R.Durbin, L.Falquet, W.Fleischman, J.Gouzy, H.Hermjakob, N.Hulo, I.Jonassen, D.Kahn, A.Kanapin, Y.Karavidopoulou, R.Lopez, B.Marx, N.J.Mulder, T.M.Oinn, M.Pagni, F.Servant, C.J.A.Sigrist and E.M.Zdobnov. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Research, Vol. 29, No. 1: 37-40, 2001
[10] InterPro
[http://www.ebi.ac.uk/interpro/]
[11] Jorja G.Henikoff, Elizabeth A.Greene, Shmuel Pietrokovski and Steven Henikoff. Increased coverage of protein families with the BLOCKS database servers. Nucleic Acids Research, Vol. 28, No. 1: 228-230, 2000
[12] Steven Henikoff, Jorja G.Henikoff. Protein family classification based on searching a database of blocks
[13] BLOCKS
[http://blocks.fhcrc.org/blocks/]
[14] PRINTS
[http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/]
[15] Alex Bateman, Lachlan Coin, Richard Durbin, Robert D.Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik L.L.Sonnhammer, David J.Studholme, Corin Yeats and Sean R.Eddy. The Pfam protein families database. Nucleic Acids Research, Vol. 32: 138-141, 2004
[16] Pfam
[http://www.sanger.ac.uk/Software/Pfam/]
[17] European Bioinformatics Institute
[http://www.ebi.ac.uk/
[18] SRS [http://srs.ebi.ac.uk]