PDBSITE, PDBLIGAND AND PDBSITESCAN: A COMPUTATIONAL WORKBENCH FOR THE RECOGNITION OF THE STRUCTURAL AND FUNCTIONAL DETERMINANTS IN PROTEIN TERTIARY STRUCTURES COMBINED WITH PROTEIN DRAFT DOCKING

Ivanisenko V.A.\(^1\)*, Pintus S.S.\(^2\), Krestyanova M.A.\(^2\), Demenkov P.S.\(^2\), Znobisheva E.K.\(^2\), Ivanov E.E.\(^2\), Grigorovich D.A.\(^1\)

\(^1\) Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; \(^2\) Novosibirsk State University, Novosibirsk, Russia

* Corresponding author: e-mail: salix@bionet.nsc.ru

Keywords: protein functional sites, site recognition, protein tertiary structure, draft docking, transcription factors classification, hepatitis C virus

Summary

Motivation: The recognition of the structural/functional determinants in proteins has broad implications for structural genomics. A better understanding of the structural/functional determinants of proteins, such as protein-protein, protein-DNA, and protein-RNA interaction sites would provide insight into protein functions.

Results: A computational workbench for recognizing functional sites in protein tertiary structure combined with molecular draft docking was developed. Here, we illustrate the capabilities of the workbench by providing examples of search for interactions of the hepatitis C virus proteins with the human proteins and also of structural protein classification based on the structural similarity to the functional sites. The program PDBSiteScan is available at http://wwwmgs.bionet.nsc.ru/mgs/systems/fastprot/pdbsitescan.html. The PDBSite database is available at http://srs6.bionet.nsc.ru/srs6bin/cgi-bin/wgetz?-page+LibInfo+-newId+-lib+PDBSite.

Introduction

Experimental data on protein tertiary structure are growing at a rapid pace (Westbrook et al., 2003). The body of literature is extensive. With the advent of methodologies for the recognition of functional sites in primary structure (Bairoch, Bucher, 1994), tools for site recognition in tertiary structure based on structural data on it alone (Ondrechen et al., 2001; Liang et al., 2003; Gutteridge et al., 2003), as well as on structural similarity to related proteins of known function (Wallace et al., 1997; Fetrow, Skolnick, 1998; Jones et al., 2003) were developed. There is now a repertoire of tools for the search of functional sites using databases containing structural data on protein-ligand interactions (Hendlich, 2003).

Research increasingly focuses on proteomics in efforts to clarify how ligand-protein binding sites may be recognized and to generate their complexes. Thus, the concept of molecular docking became popular (for an overview, see Schneidman-Duhovny et al., 2004).

We have developed the PDBSite database for the spatial structures of the protein functional sites, including the posttranslational modification and binding sites, the active enzyme centers. The created PDBSiteScan program provides search on the PDBSite database using pairwise protein-site structure alignment. Good recognition accuracy of the functional sites by screening of protein tertiary structure on the PDBSite database has been illustrated by the active enzyme centers (Ivanisenko et al., 2004).

Here, we extend and improve PDBSite by developing a PDBLigand database and a draft molecular docking module to further combine them. The draft docking module can help to solve an important
aspect of the molecular docking problem, the initial disposition of interacting molecules with respect to each other in space.

An approach to automated structural and functional classification of proteins on the basis of their structural similarity to the functional sites from the PDBSite database is proposed. The resulting classification of the representatives of the main transcription factor families agreed well with the standard manual classification (Heinemeyer et al., 1999).

An example is provided to illustrate the benefits of combining PDBSite + PDBSiteScan + PDBLigand into a common workbench, namely evidence for the possible role of RNA-directed RNA polymerase (NS5B) hepatitis C virus (HCV) in the regulation of host immunity.

**Methods and Algorithms**

The workbench consists of the PDBSite and the PDBLigand databases, PDBSiteScan program, and the draft docking module. The PDBSite database stores the data for the functional protein sites, the PDBLigand database those for the ligands of the sites.

A brief description of the PDBSite structure and the PDBSiteScan program follows (for details, see Ivanisenko et al., 2004; Ivanisenko et al., 2002). PDBSite contains more than 8,000 sites, including catalytically active centers of various enzymes, the sites of posttranslational protein modification, the sites of ion metal binding, the sites of binding organic/inorganic compounds, the sites of drug binding, the sites of protein-protein, protein-DNA and protein-RNA interactions. The data extracted from the PDB databank (Berman et al., 2000) on the basis of information in the SITE field of PDB indicating the amino acid residues of the functional sites; the sites of protein-protein, protein-DNA and protein-RNA interactions were identified by analysis of the atom coordinates in their heterocomplexes. The sites included the amino acid residues that are in contact with the ligand (protein, RNA or DNA). A residue was accepted as contact if it had at least three atoms whose distance from any atom of the partner chain was smaller than 5 Å.

The PDBLigand database contains data on the low molecular ligands, proteins, DNA and RNA, which bind to the sites from PDBSite. The PDBLigand database includes the atom coordinates of the ligands, also their functional description extracted from the PDB databank. Every entry of the PDBLigand database contains information on a particular ligand links to an entry of the PDBSite database providing information on the binding site of the ligand.

The PDBSite database is integrated with the PDBSiteScan program for recognizing the functional sites in protein tertiary structures. PDBSiteScan provides automated search of the spatial fragments in protein tertiary structure similar in structure to the functional sites from the PDBSite database. The draft docking module works as follows. The PDBSite database contains the site-templates with known atom coordinates of their complexes with the ligands from the PDBLigand database. Draft docking is done by transfer of the ligand together with the site-template during the structural alignment of the site-template to protein. The generated draft protein-ligand complex can be accepted as a start approximation for the further docking or molecular dynamics analysis.

**Implementation and Results**

*Search for the potential interactions between HCV and human proteins.* The RNA dependent RNA polymerase NS5B is a 65 kDa protein that resembles other viral RNA polymerases (Lohmann et al., 1997). HCV replication is thought to occur in membrane bound replication complexes. The complexes transcribe the positive strand and the resulting minus strand is used as a template for the synthesis of genomic RNA. Search on the PDBSite database using PDBSiteScan demonstrated that NS5B contains fragments structurally similar to the binding site to the human nuclear transport factor 2 (NTF2) and to the human nuclear factor of activated T cells (NFAT). NTF2, a homodimer of approximately 14 kDa subunits, stimulates efficient nuclear import of a cargo protein (Stewart, 2000).
NFAT transcription factor family is involved in the expression of the cytokines IL-2, IL-3, IL-4, IL-5, granulocyte-macrophage colony-stimulating factor, and tumor necrosis factor-alpha, as well as several cell-surface molecules, such as CD40L and FasL. NFAT proteins are also expressed in B cells, mast cells, basophils and natural killer cells, as well as in a variety of non-immune cell types and tissues, such as skeletal muscle, neurons, heart and adipocytes (Porter et al., 2000).

The potential complexes of NS5B with NTF2 and NFAT generated using the draft docking module are shown in Fig. 1. It is seen that two loops are involved in the interaction of NS5B with NTF2 (Fig. 1a) and, hence, the contact might be close. Further calculations in terms of molecular dynamics, for example, are required to estimate contact affinity. The second complex results from contacts between only four residues at each site (Fig. 1b). However, NS5B can contact with the DNA bound to NFAT. It is suggested that the double contact of NS5B with NFAT and DNA can establish a stable complex. Molecular modeling is required to prove this.

Classification of transcription factors. Seventeen families of transcription factors were chosen for classification. Structural similarity to the functional sites from the PDBSite database were searched for every protein. The total number of functional site types examined was 88. The maximum distance mismatch (MDM) and amino acid type match were calculated to express the similarity between a protein fragment and a site (see Ivanisenko et al., 2004). A site and a protein fragment were accepted as structurally similar if the MDM value was less then 2 Å. The fragments structurally similar to the sites were further divided into four classes: 1) completely matching the amino acids; 2) one mismatch; 3) two mismatches and 4) three or more mismatches.

The distance between a pair of protein tertiary structures was calculated from

$$D_{ij} = \sqrt{\sum_{k=1}^{88} \sum_{m=1}^{4} (x_{km}^i - x_{km}^j)^2}$$

where $i, j$ are the indices of protein tertiary structure, $x_{km}^i$ is the variable indicating whether or not at least one fragment, structurally similar to a functional site of the $k$-th type and assigned to the $m$ mismatch class, is present in the $i$-th protein structure; the values assigned to were either 1 or 0; 1 was assigned to a particular site type if at least one fragment in the protein structure was found to be similar to at least one site from PDBSite of this type; otherwise 0 was assigned to this site type. The UPGMA method was used for clustering, and the PHYLIP package (Lim, Zhang, 1999) to construct the hierarchical tree (Fig. 2).

Discussion

The developed workbench was designed for addressing problems related to the functional annotation and draft docking of proteins with low molecular weight ligands, proteins, RNA and DNA. The workbench was applied to the analysis of interactions of the HCV-human proteins. As a result, we identified the potential binding site NS5B of HCV to the human nuclear transport factor 2 (NTF2) and also to the DNA binding domain of the human transcription factor NFAT. The results suggest that the NS5B-NTF2 interaction provides NS5B transport into the cell nucleas where it interacts with NFAT and DNA, participates in the regulation of gene expression, thereby suppressing antiviral immunity. It should be noted that the assumption requires support: modeling of the NS5B-NTF2, NS5B-DNA-NFAT draft complexes.

Although the structural classification of proteins is a powerful clue to problems in proteomics, there are no universal algorithms. Structure alignment methods are difficult to implement because of the vagueness of their global similarity measures. The structure alignment methods often measure similarity by the root-mean-square-deviation (RMSD) between the aligned atoms. Rogen and Fain have indicated that the RMSD of aligned atom coordinates is a perfect measure of similarity for two shapes that are nearly identical (Rogen, Fain, 2003). However, the RMSD is a poor measure when the two shapes compared differ significantly. As a result, automated classification of proteins
remains an open issue. We suggested an approach to automated protein classification based on search for structural similarity between protein fragments and functional sites from the PDBSite database. The approach was applied to the classification of the representatives of the main classes of transcription factors. The resulting classification agrees well with the one obtained by manual analysis. The proposed workbench has already proven itself to be useful in analysis. Further integration with other computational tools is possible and beneficial.

**Fig. 1.** The potential complexes of NS5B with NTF2 (a) and NFAT (b). The NS5B structure is dark grey, the NTF2 and NFAT structures are light grey. In the NS5B-NFAT complex, a fragment of double stranded DNA, which interacts with NFAT and presumably with NS5B, is depicted. The atom coordinates of NS5B, NTF2 and NFAT in complex with DNA were extracted from PDB 1QUV, 1A2K and 1A02, respectively.

**Fig. 2.** A hierarchical tree for a classification of the representatives of the main classes of transcription factors. The tree was built on the basis of search for the structural homology of the DNA-binding domains of these factors with the functional sites from the PDBSite database. The name of the class, the PDB ID and a schematic representation of tertiary structure are given for every domain.
Acknowledgements

The work was supported by the Russian Foundation for Basic Research (02-07-90355, 03-07-96833-p2003, 03-07-96833, 03-07-90181-n, 02-04-48802-a, 03-04-48829, 03-04-48555-a); Project No. 10.4 of the RAS Presidium Program “Molecular and Cellular Biology”; the Siberian Branch of the Russian Academy of Sciences (integration project No. 119); Russian Ministry of Industry, Science, and Technologies (grants Nos. 43.073.1.1.1501); the U.S. Civilian Research & Development Foundation for the Independent States of the Former Soviet Union (CRDF) No. NO-008-X1.

References