SeSAW: balancing sequence and structural information in protein functional mapping

Daron M. Standley1,∗, Reiko Yamashita2, Akira R. Kinjo2, Hiroyuki Toh3 and Haruki Nakamura2

1WPI Immunology Frontier Research Center (IFReC), Osaka University, 3-1 Yamadaoka, 2Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871 and 3Medical Institute of Bioregulation, Kyushu University, 3-1-1, Maidashi, Higashi-ku, Fukuoka, Fukuoka 812-8582, Japan

Abstract

Motivation: Functional similarity between proteins is evident at both the sequence and structure levels. SeSAW is a web-based program for identifying functionally or evolutionarily conserved motifs in protein structures by locating sequence and structural similarities, and quantifying these at the level of individual residues. Results can be visualized in 2D, as annotated alignments, or in 3D, as structural superpositions. An example is given for both an experimentally determined query structure and a homology model.

Availability and Implementation: The web server is located at http://www.pdbj.org/SeSAW/

Received on December 14, 2009; revised on February 26 2010; accepted on March 13, 2010

1 INTRODUCTION

Sequence alignment and structural alignment are widely used techniques for inferring functional or evolutionary relationships between proteins. However, most alignment methods do not integrate sequence and structural information into one measure of similarity or describe the similarity at the level of individual residues. We recently introduced a sequence and structure-based scoring method that employs sequence profile–profile comparisons, but is anchored by structural alignments and showed that the functional information associated with the top-scoring hits found by the method agreed well with expert annotations published in the literature (Standley et al., 2008b). Subsequently, we have shown that this approach can be used to identify functional sites in remote (e.g. 10–20% sequence identity) homology models, even when the structural template used to build the model is itself un-annotated (Standley et al., 2008a). That is, a structure without a known function (e.g. a structural genomics target) can be used as an intermediate template to subsequently locate a functionally characterized structure, and thus map putative functional sites onto a distantly related query sequence. Here, we describe a web-based implementation of the method called SeSAW (sequence-derived structural alignment weights) that can automatically perform putative functional residue mapping. We emphasize that such mapping is intended to guide subsequent experiments rather than to serve as a substitute for experimental annotations.

Fig. 1. Outline of the server. The rectangles on the left indicate major steps that are performed in real-time. Those on the right indicate steps that are done offline. Ovals in the center represent external software used for both types of calculations. Colored lines indicate their interconnection.

2 ALGORITHM

SeSAW takes as input a PDB-formatted query file, chain ID, and, in the case of a template-based model, the PDB ID and chain ID of the template. As illustrated in Figure 1, a PSI-BLAST position specific scoring matrix (PSSM) for the query is retrieved or computed as necessary (we maintain a database of PSSMs for every unique PDB chain). Subsequently, we have shown that this approach can be used to identify functional sites in remote (e.g. 10–20% sequence identity) homology models, even when the structural template used to build the model is itself un-annotated (Standley et al., 2008a). That is, a structure without a known function (e.g. a structural genomics target) can be used as an intermediate template to subsequently locate a functionally characterized structure, and thus map putative functional sites onto a distantly related query sequence. Here, we describe a web-based implementation of the method called SeSAW (sequence-derived structural alignment weights) that can automatically perform putative functional residue mapping. We emphasize that such mapping is intended to guide subsequent experiments rather than to serve as a substitute for experimental annotations.

© The Author(s) 2010. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
The score is given by adding the ASH structure alignment score to the sum of a per-residue similarity score:

\[ S_{\text{SeSAW}} = S_{\text{ASH}} + \sum_{i} S_{i} \]  

where per-residue similarity score \( S_{i} \) is defined as:

\[ S_{i} = e^{-\left( \frac{d_{i}}{d_{\text{max}}} \right)^{2} \left( w_{B} S_{B}[a_i,a_j] + w_{P} \text{MAX} \left[ S_{Q}[a_i], S_{T}[a_j] \right] \right) / 2} \]

Here, \( d_{i} \) is the distance between \( C_{\alpha} \) atoms in the two aligned residues (after superposition of the query and template), \( d_{\text{max}} \) is a reference distance (4 Å used in all calculations), \( w_{B} \) is a scalar weight (0.8 used in all calculations), \( S_{B} \) is the bit BLOSUM62 matrix, \( S_{Q} \) and \( S_{T} \) are the amino acid types of the query and template, respectively, \( w_{P} \) is a scalar weight (1.5 used in all calculations), and \( S_{Q} \) and \( S_{T} \) are the odds column vectors of the query and template PSSMs, respectively.

The SeSAW score is reported, along with a P-value computed by randomization and a significance level, and ultimately, biochemical experimentation. Nevertheless, SeSAW is a significant improvement over running structural and sequence analysis separately (Standley et al., 2008b), and can thus play an important role in automated functional annotation of structural genomics targets or homology models.

ACKNOWLEDGEMENTS

The authors would like to thank A. Yoshihara for technical assistance.

Funding: This work was supported by a kakenhi grant 21570169: Grant-in-Aid for Scientific Research (C) from the Japan Society for the Promotion of Science (JSPS).

Conflict of Interest: none declared.

REFERENCES

Alexandrov,N. and Shindyalov,I. (2003) PDP: protein domain parser. Bioinformatics, 19, 429–430.

Arazi.R. et al. (2009) Crystal structure of MapD (TTHA1568), a menaquinone biosynthetic enzyme from Thermus thermophilus HB8. J. Struct. Biol., 168, 575–581.

Hiratsuka,T. et al. (2008) An alternative menaquinone biosynthetic pathway operating in microorganisms. Science, 321, 1670–1673.

Matsushita,K. et al. (2009) Zn(II)12a is an RNAase essential for controlling immune responses by regulating mRNA decay. Nature, 458, 1185–1190.

Murzin,A.G et al. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol., 247, 536–540.

Pearl,F. et al. (2005) The CATH Domain Structure Database and related resources Gen3D and DIDS provide comprehensive domain family information for genome analysis. Nucleic Acids Res., 33, D247–D251.

Standley,D.M et al. (2007) ASH structure alignment package: sensitivity and selectivity in domain classification. BMC Bioinformatics, 8, 116.

Standley,D.M et al. (2008a) Structure-based functional annotation of protein sequences guided by comparative models. In Zhang, X.S., Chen, L., Wu, L.Y . and Wang, Y . (eds), The Second International Symposium on Optimization and Systems Biology. Lijiang, China, pp. 395–403.

Standley,D.M et al. (2008b) Functional annotation by sequence-weighted structure alignments: statistical analysis and case studies from the Protein 3000 structural genomics project in Japan. Proteins, 72, 1333–1351.

Waterhouse,A.M et al. (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics, 25, 1189–1191.

Yamamoto,M. et al. (2009) A single polymorphic amino acid on Toxoplasma gondii kinase ROP16 determines the direct and strain-specific activation of Stat3. J. Exp. Med., 206, 2747–2760.