Function and Interaction Prediction using multiple motif descriptors for classified domain-domain interactions and ligand binding sites.

Andreas Henschel*, Christof Winter, Wan Kyu Kim and Michael Schroeder
Biotechnological Center, TU Dresden, Tatzberg 47-51, 01307 Dresden, Germany
*To whom correspondence should be addressed: ah@biotec.tu-dresden.de

1. INTRODUCTION

Many protein sequences are still poorly annotated. Functional characterization of a protein often is improved by the identification of novel interaction partners. Here, we aim to create descriptors for all relevant sequence parts of structurally known protein-protein and protein-ligand binding sites. These binding sites are often well-conserved (1). In contrast, the rest of the surface seems to be variable (see Figure 1A) which impedes sequence similarity searches for functionally equivalent or similar proteins. Descriptors previously used for conserved domains and interface motifs are regular expressions, weight matrices and Hidden Markov Models (HMMs), covering either sequentially consecutive stretches (2-4) or full length domains (5). In particular, HMMs were successfully employed in many sequence similarity search tools (5, 6, 7).
Based on the family level of the Structural Classification of Proteins, SCOP (8), it is possible to extract and classify all domain-domain interactions found in the Protein Data Bank, PDB (9). This classification is available in the SCOPPI database (10). SCOPPI clusters similar interfaces into interface types. As pointed out by Kim and Ison, even homologous domain pairs can associate in geometrically different ways by employing different sets of residues to form interfaces (11). Consequently, the corresponding interface profiles would differ substantially which makes profile merging meaningless. However, often a number of domain-domain interactions expose striking similarities and it is desirable to collect all instances of one interface type for the calculation of the respective interface profile. We therefore compose descriptors for all interface types in SCOPPI by merging all interface profiles describing that interface type. When data for interface types is sparse, we utilize sequence data provided by HSSP (12).
Often several sequentially remote segments contribute to a binding site (exemplified in Figure 1B). To accommodate for this phenomenon, we adopt the multiple-motif approach from PRINTS (13) to represent binding sites as a collection of small HMMs for one local binding motif thus describing only the important sequence parts that form a structural feature. Each collection member gives rise to an individual sequence similarity search using the HMMer package. The P-score of the sum of the individual search result scores can be calculated using Karlin-Altschul's sum statistics for multiple high scoring sequence segments (14, formula [5]).

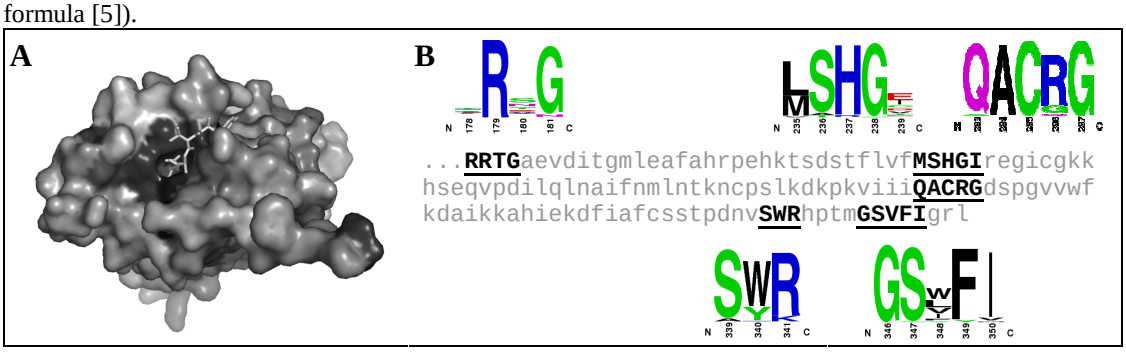

Fig.1: Constructing a set of sequence profiles to represent a conserved structural feature. A: Caspase's active site is highly conserved (1ICE, conservation levels are calculated using the von Neumann entropy and displayed in shades of gray, the darker the better conserved). Conserved residues in close vicinity of the tetrapeptide inhibitor largely define the catalytic site environment. **B:** Caspase residues within 5Å of the inhibitor are underlined. Segments are patched and those with low conservation are discarded to avoid insignificant hits. We add amino acid distribution from HSSP data for each site of the remaining segments. It is thus possible to construct HMMs and visualize the profiles as sequence logos (15).

2. RESULTS

We compiled a comprehensive database that comprises descriptors (interface profiles) for each interface type in SCOPPI and ligand binding sites in the PDB totaling more than 3000 interface profiles. These interface profiles characterize an interaction/ligand binding site on sequence level. Hence, given a query sequence of interest, it is possible to compare it to each interface profile thus identifying possible interaction partners including ligands. Profiles for domain-domain interactions have the advantage that both interfaces can be considered. Double sided hits increase significance, i.e. given two candidate sequences, double sided hits from an interface profile pair with respective P-scores p_1 and p_2 yield a joint probability of $p_1 \cdot p_2$. Finally, Gene Ontology (16) annotations are linked to each interface profile from the original PDB entries that were used to construct this profile. The complete list of HMMs is freely available for academics upon request.

3. REFERENCES

1. Saeed R, Deane CM. 2006. Protein protein interactions, evolutionary rate, abundance and age. *BMC Bioinformatics*. 7:128.
2. Bairoch A. 1992. PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res*. 20(Suppl): 2013–2018.
3. Espadaler J, Romero-Isart O, Jackson RM, Oliva B. 2005. Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. *Bioinformatics*. 21(16):3360-8.
4. Li H, Li J. 2005. Discovery of stable and significant binding motif pairs from PDB complexes and protein interaction datasets. *Bioinformatics*. 21(3):314-24.
5. Bateman A, Haft DH. 2002. *Brief Bioinform*. HMM-based databases in InterPro. 3(3):236-45.
6. Eddy SR: Profile hidden Markov models. *Bioinformatics* 1998, 14:755-763.
7. Zdobnov EM, Apweiler R. 2001. InterProScan-an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. 17(9):847-8.
8. Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 247(4):536-40.
9. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28(1):235-42.
10. Winter C, Henschel A, Kim WK, Schroeder M. 2006. SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res.* 34(Database issue):D310-4.
11. Kim WK, Ison JC. 2005. Survey of the geometric association of domain-domain interfaces. *Proteins*. 61(4):1075-88.
12. Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*. 9(1):56-68.
13. Scordis P, Flower DR, Attwood TK. 1999. FingerPRINTScan: intelligent searching of the PRINTS motif database. *Bioinformatics*. 15(10):799-806.
14. Karlin S, Altschul SF. 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. Proc Natl Acad Sci U S A. 90(12):5873-7.
15. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14(6):1188-90.
16. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 25(1):25-9.