BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 12 2010, pages 1564–1565
doi:10.1093/bioinformatics/btq208

Sequence analysis

Advance access publication April 22, 2010

HangOut: generating clean PSI-BLAST profiles for domains with long insertions

Bong-Hyun Kim[1,*], Qian Cong[1] and Nick V. Grishin[1,2,*]

[1]Department of Biochemistry and [2]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390, USA

Associate Editor: Burkhard Rost

ABSTRACT

Summary: Profile-based similarity search is an essential step in structure-function studies of proteins. However, inclusion of non-homologous sequence segments into a profile causes its corruption and results in false positives. Profile corruption is common in multidomain proteins, and single domains with long insertions are a significant source of errors. We developed a procedure (HangOut) that, for a single domain with specified insertion position, cleans erroneously extended PSI-BLAST alignments to generate better profiles.

Availability: HangOut is implemented in Python 2.3 and runs on all Unix-compatible platforms. The source code is available under the GNU GPL license at http://prodata.swmed.edu/HangOut/

Contact: kim@chop.swmed.edu; grishin@chop.swmed.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 1, 2010; revised on March 31, 2010; accepted on April 16, 2010

1 INTRODUCTION

PSI-BLAST (Altschul *et al.*, 1997) is an indispensable tool for remote homology inference and structure-function predictions (Devos and Valencia, 2000; Friedberg, 2006; Grishin, 2001; Hegyi and Gerstein, 2001). However, false positives in PSI-BLAST can cause errors in automated annotations (Bork and Koonin, 1998). One major source for such false positives is profile corruption, usually resulting from extension of alignments over non-homologous sequence regions (Galperin and Koonin, 1998). For instance, for two 2-domain proteins, AB and A′C, PSI-BLAST may extend a correct alignment of the homologous domains A and A′ to include sequences from the non-homologous domains B and C. Despite significant effort devoted to this multidomain problem, no satisfactory solution exists (Gonzalez and Pearson, 2010; Galzitskaya and Melnik, 2003; George and Heringa, 2002; Nagarajan and Yona, 2004). Currently, the best approach is to start PSI-BLAST with precisely defined query sequence bounds (Corpet *et al.*, 2000; Wheeler *et al.*, 2001).

However, we found that even a single, well-defined domain does not guarantee a corruption-free profile. Domains hosting insertions, which represent close to 5% of domains in the structural classification of proteins (SCOP) 1.75 database (Murzin *et al.*, 1995), may generate a corrupted PSI-BLAST profile due to incorrect alignment extension around the insertion position. Our analysis shows that the N- and C-terminal segments of the host domain are frequently aligned as separate PSI-BLAST high scoring pairs (HSPs), and the two HSPs overlap when mapped onto the query sequence. Each alignment can be divided into two segments: (i) correctly aligned and (ii) incorrectly aligned or extended (Fig. 1a and Supplementary Fig. S1). These incorrectly aligned 'overhangs' are detected and removed by the HangOut program to clean the profile and prepare it for consequent remote homology searches with various tools, such as PSI-BLAST and HHsearch.

2 METHODS

The HangOut input is a single domain query sequence with the insertion boundary specified. The HangOut algorithm proceeds as follows (Fig. 1a): (1) Run BLAST with the input sequence against the NCBI non-redundant database with *e*-value threshold 0.001. (2) Detect and remove lower-scoring (see second half of this paragraph for clarification) regions from HSPs and regions matching a PSI-BLAST profile of the inserted domain (see Supplementary Figures S2 and S3 for rationale). (3) Terminate upon convergence or iteration limit. Otherwise, repeat Steps 1 to 3 with the following modifications: (i) PSI-BLAST replaces BLAST, seeded (-B option) with the cleaned profile from Step 2 and (ii) profile scores (PSSM) replace BLOSUM62 scores (for HSP removal). Thus, HangOut builds multiple sequence alignments similarly to PSI-BLAST, but has a 'clean-up' step after each iteration intended to remove incorrect extensions. HangOut is based on two assumptions: (i) each HSP contains at least one correctly aligned region, and (ii) incorrectly extended regions exist in every HSP that crosses the insertion point. Based on these assumptions, HangOut splits all local alignments into two segments with a boundary at the insertion point and selects the best scoring (BLOSUM62 or PSSM) segment out of each split pair. The lower scoring segment is removed as a possibly erroneous extension. In addition to this HangOut procedure, we applied RemoveHit, a simpler method that does not require a defined insertion point and removes entire alignments for hits with two overlapping HSPs (Supplementary Fig. S4).

HangOut was tested on a set of 40% representative SCOP 1.75 domains defined to contain insertions (302 domains, see Supplementary Table 1 for the list) to measure the number of corrupted profiles (false positives) and the number of correct homologs found by each discontinuous query domain sequence (with insertion sequence removed). The 302 hidden Markov Models (HHMs) built from each PSI-BLAST profiles, HangOut profiles or RemoveHit profiles were compared to HHMs built from all 9528 SCOP 1.75 40% representative domains (Murzin *et al.*, 1995) using HHsearch ver. 1.5.1 (Soding, 2005). The number of corrupted profiles was increased by one if HHsearch found homologs of inserted domains with probability higher than 0.9. The number of homologs found are counted by the number of hits that have strong profile similarity (HHsearch probability above 0.9) and overall structural similarity (DaliLite *Z*-score higher than 4) (Holm and Park, 2000) or belonged to the same SCOP superfamilies as the query domains.

*To whom correspondence should be addressed.

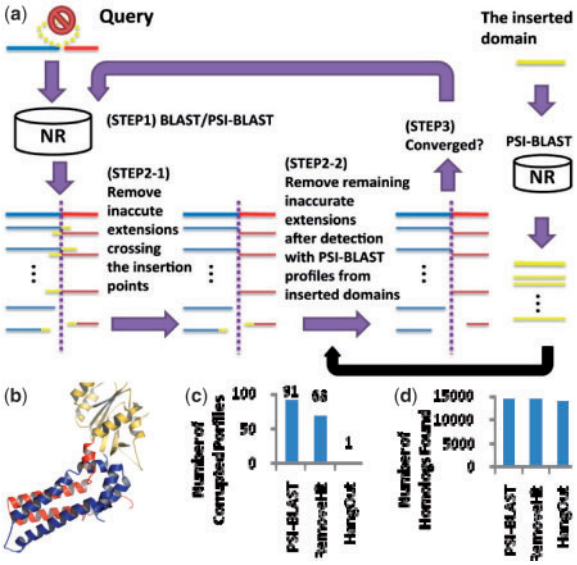

Fig. 1. HangOut method to clean PSI-BLAST profiles. (**a**) HangOut flowchart. Starting from a query domain (blue–red, with inserted domain in yellow removed), a PSI-BLAST search of the NCBI non-redundant database (NR) is performed (Step 1) to produce alignments. Erroneously extended regions (yellow) that cross the insertion boundary (vertical dotted line) are removed to produce a 'cleaned' alignment (Step 2–1). Remaining contaminants not crossing the domain boundary are removed by PSI-BLAST profiles built from the long insertion (Step 2–2). The PSI-BLAST result is checked for convergence (Step 3), and possibly continued from Step 1. (**b**) Structure of two middle domains of the TrmE GTP-binding protein (PDB ID 1xzp). The α/β P-loop hydrolase domain (yellow, SCOP ID d1xzpa1, chain A: 118–211, 372–450) is inserted into an α-helical bundle (N- and C-terminal segments are colored blue and red, respectively; SCOP ID d1xzpa2, chain A: 212–371). (**c** and **d**) HangOut performance test showing the number of corrupted profiles and the number of found homologs, respectively. Performances, of PSI-BLAST and RemoveHit, are also shown. RemoveHit removes all alignments for hits with two overlapping HSPs as in Supplementary Figure S1b. The HangOut profiles show high accuracy with only one case of possible corruption (c) and without losing sensitivity (d). Color version of the figure is available at *Bioinformatics* online.

3 RESULTS

HangOut is intended to clean PSI-BLAST generated profiles of erroneous extensions caused by domain insertions. One typical example of this domain problem is shown in Figure 1b: an α/β P-loop hydrolase (yellow in Fig. 1b) is inserted into an α-helical bundle (blue and red in Fig. 1b). Corruption of the PSI-BLAST alignment built from hits to the α-helical bundle is evidenced by a profile-based similarity search (HHsearch), which finds the α/β P-loop hydrolase domain with probability 98%. Since the query α-helical bundle does not share any sequence or structural similarities with the hydrolase domain, the high HHsearch probability results from profile corruption (for details see Supplementary Fig. S2).

Given the success of this example, we tested the ability of HangOut to clean profiles of all SCOP domains with defined insertions (302 domains). As a basis for comparison, 91 PSI-BLAST profiles (30%) were corrupted. RemoveHit cleans only 23 of these profiles, while HangOut cleans all but one (Fig. 1c). The single exception is probably due to distant homology, since both the host and inserted domain represent similar doubly wound Rossmann folds (Supplementary Fig. S5). Because the removal of sequence segments from alignments may deprive the profile, we also checked for the loss of true hits. Surprisingly, cleaned HangOut profiles retained ~98% of the homologs found by PSI-BLAST profiles (99.6% for RemoveHit), suggesting that useful information is not lost from the profiles. Compared to RemoveHit, the complexities of HangOut that use domain boundary information are apparently needed to clean corrupted profiles. The presence of overlapping HSPs (removed by RemoveHit) does not sufficiently indicate corrupted segments. For remote homologs, only a single HSP may be found and incorrectly extended to cover part of the insertion. Although our current HangOut procedure does not offer a comprehensive solution to the multidomain problem, it addresses a special case of domains with insertions that represent the major source of profile corruption when PSI-BLAST is initiated with single, discontinuous domain queries. HangOut will be especially useful for large-scale bioinformatics efforts that are initiated from defined structure domains and require uncorrupt sequence profiles for subsequent analysis. Additional work will be done to offer a general solution without prior knowledge of domain boundaries.

ACKNOWLEDGEMENTS

The authors thank Lisa N. Kinch, Jimin Pei and Jeremy Semeiks for helpful comments.

Funding: Welch Foundation I1505 (to N.V.G.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bork,P. and Koonin,E.V. (1998) Predicting functions from protein sequences–where are the bottlenecks? *Nat. Genet.*, **18**, 313–318.

Corpet,F. *et al.* (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, **28**, 267–269.

Devos,D. and Valencia,A. (2000) Practical limits of function prediction. *Proteins*, **41**, 98–107.

Friedberg,I. (2006) Automated protein function prediction - the genomic challenge. *Brief. Bioinform.*, **7**, 225–242.

Galperin,M.Y. and Koonin,E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol.*, **1**, 55–67.

Galzitskaya,O.V. and Melnik,B.S. (2003) Prediction of protein domain boundaries from sequence alone. *Protein Sci.*, **12**, 696–701.

George,R.A. and Heringa,J. (2002) SnapDRAGON: a method to delineate protein structural domains from sequence data. *J. Mol. Biol.*, **316**, 839–851.

Gonzalez,M.W. and Pearson,W.R. (2010) Homologous over-extension: a challenge for iterative similarity searches. *Nucleic Acids Res.*, **38**, 2177–2189.

Grishin,N.V. (2001) Fold change in evolution of protein structures. *J. Struct. Biol.*, **134**, 167–185.

Hegyi,H. and Gerstein,M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.*, **11**, 1632–1640.

Holm,L. and Park,J. (2000) DaliLite workbench for protein structure comparison. *Bioinformatics*, **16**, 566–567.

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.

Nagarajan,N. and Yona,G. (2004) Automatic prediction of protein domains from sequence information using a hybrid learning system. *Bioinformatics*, **20**, 1335–1360.

Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Wheeler,D.L. *et al.* (2001) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **29**, 11–16.