W48–W52 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* *Published online 6 May 2009*
doi:10.1093/nar/gkp279

webPRC: the Profile Comparer for alignment-based searching of public domain databases

Bernd W. Brandt* and Jaap Heringa

Centre for Integrative Bioinformatics (IBIVU), VU University Amsterdam, The Netherlands

Received January 30, 2009; Revised April 3, 2009; Accepted April 14, 2009

ABSTRACT

Profile–profile methods are well suited to detect remote evolutionary relationships between protein families. Profile Comparer (PRC) is an existing stand-alone program for scoring and aligning hidden Markov models (HMMs), which are based on multiple sequence alignments. Since PRC compares profile HMMs instead of sequences, it can be used to find distant homologues. For this purpose, PRC is used by, for example, the CATH and Pfam-domain databases. As PRC is a profile comparer, it only reports profile HMM alignments and does not produce multiple sequence alignments. We have developed webPRC server, which makes it straightforward to search for distant homologues or similar alignments in a number of domain databases. In addition, it provides the results both as multiple sequence alignments and aligned HMMs. Furthermore, the user can view the domain annotation, evaluate the PRC hits with the Jalview multiple alignment editor and generate logos from the aligned HMMs or the aligned multiple alignments. Thus, this server assists in detecting distant homologues with PRC as well as in evaluating and using the results. The webPRC interface is available at http://www.ibi.vu.nl/programs/prcwww/.

INTRODUCTION

Sequence-alignment techniques are essential in providing predictions of protein function and evolution. The introduction of sequence–profile methods, such as hmmpfam, hmmsearch (1) and PSI-BLAST (2,3), increased the detection of homologous sequences considerably compared to sequence-sequence methods [e.g. (4)], such as BLAST (3). A profile numerically encodes a multiple sequence alignment and its amino acid diversity by counting the amino acids in each column. Profile hidden Markov models (HMMs), or (profile) HMMs, are statistically more advanced than numerical profiles and allow for variable gap penalties (1). Clearly, profiles, based on an alignment, contain more information than a single sequence. Indeed, including distant but true homologues in the alignment, further increases the chance of detecting of similar families (5). We here use the word 'profiles' to refer to both numerical profiles and profile HMMs.

The last decade the sequence–profile methods have been advanced to profile–profile methods. Profile–profile methods provide a more sensitive (6–9) way to find distant homologies between proteins. Using profiles for both query and subject (domain database), has been shown to lead to more sensitive detection of evolutionary remote relationships [e.g. (9,10)]. Different profile–profile methods have been developed, including *prof_sim* (9) and FFAS (11). We here focus on three widely used state-of-the-art profile–profile programs: Profile Comparer [first released in 2002 (12)], COMPASS [COmparison of Multiple Protein sequence Alignments with assessment of Statistical Significance (6,13)] and HHsearch (7,14). Profile Comparer [PRC, (12)] is a stand-alone program for scoring and aligning HMM and is routinely used by, for example, the CATH (15) and Pfam (16,17) domain databases. The CATH pipeline uses PRC to detect extremely remote homologues and group them in superfamilies [http://www.cathdb.info/wiki/doku.php?id = about:intro, (15)]. Initially, Pfam used only PRC to detect similar domains (16), but now also uses HHsearch (14) [and SCOOP (18)] to establish Pfam clans (17). In addition, internal links from one Pfam family to another are generated with PRC and SCOOP.

In contrast to HHsearch and COMPASS (7,13), PRC did not have a web interface available yet. We therefore have implemented webPRC, a server for searching several public domain databases with additional functionality, including HMM-to-alignment translation, as compared to stand-alone PRC.

METHODS

Database construction

Several major domain databases are provided: Pfam (17), NCBI's Conserved Domain Database (19), KOG (20), TIGRFAMs (21), CATH (22) and SUPERFAMILY

*To whom correspondence should be addressed. Tel: +31 20 59 87816; Fax: +31 20 59 87653; Email: bwbrandt@few.vu.nl

(23). We briefly indicate how the profile HMMs and their seed alignments were obtained.

Pfam-A: The Pfam-A (16,17) profile HMMs have been rebuilt locally using the seed alignments downloaded from the Pfam FTP site (http://pfam.sanger.ac.uk) and the hmmbuild options provided therein. When building the HMMs the starting alignment, also for CDD/KOG and TIGRFAMs, was re-saved by hmmbuild (HMMER v2.3.2; http://hmmer.janelia.org/). This re-saved alignment includes an 'RF' line that indicates which alignment columns are absent from the HMM. This line is used to translate the HMM coordinates of the PRC results back to the alignment coordinates.

CDD/KOG: NCBI's Conserved Domain Database [CDD (19)] and KOG (20) HMMs have been built from the seed alignments downloaded via the CDD site (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). As there can be multiple identical sequence identifiers in CDD alignments, the sequence identifiers in the re-saved alignments were made unique by prepending a number to the entire identifier for reoccurring identifiers only.

TIGRFAMs: The TIGRFAMs (21) HMMs have been rebuilt locally from the seed alignments and the hmmbuild options provided in the TIGRFAMs HMM files.

CATH: The CATH (15,22) HMMs have been obtained from the CATH web site (http://www.cathdb.info). These models are not based on Pfam-like seed alignments, but are produced iteratively starting from a single sequence (24). This can result in huge alignments with high gap content (up to about 80 000 sequences, >50 000 columns, or 680 Mb for a single alignment). For this reason, the CATH models are used directly. Their underlying alignments have been processed to include an 'RF' line and a maximum of the first 200 sequences are included in the alignment output.

SUPERFAMILY: The SUPERFAMILY (23) models were retrieved from http://supfam.org.

User input

The user can provide a single protein sequence or multiple sequence alignment via the paste or upload field. A variety of alignment formats is accepted (ClustalW, FASTA, GCG MSF, Stockholm and SELEX). The user may configure the following search parameters: the domain database, PSI-BLAST options, several PRC options, the number of unique hits to be visualized in the hit graphic, and the use of the hmmbuild '–hand' option. This option can be used to mark regions of the alignment that should be absent from the HMM produced by hmmbuild, which is useful for searching with discontinuous domains. The 'RF' annotation line, required for the optional '–hand' option, is supported for the SELEX (#= RF) and Stockholm (#=GC RF) formats. Finally, the user may choose to generate logos from the HMM alignments or from the aligned multiple sequence alignments [with LogoMat-P (25) and Two Sample Logo (26), respectively] to visualize the alignments. Example input and output are provided, including the possibility to regenerate the example output ('rerun the example').

Alignment calculation

The webPRC searches run on a 64-CPU computer cluster. The processing scripts are coded in Perl, Bioperl [Bio::Graphics and Bio::SimpleAlign; (27)], PHP and Javascript. PRC is run with the selected domain library and domain descriptions of the hits are parsed from the chosen domain database. Since PRC results are reported in profile HMM space, both the PRC alignment output and the re-saved alignment files, produced by hmmbuild, are processed to provide a mapping of PRC results to the query and hit multiple sequence alignments. Then, these alignments are sliced according to the calculated alignment coordinates and joined in one alignment. The IDs of the hit alignment in this combined alignment file are prepended with 'Hit:'. In addition, an 'aligned alignments' view is constructed which contains the first sequence and the consensus sequence from each alignment. For viewing the alignment interactively, an extended version of Jalview (28) is used that supports regular expressions to parse sequence identifiers for its linkUrl parameters.

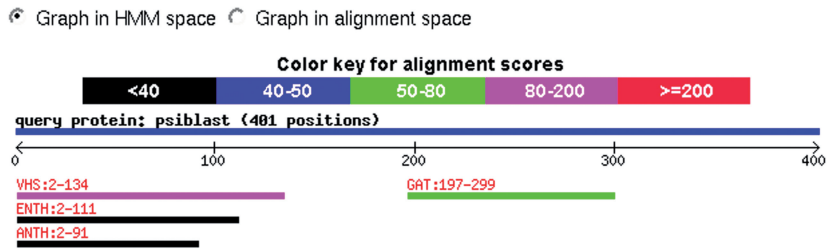
Logo generation

The logos are generated with local installations of LogoMat-P (25) and Two Sample Logo (26). LogoMat-P was adapted such that the generated logos correspond exactly to the HMM alignments reported by PRC. Thus, LogoMat-P is not executing a new pair-wise PRC search to find an HMM alignment between the query and the single subject HMM, but now directly uses the alignment produced by the PRC library run against the domain databases.

RESULTS AND DISCUSSION

The webPRC server facilitates the use of PRC for finding domains related to a query alignment. Besides the possibility to run PRC against different domain databases, webPRC offers additional functionality not available with a PRC stand-alone run.

After completion of a PRC search, the raw PRC output is reformatted into a BLAST-like report, which includes a domain hit distribution graphic and a hit table (Figure 1). This makes interpreting PRC output as straightforward as reading a BLAST report. The reformatted PRC alignments now include the match, insert, and delete percentages (Figure 2). In addition, several other features aiding the evaluation of the hits are included in the report: hits in the table are linked to the source domain database and include a description from the selected domain database. The alignments section contains links to the optionally produced logos. These logos are graphical representations of the aligned HMMs or the aligned alignments and can help in the evaluation of the found domains. LogoMat-P (25) produces pair-wise HMM logos based on the reported PRC alignment. These HMM logos are related to the HMM logos (29) used to visualize the HMMs of protein families in Pfam (17). In addition, Two Sample Logos are produced. These logos are based on two multiple sequence alignments and show the positions that are significantly different between the alignments (26). Furthermore, the



PRC hit table

Download: PRC scores file and PRC alignments file.
PSI-BLAST: results and generated alignment.

Results of search against Pfam 23.0 - July 2008 (10340 profiles):

hit (hmm2)	description		co-emis	simple	reverse	E-value
VHS	VHS domain: Domain present in VPS-27, Hrs and STAM.		115.9	115.5	99.3	4.7e-44
GAT	GAT domain: The GAT domain is responsible for bindi...	>>	74.9	74.7	56.9	8.4e-24
ENTH	ENTH domain: The ENTH (Epsin N-terminal homology) d...	>>	30.0	28.9	16.2	0.00022
ANTH	ANTH domain: AP180 is an endocytotic accessory prot...	>>	25.3	23.4	13.0	0.0072

Figure 1. An example of the webPRC domain graphic and hit table section for GGA1_HUMAN run against Pfam (after running PSI-BLAST). The graph can be viewed in HMM or alignment space and the hits are hyperlinked to the alignments. The PRC hit table provides links to the original PRC and PSI-BLAST output and shows a table with annotated hits, including the name and, after clicking on '>>', the description from the domain database. The hits are hyperlinked to the source database and *E*-values are hyperlinked to the alignments. Co-emission, simple and reverse scores are calculated by PRC [cf. (12)]. The *E*-value is calculated from the reverse score.

>Hit (#1): VHS Show description View alignment View HMM-Logo View TS-Logo Download
Length=153

```
 Score  = 99.3, Expect = 4.7e-44
 Match = 133/149 (89%), Insert = 2/149 (1%), Delete = 14/149 (9%)

       Query    2    MMMMMMMMMMMMMMMMMM~~MMMMMMMMMMMMMMMM~~~MMMMMMMMMMMMMMMMM~MMMM     55
       Sbjct    5    MMMMMMMMMMMMMMMMMMDDMMMMMMMMMMMMMMMMMDDDMMMMMMMMMMMMMMMMMDMMMM     64

       Query   56    MMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMIMMMMM~~MMMMMMMMMM     112
       Sbjct   65    MMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMDMMMMMMMMMM     124

       Query  113    MMMMM~~~~~~IMMMMMMMMMMMMMMMMM                                    134
       Sbjct  125    MMMMMDDDDDDMMMMMMMMMMMMMMMMM                                    153


Aligned alignments:

       QUERY    8    ETLE.ARINR..ATNPLN.KEL~~DWASINGFCEQLNED~~:~F.EG.PPLATRLLAHKI     55
   Consensus   12    TPLGfQRIEKkiATDPSLlQSE~~DWALNMEICDIINET~~:~EgEGaPKDAVRALKKRI     65
                     ++++ +++++  ++++++ +++  ++++++++++++++    + ++ ++++++++++++
   Consensus    5    SPLE:RLIDK::ATDPSL:PEEDEDWSLILDICDLINEKIYkQG:AG:PKEAVRAIKKRI     58
   HGS_HUMAN    7    T-FE:RLLDK::ATSQLL:LET--DWESILQICDLIRQG--.-D:TQ:AKYAVNSIKKKV     53

       QUERY   56    ...Q...S...P.QEWEA.........I.QA:L...........................     67
   Consensus   66    hnvQqngSnagPgNEWEAtlahsarrhMqLA:Ltvrrgeatrqrrscfqkrtirpppcdd     124
                        +   +   + +++++         + ++ +
   Consensus   59    :::N:::SN::K:NPHVA:::::::::L:LAsL:::::::::::::::::::::::::::     72
   HGS_HUMAN   54    :::N:::D-::K:NPHVA:::::::::L:YA.L:::::::::::::::::::::::::::     65
```

Figure 2. An example alignment showing hit number (#1), links, PRC alignment and aligned alignments (truncated). The original PRC HMM alignment is formatted in a BLAST-like style and now includes the counts and percentages of the Match, Insert and Delete states (M–M, M–I, D–~ pairs, respectively). The aligned alignments view shows the PRC result in multiple sequence alignment space and includes the first sequence of the query and hit alignment as well as their consensus sequences. The alignments are separated by a mid-line that indicates the PRC match states (M) with a '+'. Gaps present in the seed alignments are indicated by '–', gaps introduced by PRC by '~' and positions corresponding to columns missing from the HMM by ':'. The entire (aligned) alignments can be viewed with Jalview or downloaded by clicking on 'View alignment' or 'Download', respectively.

alignments section contains an 'aligned alignments' presentation. Specifically, this translation of 'raw' PRC results to query and hit alignments facilitates the identification of conserved residues. The combined multiple sequence alignments can be viewed in Jalview (28). The sequence labels in the Jalview applet are linked to several sequence databases, including UniProt and Entrez Protein, to facilitate the retrieval of sequence annotations.

Finally, the alignments can be downloaded for additional analyses. For example, Sequence Harmony can be used to predict specificity-determining residues from these alignments (30).

The translation from HMM alignments to sequence alignments is provided for most databases. However, the sequence alignments resulting from searches against CATH generally include a large number of gaps (indicated with ':' in the web output). Many alignment columns are indeed absent from their corresponding HMMs due to the high gap content of the seed alignments: for the entire CATH database only 15% of all alignment columns are represented in the HMMs as opposed to 91% for Pfam-A.

Figures 1 and 2 illustrate the webPRC output of a search with ADP-ribosylation factor-binding protein GGA1 (UniProt: GGA1_HUMAN) against Pfam and explain the aligned alignments view. A search with the single sequence indeed finds the known domains: VHS, GAT, and GAE (cf. UniProt). PSI-BLAST was run on this sequence to build an alignment (three iterations, *E*-value 0.0005, NCBI's NR database). Now, not only the VHS, but also the ENTH and ANTH domains are detected, while the GAE domain is not detected anymore. Indeed, the VHS, ENTH and ANTH domains are related, though in general, especially an *E*-value like that for the ANTH match (0.007) would require further data to state a homologous relationship. In addition to further profile–profile based searching, it is worthwhile to check the Pfam and CDD databases for information on the retrieved hits: CDD contains superfamilies and Pfam groups related families into clans and also provides 'internal database links'. Pfam and CDD provide information on this VHS/ENTH/ANTH cluster. Hence, webPRC can be used to easily find such clusters and links for any query alignment.

E-values can be used to judge the significance of the hits returned by PRC. However, they are accurate only if the library contains more than 1000 profile HMMs (12). The author of PRC indicated that 'for libraries of sufficient size, $E < 0.003$ can be taken as indicative of homology and $E < 10^{-5}$ as a strong match' (12). For profile–profile comparisons, Pfam uses an $E < 0.001$ as an indication of a significant match and *E*-values between 0.1 and 0.001 as an indication of a true relationship (16).

We here describe our PRC web interface and refrain from including another PRC validation. We would like to refer the reader to several benchmarking studies that report on the performance of PRC [(8,12,14,18), http://toolkit.tuebingen.mpg.de/hhpred/help_ov]. Reid *et al.* (24) benchmarked profile–profile and profile-sequence methods, including PRC, COMPASS, HHsearch, and concluded that PRC is the best method for distinguishing homologous from non-homologous domains. Depending on the specific benchmarking study, PRC performs better or worse than HHsearch, but generally better than COMPASS. We encourage prospective webPRC users to have a look at these benchmarking studies as well as the COMPASS (13) and HHsearch web servers (7).

CONCLUSION

The webPRC server provides a web-based front end to PRC, one of the state-of the-art methods for detecting remote homology, to carry out similarity searches against well-established domain databases. Since the input is a single sequence or an alignment, users need not build an HMM themselves. In addition to the domain hit distribution graphic and logo visualizations, webPRC features the translation of the PRC HMM alignments to multiple sequence alignments. This supports evaluation of a hit based on multiple sequence alignments. To this end, the Jalview applet is implemented. Furthermore, the hit, query and combined alignments can be downloaded for additional analyses.

ACKNOWLEDGEMENTS

We would like to thank Dr James Procter for extending the Jalview alignment editor with regular expression based link parsing.

FUNDING

ENFIN, a Network of Excellence funded by the European Commission within its FP6 Programme, under the thematic area 'Life sciences, genomics and biotechnology for health', contract number LSHG-CT-2005-518254. Funding for open access charge: ENFIN.

Conflict of interest statement. None declared.

REFERENCES

1. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
2. Schäffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
3. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
4. Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, **284**, 1201–1210.
5. Sadreyev,R.I. and Grishin,N.V. (2004) Quality of alignment comparison by COMPASS improves with inclusion of diverse confident homologs. *Bioinformatics*, **20**, 818–828.
6. Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.
7. Söding,J., Biegert,A. and Lupas,A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–W248.
8. Sadreyev,R.I. and Grishin,N.V. (2008) Accurate statistical model of comparison between multiple sequence alignments. *Nucleic Acids Res.*, **36**, 2240–2248.
9. Yona,G. and Levitt,M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.*, **315**, 1257–1275.
10. Madera,M. and Gough,J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.*, **30**, 4321–4328.

11. Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.
12. Madera,M. (2008) Profile Comparer: a program for scoring and aligning profile hidden Markov models. *Bioinformatics*, **24**, 2630–2631.
13. Sadreyev,R.I., Tang,M., Kim,B.H. and Grishin,N.V. (2007) COMPASS server for remote homology inference. *Nucleic Acids Res.*, **35**, W653–W658.
14. Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
15. Greene,L.H., Lewis,T.E., Addou,S., Cuff,A., Dallman,T., Dibley,M., Redfern,O., Pearl,F., Nambudiry,R., Reid,A. *et al.* (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.*, **35**, D291–D297.
16. Finn,R.D., Mistry,J., Schuster-Böckler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
17. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–288.
18. Bateman,A. and Finn,R.D. (2007) SCOOP: a simple method for identification of novel protein superfamily relationships. *Bioinformatics*, **23**, 809–814.
19. Marchler-Bauer,A., Anderson,J.B., Chitsaz,F., Derbyshire,M.K., DeWeese-Scott,C., Fong,J.H., Geer,L.Y., Geer,R.C., Gonzales,N.R., Gwadz,M. *et al.* (2009) CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res.*, **37**, D205–D210.
20. Koonin,E.V., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Krylov,D.M., Makarova,K.S., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N., Rao,B.S. *et al.* (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.*, **5**, R7.
21. Selengut,J.D., Haft,D.H., Davidsen,T., Ganapathy,A., Gwinn-Giglio,M., Nelson,W.C., Richter,A.R. and White,O. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.*, **35**, D260–D264.
22. Cuff,A.L., Sillitoe,I., Lewis,T., Redfern,O.C., Garratt,R., Thornton,J. and Orengo,C.A. (2009) The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.*, **37**, D310–D314.
23. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
24. Reid,A.J., Yeats,C. and Orengo,C.A. (2007) Methods of remote homology detection can be combined to increase coverage by 10% in the midnight zone. *Bioinformatics*, **23**, 2353–2360.
25. Schuster-Böckler,B. and Bateman,A. (2005) Visualizing profile-profile alignment: pairwise HMM logos. *Bioinformatics*, **21**, 2912–2913.
26. Vacic,V., Iakoucheva,L.M. and Radivojac,P. (2006) Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*, **22**, 1536–1537.
27. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G.R., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
28. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
29. Schuster-Böckler,B., Schultz,J. and Rahmann,S. (2004) HMM Logos for visualization of protein families. *BMC Bioinformatics*, **5**, 7.
30. Feenstra,K.A., Pirovano,W., Krab,K. and Heringa,J. (2007) Sequence harmony: detecting functional specificity from alignments. *Nucleic Acids Res.*, **35**, W495–W498.