DOI: 10.1002/minf.201800057

RadAA: A Command-line Tool for Identification of Radical Amino Acid Changes in Multiple Sequence Alignments

Inge Seim,*[a, b] Andrew M. Baker,[c] and Lisa K. Chopin[a]

Abstract: High-throughput sequencing has revolutionised biology and medicine. Numerous genomes and transcriptome assemblies are now available, and these genomic data sets lend themselves to comparisons between species, strains, and other strata. Researchers often need to rapidly identify changes, in particular amino acid substitutions that could confer biological function in their system of interest. However, we are not aware of an easy-to-use tool that can be used to detect such changes, and researchers currently rely on idiosyncratic computer code. We present RadAA, a command-line tool which screens multiple sequence alignments for radical amino acid changes in a stratum/strata by classifying residues into groups by charge (with cysteine in its own group). RadAA is easy to use, even for researchers with little experience in computational biology. It can be run on most operating systems – including MacOS, Windows, and Linux – and integrated into high-performance computing environments. The RadAA source code and executable binaries are freely available at https://github.com/sciseim/RadAA.

Keywords: Multiple sequence alignment · Genomic data sets · Protein sequences · Amino acid changes · Command-line tool

In recent years, high-throughput sequencing technologies have enabled the sequencing, assembly, and annotation of dozens of animal genomes. Moreover, tools such as Trinity[1] and BinPacker[2] allow assembly and downstream annotation of transcriptomes, offering affordable gene set acquisition when a reference genome is not available. These genomic data provide an avenue for cross-species analyses previously not possible.[3]

While there are numerous bioinformatics tools for phylogenetic analyses, including the identification of sites under parallel or convergent evolution,[4] such analyses are not always desired. For example, an exploratory tool allowing rapid identification of amino acid residue changes that could alter the function of a protein would be useful for researchers armed with annotated genome(s) or transcriptomes. The proteome is encoded by 21 amino acids, including selenocysteine, a residue co-translationally incorporated into selenium-containing proteins.[5] Some amino acids have very different physical and chemical properties, allowing them to be compared in various ways. In the literature the umbrella term 'conservative' has been used to denote changes within amino acid groups and the term 'radical' to describe changes between amino acid groups.[6] Radical amino acid changes are often reflected by functional changes. Examples of this include genes associated with male reproduction in the human lineage[7] and blonde hair in Pacific Islanders caused by an Arg93Cys change in tyrosinase-related protein 1 (*TYRP1*).[8]

We are not aware of a published software tool that enables the user to quickly identify amino acid changes in multiple sequence alignments. Rather, it appears that researchers keep 'reinventing the wheel', developing computer code that is often not shared beyond a research team or project. To this end, we report on RadAA, a tool to identify radical amino acid changes in multiple sequence alignments when residues are classified by charge, with cysteine in its own group.

The source code for RadAA (v2.0) was written and tested using Perl v5.18.2 on an Ubuntu Linux desktop. RadAA is a command-line tool that can be executed under any Unix-based operating system, as well as Microsoft Windows. The standalone executable, which will work on systems that do not have Perl installed or do not allow installation of Perl

[a] *I. Seim, L. K. Chopin*
Comparative and Endocrine Biology Laboratory, Translational Research Institute - Institute of Health and Biomedical Innovation, School of Biomedical Sciences, Queensland University of Technology, 37 Kent St, 4102, Woolloongabba, Australia
E-mail: i.seim@qut.edu.au

[b] *I. Seim*
Integrative Biology Laboratory, College of Life Sciences, Nanjing Normal University, 1 Wenyuan Road, 210023, Nanjing, China

[c] *A. M. Baker*
School of Earth, Environmental and Biological Sciences, Science and Engineering Faculty, Queensland University of Technology, 2 George St, 4001, Brisbane, Australia

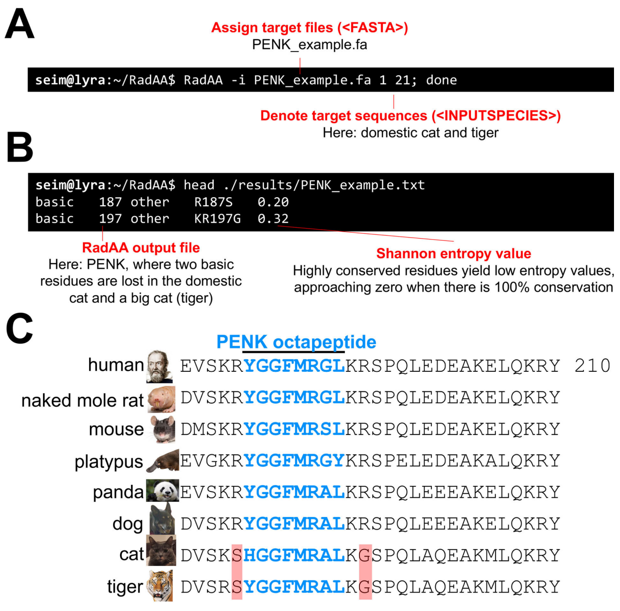
Supporting information for this article is available on the WWW under https://doi.org/10.1002/minf.201800057

modules, was generated using the PAR module (PAR::Packer; available from the CPAN archive; http:cpan.org) in Ubuntu 17.10, Darwin v17.4.10 (macOS High Sierra), and Windows 10 (64-bit). RadAA is available under a GNU General Public License (GPLv3 +, granting freedom to use the software, guaranteeing included source code, allowing modifications, and allowing free redistribution. The standalone tool and the source Perl script can be downloaded from GitHub (https://github.com/sciseim/RadAA).

RadAA accepts FASTA files containing two or more protein sequences, aligned using softwares such as Clustal-O,[9] MUSCLE,[10] or T-Coffee.[11]

In RadAA amino acids are classified by charge, with the exception of cysteine which is classified into a distinct group to reflect its critical functional properties:[12] negative (ED), positive (HRK), neutral (STYNQGAVLIFPMW), and cysteine (C). To illustrate, in zoology the user scenario is typically the protein-coding sequences of a large number of species (orthologous proteins). For example, RadAA identifies cases where two animal species harbor E or D, whilst the other species contain positive, cysteine, or 'other' residues at that particular site.

We present a case study where we wanted to identify radical amino acid changes unique to cats (family Felidae). Multiple sequence alignments of 62 mammalian species, including the domestic cat (*Felis catus*), were downloaded via the UCSC Human Genome Browser[13] (see 'Sequence retrieval' in Supplementary file 1 for overview of the species examined). These alignments were merged with predicted proteins from the tiger, *Panthera tigris*, genome.[14] This typical RadAA user scenario, comparing a range of species, is shown in Figure 1.


A
Assign target files (<FASTA>)
PENK_example.fa
Denote target sequences (<INPUTSPECIES>)
Here: domestic cat and tiger
B
RadAA output file
Here: PENK, where two basic residues are lost in the domestic cat and a big cat (tiger)
Shannon entropy value
Highly conserved residues yield low entropy values, approaching zero when there is 100% conservation
C
PENK octapeptide
human
naked mole rat
mouse
platypus
panda
dog
cat
tiger

```
seim@lyra:~/RadAA$ RadAA -i PENK_example.fa 1 21; done
```

```
seim@lyra:~/RadAA$ head ./results/PENK_example.txt
basic   187 other   R187S   0.20
basic   197 other   KR197G  0.32
```

```
EVSKRYGGFMRGLKRSPQLEDEAKELQKRY 210
DVSKRYGGFMRGLKRSPQLEDEAKELQKRY
DMSKRYGGFMRSLKRSPQLEDEAKELQKRY
EVGKRYGGFMRGYKRSPELEDEAKALQKRY
EVSKRYGGFMRALKRSPQLEEEAKELQKRY
DVSKRYGGFMRALKRSPQLEEEAKELQKRY
DVSKSHGGFMRALKGSPQLAQEAKMLQKRY
DVSRSYGGFMRALKGSPQLAQEAKMLQKRY
```

Figure 1. (A) Screenshot illustrating RadAA usage and (B) the anticipated summary file. In this example RadAA is used to identify radical amino acid residue changes in proenkephalin A sequence (PENK unique to the domestic cat (*Felis catus*) and the tiger (*Panthera tigris*). (C) Multiple sequence alignment showing the radical amino acid changes (highlighted in salmon) at the cat proenkephalin A (PENK) octapeptide processing sites. Sequences were aligned using MUSCLE[10].

The RadAA tool is controlled by the following parameters (Figure 1A): RadAA -i < FASTA > < INPUTSPECIES >, in which '-i' is the input file argument and '< FASTA >' specifies a target multiple-sequence alignment flat file (FASTA format). '< INPUTSPECIES >' refers to the FASTA header number of the species examined (one or more species of interest), which can be determined using the companion tool FASTAheadernumber. The RadAA tool can be used to query multiple FASTA sequences by invoking core UNIX shell features (illustrated in Supplementary file 1). Because RadAA is a stand-alone executable, it can also be implemented into high-performance computing environments. Its memory footprint is typically less than 150 MB (138 MB in the case of a 26,926-amino acid isoform of titin (*TTN*), the largest protein known[15]).

In our data set RadAA parsed ~170 amino acid residues per second on a 2015 MacBook Pro with 16 GB of RAM and on a virtual machine running Ubuntu 17.10 allocated 2 GB of RAM (on a Synology DS2415 + network-attached storage device). The Windows version of RadAA is considerably slower at ~0.5 amino acid residues per second on a computer with 64 GB of RAM). We are providing this tool for users with more modest data sets, however. On a high-performance computing cluster (assigned 1 CPU and 1 GB RAM), the CPU time for a RadAA job of 37,531 FASTA files (62-species UCSC RefSeq multiway alignments) was 1.5 days. Taken together, a typical genome (~25,000 annotated proteins) can be interrogated using RadAA on a Unix-based system in less than 24 hours on a typical laptop or desktop computer.

The tool outputs one tabulated file per FASTA input file in a folder entitled 'results' (see Figure 1B). These tab-delimited files can be further parsed, visualised, and analysed using tools such as the R programming language.[16] To allow users to further filter the output, RadAA also outputs the Shannon entropy[17] for each residue (see Supplementary file 1 for details). In our data set RadAA revealed two radical amino acid changes, Arg187Ser and Lys/Arg197Gly, in domestic cat and tiger proenkephalin-A (encoded by *PENK*) (Figure 1B–C), validating and extending previous studies on this peptide hormone in the domestic cat.[18] The amino acid changes in cat PENK removes the dibasic proteolytic cleavage sites, of the opioid peptide Met-enkephalin-Arg-Gly-Leu. We speculate that the changes to cat PENK may reflect an evolved improved tolerance to pain or stress in all 37 species[19] in the family Felidae.

In conclusion, we have presented the command-line tool RadAA and demonstrated its utility. It runs on most operating systems, including Linux, MacOS (Darwin), and Windows. RadAA interrogates multiple sequence alignments and identifies radical amino acid changes, in one or more species, which could alter the biological function of a protein. The tool is easy to use, even for scientists with little experience in computational biology. We believe that our exploratory tool can help researchers in diverse scientific fields, including projects where new genomes and transcriptomes have been assembled and annotated, providing a wealth of additional analytical possibilities.

Author Contributions

I.S. initiated the project and wrote the code. All authors read and approved the final manuscript.

Conflict of Interest

None declared.

Acknowledgements

This work was supported by the National Health and Medical Research Council Australia (to L.K.C. and I.S.), the Cancer Council Queensland (to L.K.C. and I.S.), a QUT Vice-Chancellor's Senior Research Fellowship (to I.S.), the Movember Foundation and the Prostate Cancer Foundation of Australia through a Movember Revolutionary Team Award, the Australian Government Department of Health, and the Australian Prostate Cancer Research Center, Queensland (to L.K.C.). We acknowledge the use of the high-performance computational facilities (lyra.qut.edu.au) at the Queensland University of Technology.

References

[1] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, *Nat. Biotechnol.* **2011**, *29*, 644–652.

[2] J. Liu, G. Li, Z. Chang, T. Yu, B. Liu, R. McMullen, P. Chen, X. Huang, *PLoS Comput. Biol.* **2016**, *12*, e1004772.

[3] a) M. Keane, T. Craig, J. Alfoldi, A. M. Berlin, J. Johnson, A. Seluanov, V. Gorbunova, F. Di Palma, K. Lindblad-Toh, G. M. Church, J. P. de Magalhaes, *Bioinformatics* **2014**, *30*, 3558–3560; b) I. Seim, X. Fang, Z. Xiong, A. V. Lobanov, Z. Huang, S. Ma, Y. Feng, A. A. Turanov, Y. Zhu, T. L. Lenz, M. V. Gerashchenko, D. Fan, S. Hee Yim, X. Yao, D. Jordan, Y. Xiong, Y. Ma, A. N. Lyapunov, G. Chen, O. I. Kulakova, Y. Sun, S. G. Lee, R. T. Bronson, A. A. Moskalev, S. R. Sunyaev, G. Zhang, A. Krogh, J. Wang, V. N. Gladyshev, *Nat. Commun.* **2013**, *4*, 2212; c) X. Fang, I. Seim, Z. Huang, M. V. Gerashchenko, Z. Xiong, A. A. Turanov, Y. Zhu, A. V. Lobanov, D. Fan, S. H. Yim, X. Yao, S. Ma, L. Yang, S. G. Lee, E. B. Kim, R. T. Bronson, R. Sumbera, R. Buffenstein, X. Zhou, A. Krogh, T. J. Park, G. Zhang, J. Wang, V. N. Gladyshev, *Cell Rep.* **2014**, *8*, 1354–1364.

[4] a) Z. Yang, *Mol. Biol. Evol.* **2007**, *24*, 1586–1591; b) H. Ashkenazy, E. Erez, E. Martz, T. Pupko, N. Ben-Tal, *Nucleic Acids Res.* **2010**, *38*, W529-533; c) J. Parker, G. Tsagkogeorga, J. A. Cotton, Y. Liu, P. Provero, E. Stupka, S. J. Rossiter, *Nature* **2013**, *502*, 228–231.

[5] X. M. Xu, A. A. Turanov, B. A. Carlson, M. H. Yoo, R. A. Everley, R. Nandakumar, I. Sorokina, S. P. Gygi, V. N. Gladyshev, D. L. Hatfield, *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 21430–21434.

[6] a) P. H. Sneath, *J. Theor. Biol.* **1966**, 12, 157–195; b) B. Clarke, *Nature* **1970**, 228, 159–160; c) R. Grantham, *Science* **1974**, 185, 862–864; d) T. Miyata, S. Miyazawa, T. Yasunaga, *J. Mol. Evol.* **1979**, 12, 219–236; e) J. Majewski, J. Ott, *Gene* **2003**, 305, 167–173; f) K. Hanada, T. Gojobori, W. H. Li, *Gene* **2006**, 385, 83–88.

[7] G. J. Wyckoff, W. Wang, C. I. Wu, *Nature* **2000**, *403*, 304–309.

[8] E. E. Kenny, N. J. Timpson, M. Sikora, M. C. Yee, A. Moreno-Estrada, C. Eng, S. Huntsman, E. G. Burchard, M. Stoneking, C. D. Bustamante, S. Myles, *Science* **2012**, *336*, 554.

[9] F. Sievers, D. G. Higgins, *Methods Mol. Biol.* **2014**, *1079*, 105–116.

[10] R. C. Edgar, *Nucleic Acids Res.* **2004**, *32*, 1792–1797.

[11] C. Notredame, D. G. Higgins, J. Heringa, *J. Mol. Biol.* **2000**, *302*, 205–217.

[12] a) S. M. Marino, V. N. Gladyshev, *J. Mol. Biol.* **2010**, *404*, 902–916; b) E. Weerapana, C. Wang, G. M. Simon, F. Richter, S. Khare, M. B. Dillon, D. A. Bachovchin, K. Mowen, D. Baker, B. F. Cravatt, *Nature* **2010**, *468*, 790–795.

[13] W. Miller, K. Rosenbloom, R. C. Hardison, M. Hou, J. Taylor, B. Raney, R. Burhans, D. C. King, R. Baertsch, D. Blankenberg, S. L. Kosakovsky Pond, A. Nekrutenko, B. Giardine, R. S. Harris, S. Tyekucheva, M. Diekhans, T. H. Pringle, W. J. Murphy, A. Lesk, G. M. Weinstock, K. Lindblad-Toh, R. A. Gibbs, E. S. Lander, A. Siepel, D. Haussler, W. J. Kent, *Genome Res.* **2007**, *17*, 1797–1808.

[14] Y. S. Cho, L. Hu, H. Hou, H. Lee, J. Xu, S. Kwon, S. Oh, H. M. Kim, S. Jho, S. Kim, Y. A. Shin, B. C. Kim, H. Kim, C. U. Kim, S. J. Luo, W. E. Johnson, K. P. Koepfli, A. Schmidt-Kuntzel, J. A. Turner, L. Marker, C. Harper, S. M. Miller, W. Jacobs, L. D. Bertola, T. H. Kim, S. Lee, Q. Zhou, H. J. Jung, X. Xu, P. Gadhvi, P. Xu, Y. Xiong, Y. Luo, S. Pan, C. Gou, X. Chu, J. Zhang, S. Liu, J. He, Y. Chen, L. Yang, Y. Yang, J. He, S. Liu, J. Wang, C. H. Kim, H. Kwak, J. S. Kim, S. Hwang, J. Ko, C. B. Kim, S. Kim, D. Bayarlkhagva, W. K. Paek, S. J. Kim, S. J. O'Brien, J. Wang, J. Bhak, *Nat. Commun.* **2013**, *4*, 2433.

[15] M. L. Bang, T. Centner, F. Fornoff, A. J. Geach, M. Gotthardt, M. McNabb, C. C. Witt, D. Labeit, C. C. Gregorio, H. Granzier, S. Labeit, *Circ. Res.* **2001**, *89*, 1065–1072.

[16] R Core Team, **2013**.

[17] C. E. Shannon, *Bell Syst. Tech* **1948**, *27*, 379–423.

[18] a) M. Chaminade, E. Chelot, L. Prado de Carvalho, P. Bochet, J. Rossier, *Neurochem. Int.* **1996**, *28*, 155–160; b) A. Cupo, L. Leger, Y. Charnay, O. Fourrier, T. Jarry, F. Masmejean, *Neuropeptides* **1990**, *17*, 171–176.

[19] S. J. O'Brien, W. E. Johnson, *Sci. Am.* **2007**, *297*, 68–75.

Received: April 28, 2018
Accepted: June 24, 2018
Published online on July 17, 2018