# Degen1 v1.2

April Hussey, Andreas Zwick & Jerome C. Regier
University of Maryland (AH), USA
University of Maryland Biotechnology Institute (AZ, JCR), USA

Comments or questions about this script should be sent to
Dr. Andreas Zwick at degen1@phylotools.com.

README file for Degen1.pl, version 1.2
Last updated: 20 NOV 2009
Latest version available at: http://www.phylotools.com

```
*********************************************************************
*  Please acknowledge the use of this script in your publications   *
*     by citing:                                                    *
*  Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball,  B.,    *
*     Wetzer, R. Martin, J.W. & Cunningham, C.W. (2010).            *
*     "Arthropod relationships revealed by phylogenomic analysis    *
*     of nuclear protein-coding sequences". Nature (in press)       *
*********************************************************************
```

CONTENTS:

## 1) INTRODUCTION

================

The reconstruction of evolutionary relationships between ancient lineages using highly divergent, protein-coding DNA sequence data can be hampered by nucleotide compositional heterogeneity. Current versions of software frequently used to infer molecular phylogenies (e.g., MrBayes, RAxML, Garli, PAUP*) do not account for compositional heterogeneity across taxa, with divergences from homogeneity resulting in signals that can conflict, but occasionally concur, with the phylogenetic signal inferred from the sequence of nucleotides.

The PERL script "Degen1" aids in the phylogenetic analysis of highly divergent DNA sequence data by greatly reducing nucleotide compositional heterogeneity between taxa. The key observation that "Degen1" exploits is that most compositional heterogeneity resides in sites that undergo synonymous change. "Degen1" operates by degenerating nucleotides at all sites that can potentially undergo synonymous change in any and all pairwise comparisons of sequences in the data matrix, thereby making synonymous change largely invisible and reducing compositional heterogeneity but leaving the inference of nonsynonymous change largely intact. The "Degen1" script fully degenerates all codons that encode single amino acids by substituting one of the four standard nucleotides (A,C,G,T) with IUPAC ambiguity codes (M,R,W,S,Y,K,V,H,D,B,N) that allow for all possible synonymous change for that amino acid. In the current version of "Degen1" (v1.2), only the "standard genetic code" for nuclear, protein-coding genes in animals is implemented, but an updated version that can handle different genetic codes is in preparation, as well as a web-based script that will allow users to directly transform their input data. For more background information and details, see section 4, "HOW IT WORKS".

2) REQUIREMENTS & INSTALLATION
===============================

Data requirements:

- nucleotide sequences that conform to "standard genetic code" [nuclear, protein-coding genes in animals]: http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi
- sequences have to be in first open reading frame and consist only of complete codons<br>[sequence begins with 1st and ends with 3rd codon position]
- any indels, represented by dashes in the data matrix, must be triplets or their multiples and in-frame, i.e., located between nt3 and nt1 of different codons
- sequence data are in non-interleaved FASTA or FLAT file format [sequence identifier preceded by ">" or "#", entire nucleotide sequence on next line]
- typically, but not necessarily, the data file is a multi-sequence alignment

System requirements:

- any operating system (e.g., Linux/UNIX, Mac OS-X, Windows) with a functional installation of a PERL language interpreter (e.g., http://www.activestate.com/activeperl/); type "perl -V" in a shell to check for a PERL installation

There is no need to install the script other than to copy it to a directory of your choice. To use the script directly from any directory, place the script in a location that is included in your PATH variable or adjust the PATH variable accordingly.

3) USAGE
=========

The script expects only a single command-line parameter, namely the data input file:

```
                    Degen1_v1_2.pl <datafilename>
```

It is best called with your PERL interpreter, but the actual command might vary depending on your operating system. A ".fasta" or ".flat" suffix is not needed for the script to operate correctly. To reduce the run-time of the script for particularly large data files, re-direct the screen output to a log file.

An example of how to use the script correctly under Linux with re-directing screen output:

```
          perl ./Degen1_v1_2.pl mydata.fasta > mylog.txt
```

The output doesn't overwrite the original input file, but consists of three files in a single folder ["Degen1_datafilename"]. These files are the new, degenerated matrix in FASTA format ["Degen1_datafilename.fasta"], a list of each and every nucleotide transformation ["HashRegEx1_datafilename.txt"] and a list of sequence lengths and potentially problematic nucleotide triplets such as termination codons, out-of-frame indels and unexpected characters ["Warnings1_datafilename.txt"].

4) HOW IT WORKS
================

The "Degen1" script reads individual DNA sequences as strings of codons, in which there are three sequential nucleotides per codon (nt1 nt2 nt3). It then replaces every codon with a fully "degenerated" codon, using IUPAC nomenclature of polymorphic nucleotides (e.g., C+T = Y) for those nucleotides that can be variable, yet encode for the same amino acid. Hence, whenever there are multiple codons that encode the same amino acid, the original nucleotides of the input sequence are expanded to match all such codons.

At nt2, nucleotides are left unaltered, since no synonymous differences based on single nucleotide substitutions are possible at that position. Most of the expansions occur at the highly variable nt3, for example, GGG --> GGN (glycine), ATT --> ATH (isoleucine), GAT --> GAY (aspartic acid) but ATG (methionine) remains as is.

In addition to degenerating synonymous differences, the script also modifies already polymorphic codons that encode more than one amino acid (typically not commonplace). Triplets that encode leucine + phenylalanine are converted to YTN. Triplets that encode arginine + serine2 are converted to MGN. Triplets that encode histidine + glutamine, asparagine + lysine and aspartic acid + glutamic acid are converted to CAN, AAN and GAN, respectively. All other nucleotide triplets that encode such non-synonymous polymorphisms are converted to NNN.

Any indels in the sequence, represented by (multiples of) three dashes (---), are not modified by the script. In contrast, missing data, represented by question marks, are converted to N's.

For leucine and arginine codons, of which there are six each, nt3 is converted to "N". For nt1, however, there is no fully satisfactory manner to degenerate leucine- and arginine-encoding nucleotides using single, informative tokens without also being slightly misinformative or without eliminating synonymous and nonsynonymous differences. We have explored three approaches (see below) and find that the first one, which is implemented in "Degen1", is the most effective.

This first approach is to convert all leucine-encoding codons to YTN, and all arginine-encoding sequences to MGN. By this approach the conversion of, e.g., TTG and CTG (both leucine) to YTN encodes not only leucine, but also phenylalanine (TTT and TTC). Likewise, the conversion of, e.g., AGA and CGA (both arginine) to MGN encodes not only arginine, but also serine2 (AGT and AGC = AGY, as compared to serine1, which is encoded by TCN).

A second approach, which would eliminate the just-mentioned "phenylalanine" and "serine2" miscoding problems, is to leave the nt1 codings of leucine and arginine unaltered. For example, by converting leucine TTG --> TTR (leucine) and CTG --> CTN (leucine), but this would retain synonymous differences at nt1 across all leucine codons (T <--> C). The equivalent would be the case for arginine (AGG --> AGR and CGG --> CGN), which retains synonymous differences at nt1 across all arginine codons (A <--> C).

A third approach is to convert all nt1 characters that encode either for leucine / arginine or for phenylalanine / serine2 to "N", but this results in a substantial loss of information. These coding issues summarized above are to be the subject of a future manuscript (in preparation).

The effect of "Degen1" coding is to minimize, but not necessarily to eliminate, the detection of synonymous change through the introduction of an analytical criterion for data set modification. This criterion is that the input data matrix be degenerated such that when inferring change in a parsimony framework between any and all pairs of terminal sequences, synonymous change must become invisible (undetectable) while nonsynonymous change be left largely intact. There are two sequelae of particular note. Firstly, because "Degen1" degenerates leucine to YTN and arginine to MGN, it actually reinterprets leucine as part leucine and part phenylalanine, and arginine as part arginine and part serine2 (see above). However, this affects only selected nt1 characters, as it is nucleotides and not codons that are analyzed. We note that currently available software cannot analyze a "Degen1" matrix under a codon model due to the use of ambiguity codes. Secondly, because "Degen1" degenerates selected sites, it changes the estimate of the average base composition in analysis software, such as PAUP* and Garli.

5) Versions
============

The latest version of the "Degen1" script and this README file can be downloaded at http://www.phylotools.com

Future versions of the "Degen1" script will include the ability to select alternative genetic codes and to handle more input and output file formats. Further, we are in the process of setting up a web service that will allow users to generate "Degen1" matrices online.

Version 1.2: 14 MAY 2009
Version 1.0: 26 AUG 2008

6) REFERENCES
==============

The original "Degen1" script was written by April Hussey and subsequently modified by Andreas Zwick. It was first published in:

Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W. and Cunningham, C.W. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature (in press).

A more detailed description of the script, coding issues and comparison of its performance to other approaches is in preparation.

7) License
===========

This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version. This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.