Molecular Phylogenetics and Evolution 115 (2017) 50–57

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

An improved and general streamlined phylogenetic protocol applied to the fatty acid desaturase family

Matthew Wilding[a,*], Matthias Nachtschatt[a,b], Robert Speight[b], Colin Scott[a]

[a] CSIRO Land and Water, Black Mountain, Canberra, ACT 2601, Australia
[b] Queensland University of Technology (QUT), Brisbane, QLD 4001, Australia

ARTICLE INFO

Keywords:
Phylogenetics
Desaturase
Peptide pattern recognition
Multiple sequence alignment

ABSTRACT

Numerous tools to generate phylogenetic estimates are available, but there is no single protocol that will produce an accurate phylogenetic tree for any dataset. Here, we investigated some of those tools, paying particular attention to different alignment algorithms, in order to produce a phylogeny for the integral membrane fatty acid desaturase (FAD) family. Herein, we report a novel streamlined protocol which utilises peptide pattern recognition (PPR). This protocol can theoretically be applied universally to generate accurate multiple sequence alignments and improve downstream phylogenetic analyses. Applied to the desaturases, the protocol generated the first detailed phylogenetic estimates for the family since 2003, which suggested they may have evolved from three functionally distinct desaturases and further, that desaturases evolved first in cyanobacteria. In addition to the phylogenetic outputs, we mapped PPR sequence motifs onto an X-ray protein structure to provide insights into biochemical function and demonstrate the complementarity of PPR and phylogenetics.

1. Introduction

Although true phylogenetic relationships cannot be known with absolute certainty, accurate phylogenetic estimates are still possible; albeit they depend on the quality and availability of sequence data, as well as supporting biochemical information. Reliably inferring accurate phylogenetic estimates is essential for the interpretation of a phylogenetic tree and the subsequent conjecture based upon it. A variety of different approaches and tools are available, but ultimately, identifying the route to an optimal tree can be challenging. A simple phylogenetic protocol can be broken down into four steps (Hall, 2013): (1) Identify the nucleotide or amino acid sequences for analysis; (2) perform a multiple sequence alignment (MSA); (3) estimate a phylogenetic tree based on these alignments, and (4) present the tree and infer meaning from it. Tools for each stage have been developed over several decades but no one protocol is suitable for every dataset. As such, trying different tools and repeating analyses multiple times is often the best way to obtain the most accurate phylogenetic estimates (Mount, 2008).

In addition to these four steps, tools have been developed to aid the handling and management of data. For example, GBLOCKS (Castresana, 2000) can identify erroneous regions in sequence alignments, and Alistat (Misof et al., 2014) shows statistics for MSAs and can be used to remove alignment sites based on occupancy. Both have been shown to improve the quality of MSAs. Since these programs can remove sites from the alignment however, notably including aligned insertion and deletion events, determining whether the removed data are “noise” or valuable information is not always clear and the resulting alignment may be less accurate and informative than the original (Löytynoja and Goldman, 2008). MSAs are fundamental to many aspects of bioinformatics, including phylogenetics, as well as overlapping fields such as biocatalysis and structural biology where they are used to infer important amino acid positions, and as such, the accuracy of the MSA is paramount.

Once a satisfactory MSA has been produced, generating phylogenetic estimates requires identification of the phylogenetic model that best suits the data and then generation of a tree using that model. Although bootstrapping gives an indication of reproducibility, both model selection and tree production can result in a sub-optimal tree and refinement of the process is often required to assess tree space (all possible trees for a set of sequences) and establish whether the generated tree can be improved upon (Whelan, 2007). By this point, a specific combination of all the aforementioned options, and many more that have not been discussed herein, will yield the most accurate phylogenetic tree.

Recently we sought to produce a phylogenetic tree for the integral membrane fatty acid desaturase (FAD) family. There are several nomenclatures used in the literature when referring to this family, but we will simply refer to them herein as FADs. This protein family was

[*] Corresponding author.
E-mail address: matt.wilding@csiro.au (M. Wilding).

http://dx.doi.org/10.1016/j.ympev.2017.07.012
Received 2 May 2017; Received in revised form 13 July 2017; Accepted 19 July 2017
Available online 21 July 2017

identified for analysis because it contains a range of biochemical functions (including desaturases, hydroxylases, conjugases, epoxidases and acetylases), and despite some catalytically conserved sequence motifs, the sequence : activity relationships for the family remain unresolved. In 2015, protein structures for two distinct FAD family members were solved independently (Bai et al., 2015; Zhu et al., 2015), and have provided the first structural information for the family. With these structures comes the opportunity to construct more reliable homology models and interpret a portion of the available sequence data, and as such a renewed interest in the protein family is likely to follow. We therefore decided to investigate the FAD family and the processes required to generate an accurate and robust phylogenetic tree for this family. We hoped that the resulting analysis would generate insights into the sequence: activity relationship for the family, and act as a resource for the development of FAD enzymes. In addition, to our knowledge, a comprehensive phylogenetic analysis for the FAD family has not been reported in over a decade (López Alonso et al., 2003; Sperling et al., 2003), and in that time more sequences have been deposited and subsequently characterised. The availability of new bioinformatics software, improvements to existing tools and models for estimates, as well as changes to the way that data is handled and interpreted have also been improved in this time. Herein we report a comprehensive phylogenetic analysis of the FAD family, comparing several different methods in order to generate accurate estimates.

2. Materials and methods

2.1. Sequence identification

The solved structure of a FAD from *Mus musculus* (pdb accession number: 4YMK) was selected as a starting point for analysis. Based on the family Hidden Markov Model (HMM), the Pfam database for the protein (PF00487) contained 52 architectures and an alignment of > 5000 sequences. A larger alignment of 22,100 sequences, based on the same HMM from the UniProtKB sequence database, was used as the initial sequence set for analysis. Of those 22,100 the majority were uncharacterised and as such the sequences were filtered down to 122 that had confirmed activity.

2.2. Multiple sequence alignment

MSAs for the amino-acid sequences were inferred using MAFFT (Katoh and Standley, 2013) (v. 7.301b; MAFFT was chosen because of its accuracy (Blackburne and Whelan, 2012; Golubchik et al., 2007)), and Seaview was used to visualise the alignments (v. 4.4.1) (Gouy et al., 2010). MSAs were inferred using the L-INS-i or E-INS-i options of MAFFT. Merged alignments were performed first on the sequence subgroups using a L-INS-i alignment for each with the --maxiterate 1000 and --localpair options invoked. A merged alignment using E-INS-i was then performed with the --genafpair --maxiterate 1000 options invoked.

2.3. Alignment Masking

Analysis of the MSAs showed that the majority of the sites in the alignment were largely unoccupied (see Fig. 2; < 5% of sequences were represented for a given site). To remove these under-represented sites, Alistat (v1.6, -m 0.05 option invoked) was used to generate masked alignments.

2.4. Identifying optimal models for sequence evolution

The optimal model of sequence evolution for the master alignment was identified by IQ-TREE (v. 1.4.3) using the -m TESTNEWONLY option with IQ-TREE invoked. This included FreeRate modelling (to account for heterogeneous changes in evolution; -cmax = 20). The optimal model was found to be the LG model with varying free rate model categories (R values). Model selection was carried out one hundred times for each tree and compared to determine whether a consistent method was identified as optimal for the data.

2.5. Phylogenetic analysis

For each alignment, the most likely tree was inferred using the previously identified optimal model of sequence evolution. In addition, a bootstrap analysis (using the UFBoot method with 10,000 replicates) was done to determine the consistency of the data in each case. Both procedures were executed using IQ-TREE. As with model selection, tree production was repeated one hundred times and the resulting trees were ranked according to their Bayesian Information Criterion (BIC) scores. The best scoring tree from each protocol was taken as the optimum tree from that alignment.

2.6. PPR analysis

PPR software was obtained from http://vbn.aau.dk/en/publications/peptide-pattern-recognition(1400c5df-fa69-4701-8d67-ec5c38cc963b).html. The 122 unaligned sequences were input and parameters for peptide length, number of peptides and cut off were varied to maximise the number of groups and sequences retained in the analyses. The results obtained are detailed in Supplementary materials and an overview of the results is illustrated in Fig. 3.

3. Results and discussion

3.1. 1Sequence selection

Of the > 20,000 sequences identified in the UniProtKB sequence database which shared a common Hidden Markov Model (HMM), only a small number were accompanied by supporting literature that detailed functional characterisation. Ultimately confirmation of activity was obtained for 122 non-identical sequences, which were selected for further analysis (sequences detailed in Supplementary materials). Phylogenetic trees were constructed using several protocols (illustrated in Fig. 1), which are described in more detail below and compared against one-another.

3.2. Initial sequence alignments

Although various tools exist to facilitate MSAs (Edgar and Batzoglou, 2006), for this investigation we decided to use the MAFFT (Katoh and Standley, 2013) package and vary the alignment method within the package. MAFFT was selected based on previous reports of consistency and accuracy compared with other methods (Ahola et al., 2006; Dessimoz and Gil, 2010; Letsch et al., 2010; Nuin et al., 2006). The protocols described herein used two alignment methods. The first used the L-INS-i algorithm, which was selected by MAFFT as the optimal method in each case when automatic method selection was invoked. The second method used the E-INS-i algorithm, and was chosen because of its accuracy with sequences containing multiple conserved domains to generate MSAs (Katoh and Toh, 2008). Although the L-INS-i method was preferred by the software, analysis of the sequences and an understanding of the multiple conserved active site histidine box motifs, which are known to be essential for catalysis, suggested that the E-INS-i method may be more suitable for the FAD family.

3.3. Phylogenetic method optimisation

The MSAs generated in MAFFT were used to produce the first phylogenetic trees for the FAD family. Using IQTree (Minh et al., 2013) for all phylogenetic analyses, the optimal model was determined for each alignment, and in both cases shown to be the LG + R6 model,