Infection, Genetics and Evolution 36 (2015) 275–286

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

A deep phylogeny of viral and cellular right-hand polymerases

Jiří Černý [a,b,*], Barbora Černá Bolfíková [c], Paolo M. de A. Zanotto [d], Libor Grubhoffer [a,b], Daniel Růžek [a,e]

[a] Institute of Parasitology, Biology Centre of the Czech Academy of Sciences, Branišovská 31, 370 05 České Budějovice, Czech Republic
[b] Faculty of Science, University of South Bohemia in České Budějovice, Branišovská 31, 370 05 České Budějovice, Czech Republic
[c] Faculty of Tropical AgriSciences, Czech University of Life Sciences Prague, Kamýcká 126, Suchdol, 165 21 Prague 6, Czech Republic
[d] Department of Microbiology, Biomedical Sciences Institute, ICB II University of Sao Paulo, 05508-000 Sao Paulo, Brazil
[e] Veterinary Research Institute, Hudcova 296/70, 621 00 Brno, Czech Republic

ARTICLE INFO

Article history:
Received 12 June 2015
Received in revised form 22 September 2015
Accepted 28 September 2015
Available online 22 October 2015

Keywords:
Right-hand polymerase
Polymerase evolution
Virus evolution
Structural evolution
Protein tertiary structure

ABSTRACT

Right-hand polymerases are important players in genome replication and repair in cellular organisms as well as in viruses. All right-hand polymerases are grouped into seven related protein families: viral RNA-dependent RNA polymerases, reverse transcriptases, single-subunit RNA polymerases, and DNA polymerase families A, B, D, and Y. Although the evolutionary relationships of right-hand polymerases *within* each family have been proposed, evolutionary relationships *between* families remain elusive because their sequence similarity is too low to allow classical phylogenetic analyses. The structure of viral RNA-dependent RNA polymerases recently was shown to be useful in inferring their evolution. Here, we address evolutionary relationships between right-hand polymerase families by combining sequence and structure information. We used a set of 22 viral and cellular polymerases representing all right-hand polymerase families with known protein structure. In contrast to previous studies, which focused only on the evolution of particular families, the current approach allowed us to present the first robust phylogenetic analysis unifying evolution of all right-hand polymerase families. All polymerase families branched into discrete lineages, following a fairly robust adjacency pattern. Only single-subunit RNA polymerases formed an inner group within DNA polymerase family A. RNA-dependent RNA polymerases of RNA viruses and reverse transcriptases of retroviruses formed two sister groups and were distinguishable from all other polymerases. DNA polymerases of DNA bacteriophages did not form a monophyletic group and are phylogenetically mixed with cellular DNA polymerase families A and B. Based on the highest genetic variability and structural simplicity, we assume that RNA-dependent RNA polymerases are the most ancient group of right-hand polymerases, in agreement with the RNA World hypothesis, because RNA-dependent RNA polymerases are enzymes that could serve in replication of RNA genomes. Moreover, our results show that protein structure can be used in phylogenetic analyses of distantly related proteins that share only limited sequence similarity.

1. Introduction

Right-hand polymerases are important players in genome replication and repair in *Eubacteria*, *Archaea*, *Eukarya*, and viruses. Genes coding for right-hand polymerases are present in genomes of all cellular life forms and in the vast majority of viruses (Koonin, 2006). Right-hand polymerases are a monophyletic group that evolved from one common ancestor in the very early stages of life evolution (Delarue et al., 1990). Nevertheless, it is not known whether the common ancestor was a processive polymerase or a non-processive nucleotidyl transferase. According to the Structural Classification of Proteins (SCOP) database (Murzin et al., 1995), the superfamily of right-hand polymerases consists of six families: i) viral RNA-dependent RNA polymerases, which are responsible for replication and transcription of viral genomes (Ferrer-Orta et al., 2006); ii) reverse transcriptases, involved in replication of reverse-transcribing viruses (Miller and Robinson, 1986); iii) single-subunit RNA polymerases, important for transcription in T-odd phages, α-Proteobacteria, and mitochondria (Cermakian et al., 1997; Shutt and Gray, 2006); iv) DNA polymerase family A, involved in replication of T-odd phages or in repair of cellular DNA (Shutt and Gray, 2006); v) DNA polymerase family B, important for replication in the vast majority of DNA viruses as well as eukaryotes (Zhu and Ito, 1994); and vi) DNA polymerase family Y, involved in repair of eukaryotic DNA (Sale et al., 2012).

Apart from the right-hand polymerases, many life forms also use evolutionarily unrelated polymerases, such as i) multi-subunit RNA polymerases, which are involved in RNA transcription; ii) barrel-shaped cellular RNA-dependent RNA polymerases, involved in RNA interference (Cramer, 2002; Salgado et al., 2006); iii) bacterial DNA polymerase family C, major players in bacterial genome replication (Timinskas et al., 2014); and iv) the DNA polymerase family X, such as

* Corresponding author at: Institute of Parasitology, Biology Centre of the Czech Academy of Sciences, Branišovská 31, 370 05 České Budějovice, Czech Republic.
E-mail address: cerny@paru.cas.cz (J. Černý).

http://dx.doi.org/10.1016/j.meegid.2015.09.026
1567-1348/

DNA polymerase β, which are important for DNA repair (Pelletier et al., 1994; Sawaya et al., 1994).

All right-hand polymerases fold into a right hand-resembling structure containing three subdomains called fingers, palm, and thumb (Hansen et al., 1997; Kohlstaedt et al., 1992; Ollis et al., 1985; Sousa et al., 1993). The conserved protein core, responsible for nucleotide polymerization, is formed by the palm subdomain. It folds into an RNA recognition motif (RRM) containing four conserved sequence motifs (A, B, C, and D) (Lang et al., 2013). The thumb and fingers subdomains are variable, and they can be aligned only among closely related polymerases (Lang et al., 2013).

Evolutionary relationships within each of the seven families of right-hand polymerases have been extensively studied, and partial phylogenies for some of them have been obtained (Černý et al., 2014; Filée et al., 2002; Koonin, 1991; Villarreal and DeFilippis, 2000). Nevertheless, evolutionary relationships between the individual polymerase families within the right-hand polymerase superfamily are not fully understood, primarily because sequence differences between homologous but highly diverged polymerases are too high to allow for classical distance-based phylogenetic studies (Zanotto et al., 1996). Recently, Mönttinen and colleagues (Mönttinen et al., 2014) inferred the evolutionary relationships between right-hand polymerase families using the HSF program, which performs comparison and classification of protein structures (Ravantti et al., 2013). This approach allowed proposing evolutionary relationships among polymerases with known structure, giving particularly reliable phylogenies for polymerases within each family. Nevertheless, the statistical support for inter-family associations was still quite low (Mönttinen et al., 2014).

In contrast to protein sequence, which may diverge considerably over time, protein structure changes much more slowly (Holm and Sander, 1996). It is maintained by the high plasticity of interactions among several amino acid residues. Particular intra- and inter-chain interactions are achieved in a variety of ways (hydrogen bonding, stacking interactions of aromatic residues, hydrophobic interactions, etc.) without substantial changes in the protein fold, despite extensive sequence divergence (Illergård et al., 2009). The protein core is the most conserved part of all proteins. Amino acid residues involved in important contacts are usually not only well conserved but also are located at the same positions of the conserved folds (Illergård et al., 2009). The protein core is surrounded by less conserved region, which show higher sequence similarity only among closely related proteins. Changes in these domains lead to changes in enzyme specificity or to changes in protein interacting partners (Lu et al., 2013). Nevertheless, conserved residues present in highly divergent proteins may not convey sufficient phylogenetic signal to unveil deeper ancestral relationships among organisms (Zanotto et al., 1996). For this reason, the evolutionary stability of protein tertiary structures can be used to reconstruct the evolutionary relationships of distantly related proteins.

One of the approaches to increasing phylogenetic evidence is to create a character matrix quantifying the morphological features of the studied proteins. Such a matrix can then be combined with protein sequence alignment during phylogenetic inference to increase the amount of available useful information (Scheeff and Bourne, 2005).

In this study, we present the first robust phylogenetic tree to describe evolutionary relationships among right-hand polymerases based on comparison of both their structure and sequence. The resulting tree allowed us to speculate about the evolutionary history of right-hand polymerases and their role in the evolution of life.

2. Materials and methods

2.1. Selection of right-hand polymerase representatives

The polymerases were selected from the SCOP database (Murzin et al., 1995) superfamily of RNA/DNA polymerases (e.8.1). This condition leads to quite a narrow definition of right-hand polymerases because it includes only polymerases with known tertiary protein structure while excluding, for example, all eukaryote-infecting DNA virus polymerases for which structural information is missing. Some polymerases are not listed in the SCOP superfamily e.8.1, despite apparently being members of it, as is the case with Qβ phage polymerase (PDB ID 3AVX) (Takeshita and Tomita, 2010), which was arbitrarily added to our list despite not being listed in the e.8.1 superfamily.

Selected polymerases were clustered via BLASTCLUST (Altschul et al., 1997) to allow grouping using an identity cut-off of 40%. Proteins with higher sequence identity can be easily aligned using only sequence information (Elofsson, 2002; Illergård et al., 2009). The representatives of polymerase groups created by BLASTCLUST were selected manually. Structures with a bound template, substrate, and/or primer, structures of non-mutated proteins, high-resolution structures, and structures with maximal solved protein chain length were preferred to minimize differences arising from conformational changes in polymerases at different steps of the enzymatic cycle.

2.2. Comparison of right-hand polymerase structures and sequences

Structural superposition of selected right-hand polymerases was calculated using the DALI server (Holm and Rosenström, 2010). The structure-based sequence alignment of the polymerase palm subdomain sequences was generated using an automatic algorithm implemented in T-Coffee Expresso (Armougom et al., 2006). The known tertiary structure of selected polymerases was used to improve the final alignment (Armougom et al., 2006).

A character matrix describing structural features of selected right-hand polymerases was constructed manually. Individual quantified protein features were selected on an empirical basis by comparing the structural and functional features used previously for the description of these enzymes (Černý et al., 2014; Gong and Peersen, 2010; Hansen et al., 1997; Lang et al., 2013; Sousa et al., 1993; Steitz, 1999). Each of the matrix columns represents a single selected character typical for at least one but not all viral RNA-dependent RNA polymerases (RdRPs) while the matrix rows represent each evaluated polymerase. The structural characters were coded for subsequent analysis in MrBayes as standard data (0–9). Their character was set as unordered, allowing them to move freely from one state to another (e.g., a character designated as “0” can change to “2” without passing “1”).

2.3. Phylogenetic analyses

The best-fitting model of amino acid residue substitutions was tested in PROTTEST 2.4 (Abascal et al., 2005). The BLOSUM matrix, with a proportion of invariable sites and a gamma-shaped distribution of rates across sites (Yang, 1994), was chosen. Phylogenetic analysis was performed using MrBayes v3.1.2 (Ronquist and Huelsenbeck, 2003). MrBayes was selected for analysis because it is the best currently available method for reconstruction of distant evolutionary relationships that is less prone to attracting long branches using proper model and appropriate taxon sampling (Glenner et al., 2004; Huelsenbeck and Ronquist, 2001). The analysis was run using a mixed dataset including both sequence and structural features (datatype = mixed). The analysis consisted of two runs with four chains (one cold and three heated) and was run for 10 million generations and sampled every 100 generations. The first 25% of the samples were discarded as a burn-in period. The average standard deviation of the split frequencies was significantly below 0.01. Chain convergence was verified with the AWTY system (Wilgenbusch et al., 2004). The equal settings were used in analyses of phylogenetic tree stability. Moreover, datasets with (i) excluded individual conserved motifs or (ii) excluded individual representatives of all polymerase families were used to verify the robustness of the phylogenetic tree topology. This verification allowed us to detect possible systematic sources of error during the inferential process. The first approach is intended to evaluate the variation in the contribution of