Computer System for Analysis of Molecular Evolution Modes (SAMEM): Analysis of molecular evolution modes at deep inner branches of the phylogenetic tree

Konstantin V. Gunbin, Valentin V. Suslov, Mikhail A. Genaev and Dmitry A. Afonnikov

Abstract. SAMEM (System for Analysis of Molecular Evolution Modes), a web-based pipeline system for inferring modes of molecular evolution in genes and proteins (http://pixie.bionet.nsc.ru/samem/), is presented. Pipeline 1 performs analyses of protein-coding gene evolution; pipeline 2 performs analyses of protein evolution; pipeline 3 prepares datasets of genes and/or proteins, performs their primary analysis, and builds BLOSUM matrices; pipeline 4 checks if these genes really are protein-coding. Pipeline 1 has an all-new feature, which allows the user to obtain $K_r/K_c$ estimates using several different methods. An important feature of pipeline 2 is an original method for analyzing the rates of amino acid substitutions at the branches of a phylogenetic tree. The method is based on Markov modeling and a non-parametric permutation test, which compares expected and observed frequencies of amino acid substitutions, and infers the modes of molecular evolution at deep inner branches.

Keywords: Molecular evolution, Markov modeling, permutation test, radical and conservative amino acid substitutions, web-based application

1. Introduction

Analysis of the evolution of proteins and protein-coding genes typically involves the following stages: (1) retrieval of sequences of gene/protein families from data banks; (2) a multiple sequence alignment; (3) development of a gene or protein evolutionary model that describes the nature of evolutionary changes in the set of sequences; (4) construction of phylogenetic trees based on the evolutionary model and multiple sequence alignment (this tree tells about phylogenetic relationships between gene/protein molecular evolution and evolution of species); (5) reconstruction of ancestral sequences in each internal node of the phylogenetic tree (this reconstruction is based on the evolutionary model developed at Stage 3 and allows the most likely DNA or protein sequences of transitional forms, which have become extinct in the course of evolution, to be inferred); (6) analysis of modes in gene and/or protein evolution (this analysis allows selection pressure on protein structure to be estimated, selectable structural and functional characteristics to be identified, and tells whether selection is neutral, positive or negative); (7) identification of positive, negative or neutral evolutionary mode at each branch of the phylogenetic tree (this helps to understand relationships between geological/ecological events and molecular evolutionary events); (8) a search for correlations between events in the molecular and phenotypic evolution of organisms (this search helps to understand relationships between the evolution of molecular characteristics of
organisms and their physiological, ecological, and morphological evolution).

A large variety of free software and algorithms that can handle some of the above stages are available; however, only two titles can handle more, and have a simple common user interface: MEGA 5, an integrated tool [72], which handles stages 2–7; and Phylemon 2, a suite of web-tools [63], which handles stages 2–4 and 6–7. We present SAMEM, an original development, which combines all the existing advantages with useful additions.

The problem of identification of modes of evolution in protein-coding genes was stated in the mid-twentieth century as the first genes and proteins were sequenced [17; 38]. Until recently, positively selected genes were largely searched for relying on non-synonymous-to-synonymous substitution ratios (Ks/Ka) [61; 53]. Over the last decade, a wealth of data has been gathered in support of the existence of a universal relationship between the level of gene expression and the rates of synonymous substitutions [13; 11; 12; 22]. In 2008, it was shown that this relationship exists because of selection against protein misfolding due to translation and in favor of (1) reducing the number of translation errors, (2) decreasing the probability of protein misfolding due to translation, and (3) reducing the probability of protein misfolding and denaturation [11; 12]. Protein structure, function and folding are determined by a combination of amino acid properties. Obviously, selection pressure exerted on items (2) and (3) in this listing is associated with amino acid substitutions that have different effects on protein structure, which makes us revise our views on the nature of positive selection events in protein-coding genes. Therefore, analysis of amino acid substitutions rates is likely to hold promise [11; 12]. Recently, more than 500 amino acid properties have been identified [35], which allow a systematic study of protein-coding gene evolution to be conducted. The two approaches intended for inferring the rates of change in certain properties of amino acids are analysis of radical-to-conservative substitution ratios (Kr/Kc) [82; 71] and analysis of the rates of change in physicochemical properties of amino acids (VPC) in the course of their evolution [56].

SAMEM (System for Analysis of Molecular Evolution Modes), a free web-based system of pipelines, implements improved modifications of both approaches. With SAMEM, Kr/Kc estimates are obtained from a unique division of amino acid substitutions into conservative and radical using all currently available data on all proteins in any protein family under study. The rates of protein evolution are estimated under a Markov model that takes into account all known structural and functional properties of the protein family (the rate of amino acid substitutions, the amino acid composition, and the clade-specific rate of protein evolution). Additionally, SAMEM does a statistical comparison of the rates of change in the physicochemical properties of proteins (VPC) and the numerical values of phenotypic traits of organisms, which allows the user to make inferences as to whether a particular molecular evolutionary event and a particular adaptive evolutionary are correlated.

2. SAMEM architecture

SAMEM is a system of four pipelines (Fig. 1). Pipeline 1 performs analyses of protein-coding gene evolution; Pipeline 2 performs analyses of protein evolution; Pipeline 3 prepares datasets of genes and/or proteins, performs their primary analysis, and builds BLOSUM matrices [30]; Pipeline 4 checks if these genes really are protein-coding. Pipelines 1 and 2 do the main stages of data processing using various methods of multiple sequence alignment, phylogenetic tree reconstruction and reconstruction of ancestral sequences (Fig. 2).

Multiple sequence alignments are generated using MAFFT 6.717 [34] and KALIGN 2.04 [41]; phylogenograms (phylogenetic trees, in which different root-to-node branch lengths reflect different substitution rates) are constructed using FastTree 2.1.1 [52] and PhyML 3.0 [24]. Each phylogram is converted to a chronogram (a phylogenetic tree constructed using

![Fig. 1. The SAMEM start page.](image-url)
relaxed molecular clocks) using r8s 1.71 [64] with user-specified divergence dates. Ancestral protein sequences are reconstructed based on gapless alignments of protein using programs ANCESCON [4], FASTML (server version) [55] and codeml (a program in the PAML 4.4 package) [78]; ancestral gene sequences are reconstructed based on gapless alignments of codons using ANC-GENE [80], FASTML (server version) [55] and codeml (a program in the PAML 4.4 package) [78]. The parameters of the evolutionary model of a particular protein family can be estimated using MODELESTIMATOR 1.1 [2] or generalized evolutionary models.

The key feature of Pipeline 1 is that it has an option to use various methods for obtaining $K_r/K_c$ estimates: Zhang’s method (the HON-NEW program) [82],...
Smith’s method [71] and its improved modification (this work). With Zhang’s and/or Smith’s methods, the user can specify the number of groups into which to classify 20 amino acids. For each of 531 properties [35], amino acids are divided into classes by \( k \)-means clustering using R [57]. The key feature of Pipeline 2 is that it implements a new original method for analysis of the evolutionary rates of various classes of amino acid substitutions based on the same 531 properties.

SAMEM has a module that does a statistical comparison of the rate of change in the properties of amino acids and the phenotypic traits is implemented in R (ape package) and implements three groups of statistical data analysis techniques [50] (GEE, Generalized Estimating Equations; the Lynch method or Variance Partitioning; GLS, Generalized Least Squares), which take into account phylogenetic inertia (that is, any correlation that arises solely because organisms are related).

Each pipeline program is provided with a hyperlinks to view a detailed description of the method being applied. To make the user experience with particular computation tasks easier, SAMEM has example input data. Additionally, the SAMEM user has an option to specify what computation tasks or pipeline programs to run. Computational modules of the SAMEM package are available both as web-pages and as web-services.

3. Early approaches for inferring evolutionary modes in proteins and protein-coding genes

During the evolution of proteins, two classes of amino acid substitutions occur [38; 48; 58]: neutral
substitutions \( (V_a) \) and functionally significant substitutions \( (V_c) \), the latter normally lowering viability or giving the species some selective advantages. Analysis of \( V_a/V_c \) values reveals evolutionary periods of intensive change in protein function \( (V_a/V_c > 1) \).

Canonical amino acids can be grouped based on the amount of similarity between their respective physicochemical properties. Amino acid substitutions within each of these groups are called conservative, while the substitutions that occur between groups are called radical. As is known, if amino acids are similar in charge and polarity, substitutions in protein evolution will largely be conservative [84; 14; 6; 9]. It has also been shown that the genetic code has such a structure that a random nucleotide substitution is rather conservative than radical in most cases [14; 19; 15]. Consequently, a simplification was proposed to enable analysis of the modes of evolution in proteins without reference to their spatial structure or function. In 1990, Hughes and the co-workers developed a method, the key of which is the calculation of ratios between the rates of radical and conservative nonsynonymous substitutions \( (K_{R}/K_{C}) \) [31]. Counting the number of conservative and radical nonsynonymous sites and the number of observed conservative and radical nonsynonymous substitutions per site [31] was carried out as in the method proposed by Nei and Gojobori [46]. Later, several improvements to Hughes and co-workers’ method were proposed. Zhang’s approach [82] considers the fact that the number of transitions is not the same as the number of transversions [81]; Smith’s approach [71] does the same and, additionally, considers alternative ways of substitutions, if substitutions are multiple, and considers the frequencies of occurrence of codons. To consider alternative ways of substitutions in Smith’s method [71], it is required that transition-to-transversion ratios, nonsynonymous-to-synonymous substitution ratios \( (K_R/K_S) \) and radical-to-conservative nonsynonymous substitution ratios \( (K_R/K_C) \) be estimated simultaneously. The iterative method proposed by Yang and Nielsen [77] based on approximate maximum likelihood estimation was used to this end.

The \( K_{R}/K_{C} \) ratio is useful for detecting selection pressure on protein globule [25] and for identification of relaxed selection pressure in protein-coding genes [29]. Nevertheless, analysis of evolutionary modes in genes with reliance on \( K_R/K_C \) estimation has one major drawback: the researcher decides what amino acid substitutions are conservative and what are radical based on his or her subjective opinion [29]. In 2001–2005, McClellan and co-workers [44; 43] proposed a simple objective measure: analysis of the deviations of the observed changes in physicochemical properties from the expectation. In this case and in the case of \( K_{R}/K_{C} \) estimation, the reconstruction of ancestral gene sequences is due. However, at the deepest branches of the phylogenetic tree, where large taxonomic groups of organisms reside, synonymous substitutions achieve a high level of saturation [76; 40; 26], which prevents accurate reconstruction of ancestral DNA. Hopefully, the reconstruction of ancestral proteins is a more reliable choice, and so we can explore the evolutionary modes of protein-coding genes even at the deepest branches. Pupko and co-workers [56] proposed an analysis of evolutionary modes of proteins by reconstructing sets of ancestral sequences of proteins using a maximum likelihood method, which considers the tree topology, and using a generalized matrix of the relative frequencies of amino acid substitutions, which tells something (but not enough) about protein evolution [56]. In their approach, a Markov model of amino acid substitutions is used, the distribution of the physicochemical distances between the ancestral and descendant proteins at each branch is calculated and so are the mean physicochemical distances. Their approach examines the probability of the deviation of the observed (most likely) physicochemical distance from the average distance at a particular branch of the tree.

4. SAMEM-implemented approaches for analysis of protein and protein-coding gene evolution

4.1. Modification of Smith’s method for estimation of \( K_{R}/K_{C} \)

The common way to divide amino acid substitutions into radical and conservative is to make decisions by looking at their physicochemical properties [82; 71; 25]. In our opinion, a better way to perform this discrimination is by analysis of elements in a matrix of substitution rates for each particular protein family. For example, in log odds BLOSUM series of matrices [30], values greater than zero are common for the classes of amino acid substitutions that occur frequently in homologous proteins [1]. It should also be noted that in the \( K_{R}/K_{C} \) estimation methods proposed by Zhang [82] and Smith [71], no division of amino acids by their property is used, and what is used is the fact whether the classes in each pair of exchanged amino acids match or mismatch.
Consequently, we propose to use BLOSUM matrices. If the value of an element corresponding to a pair of exchanged amino acids is non-negative, then the substitution is conservative, else radical. An additional advantage of our approach is that the BLOSUM series can be generated based on amino acid sequence sets that are specific for the protein family being studied. In SAMEM, the BLIMPS package [30] serves this purpose. The $K_R/K_C$ estimates so obtained are not dependent of the researcher’s subjective opinion, but reflect the frequencies of amino acids and the probabilities of their substitution.

4.2. Identification of positive selection by $K_R/K_C$ estimation

As is known, selection pressure is not the only factor operating in molecular evolution. The other factors affecting $K_R/K_C$ values are as diverse as the frequency of codons, transitions-to-transversions ratios, the rates of gene evolution at different branches of the tree, non-synonymous-to-synonymous substitution ratios [71]. The contribution of these factors cannot be expressed in analytical form when assessing the significance of the deviation of the observed $K_R/K_C$ values from those specific for neutral evolution. To eliminate the effects of these factors on identification quality, SAMEM uses Markov modeling using the INDELible 1.03 package [16]. The molecular evolution of gene sequences is simulated with the parameter values as obtained from the dataset being processed. Whenever the distribution of model $K_R/K_C$ values is normal, the deviation of the observed $K_R/K_C$ values from the mean model $K_R/K_C$ equal to or greater than 3 standard deviations implies that selection is either negative or positive, but never neutral. Thus the SAMEM user can evaluate the bias in $K_R/K_C$ estimation due to the above factors and identify cases of positive selection pressure on the rate of radical substitutions.

4.3. Identification of atypical amino acid substitutions in protein evolution

Any study of the molecular evolutionary modes in protein-coding genes early in the evolution of organisms is only possible with the use of Markov modeling of protein evolution. To develop our method for searching for statistically rare (atypical) amino acid substitutions, we followed in the steps of Pupko and co-workers [56]. Our method is based on Markov modeling of protein evolution using the INDELible 1.03 package [16] and the non-parametric permutation test [25; 27]. Importantly, INDELible 1.03 [16] allows the researcher to introduce to the model as many parameters describing protein evolution as available (for example, differences in the frequencies of amino acids, changes in the frequency of amino acids in paralogous groups within the same protein family, and evolutionary parameters at separate branches of the tree). The permutation test [25; 27] is applied to identify, at each branch of the phylogenetic tree, statistically rare classes of amino acid substitutions (atypical changes) with frequencies significantly higher than in the model sequences. Additionally, with the permutation test, no correction needs to be applied to the significance level when multiple hypotheses are being tested simultaneously [21]. Our method works in three steps. Step (1): The observed amino acid substitutions of all possible classes (190) are counted ($n_{\text{real(Class)}}$) at each branch of the tree. For internal branches, the number is inferred from a pairwise comparison of each reconstructed ancestral and each descendant sequence, taking into account the probability of each reconstructed ancestral amino acid. For terminal branches, the number ($n_{\text{real(Class)}}$) of amino acid substitutions of all the classes is inferred from a pairwise comparison of the reconstructed ancestor and descendant sequences in the modern organism. Step (2): The expected number of amino acid substitutions of all the classes is calculated for each branch of the tree based on the user-specified number of simulation runs (in this work, this number was set at 1000). To this end, model protein alignments are generated using INDELible 1.03 [16] with input data as specified by the user (e.g. an evolutionary model, and a phylogram). Step (3): the permutation test is applied to compare the expected and observed frequencies of each class of the amino acid substitution at each branch of the tree. Given a sufficient number of simulation runs, any set of observed amino acid substitutions is always a subset of expected amino acid substitutions. Therefore, any user-specified number ($N$) of sets of expected amino acid substitutions (in this work, $N$ was set at 10000) such that each of these sets contains as many substitutions as the set of observed amino acid substitutions can be generated by random permutation. For each of these random sets, the test obtains the number ($n_{\text{rand(Type)}}$) of amino acid substitutions of each particular class. Furthermore, random sets with $n_{\text{rand(Type)}} > n_{\text{real(Type)}}$ are counted ($M$).
$M/N$ is equal to the probability ($p$) of substitutions of any particular class occurring randomly. In this work, we considered only tree branches with atypical substitutions at $p \leq 0.01$.

5. Results and discussion

We implemented a modification of Smith’s original method for $K_R/K_C$ estimation [71]: we used his parameter estimation procedure, but based on maximum likelihood as described by Goldman and Yang in 1994 [20]. We compared the results of our modification with those of the original method using tests described by Smith in 2003 [71]. The results of this comparison are presented in Table 1. As can be seen, the modified method is as accurate as the original.

To demonstrate SAMEM relevance for the analysis of molecular evolutionary modes in protein-coding nucleotide sequences, we ran it on primate lysozymes (the matrix used was BLOSUM 60). The sequences and tree topology were the same as in the work byMessier and co-workers [45]. The results of the analysis are presented in Fig. 3. As can be seen from this figure, the absolute $K_R/K_C$ values are not correlated with the $K_S/K_S$ values. At the same time, the inferences made from modeling are consistent with those made by Messier and co-workers [45]: all the values in the figure are greater than five standard deviations from the mean $K_R/K_C$ value inferred from analysis of 300 generated (modeled) multiple alignments. Interestingly, the primate branch that leads to the colobine monkeys is characterized by the strongest departures from mean model $K_R/K_C$ value, which is very much consistent with the fact that these monkeys have the most specialized diet among the Old World monkeys. Lysozyme is an antibacterial enzyme. Colobine monkeys are a group of the “most” phyllophagous monkeys. To be able to benefit from diets poor in nutrients, they entered into a symbiosis with bacteria and have evolved a complex, which can be described as “large salivary glands and a complex sacculated stomach as a fermentation chamber” [18], but still have retained the ability to benefit from animal proteins (e.g. eggs, insects). Additionally, with the method being proposed, previously undescribed features in gene evolution can be uncovered.

To demonstrate SAMEM relevance to analysis of molecular evolutionary modes in proteins, we ran it on orthologous proteins (extracted from the MetaPhors database [54] (Table 2)) in the Vertebrata (with Cnidaria and Tunicata as an outgroup). These proteins perform a similar function in all modern vertebrates (orthologous proteins). The phylogenetic tree branching order was taken from the Ensembl (rel. 65) species tree (http://www.ensembl.org/info/docs/compara/index.html). In this paper we focus on the molecular evolution of two stress-response protein families, RAD1 and ubiquilin, and a family of proteins involved in tissue formation, connexin 31.1.

The RAD1 gene is involved in dark repair of ultraviolet (UV) and gamma radiation-induced damage [79]. An increased frequency of atypical amino acid changes in the RAD1 gene and its orthologs was identified in the lineage extending from the common ancestor of the Bilateria towards the common ancestor of the Chordata (Fig. 4). The importance of being protected from radiation might be associated with factors of cosmic origin (such as increased UV and gamma radiation levels [59; 73]) and the exploration of new ecological niches: during the Cambrian explosion and Ordovician adaptive radiation, the chordates (and the related conodonts) were one of the few groups that became nektonic. By contrast, the rise of arthropods, with the trilobites as the dominant group, is associated with detritus feeding as their benthic habitats were protected against UV by depth and mud [33].

The protein ubiquilin is the product of the gene that controls autophagy [62]. Ubiquilin dysfunction is associated with impaired neurogenesis and neurodegenerative diseases [75; 28], in particular, through the protein presenilin-1 (PS1) (ubiquilin regulates its function by binding to this protein) [42] and through hypoxia/brain ischemic stress-mediated neuronal apoptosis [47; 39]. Ubiquilin orthologs are rapidly evolving in the lineage extending from the common ancestor of the Chordata towards the fish-tetrapod split (Fig. 5). Furthermore, the rate of atypical amino acid changes is increased in the common ancestors of all mammals (Fig. 5). The increasing complexity of the brain is one of the main trends in mammalian evolution. This process would not be possible if it had not been for the evolution of cerebral metabolic pathways as part of the metabolism of an organism [66; 60]. Curiously, in the placental lineage, increased rates of atypical changes are observed in the hedgehog tenrec and armadillo with a slow metabolism, and in artiodactyls, rodents and rabbit with a fast metabolism (Fig. 5). It should be noted that, except for artiodactyls, all these taxonomic entities are small size-rank animals.

The gap junction genes connexins (Cnx31.1) are directly involved in tissue formation. Studies of the
<table>
<thead>
<tr>
<th>Simulation set*</th>
<th>Codon frequencies**</th>
<th>$K_A/K_S$</th>
<th>$D^{***}$</th>
<th>$K_A/K_C$ (polarity and volume)</th>
<th>$K_A/K_C$ (charge)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>after Zhang</td>
<td>after Smith</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PAML 3.12</td>
<td>PAML 4.13</td>
</tr>
<tr>
<td>1</td>
<td>Unbiased</td>
<td>0.3</td>
<td>0.25</td>
<td>1</td>
<td>0.971</td>
</tr>
<tr>
<td>2</td>
<td>Unbiased</td>
<td>0.3</td>
<td>0.25</td>
<td>5</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td>Unbiased</td>
<td>0.3</td>
<td>0.25</td>
<td>10</td>
<td>0.974</td>
</tr>
<tr>
<td>4</td>
<td>Unbiased</td>
<td>0.3</td>
<td>0.5</td>
<td>1</td>
<td>0.972</td>
</tr>
<tr>
<td>5</td>
<td>Unbiased</td>
<td>0.3</td>
<td>0.5</td>
<td>5</td>
<td>0.971</td>
</tr>
<tr>
<td>6</td>
<td>Unbiased</td>
<td>0.3</td>
<td>0.5</td>
<td>10</td>
<td>0.963</td>
</tr>
<tr>
<td>7</td>
<td>Unbiased</td>
<td>0.6</td>
<td>0.25</td>
<td>1</td>
<td>0.965</td>
</tr>
<tr>
<td>8</td>
<td>Unbiased</td>
<td>0.6</td>
<td>0.25</td>
<td>5</td>
<td>0.959</td>
</tr>
<tr>
<td>9</td>
<td>Unbiased</td>
<td>0.6</td>
<td>0.25</td>
<td>10</td>
<td>0.961</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Means for unbiased data</strong></td>
<td></td>
<td><strong>0.97</strong></td>
<td><strong>1.01</strong></td>
<td><strong>0.97</strong></td>
<td><strong>0.98</strong></td>
</tr>
<tr>
<td>10</td>
<td>GC-bias</td>
<td>0.3</td>
<td>0.25</td>
<td>1</td>
<td>0.798</td>
</tr>
<tr>
<td>11</td>
<td>GC-bias</td>
<td>0.3</td>
<td>0.25</td>
<td>5</td>
<td>0.885</td>
</tr>
<tr>
<td>12</td>
<td>GC-bias</td>
<td>0.3</td>
<td>0.25</td>
<td>10</td>
<td>0.91</td>
</tr>
<tr>
<td>13</td>
<td>GC-bias</td>
<td>0.3</td>
<td>0.5</td>
<td>1</td>
<td>0.8</td>
</tr>
<tr>
<td>14</td>
<td>GC-bias</td>
<td>0.3</td>
<td>0.5</td>
<td>5</td>
<td>0.874</td>
</tr>
<tr>
<td>15</td>
<td>GC-bias</td>
<td>0.3</td>
<td>0.5</td>
<td>10</td>
<td>0.904</td>
</tr>
<tr>
<td>16</td>
<td>GC-bias</td>
<td>0.6</td>
<td>0.25</td>
<td>1</td>
<td>0.808</td>
</tr>
<tr>
<td>17</td>
<td>GC-bias</td>
<td>0.6</td>
<td>0.25</td>
<td>5</td>
<td>0.871</td>
</tr>
<tr>
<td>18</td>
<td>GC-bias</td>
<td>0.6</td>
<td>0.25</td>
<td>10</td>
<td>0.896</td>
</tr>
<tr>
<td><strong>Means for GC-biased data</strong></td>
<td></td>
<td><strong>0.86</strong></td>
<td><strong>0.98</strong></td>
<td><strong>0.95</strong></td>
<td><strong>0.95</strong></td>
</tr>
<tr>
<td>19</td>
<td>AT-bias</td>
<td>0.3</td>
<td>0.25</td>
<td>1</td>
<td>1.061</td>
</tr>
<tr>
<td>20</td>
<td>AT-bias</td>
<td>0.3</td>
<td>0.25</td>
<td>5</td>
<td>1.074</td>
</tr>
<tr>
<td>21</td>
<td>AT-bias</td>
<td>0.3</td>
<td>0.25</td>
<td>10</td>
<td>1.04</td>
</tr>
<tr>
<td>22</td>
<td>AT-bias</td>
<td>0.3</td>
<td>0.5</td>
<td>1</td>
<td>1.061</td>
</tr>
<tr>
<td>23</td>
<td>AT-bias</td>
<td>0.3</td>
<td>0.5</td>
<td>5</td>
<td>1.059</td>
</tr>
<tr>
<td>24</td>
<td>AT-bias</td>
<td>0.3</td>
<td>0.5</td>
<td>10</td>
<td>1.046</td>
</tr>
<tr>
<td>25</td>
<td>AT-bias</td>
<td>0.6</td>
<td>0.25</td>
<td>1</td>
<td>1.078</td>
</tr>
<tr>
<td>26</td>
<td>AT-bias</td>
<td>0.6</td>
<td>0.25</td>
<td>5</td>
<td>1.063</td>
</tr>
<tr>
<td>27</td>
<td>AT-bias</td>
<td>0.6</td>
<td>0.25</td>
<td>10</td>
<td>1.007</td>
</tr>
<tr>
<td><strong>Means for AT-biased data</strong></td>
<td></td>
<td><strong>1.05</strong></td>
<td><strong>1.02</strong></td>
<td><strong>1.04</strong></td>
<td><strong>1.04</strong></td>
</tr>
</tbody>
</table>

*Each set represents 25 simulated sequence pairs (each sequence is 30000 codons in length).
**unbiased (A=T=G=C=0.25), GC-bias (A=T=0.1; G=C=0.4), AT-bias (G=C=0.1; A=T=0.4);
***the distance between pair of sequences;
****the transition/transversion ratio.
Fig. 3. A phylogenetic tree for lysozyme genes. Figures above branches: $K_D/K_C$ values and, in parentheses, the excess of actual $K_D/K_C$ values over the mean model $K_D/K_C$ values in terms of standard deviations, n.d. – not determined. Figures below branches: $K_A/K_S$ values as in the work by from Messier and the co-workers, n.d. – not determined [45]. The branches, for which the differences between the actual and mean model $K_D/K_C$ values are equal to or greater than 5 standard deviations, are in bold.

Table 2

<table>
<thead>
<tr>
<th>Protein family</th>
<th>MetaPhOrs accessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAD1</td>
<td>Phy000CIMX_RAT; Phy002OCIU_HORSE; Phy002U0OZ1_OTOGA; Phy002S55H_GASAC; PhyZ004KMW_PANTR; PhyZ001CHP_CANFA; PhyZ002C43_CHICK; PhyZ006HZ1_TETNG; Phy002UWH7_ERIEU; Phy006GYH_DANRE; Phy002V8CF_ORNAN; Phy0001ZN0_BOVIN; PhyZ001H2C_MONDO; Phy002YDS_LOXAF; Phy002QUVO_CHOHO; PhyZ00702W_XENTR; Phy002PMYD_PTEVA; PhyZ00332W_MACMU; PhyZ001H28_CIOIN; Phy002IQQF_ANOCA; Phy002SEOR_RABIT; Phy000WI13_NEMVE</td>
</tr>
<tr>
<td>Ubiquilins</td>
<td>Phy000C7JF_RAT; Phy002O90Z_HORSE; Phy002UP34_OTOGA; Phy002KNMR_ORYLA; Phy0008K8Z_HUMAN; PhyZ0018QH_CANFA; Phy002Z876_CHICK; PhyZ006ADI_TETNG; Phy002UYCB_ERIEU; PhyZ002SN0_DANRE; Phy002V17M_ORNAN; Phy001QCLM_BOVIN; PhyZ003KM6_MONDO; Phy002NYTE_ECHTE; Phy002IE8N_DASNO; PhyZ006SAL_XENTR; Phy002PPBF_PTEVA; PhyZ002UG5_MACMU; PhyZ001MZA_CIOIN; Phy002IMMZ_ANOCA; Phy002SBOL_RABIT; Phy000WAQX_NEMVE</td>
</tr>
<tr>
<td>Connexins</td>
<td>Phy000COYW_RAT; Phy002Z90Z_HORSE; Phy002TILT_MICMU; Phy002RY2J_GASAC; Phy0007XVF_HUMAN; Phy0003JYP_CANFA; Phy0007D06X_CHICK; PhyZ006CFB_TETNG; Phy002V162_ERIEU; PhyZ00272I_DANRE; Phy002VC7X_ORNAN; PhyZ000D06_BOVIN; PhyZ003TOQ_MONDO; Phy0020T7_ECHTE; Phy002ICAN_DASNO; Phy000E100_XENTR; Phy002L6G6_MYOLU; Phy000AAV7_MACMU; Phy002QH1J_CIOSA; Phy002IUEA_ANOCA; Phy002VPWK_OCHPR; Phy000WI13_NEMVE</td>
</tr>
</tbody>
</table>
Tunicata species *Ciona intestinalis* and *Halocynthia pyriformis* indicate that all their connexins are strongly different from vertebrate connexins, with cytoplasmic domains as the most divergent [7] (Fig. 6). Interestingly, connexins are not so diverse in the Tunicata (17 genes) as in the Vertebrata (about 20 genes). Connexins may play an important role in the physiology of the nervous system and the muscular system in the Chordata and the reception of signals from the environment [74; 65]. These physiological systems are different between the mobile and active Chordata and the sessile Tunicata. In the Chordata lineage, increased rates of atypical amino acid substitutions were found at the branch extending from the common ancestor of fish towards amphibians (Fig. 6). *Cx31.1* deficiency reduces the photostability of the cornea and lenses [3; 32; 5], which is consistent with a very unusual evolution of this gene at the time when terrestrial organisms emerged. Furthermore, some of the duplications in the connexins had occurred before the fish-tetrapod split and some after the split, yet in the tetrapods only [7]. As is known, mutations in *Cx43* (this gene resulted from a duplication) in *Danio rerio* [70] led to a change in the morphology of the fin-ray joints. In mice, *Cx40* mutations disrupt forelimb morphogenesis [51].

Because land was being invaded by species in which connexins were not so diverse as in the modern Tetrapoda or Actinopterygii, the increased rates of atypical amino acid substitutions at Amphibia branch may have affected limb morphogenesis. In the tetrapod lineage, the rates of atypical changes in Amphibia branch may have affected limb morphogenesis. Once again, the rates of atypical amino acid substitutions are increased at the inner Chordata tree branches leading to the Metatheria (Fig. 6). Additionally, mice with mutant *Cx31.1* are noted for placental insufficiency [36; 37] and disorders of nervous activity (memory impairment, deafness) [10; 31]. Thus, the adaptive evolution of *Cx31.1* and its orthologs in higher mammals is probably associated with the evolution of the placenta.

Thus, our study of the evolutionary modes in stress-response protein orthologs suggests that the rates of atypical changes in amino acids can be related to arthropores, that is, cases of adaptation that allow a taxon to leave and/or expand its habitat and/or ecological niche [67; 83]. Leaving habitats (the individuals that do so will normally come back soon, but sometimes can stay out for long) is neither very popular nor is it a rare occasion [8; 49; 69; 23]. In new, unfriendly habitats or niches, viability is supported by stress, which is, as
H. Selye put it, the general adaptation syndrome [68]. Consequently, an aromorphosis can begin with the onset of stress-response gene evolution towards stress tolerance and results in a change in tissue morphogenesis.

6. Conclusion

SAMEM, a free web-based system of pipelines for analysis of molecular evolutionary modes in genes and proteins, has been developed. SAMEM implements both previous and original methods. Analysis of gene evolution is based on an improved method for estimating ratios between the rates of radical nonsynonymous and conservative nonsynonymous substitutions \((K_R/K_C)\). Analysis of protein evolution is based on an original method for calculating the frequencies of atypical amino acid substitutions, which allows the modes of molecular evolution to be analyzed at deep inner branches of phylogenetic trees. Using this method, the molecular evolution of three families of orthologous proteins in vertebrates has been analyzed. Molecular evolutionary events corresponding to known aromorphoses in vertebrate evolution have been identified.
**Acknowledgements**

This work was supported by the following grants: RFBR grant No. 09-04-01641-a; SB RAS projects No. 26, 113, 119; RAS programs B.25 and A.II.6; Russian State Contracts No. P857, 07.514.11.4011 and 07.514.11.4003. The authors are grateful to Sofia Mamaeva and Vladimir Filonenko for technical assistance in the preparation of the English version of the manuscript.

**Author contributions**

KVG performed all analyses, designed and implemented the pipelines, made and implemented method for atypical amino acid substitutions detection, participated in Smith’s method modification. VVS suggested the biological interpretations for the results obtained. DAA modified Smith’s method for $K_R/K_C$ calculation and implemented it. MAG developed pipeline system and user’s web-interface.
of the SAMEM. KVG, VVS and DAA contributed equally to the writing of this paper.

References