ISSN 0206-5657. Вісник Львівського університету. Серія біологічна. 2016. Випуск 72. С. 75–81
Visnyk of the Lviv University. Series Biology. 2016. Issue 72. P. 75–81

УДК 579.873.1:577.181.4

OPTIMAL MODELS OF NUCLEOTIDE AND AMINO ACID SUBSTITUTION FOR SEQUENCES DERIVED FROM ACTINOBACTERIAL GENERA

I. Rokytskyy, B. Ostash

Ivan Franko National University of Lviv
4, Hrushevskyi St., Lviv 79005, Ukraine
e-mail: b.ostash@lnu.edu.ua

Bacteria of genus *Streptomyces* constitute one of the largest and most extensively studied taxon because of their prominent role as a source of small molecules, showing activity against bacteria, worms, tumors etc. Currently over hundred streptomycete and over 1000 actinobacterial genomes are available through GenBank, and this number is growing rapidly. This wealth of sequence data opens the door for all kinds of molecular evolutionary analyses within the genus. Nevertheless, to be correct and meaningful, these analyses should be based on adequate substitution models – sets of rules (in the form of tables) that explain how the characters (nucleotide or aminoacid residues) substitute each other over the time. Selection of optimal substitution models for GC-rich actinobacterial genomes has not reported so far, and this work aims to fill this gap. Here we compiled several different gene and protein sets and performed the search of substitution models that best fit our dataset. The results of our findings are reported here.

Keywords: *Streptomyces*, substitution models, evolution.

The process in which a sequence of characters changes over the time is described by substitution model. In the study of molecular evolution, it is important to know nucleotide substitutions between DNA sequences. Substitution models differ in terms of the parameters used to describe the rates at which one nucleotide replaces another during evolution. Models of DNA evolution were first proposed by Jukes and Cantor [9], assuming equal transition rates as well as equal equilibrium frequencies for all bases. Kimura [11] introduced a model with two parameters: one for the transition and one for the transversion rate. Later he improved it to three parametric model – K3Pu [12]. Felsenstein [4] proposed a four-parameter model in which the substitution rate corresponds to the equilibrium frequency of the target nucleotide. Hasegawa, Kishino and Yano [7] unified the two last models to a five-parameter HKY model. T92 [16] is a simple mathematical method developed to estimate the number of nucleotide substitutions per site between two DNA sequences, by extending Kimura's two-parameter method to the case where a G+C-content bias exists. This method will be useful when there are strong transition-transversion and G+C-content biases. The TN93 model [17] distinguishes between the two different types of transition. Transversions are all assumed to occur at the same rate, but that rate is allowed to be different from both of the rates for transitions. The small number of parameters to estimate make DNA substitution models belong to mechanistic ones. The general time-reversible (GTR) model [18] the most general neutral, independent, finite-sites, time-reversible model possible. Its known by variable base frequencies and symmetrical substitution matrix. GTR usually with the addition of invariant sites and a gamma distribution of rates across sites, is currently the most commonly selected model for phylogenetic inference.

For analyses performed over longer evolutionary distances, the evolution is modeled on the amino acid level. Since not all DNA substitutions alter the encoded amino acid, information is lost when looking at amino acids instead of nucleotide bases. However, several advantages speak in favor of using the amino acid information [5]: DNA is much more inclined to show compositional bias than amino acids, not all positions in the DNA evolve at the same speed, for-example non-synonymous mutations are more likely to become fixed in the population than synonymous, but probably most important, because of those fast evolving positions and the limited alphabet size, the DNA suffers much more from back substitutions, making it difficult to accurately estimate longer distances [6].

Unlike the DNA models, amino acid models traditionally are empirical models. They were pioneered by Dayhoff and co-workers [1], by estimating replacement rates from protein alignments with at least 85% identity. This minimized the chances of observing multiple substitutions at a site. The Dayhoff model was used to assess the significance of homology search results, but also for phylogenetic analysis. The Dayhoff PAM matrices were based on relatively few alignments, but new matrices were estimated using almost the same methodology, but based on the large protein databases available then, the latter being known as "JTT" matrices [8]. "WAG" – empirical amino acid replacement matrices from large databases of aligned protein sequence families, the evolutionary relationships within the families are taken to account [19]. Later this approach was improved to general amino acid replacement matrix "LG" by Le and Gascuel [14].

In phylogenetics, sequences are often obtained by firstly obtaining a nucleotide or protein sequence alignment, and then taking the bases or amino acids at corresponding positions in the alignment as the characters. Mutation (substitution) matrix, which is used to score mismatches in an alignment, was derived from alignments of an extremely small set of proteins that are very similar in sequence, and is therefore unsuitable for alignments between two proteins whose sequences are sufficiently similar to suggest that they might be homologous, but not similar enough to make homology obvious. The difficulties in constructing alignment are further complicated by the requirement that they handle deletions and insertions.

The purpose of this paper is to estimate evolution models at nucleotide and amino acid levels for various actinobacterial genera, with a focus on *Streptomyces*. We created several different datasets, based on several *Streptomyces* genes and corresponding proteins of different functional classes: transcriptional regulators, biosynthetic enzymes, transmembrane proteins. We run them through publicly available software to estimate which substitution models best describe the data. Also, we check the influence of different model types on real dataset. Our results provide reference information for all future efforts to apply phylogenetic methods to *Streptomyces* genomes.

Materials and methods

Three sets of orthologous genes, exemplified below by *Streptomyces coelicolor* (*sco*) gene, were selected for analysis. Gene *sco1728* codes for putative GntR family transcriptional regulator of YtrA type. Gene *sco2706* encodes putative glycosyltransferase, similar to many enzymes proposed to be involved in lipopolysaccharide biosynthesis and subject to *bldA* regulation. Gene *sco3894* encodes putative transmembrane protein. Using *Streptomyces* database (http://strepdb.streptomyces.org.uk/) there was created a group of orthologs for every of three selected genes: G_sco1728, G_sco2706 and G_sco3894, respectively. Also, protein products of these genes were used to select optimal aminoacid substitution models. Reciprocal best hit strategy was used to identify orthologs.

Since there are many approaches to multiple pairwise sequence alignment, alignment was conducted by several different algorithms: MAFFT, Clustal Omega, MUSCLE (http://www.ebi.ac.uk/Tools/msa/) [3, 10, 15] and ProbCons (http://probcons.stanford.edu/) [2].

To determine the evolution models of selected samples the IQ-Tree Web Service (http://iqtree.cibiv.univie.ac.at/) was used. Phylogenetic reconstructions were carried out on Phylogeny.fr server, using multiple alignment program MUSCLE.

Results and discussions

As mentioned above, there are number of difficulties in generation of reliable alignment. Therefore we have chosen four different algorithms for multiple sequence alignment to view how different approaches impact the result.

As a result of selection of evolutionary model, we revealed that different groups of orthologous genes have their own model and thus different evolutionary paths. Orthologs of GntR family transcriptional regulator (G_sco1728) and transferase (G_sco2706) share the K3Pu model [12] with three substitution types and unequal base frequencies, almost irrespective of the type of alignment algorithm taken. The only exception was that alignment by MAFFT led to TVM model [16] featuring special transversion rule (AG=CT) and unequal base frequencies for GntRfamily transcriptional regulator (G_sco1728). Group of orthologous genes for transmembrane proteins (*G*_sco3894) fits GTR model [19]. All data are summarized in Table 1.

Table 1

Nucleotide substitution models for three different sets of *Streptomyces* genes

	*G*_sco1728	*G*_sco2706	*G*_sco3894
Clustal Omega	K3Pu	K3Pu	GTR
MAFFT	TVM	K3Pu	GTR
MUSCLE	K3Pu	K3Pu	GTR
ProbCons	K3Pu	K3Pu	GTR

We further modeled substitution patterns in protein sequences encoded by the aforementioned groups of genes. According to obtained results, summarized in Table 2, all three groups of homologous protein respond to its own evolutionary model. Particularly, *P*_sco1728 dataset (GntR family transcriptional regulator) was best described by general matrix of Jones-Taylor-Thornton [8] and its revision by Kosiol and Goldman [13] – JTT and JTTDCMut. The group of glycosyltransferase *proteins (P*_sco2706) fits general matrix WAG of Whelan and Goldman [19], while orthologs of transmembrane protein (*P*_sco3894) obey general matrix LG of Le and Gascuel [14].

Table 2

Amino acid substitution models for three different groups of *Streptomyces* proteins

	*P*_sco1728	*P*_sco2706	*P*_sco3894
Clustal Omega	JTTDCMut	WAG	LG
MAFFT	JTTDCMut	WAG	LG
MUSCLE	JTT	WAG	LG
ProbCons	JTT	WAG	LG

Our results show that the different algorithmic approaches of the multiple sequence alignment slightly affect the determination of the substitution model; in case of protein sequences, major impact had the nature of the proteins being analyzed. For example, transmembrane proteins, presumably having biased amino acid composition (and codon usage) fit their own type of substitution model.

Next we decided to study how the changes in substitution model affect phylogenetic analysis of real dataset. As a test case we took a set of orthologous genes and corresponding proteins of AraC type pleiotropic regulator AdpA. First we attempted to reconstruct phylogenetic tree of *adpA* nucleotide sequences. For this purpose we created a cluster of 107 orthologous *adpA* sequences, collected mainly from genomes of *Streptomyces*, although representatives of non–streptomycete species were included (accession numbers and sequences are available from authors upon request). We used server phylogeny.fr (http://www.phylogeny.fr/), because it is possible to customize all steps of inference. Here we held all the parameters constant and varied only the substitution model. According to previous assessments, either K3Pu or GTR models would be suitable for GC-rich *Streptomyces* genes (see Table 1). Consequently, we built two alternative *apdA* gene trees, either GTR-, or HKY-based (the second model is considered non-optimal for our case). The resulting gene trees are shown on Fig. 1.

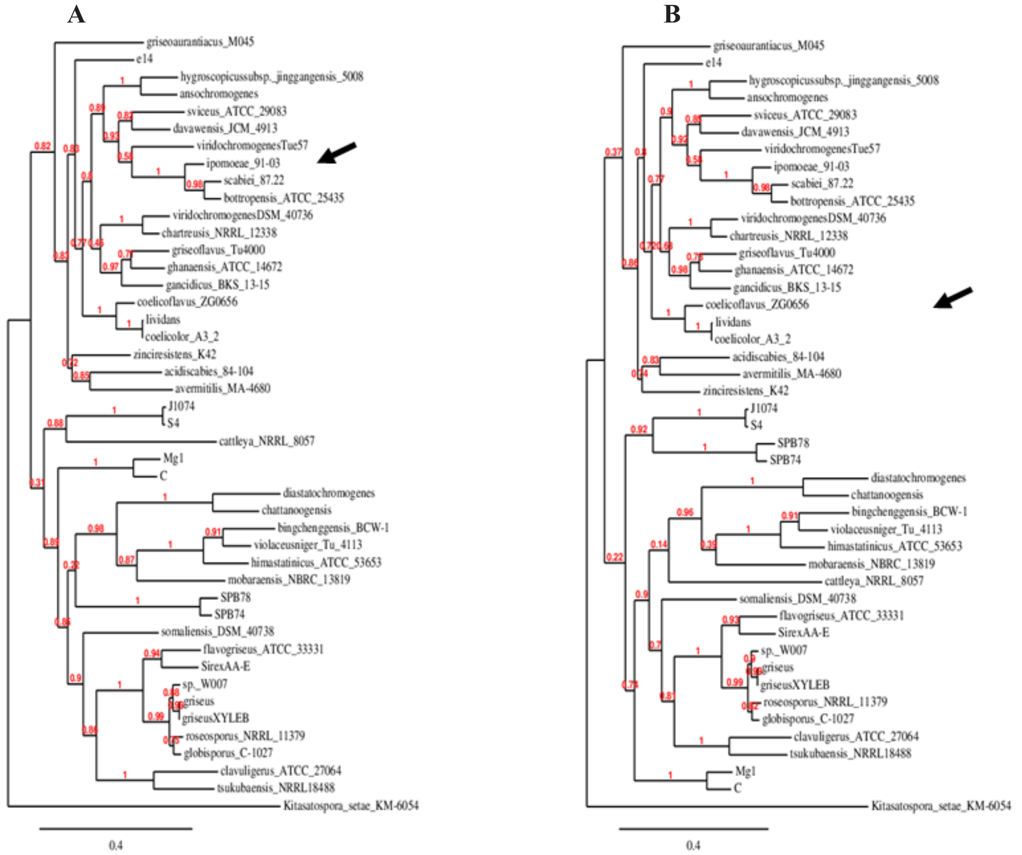

Fig. 1. Phylogenetic trees based on nucleotide (nt) sequences alignment of set of *adpA* genes, built using different substitution model, GTR (**A**) and HKY (**B**). Arrows point to *adpA* sequence that occupies different position in the tree. Bar under tree, number of nt substitutions per nt position.

Here one could see that overall topology of *adpA* tree is similar for two substitution models used: there are two major clades and adpA from *Kitasatospora* sp. was outgroup. Nevertheless, notable differences in the fine topology of the trees can be found; one of them is marked with arrow on Fig. 1. Furthermore, we noted that the tree topology built on GTR model comes with a higher statistical support values (aLRT data on the tree nodes) for major clades. Thus, different substitution models does lead to different trees, and their judicious use may improve phylogenetic inference.

Next we reconstructed AdpA protein tree from a set of 200 protein sequences using two alternative substitution models: JTT ("optimal") and WAG ("non-optimal"). The resultant trees were too big to inspect them visually, and we therefore resorted to tree manipulation web server PhyFi (http://cgi-www.cs.au.dk/cgi-chili/phyfi/go) which allows to collapse and re-draw clotted parts of the tree. Namely, over hundred *adpA* sequences from *Streptomycetaceae* genus, as well as ten members of *Frankia* genus were collapsed, as they had identical topology in both trees, shown on Fig. 2.

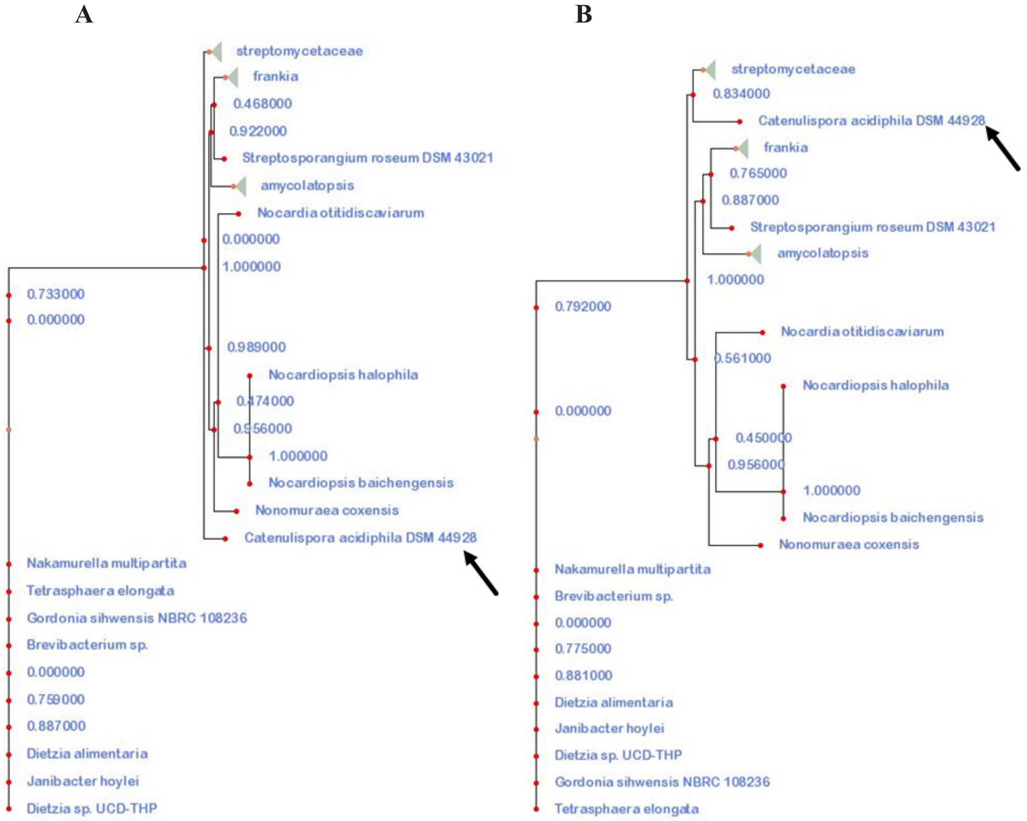

Fig. 2. Phylogenetic trees based on alignment of set of orthologous proteins of AdpA family, built using JTT (A) and WAG (B) models. Black arrow shows the position of AdpA from *Catenulispora* sp.

One can see that here, too, choice of evolutionary model had effect on protein tree topology. Namely, AdpA from recently discovered genus *Catenulispora* is placed differently in the trees. According to JTT model, the latter is situated as an outgroup with regard to the entire *Actinomycetales* clade, while in WAG model it is co-localized close to *Streptomycetaceae* clade. It is difficult to predict from this (limited) study, as to which substitution model better reflects evolution of actinomycete sequences. Nevertheless, our study provides compelling evidence that choice of evolutionary model affects phylogenies, and so this issue is worth further scrutiny. We suggest initial types of subtitution models for functionally different types of sequences and invite further exploration of suitability of these models for wider sets of actinobacterial proteins.

Acknowledgements*. Publication was partially based on the research provided by the grant support of the State Fund For Fundamental Research (project № F60/2-2015, to B.O.)*

REFERENCES

1. *Dayhoff M., Schwartz R., Orcutt B.* A model for evolutionary change in proteins. Atlas of Protein Sequence and Structure 1978. Vol. 15. P. 345–352.
2. *Do C., Mahabhashyam M., Brudno M., Batzoglou S.* PROBCONS: Probabilistic Consistency-based Multiple Sequence Alignment // Genome Res. 2005. Vol. 15 P. 330–340.

3. *Edgar R.* MUSCLE: multiple sequence alignment with high accuracy and high throughput // Nucleic Acids Res. 2004. Vol. 32. N 5. P. 1792–1797.
4. *Felsenstein J.* Evolutionary trees from DNA sequences: a maximum likelihood approach // J. Mol. Evol. 1981. Vol. 17. N 6. P. 368–276.
5. *Gonnet G., Cohen M., Benner S.* Exhaustive matching of the entire protein sequence database // Science 1993. Vol. 256. N 5062. P. 1443–1445.
6. *Halpern A., Bruno W.* Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies // Mol. Biol. Evol. 1998. Vol. 15. N 7. P. 9107.
7. *Hasegawa M., Kishino H., Yano T.* Dating of the human-ape splitting by a molecular clock of mitochondrial DNA // J. Mol. Evol. 1985. Vol. 22. N 2. P. 160–174.
8. *Jones D., Taylor W., Thornton J.* The rapid generation of mutation data matrices from protein sequences // Comput. Applic. Biosci. 1992. Vol. 8. P. 275–282.
9. *Jukes T., Cantor C.* Evolution of protein molecules. In Munro, H.N. Mammalian protein metabolism. New York: Academic Press, 1969. P. 21–123.
10. *Katoh K., Toh H.* Recent developments in the MAFFT multiple sequence alignment program // Brief. Bioinf. 2008. Vol. 9. N 4. P. 286–298.
11. *Kimura M.* A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences // J. Mol. Evol. 1980. Vol. 16. N 2. P. 111–120.
12. *Kimura M.* Estimation of evolutionary distances between homologous nucleotide sequences // Proc Natl Acad Sci U S A. 1981. Vol. 78. N 1. P. 454–458.
13. *Kosiol C., Goldman N.* Different versions of the Dayhoff rate matrix // Mol. Biol. Evol. 2005. Vol. 22. P. 193–199.
14. *Le Q., Gascuel O.* An improved general amino acid replacement matrix // Mol. Biol. Evol. 2008. Vol. 25. N 7. P. 1307–1320.
15. *Sievers F., Wilm A., Dineen D.* et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega // Mol. Syst. Biol. 2011. Vol. 7 P. 539.
16. *Tamura K.* Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases // Mol. Biol. Evol. 1992. Vol. 9. N 4. P. 678–687.
17. *Tamura K., Nei M.* Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees // Mol. Biol. Evol. 1993. Vol. 10. N 3. P. 512–526.
18. *Tavaré S.* Some probabilistic and statistical problems in the analysis of DNA sequences // Lectures on Mathematics in the Life Sciences Vol. 17. P. 57–86.
19. *Whelan S., Goldman N.* A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach // Mol. Biol. Evol. 2001. Vol. 18. P. 691–699.

Стаття: надійшла до редакції 29.01.16

доопрацьована 11.03.16

прийнята до друку 06.04.16

ОПТИМАЛЬНІ МОДЕЛІ ЗАМІЩЕННЯ НУКЛЕОТИДІВ І АМІНОКИСЛОТ У ПОСЛІДОВНОСТЯХ, ЩО ПОХОДЯТЬ З АКТИНОБАКТЕРІЙНИХ РОДІВ

І. Рокицький, Б. Осташ

Львівський національний університет імені Івана Франка
вул. Грушевського, 4, Львів 79005, Україна
e-mail: b.ostash@lnu.edu.ua

Бактерії роду *Streptomyces* становлять один із найбільших і найдетальніше вивчених таксонів. Вони відомі як джерела малих молекул, що виявляють активність проти бактерій, червів, пухлин і т. д. У даний час більше сотні стрептоміцетних і більше 1000 актинобактеріальних геномів доступні через GenBank, і це число швидко зростає. Багатство секвенованих даних відкриває двері для всіх видів молекулярно-еволюційних аналізів у межах актинобактерій. Тим не менш, щоб бути правильними і значущими, такі аналізи мають базуватися на адекватних моделях заміщення – наборі правил (у вигляді таблиць), які пояснюють, як символи (нуклеотидні або амінокислотні залишки) замінюють один одного в часі. Досі не представлено підбір оптимальних моделей заміщення для GC-багатих актинобактеріальних геномів, тому наша робота має на меті заповнити цю прогалину. Тут ми порівняли кілька вибірок різних генів і білків та провели пошук моделей заміщення, які найкраще описують наш набір даних.

Ключові слова: Streptomyces, моделі заміщення, еволюція.

ОПТИМАЛЬНЫЕ МОДЕЛИ ЗАМЕЩЕНИЯ НУКЛЕОТИДОВ И АМИНОКИСЛОТ В ПОСЛЕДОВАТЕЛЬНОСТЯХ РАЗНЫХ РОДОВ АКТИНОБАКТЕРИЙ

И. Рокицкий, Б. Осташ

Львовский национальный университет имени Ивана Франко
ул. Грушевского, 4, Львов 79005, Украина
e-mail: b.ostash@lnu.edu.ua

Бактерии рода *Streptomyces* представляют собой один из крупнейших и наиболее подробно изученных таксонов. Они известны в качестве источника малых молекул, проявляющих активность против бактерий, червей, опухолей и т.д. В настоящее время более 100 стрептомицетных и более 1000 актинобактериальных геномов доступны через GenBank, и это число быстро растет. Богатство секвенированных данных открывает двери для всех видов молекулярно-эволюционных анализов в пределах рода. Тем не менее, чтобы быть правильным и значимым, эти анализы должны базироваться на адекватных моделях замещения – наборе правил (в виде таблиц), которые объясняют, как символы (нуклеотидные или аминокислотные остатки) заменяли друг друга в течение эволюционного времени. До сих пор не представлен подбор оптимальных моделей замещения для GC-богатых актинобактериальных геномов, поэтому настоящая работа ставит своей целью заполнить этот пробел. Здесь мы сравнили несколько выборок различных генов и белков и провели поиск моделей замещения, которые лучше всего описывают наш набор данных.

Ключевые слова: *Streptomyces*, модели замещения, эволюция.