ISSN 0206-5657. Вісник Львівського університету. Серія біологічна. 2017 Випуск 75. С. 66–74
Visnyk of the Lviv University. Series Biology. 2017. Issue 75. P. 66–74

ГЕНЕТИКА

УДК 579.873.1:577.181.4

PECULIARITIES OF CODON CONTEXT AND SUBSTITUTION WITHIN STREPTOMYCETE GENOMES

I. Rokytskyy, S. Kulaha, H. Mutenko, M. Rabyk, B. Ostash

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

Actinobacteria of genus *Streptomyces* attract great interest of researchers. Their genomes encode cryptic gene clusters for as-yet-unknown antibiotics; heterologous expression of metagenomic libraries in model *Streptomyces* strains allow to discover new classes of compounds. However, it is crucial to understand the rules that govern codon usage in streptomycete genes, if we are to maximize the chances and level of expression of foreign genes in *Streptomyces*. In this study we addressed two questions related to codon usage in streptomycete genomes. First, we explored if there are patterns of dicodon usage in *Streptomyces*. Second, we searched for significant differences in patterns of codon substitution in different families of orthologous genes at different phylogenetic depth and degree of essentiality. To this end, we revealed several codon context rules, which are mainly associated with anomalous frequency of G/C downstream of C-ending codons. We developed a new bioinformatics tool, based on previously described bubble plot approach, allowing matrix-like visualization of codon substitution patterns in the dataset. Using this tool, we show that transcriptional factors of AdpA family carry significant fraction of nonsynonymous substitutions, although changes in its pattern for different actinobacterial orders (and as compared to *Streptomycetales*) do not follow simple rules.

Keywords: *Streptomyces*, codon context, codon substitution patterns

As of March 2017, there are 290 finished and over 1000 draft *Streptomyces* genomes in GenBank database. This huge body of data is expected to revitalize drug discovery through genome mining for novel gene clusters encoding as-yet-unknown metabolites [7]. One popular approach is to express *Streptomyces* metagenomic libraries in a "universal" host, saving the time spent to develop fermentation and secondary metabolism (SM) activation strategies for different actinobacteria [6]. However, full realization of this approach would require much better knowledge of factors that limit expression of foreign DNA. Most of SM gene clusters are silent (not transcribed). In addition to manipulations of transcriptional control of SM [17], optimization of codon sequences are desirable if one wants to maximize the gene expression at the level of translation [12]. *Streptomyces* genomes are GC-rich (around 70 %), and their codon usage is skewed towards preferential use of GC-rich synonymous codons. This feature, termed codon usage bias (CUB), is in the focus of intense research in model (GC-"neutral") organisms, such as *E. coli* or yeast [13]. CUB is part of a larger phenomenon of codon usage dependency (CUD) – any nonrandom association of codons in a single sequence (codon context) or in a given position of multiple sequence alignment (codon substitution/ conservation). For *Streptomyces* types, global causes and consequences of CUD are not studied. We can think of CUD within a genome as a "horizontal" axis of codon change [10]. By contrast, evolution of codon sequences (e.g codon changes observed in homologous positions of genes across different species/genera) can be considered a "vertical" axis of imaginary codon space. Evolution of codon sequences has been ex-

plored to some extent, primarily for conserved genes that can be used as phylogenetic markers for actinobacteria [4]. It is desirable to have a tool to visualize codon substitutions for large datasets. Approach was proposed that is graphically based on two-dimensional matrix where different codon substitution frequencies are shown as circles of different color and diameter (the bubble plot) [5]. To the best of our knowledge, no publicly or commercially available tool exists to compute codon-based bubble plots. In this work we took computational biology approach to address two questions related to "horizontal" and "vertical" axes of codon change in *Streptomyces*. First, does usage of certain codon influence the choice of the downstream one? Second, can we detect difference in patterns of codon substitution for large sets of homologous sequences derived from SM and primary metabolism? To this end, we provide evidence for significantly biased associations of certain codons in *Streptomyces*. In-house scripts were written as a first step towards the online program for bubble plot building. As a test case, we used the developed scripts for comparative analysis of several groups of actinobacterial genes. Main findings of this analysis, limitations and potential avenues of future use of bubble plots are discussed.

Materials and Methods

Streptomyces genome sequences were accessed via NCBI (https://www.ncbi.nlm.nih.gov/genome/). Genome and codon sequences were viewed and analyzed with UGENE bioinformatics toolkit [11]. Computational and statistical analysis of codon sequences were carried out in Mathematica software. Codon pair analysis was carried out using Anaconda software (http://bioinformatics.ua.pt/software/anaconda/) [10] Clusters of AdpA orthologous sequences of *Actinobacteria* were generated using AdpA protein of *Streptomyces clavuligerus* (SCLAV_1957) as a query and reciprocal best BLASTP hit strategy [9] for ortholog identification. Clusters of SsgA/SsgB orthologs and their coding sequences were taken from [4]. Python scripts were used to develop the bubble plot program. Bubble plot workflow is detailed in Results section. Phylogenetic relationships among actinobacterial orders were reconstructed from conservative actinobacterial proteins using PhyloPhLan phylogenetic pipeline [15]. Bioinformatics toolkit (https://toolkit.tuebingen.mpg.de/) [1] and RevTrans1.4 server (http://www.cbs.dtu.dk/services/RevTrans/) [16] were used to access various programs for multiple sequence alignment. Where appropriate, statistical significance of differences in observed and expected codon and k-mers frequencies was evaluated using bootstrap approach followed by parametric tests (t-test, ANOVA).

Results and Discussions

"Horizontal" axis: searching for codon context dependencies in *Streptomyces*. Dicodon usage is the simplest form of codon associations. Let $N_1N_2N_3|Z_1Z_2Z_3$ be dicodon sequence, where vertical bar demarcates last (third) letter (N_3) of upstream codon *N* and the first letter (Z_1) of downstream codon *Z*. Usage of codon *Z* would be considered random with regard to *N* if *Z* occurs at frequencies equal to product of background frequencies of Z_1, Z_2, Z_3 in any given genome. All significant deviations of *Z* frequency in context of *N* would be considered nonrandom. There is large body of empirical data showing that *k*-mers larger than triplet within ORFs are nonrandom [3, 14], and this observation is at the heart of several gene finding algorithms [2]. If hexamers are not random, then we should observe dependencies at a more elementary level. Particularly, we wanted to reveal codons *N* whose wobble position (N_3) determines the distribution of bases expected at Z_1. The $N_3|Z_1$ dependencies are more straightforward to infer and yet more challenging to interpret as compared to dicodon ones. If found and considered significant by some statistical measures, the $N_3|Z_1$ dependency usually implies that wobble position (often irrelevant at the level of protein) of *N* narrows down the set of possible codons *Z*, and, consequently, aminoacids we expect to observe in respective position of proteins. Existence of such dependencies in several studied cases offers new mechanistic explanation for CUB, e.g. certain codon *N* is used because it is the most preferable for occurrence of next codon *Z* [10]. We looked for such scenario in *Strep-*

tomyces genomes. The frequencies of all possible 244 quadruplets $N_1N_2N_3|Z_1$ have been calculated for the entire set of open reading frames (ORFome) of model strain *S. coelicolor* M145. Here null hypothesis would be as follows. Z_1 is random in context of $N_1N_2N_3$ when $N_1N_2N_3|Z_1$ occurs at background frequency of Z_1 in genome. The most pronounced deviations from null hypothesis were observed for C|C, C|G and A|C codon boundaries. This was found for all quadruplets having such boundaries. Examples of observed deviations are shown in Fig. 1. Interestingly, we have found that start codon ATG is nonrandomly associated with adenine (see Fig. 1, D). If this is true for other genomes, then ATG|A association can be used to map start codons.

Our data suggest nonrandom usage of certain interocodon dinucleotides. This most likely reflects nonrandom codon pair usage [10], although other factors cannot be ruled out. We proposed a statistical model to discriminate between different forces that shape the quadruplet usage. In an initial model (used to generate Fig. 1), for each codon we had four quadruplet frequencies $N_1N_2N_3|Z_1$, since Z_1 is in one of 4 mutually exclusive states (A/T/G/C). Then, for example, $P(GTT|G) = \Sigma(GTT|G)/\Sigma(GTT|N)$. Our new proposal is $P(GTT|G) = \Sigma(GTT|G)/\Sigma(GTN|N)$. (..N|N = 16 possible dinucleotides). In denominator the frequencies of 16 quadruplets are summed up (except for codon $N_1N_2N_3$ for which third letter leads to stop codon – then we will have 12 quadruplets – e.g. TGN|N). With this approach, all observed deviations of quadruplet frequencies from null (random) distribution are caused by factors other than $N_3|Z_1$ dependencies. Using parametric approaches, one can calculate *p* and standard deviations (SD) expected for random distribution of $N_1N_2N_3|Z_1$ and compare to observed data (Fig. 2).

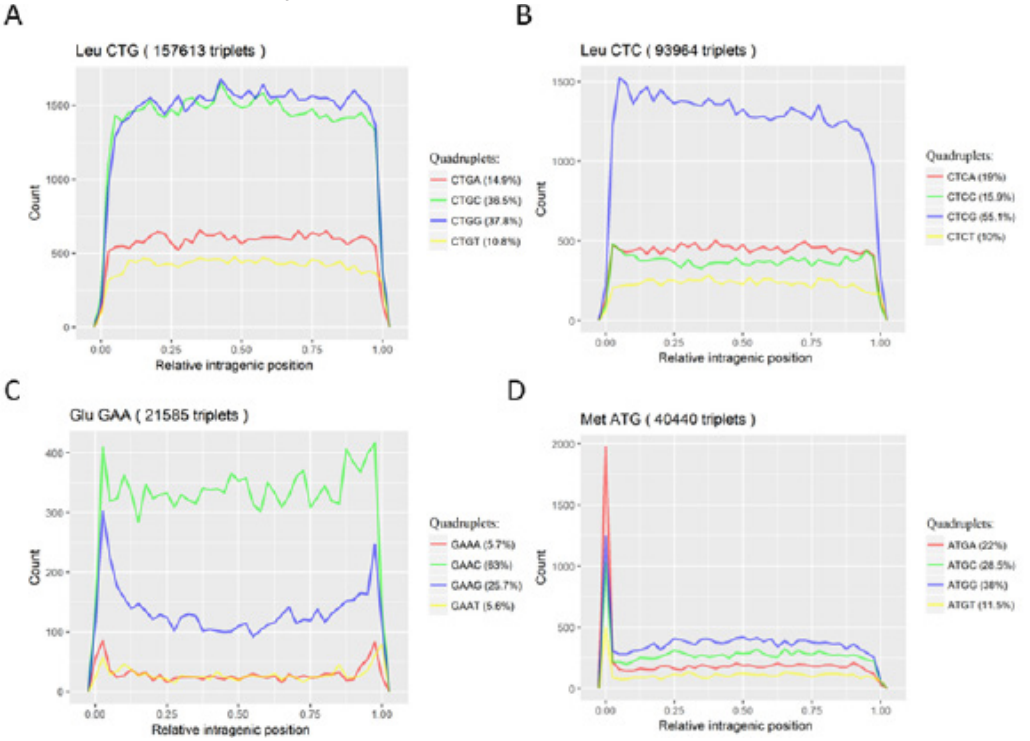

Fig. 1. Examples of observed quadruplet frequencies in *S. coelicolor* genome: occurrence at background frequencies of 4th nucleotide (**A**), biased associations for C|C, C|G, A|C (**B**, **C**), and biased ATG|A association at the beginning of the ORF (**D**). The *x* axis represents cumulative distribution of the quadruplets over relative lengths (0.0÷1.0) of all ORFs. The *y* axis represents absolute count of any given quadruplet in the ORFome

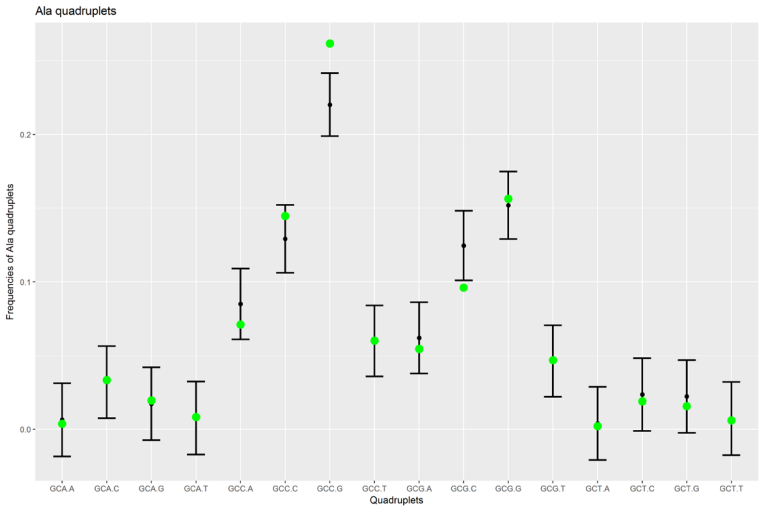

Fig. 2. Fit of observed values of quadruplet frequencies (green dots) against expected null (normalized by frequency of N3|Z1) distribution (black bars). Black dots – mean expected value. SD was calculated as described in the text. Quadruplet GCC|G is out of the null distribution

As can be seen from Fig. 2, nonrandom quadruplets can be found even after taking into account nonrandomness of $N_3|Z_1$ dinucleotides. With the new model, 11.88 % (29 out 244) of quadruplets did not fall into the null distribution (for old model 133 values were out of the null distribution). We also performed bootstrap analysis of significance of our findings [8], by randomly shuffling 145 genes of *S. coelicolor* and building 1000 permuted codon datasets. We applied both initial and new statistical models to boostrapped data. Again, C|G, C|C dinucleotides at codon boundaries deviated from null distribution, irrespective of the model (data not shown). In all, it can be concluded from our (limited) dataset that cytosine in wobble position leads to higher than random probability to find guanine in first position of next codon. Nonrandom dicodon usage contributes to the observed dependency, although other, as-yet-unknown factor(s) are at work.

Aforementioned results encouraged us to look for CUD on a scale of many *Streptomyces* genomes, where all possible codon dependencies could become more salient and general, and be converted into a set of codon context rules. For this purpose we used specialized software Anaconda [10]. Briefly, the sofware functions as virtual ribosome, that reads triplets of ORFs, records dicodon types and performs their statistical analysis. CUDs are represented by Anaconda in the form of heatmaps; an example of the latter is shown on Fig. 3. We compiled 50 *Streptomyces* genomes and subjected them to Anaconda analysis. First, we asked if there are significant differences in codon pairs between the analyzed genomes. This can be done by subtracting codon heatmaps (like the one on Fig. 3) for two different genomes; presence of any significant differences would show up as pixels that have colors other than black (no difference) or grey (no data). We revealed that the set of genomes picked up for analysis is rather homogeneous – large deviations in codon pairs were not detected (data not shown).

We built a median heatmap on basis of 50 streptomycete genomes in order to identify specific codon-pair context biases within the latter (Fig. 4). To suppress noise in the maps, values were filtered to display only those codon residuals that lie out of two SD. Green cells correspond to preferred context and red cells indicate rejected ones. All other cases were colored in black. This approach granted us clearer vision of extreme codon contexts. The main patterns are as fol-

lows: NNU|CNG (Fig. 4, A, D) and GNA|CNN (Fig. 4, C, F). There were also two patterns of dependency between non-adjacent bases: CUS|NS_2N (Fig. 4, B, E) and CUS|NW_2N (Fig. 4, E). Here, downstream of G/C-ending leucine codons (CUC and CUG) G/C nucleotides are rejected and U/A are preferred at Z_2 position ($N_1N_2N_3|Z_1Z_2Z_3$). The leucine pattern generally agrees with results discussed above (see Fig. 1, A, B).

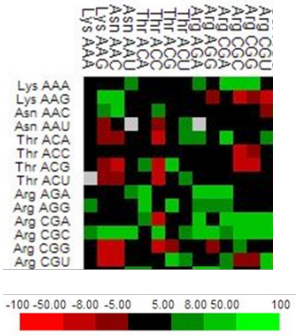

Fig. 3. Fragment of 64×64 codon heatmap that reflects all possible codon transitions for the listed (on axes) set of codons in *S. coelicolor* ORFome. Codon residuals are coded on color scale from red (rejected context) to green (preferred context). Statistically insignificant values are indicated in black. Grey pixels, no data

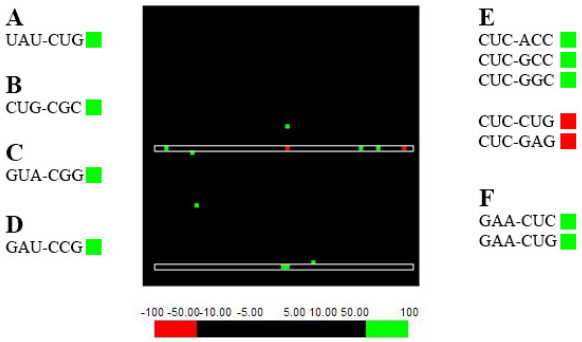

Fig. 4. Median 64×64 codon heatmap of 50 *Streptomyces* genomes. In order to identify specific codon-pair contexts, map was filtered to display codon residuals that are above 50-fold difference. Green cells correspond to preferred and red cells to rejected contexts. All other cases were colored in black. Contexts **A-D** correspond to single pixels, **E** and **F** – boxed lines of matrix

"Vertical" axis: searching for patterns of codon substitutions in different datasets of orthologous genes. We developed in-house Python scripts to automate the generation of bubble plot that depict codon substitution in alignment of multiple codon sequences. The pipeline of our analysis was as follows. First, a set of orthologous proteins has to be collected. Coding sequences of the proteins has to be identified. Second, coding sequences are subjected to amino acid sequence-guided codon alignment (program RevTrans). Third, multiple codon alignments are subjected to bubble plot analysis. Typical bubble plots are shown in Fig. 5. One of them is generated on basis of conserved sporulation-associated gene *ssgA* (Fig. 5), and it features paucity of nonsynonymous substitutions. Another graph revealed codon substitutions in a set of sequences encoding transcriptional factors of AdpA family (data not shown), master regulators of SM and morphogenesis in *Streptomyces*. Dominance of synonymous single nucleotide substitutions implies conservancy on amino acid level. The presence of other types of substitution indicates the effect of selective pressure on the sequences, as we observed for *adpA* orthologs within family *Micromonosporaceae*.

We were able to identify AdpA homologs across the entire phylogenetic tree of actinobacteria, however their functional meaning remains poorly understood. From initial analysis of a single genus (Streptomyces) it was apparent, that *adpA* coding sequences still experience evolutionary change. If so, by comparing the intensity and types of codon substitutions in phylogenetically different *adpA* groups, we might be able to pinpoint, qualitatively, those that evolve slower

or faster than *adpA* from *Streptomyces*. For this purpose, we have chosen 2 different *adpA* sets from distant actinobacteral orders *Streptomycetales* and *Streptosporangiales*. The corresponding bubble plots are shown on Fig. 6.

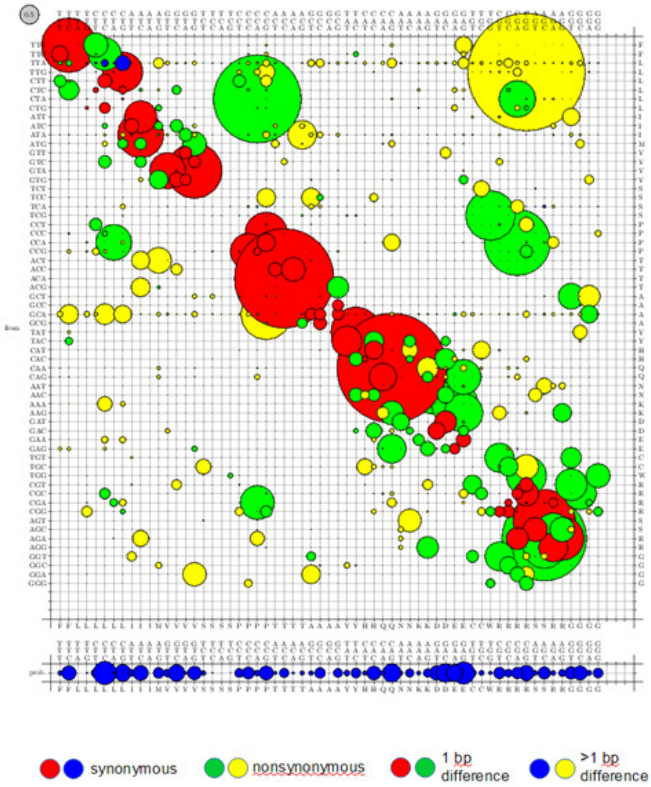

Fig. 5. Bubble plot visualizations for orthologs of highly conserved gene *ssgA* (**A**). Codons are ordered according to biochemical similarity of respective amino acids. Diameters of the bubble represent the probability of substitution one codon (rows) by another (columns). Color code labels synonymous/ nonsynonymous (red and blue/green and yellow) substitutions that differ by one (red and green) or more (blue and yellow) nucleotides. Blue diagram (below the matrix) shows overall codon usage. All plots are equally scaled (grey circle, top left corner of the plot)

It becomes apparent from our analysis that *adpA* orthologs from phylogenetically distinct orders have different evolutionary paths. Particularly, numerous synonymous substitutions (red bubbles) suggest the action of purifying selection on certain set of sequences (as seen in Fig. 6, B). In contrast, *adpA* coding sequences from order *Streptomycetales* feature surprisingly low number of synonymous changes and increased incidence of nonsynonymous substitutions (green and yellow circles, Fig. 6, A). This kind of pattern is consistent with the idea that streptomycete *adpA* genes are subject to positive selection and still exploring their sequence space. Of course, at this stage we cannot rule out the contribution of other as-yet-unknown factors to the codon sequence evolution in case of *Streptomyces*.

To conclude, we revealed several codon context dependencies within *Streptomyces* genomes. They mostly point to nonrandom use of G/C at codon boundaries, although more distant dependencies were also revealed (see Fig. 4). Python-based script was developed after bubble plot concept, which simplifies process of visualizing codon substitutions. We provided several

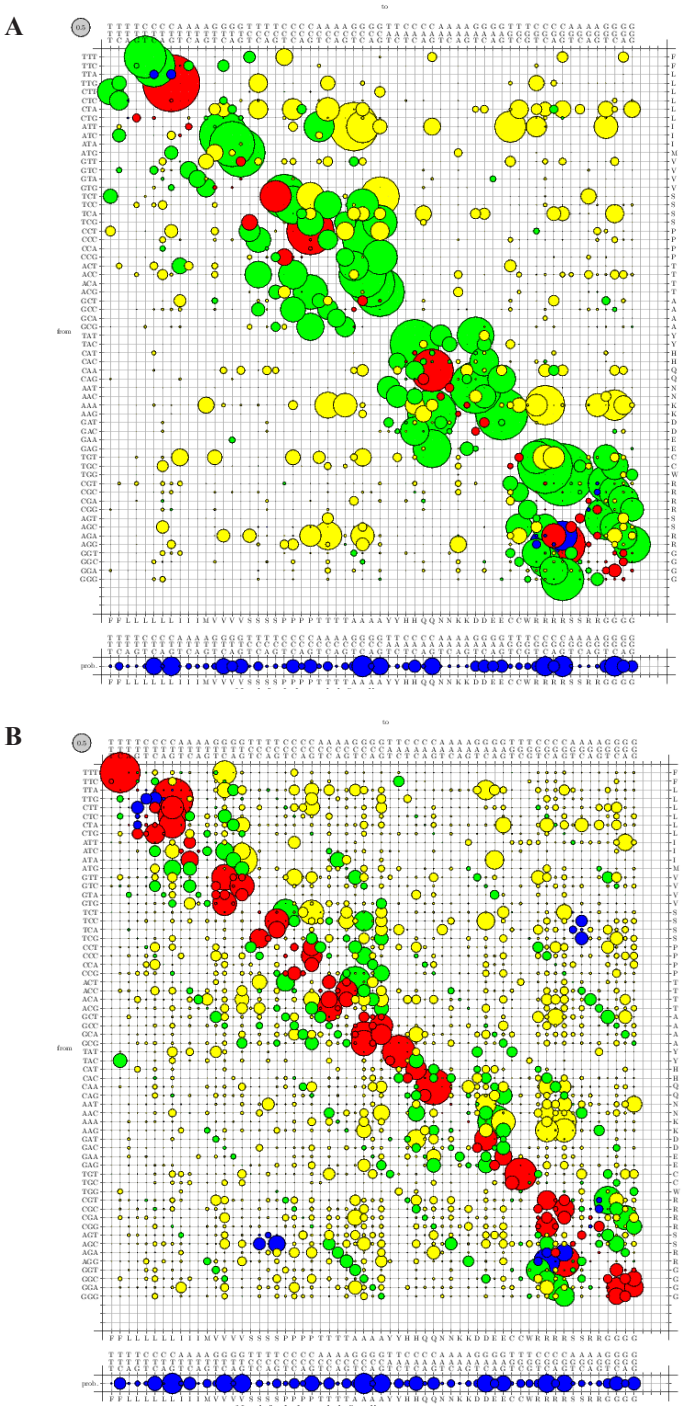

Fig. 6. Bubble plot visualisations for *adpA* orthologs within orders *Streptomycetales* (**A**) and *Streptosporangiales* (**B**). Color code, see Fig. 5

examples showing what kind of information is provided by the plots. The graphical outcome of the program resembles empirical matrices, such as PAM or BLOSUM, used for description of amino acids substitutions. Like the latter, bubble plots lose the positional information about codon changes (namely, it is not known where exactly in a codon sequence any given substitution occurs or predominates). Here we underscore that the aim of bubble plot is to convert all kinds of codon changes seen in multiple sequence alignment into a single and visually comprehensible picture, which reveals global trends (prevalence of synonymous/nonsynonymous changes) and permits to pinpoint the most interesting types of substitutions for further research. We hope that this approach will be useful to study codon sequence evolution and to define what kind of changes are the most acceptable in sets of natural genes. This information may help optimize synthetic coding sequence. Work is in progress to develop online version of bubble plot service.

Acknowledgements. *This work was supported by grant BG-41Nr from Ministry of Education and Science of Ukraine (to B.O.)*

REFERENCES

1. *Alva V., Nam S. Z., Söding J., Lupas A. N.* The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis // Nucleic Acids Res. 2016. Vol. 44. P. W410–415.
2. *Azad R. K., Borodovsky M.* Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory // Brief Bioinform. 2004. Vol. 5. P. 118–130.
3. *Fickett J. W., Tung C. S.* Assessment of protein coding measures // Nucleic Acids Res. 1992. Vol. 25. P. 6441–6450.
4. *Girard G., Traag B.A., Sangal V.* et al. A novel taxonomic marker that discriminates between morphologically complex actinomycetes // Open Biol. 2013. Vol. 3. P. 130073.
5. *Goldman N., Yang Z.* A codon-based model of nucleotide substitution for protein-coding DNA sequences // Mol. Biol. Evol. 1994. Vol. 11. P. 725–736.
6. *Iqbal H. A., Low-Beinart L., Obiajulu J. U., Brady S. F.* Natural product discovery through improved functional metagenomics in *Streptomyces* // J. Am. Chem. Soc. 2016. Vol. 138. P. 9341–9344.
7. *Katz L., Baltz R. H.* Natural product discovery: past, present, and future // J. Ind. Microbiol. Biotechnol. 2016. Vol. 43. P. 155–176.
8. *Kulesa A., Krzywinski M., Blainey P., Altman N.* Sampling distributions and the bootstrap // Nat. Methods. 2015. Vol. 12. P. 477–478.
9. *Kuzniar A., van Ham R.C., Pongor S., Leunissen J. A.* The quest for orthologs: finding the corresponding gene across genomes // Trends Genet. 2008. Vol. 24. P. 539–551.
10. *Moura G., Pinheiro M., Arrais J.* et al. Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure // PLoS One. 2007. Vol. 2. P. e847.
11. *Okonechnikov K., Golosova O., Fursov M., UGENE team.* Unipro UGENE: a unified bioinformatics toolkit // Bioinformatics. 2012. Vol. 28. P. 1166–1167.
12. *Osswald C., Zipf G., Schmidt G.* et al. Modular construction of a functional artificial epothilone polyketide pathway // ACS Synth. Biol. 2014. Vol. 3. P. 759–772.
13. *Plotkin J.B., Kudla G.* Synonymous but not the same: the causes and consequences of codon bias // Nat. Rev. Genet. 2011. Vol. 12. P. 32–42.
14. *Pride D. T., Meinersmann R. J., Wassenaar T. M., Blaser M. J.* Evolutionary implications of microbial genome tetranucleotide frequency biases // Genome Res. 2003. Vol. 3. P. 145–158.

15. *Segata N., Börnigen D., Morgan X. C., Huttenhower C.* PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes // Nat. Commun. 2013. Vol. 4. P. 2304.
16. *Wernersson R., Pedersen A.G.* RevTrans: multiple alignment of coding DNA from aligned amino acid sequences // Nucleic Acids Res. 2003. Vol. 31. P. 3537–3539.
17. *Zarins-Tutt J. S., Barberi T. T., Gao H.* et al. Prospecting for new bacterial metabolites: a glossary of approaches for inducing, activating and upregulating the biosynthesis of bacterial cryptic or silent natural products // Nat. Prod. Rep. 2016. Vol. 33. P. 54–72.

Стаття: надійшла до редакції 27.12.17

доопрацьована 04.04.17

прийнята до друку 05.04.17

ОСОБЛИВОСТІ КОНТЕКСТНОГО ВЖИВАННЯ ТА ЗАМІЩЕННЯ КОДОНІВ У ГЕНОМАХ СТРЕПТОМІЦЕТІВ

І. Рокицький, С. Кулага, Г. Мутенко, М. Рабик, Б. Осташ

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

Актинобактерії роду *Streptomyces* привертають великий інтерес дослідників, оскільки їхні геноми кодують криптичні кластери генів біосинтезу нових антибіотиків. Також гетерологічна експресія метагеномних бібліотек у модельних штамів *Streptomyces* дає змогу виявляти нові класи сполук. Важливо розуміти правила, які визначають вживання кодонів у генах стрептоміцетів, щоби підвищити шанси і рівень експресії чужорідних генів у *Streptomyces*. У цьому дослідженні ми розглянули два питання, пов'язаних із вживанням кодонів у геномах стрептоміцетів. По-перше, ми дослідили закономірності вживання дикодонів у *Streptomyces*. По-друге, ми шукали істотні відмінності в характері заміщенн кодонів у різних сімействах ортологічних генів на різних філогенетичних відстанях і різного ступеня важливості. Як результат, ми виявили кілька правил контекстного вживання кодонів, які в основному пов'язані з аномальною частотою G/C після С-кінцевого нуклеотиду попереднього кодона. Ми розробили новий інструмент біоінформатики на основі описаного раніше методу "бульбашкових" графіків, що дає змогу візуалізувати закономірності кодонних заміщень у наборі даних у вигляді матриці. Використовуючи цей інструмент, ми виявили, що гени транскрипційних факторів родини AdpA у випадку порядку *Streptomycetales* несуть частку несинонімічних замін значно більшу, ніж це спостерігається для груп *adpA* з інших порядків актинобактерій. Характер заміщень кодонів для різних порядків актинобактерій (порівняно зі *Streptomycetales*) не описується простими залежностями.

Ключові слова: Streptomyces, кодонний контекст, патерни заміщення кодонів