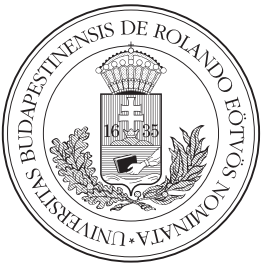
Analysis of tRNA identity with a new, in silico method

Thesis abstract

Szenes Áron



Thesis Supervisors:
Dr. Gábor Pál, Associate Professor and Dr. Éena Jakó, senior scientist

Eötvös Loránd University
Doctorate School of Biology
School Leader: Dr. Anna Erdei, Corresponding member of the Hungarian Academy of Sciences

Structural Biochemistry Doctoral Program
Program Leader: Dr. László Gráf, Member of the Hungarian Academy of Sciences

2012.

Introduction

In all organisms, the 20 aminoacyl-tRNA synthetase (aaRS) enzymes have to recognize their amino acid substrates and the corresponding tRNA molecules with high precision to produce only legitimate aminoacyl-tRNA products. This exquisite specificity is of central importance as this enables the genetic information to be faithfully translated into protein sequences by following the rules defined in the genetic code. Although principles and many fine details of this selective recognition event have already been discovered, several questions remained still unanswered. tRNA positions that have utmost roles in the selective interaction with the cognate aaRS and thus define the identity of the tRNA are denoted as identity elements. While only laboratory experiments can decisively define the identity elements, the large number of potential positions and the laborious nature of the experiments prompted a great variety of bioinformatics studies to predict such elements. These studies require large numbers of individual input tRNA sequences to locate statistically significant identity-related sequence properties. This magnitude of input data became available in the form of genomic DNA sequences, from which tRNA-detecting algorithms can identify functionally relevant tDNA sequences. Such analyses yielded numerous different tDNA databases. Several computational studies reported successful functional annotation and in silico identity element determination. Improved secondary structure-predicting algorithm-driven tRNA alignments yielded high-quality input data sets. These high-quality sets allowed for innovative sequence logo and inverse sequence logo-based analyses of tRNA features and identity element predictions. An information theory-based approach opened new frontiers in visualizing tRNA sequence features and predicting determinants and anti-determinants.

Aminoacyl tRNA synthetases are grouped into Class I and II based on primary and tertiary structure and enzyme properties suggesting two independent phylogenetic lineages. The two distinct classes exist in all three domains of life: Bacteria, Archaea and Eukarya. Analogously, tRNA molecules can also form two respective classes, based on the class membership of their corresponding aaRS. Although some aaRS–tRNA interactions are not extremely specific and require editing mechanisms to avoid misaminoacylation, most aaRS–tRNA interactions are rather specific. Thus, class-specific aaRS features could be mirrored by class-specific tRNA features. However, previous investigations failed to detect conserved class-specific nucleotides.

Aims

- To test whether class specific tRNA sequence features could be detected across the three domains of life by using a discrete mathematical as opposed to traditional statistical approaches.
- If such sequence features can indeed be detected, the second aim would be to demonstrate the same results with a non-discrete method (with an information-theory based "function sequence logo"-s) as well.
- To apply the new discrete mathematical algorithm to an even more complex study in which we assess the potential of each tRNA position to discriminate the 20 different tRNA identity sets from each other.

Methodes

Software's, programming languages and databases

For sequence data manipulation shell scripts were used along with the Emboss program package for many cases.
For more complex analysis steps Perl, Python and Java programs were applied. Statistical analyses were made by the R programming language.
The tRNA and tDNA sequence data were downloaded from the tRNAdb, the tRNAdb-CE, the SPLITSdb and from Christian Marck's tRNomics database.
The sequence logos have been drawn with the Delila package Makelogo program, while the "function logo" with the tRNALogofun program.

The Extended Consensus Partition (ECP)

As a central approach the Extended Consensus Partition (ECP) method, invented by Éena Jakó was used. The algorithm of the method is as follows:

i) Two sets of aligned sequences are provided. The first set denoted as the 'learning' set contains sequences, which represent a certain (I or II) class whereas the second set denoted as the 'mixed' set contains all the sequences from both classes.

ii) The construction of the ECP using the Class I and Class II learning sets

   a) Consider those positions and characters, where all the characters are the same at that position in the given class. These residues form the strictly present set of the ECP.

   b) Collect those positions and characters, where a given character is missing from a position in all the sequences of the class. These residues form the strictly absent set of the ECP.

iii) The selection

   a) For each sequence in the mixed set a sequence is a member of the class defined by the ECP if and only if

      1. all the elements of the strictly present set of the ECP are present; and
      2. all the elements of the strictly absent set are missing from the given sequence.

Method development: the AEV

If an element is strictly absent from a given set while it is allowed to be present in another set then that strictly absent element discriminates the two sets from one another and therefore it is defined as discriminating element (DE).

The author introduced the AEV to determine the weighted average frequency of DE at each position as follows. At each identity set and at each position, we determined how many identity set pairs are discriminated by the given position. These numbers from each set were summed up and were divided by 20 (the number of identities), resulting in the AEVs that demonstrate the discriminating potential of each tRNA position.

Mathematical description of the ECP is described below for the case when all pairs of the 20 tRNA identity sets are compared. The elements of variable $\mathcal{Y}$ are nucleotide bases; therefore, $\mathcal{Y} \in \chi$, where $\chi = \{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\}$. The value of $\mathcal{Y}_{ik}^{j}$ is the nucleotide base corresponding to position j ($j = 1, \ldots, L, L = 96$; from normal position 0 to position 73, including the extra positions from e1 to e22) of the sequence k ($k = 1, \ldots, M_i$, where the value of M_i varies for individual species) of amino acid identity i ($i = 1, \ldots, N, N = 20$). Then, we introduce the set of bases existing at the position j of identity i:

$$\mathcal{Y}_{i}^{j} := \{\mathcal{Y}_{ik}^{j} | k = 1 \ldots, M_i\} \tag{1}$$

DE of identity set i against identity set l (where $l = 1, \ldots, N, N = 20$) at position j is defined as follows:

$$A_{il}^{j} := \left(\chi \setminus \mathcal{Y}_{i}^{j}\right) \cap \mathcal{Y}_{l}^{j} \tag{2}$$

In mathematical terms, the AEV is defined by the following functions:

$$\mathbb{R}\left(A_{il}^{j}\right) := \begin{cases} 1, & \text{if } A_{il}^{j} \neq \varnothing \\ 0, & \text{if } A_{il}^{j} = \varnothing \end{cases} \tag{3}$$

$$n^{j} = \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{i=1 \\ i \neq l}}^{N} \mathbb{R}\left(A_{il}^{j}\right) \tag{4}$$

The n^j value is denoted as the AEV

Results

1. As it was demonstrated by statistical significance tests, the ECP analysis, which focuses on the set-specific strict absence of elements, clearly outperformed the previously used simple statistical analyses for all domains of life. The power of the ECP analysis was proven in finding class-specific tRNA features as opposed to the SCP method.

2. By analyzing the ECP for both Class I and II tDNA sets from 50 (13 archaeal, 30 bacterial and 7 eukaryotic) species, the author and co-workers could demonstrate that class-specific tRNA sequence features do exist, although not in terms of strictly conserved nucleotides as it had previously been anticipated.

    a) There are two universal rules that discriminate Class I from Class II. At position 35, the middle of the anticodon, G is strictly excluded from Class I sequences. At position 73 C is excluded for 47 out of the 50 species.

    b) We could not locate strict discriminating Class II features that would apply for all species. Nevertheless, from position 1–72 usually a C–G, an A–T or both C–G and A–T (M–K) pairs are excluded.

    c) The absence of G from position 35 is a strong characteristic element of Class I.

3. The "inverse function logo's" of Class I and II bacterial sequences reproduce the results of ECP analysis.

4. The positional AEVs measure sequence feature distances of functional groups and correlate with the number of already identified determinants: the ECP algorithm is capable of assessing the level of discriminating power of positions in separating functionally different sequence sets.

5. The algorithm detected positions having higher than average AEVs but containing no published determinants that would have been used in database filtering. Such positions shown below might harbor hitherto unidentified determinants:

    a) *Escherichia coli* $tRNA^{Trp}$ T31:A39

    b) Yeast $tRNA^{Met}$ T31:T39

    c) Yeast $tRNA^{Ile}$ T30:G40

    d) Yeast $tRNA^{Asp}$ G30:T40 (In this case structural evidence is available.)

Notes to the results

1. The analyses were performed on cellular tDNA sequences rather than tRNA sequences therefore the effects of posttranscriptional base modifications could not be studied.

2. Discriminatory elements (DE) could be connected to biological functions or properties other than tRNA identity as well. The class-specific (or identity-specific) absence of certain nucleotide types could be linked to properties such as stability, post-transcriptional processing, ribosome-, or elongation factor binding of the tRNA molecule, just to mention some possibilities that affect the functionality of a tRNA in the complex environment of the cell.

3. Originally we started the analyses by rigorous filtering of the tDNA databases in order to exclude sequences that might not code for functional tRNA. Then we grouped each sequence based on identity and filtered for the presence of already identified and published major, strictly present identity elements characteristic to the given amino acid identity set. As such filtering might bias the results of the analysis we performed the analyses by omitting the database filtration as well. In the case of both Bacteria and Eukarya analyses omitting the filtration still preserved the characteristic pattern of high AEV positions, which overlaps with the positions of the most important identity elements such as the anticodon bases and the discriminator base.

4. Only carefully designed mutations and laboratory experiments can assess the predicting potentials of the ECP algorithm to identify hidden determinants.

Publications related to the thesis

Áron Szenes, Gábor Pál: *Mapping Hidden Potential Identity Elements by Computing the Average Discriminating Power of Individual tRNA Positions.* DNA Res. 2012 Mar 5. doi: 10.1093/dnares/dss008

Éena Jakó, Péter Ittzés, **Áron Szenes**, Ádám Kun, Eörs Szathmáry, Gábor Pál: *In silico detection of tRNA sequence features characteristic to aminoacyl-tRNA synthetase class membership.* Nucleic Acids Res. 2007;35(16):5593-609.