Bach User's Guide

ZJU-China iGEM

October 2010

Contents

1. Introduction

Bach is a gene composer software that optimizes coding sequences and calculates corresponding RiPS to illustrate their translation rates. In sequence optimization, Bach incorporates 4 distinct approaches from which users can freely choose: SYNONYMOUS SUBSTITUTION, SLOW, FAST and MATCH. In the last section of GET RIPS, users are provided with a quantitative view of the translational behavior of the input coding sequence and the four sequences produced by the four methods as well.

Run Bach for the following research needs:

- Heterologous protein expression
- Evaluation of translation rates and behavior
- Design of more predictable and robust genetic circuits

Run Bach for the following benefits:

- Diverse optimization approaches of choice
- Quantitative Analysis of translational behavior
- Free and easy to use

2. System Requirement

Operating System:

- Recommended OS: Windows 7 32-bit Operating System
- Minimum OS: Windows XP 32-bit Operating System

Hardware Configuration:

- No strict limits.
- Recommended configuration:

  RAM ≥ 1GB

  Hard Disk ≥ 100GB

  CPU ≥ 1.66GHZ

3. Algorithm

3.1 Synonymous Substitution

Synonymous Substitution is more of a paraphrase than optimization, the goal of which is to retain and transfer the former performance of translation from the source species into the host species. The codon rank sequence is unchanged, implying the relative availability of aa-tRNA to bind to each codon site along the sequence chain is unchanged.

The input sequence is dealt with the following steps:

- The codon usage table of both the source and host species and the input sequence are loaded.
- For each amino acid, the corresponding n ($1 \leqslant n \leqslant 6$) codons are given rank from 1-n according to their relative amount in percentage.
- Each codon in the input sequence is assigned with the rank in the source species, thus producing a new sequence in terms of rank.
- Each codon in the input sequence is then substituted by codon of the same rank in the host species.

Example:

Consider a short coding sequence of AUUAUCAUA, the amino acid sequence of which is III. A is the source species, while B is the host species. The relative amount (in percentage) table of each codon for two species is as follows:

Codon	Amino Acid	Amount in A	Rank in A	Amount in B	Rank in B
AUU	I	0.47	1	0.21	3
AUC	I	0.32	2	0.47	1
AUA	I	0.21	3	0.32	2

Therefore, AUU which ranked number 1 in A should be substituted with the codon ranked number 1 in B, which is AUC. The new sequence produced by synonymous substitution then is AUCAUAAUU.

3.2 Slow

The output sequence of this phase, denoted by the phase name ‘slow’, is theoretically the slowest sequence one can obtain without altering the sequence of amino acid. As the codons adopted by the slow sequence correspond to a scarce source/concentration of its aa-tRNA, it’s very much likely that the ribosome has to wait longer for the aa-tRNA to come and attach.

The input sequence is transformed with the following procedures:

- The codon usage table of the host species and the input sequence are loaded.
- For codons encoding the same amino acid, the codon with the smallest number/percentage is designated as the slowest codon for that amino acid.
- Each codon in the input sequence is substituted with the slowest codon of the same amino acid.

Example:

Continue to consider the example in Synonymous Substitution. Among the three codons encoding amino acid I, AUU is the slowest codon in species B. Thus all the codons in the sequence encoding amino acid I should be replaced by AUU in this approach, producing the slowest sequence: AUUAUUAUU.

3.3 Fast

Exactly the opposite of SLOW, the FAST section produces the fastest sequence of the same protein one can possibly get. Such coding sequence is usually translated fastest, since all the codons used in the sequence correspond to richest sources of aa-tRNAs, thus speeding up the step of aa-tRNA attachment.

The steps for such optimization is similar to the one in SLOW:

- The codon usage table of the host species and the input sequence are loaded.
- For codons encoding the same amino acid, the codon with the largest number/percentage is designated as the fastest codon for that amino acid.
- Each codon in the input sequence is substituted with the fastest codon of the same amino acid.

Example:

Still take use of the example in the previous two sections. In this method, all codons encoding amino acid I should be replaced by the fastest codon in species B, which is AUC, thus resulting in the output sequence: AUCAUCAUC.

3.4 Match

This section aims to produce the sequence that matches the codon bias of the host species the best without largely altering the input sequence. The product of MATCH serves beyond the scope of elongation process and may be subjected to less failure when expressed in the host organism.

Computation has been made according to the following procedures:

- The codon usage table of the host species and the input sequence are loaded.
- The input sequence is converted into a new sequence of amino acid.
- According to the protein sequence and the codon bias of the host species, the number of each codon is calculated by multiplying the number of its corresponding amino acid in the sequence and the percentage of the codon of each amino acid in the host species.
- The difference between the assumed number of each codon obtained from the previous step and the original number of each codon in the input sequence is calculated.
- The replace of codon in the input sequence then proceeds from 5’-3’ according the difference obtained above, so that the output sequence has the proportion of each codon of the same amino acid approximately the same with that of the host species.

Example:

Let’s now consider a longer sequence AUUAUUAUUAUUAUUAUUAUCAUAAUAAUAGGG, in which

there're 10 amino acid I. The relative amount of each codon in species B is still the same as the previous examples. To match the codon bias in species B, we can first calculate the assumed amount of each codon encoding amino acid I in the sequence. The amount of AUU = 10*0.21≈2; the amount of AUC = 10*0.47≈5; the amount of AUA=10*0.32≈3. The difference of each codon can then be calculated as -4 for AUU, 4 for AUC and 0 for AUA. The output sequence is AUUAUUAUCAUCAUCAUCAUCAUAAUAAUAGGG, in which the number of each codon encoding amino acid I agrees with the assumed value calculated from the codon bias of the host.

3.5 Get RiPS

One of the brilliance of Bach lies in GET RIPS, in which RiPS (Ribosome per sec) of each sequence can be calculated. RiPS is a newly emerged term in synthetic biology to denote translation rates. The computation is based on the mathematical model built by ZJU-CHINA 2010 iGEM team.

Please refer to http://2010.igem.org/Team:ZJU-China/Modeling for a detailed description of the underlying model and algorithm.

4. Using Bach

4.1 Getting Started

To start running Bach, double click the icon of Bach. The window for input will appear.

4.2 Data Prep

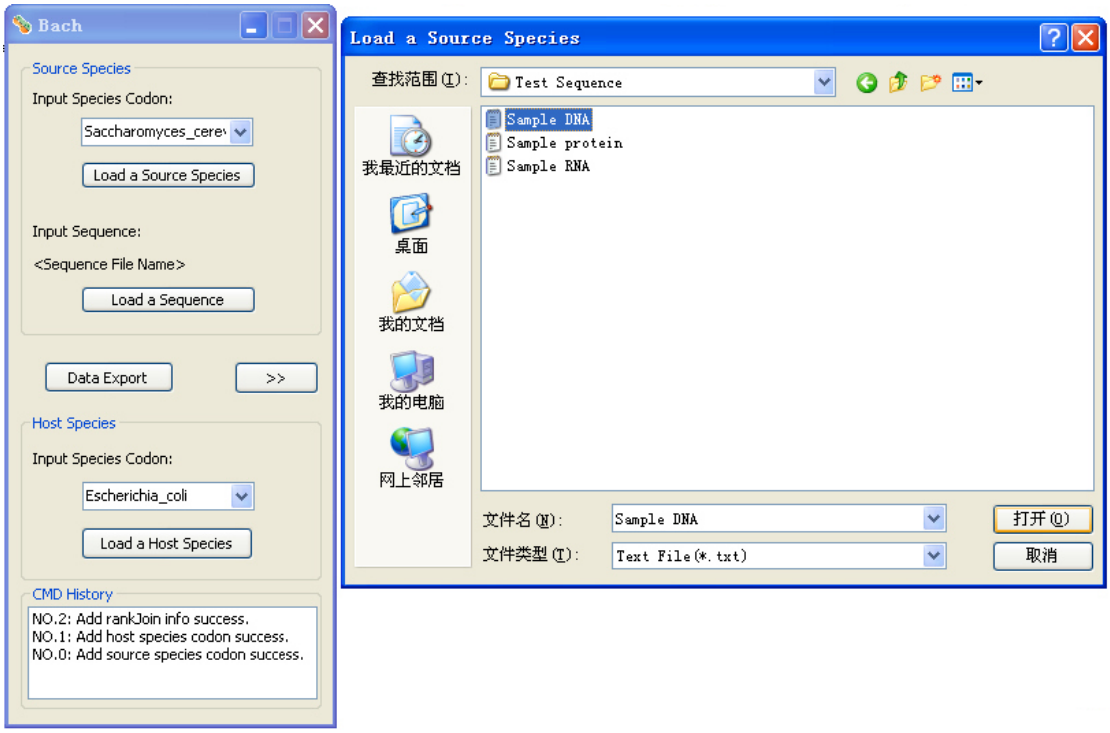
In the software package downloaded, there's a file 'Species Data' in which data of more than 1200 species is stored. The codon usage table of each species is stored in each distinct text file in 'Species Data', named by the species Latin name. Click Load a Source Species button and load the source species (the original host of the gene or the species where the gene is first discovered) from the 'Species Data' file in the software package. Then select it from the list above the Load a Source Species button.

Choosing the host species is similar. Click Load a Host Species button and load the host species (the host you want the gene to be expressed) from the 'Species Data' file in the software package. Then select it from the list above the Load a Host Species button.

Note that for both the source species and host species you've loaded, the data and history will be kept, which means you don't have to load the same species again in the future. If you want to clean the history of loading, you can manage to do so by deleting the database file in your software package. Then, you'll have to load each species data again.

If you either are unsatisfied or simply can't find the species you desire in the 'Species Data' we've

provided, you can make your own data file and load it into the software, as long as the form of the data file is exactly the same as those we provide. More information on how to make eligible data file can be found in the Appendix.



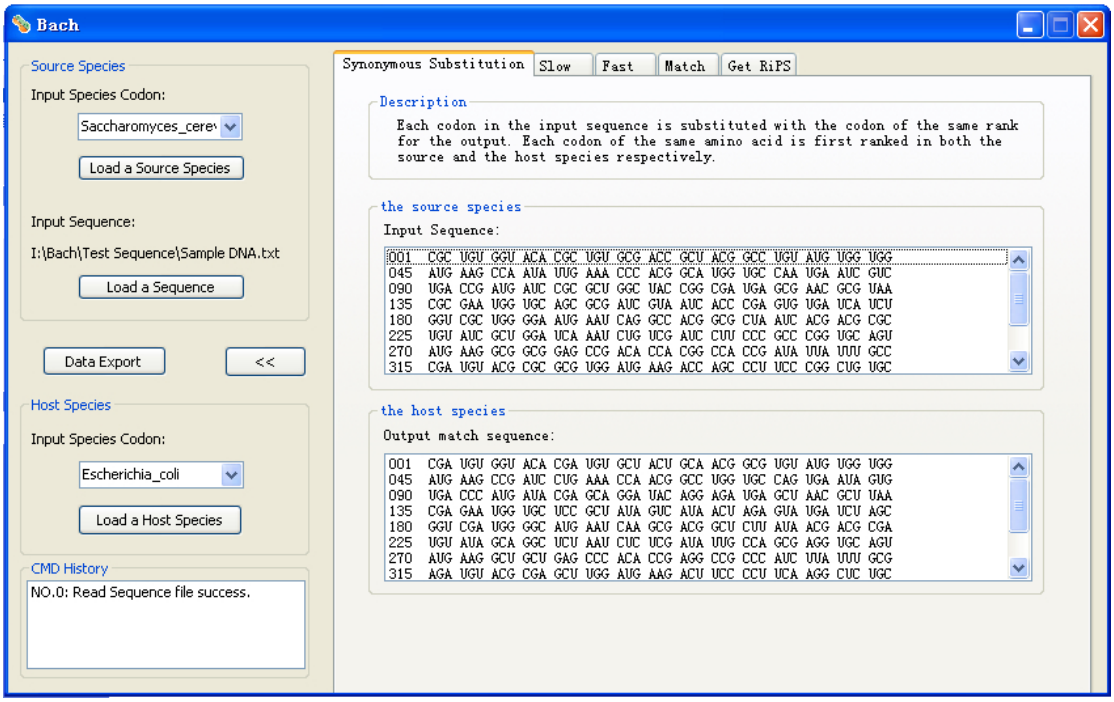
Click Load a Sequence button and load the sequence file for optimization. In the software package, there's a 'Test Sequence' file for tryout if you don't have any gene sequence at hand. If you do have a sequence, please make your sequence file eligible for Bach to recognize:

- State the nature of the input sequence (DNA, RNA or protein) in the first line.
- Paste your sequence in the second line. The sequence can be acknowledged and extracted by Bach as long as there're no characters other than numbers and English letters.
- As Bach can't distinguish between coding sequences and non-coding sequences, please make sure the sequence you've submitted is from the coding region, otherwise Bach wouldn't be of much help in the optimization and calculation of RiPS.

4.3 Optimization

Click >> button and the sequence optimization results are displayed on the right. You can switch to view sequences optimized by different methods by clicking on the different tags in the upper window. If you don't have any idea what these sequences stand for or how these sequences are obtained, you can refer to the previous section of this manual on Algorithm.

In each window, there's a description to better guide you through each window and find what you need. The box in the middle displays the input sequence you've loaded. The optimized sequence is displayed in the box underneath. The sequence is displayed in the unit of codon to better illustrate the change of codon and the number on the left helps you to better track the length of the sequence.



4.4 Calculation of RiPS

To get RiPS of each sequence, click the tag Get RiPS. For a more precise calculation of RiPS, you need to first change the values of parameters.

The parameters on the left in the RiPS Parameters box relate to the Monte Carlo simulation itself and are species independent:

- Walpha: The parameter denoting the initiation rate of translation, the value of which may be influenced by the RBS sequence and other nonsense sequences in both the 5' and 3' tails of the mRNA.
- T: Time for Monte Carlo simulation in terms of 0.001s. The more times Monte Carlo run, the more accurate value of RiPS you'll get. Values more than 100 for T is highly recommended, otherwise the calculation results will display a much more random fashion.
- Length: The length of ribosome occupation in terms of codon. The value of 11 is recommended according to published researches. The range of the parameter is set as $1 \leqslant$ Length $\leqslant$ Total Length.
- Look Index: The site along the sequence where RiPS is calculated. Look Index of 20 indicates that RiPS is calculated by counting the number of ribosomes passing the 20^{th} codon (5'->3') per second. As the first 0-20 codons constitute a zone which performs quite differently from the latter part of the sequence, Look Index of values from 20-n is highly recommended, where n is the total length of the sequence in terms of codon displayed in the box on the right.

The parameters on the right in the Advanced Parameters box may be species dependent. The default values have been obtained from experiments for Escherichia coli. If you are setting E.coli as the host species or want to calculate the RiPS of each sequence when expressed in E.coli, no

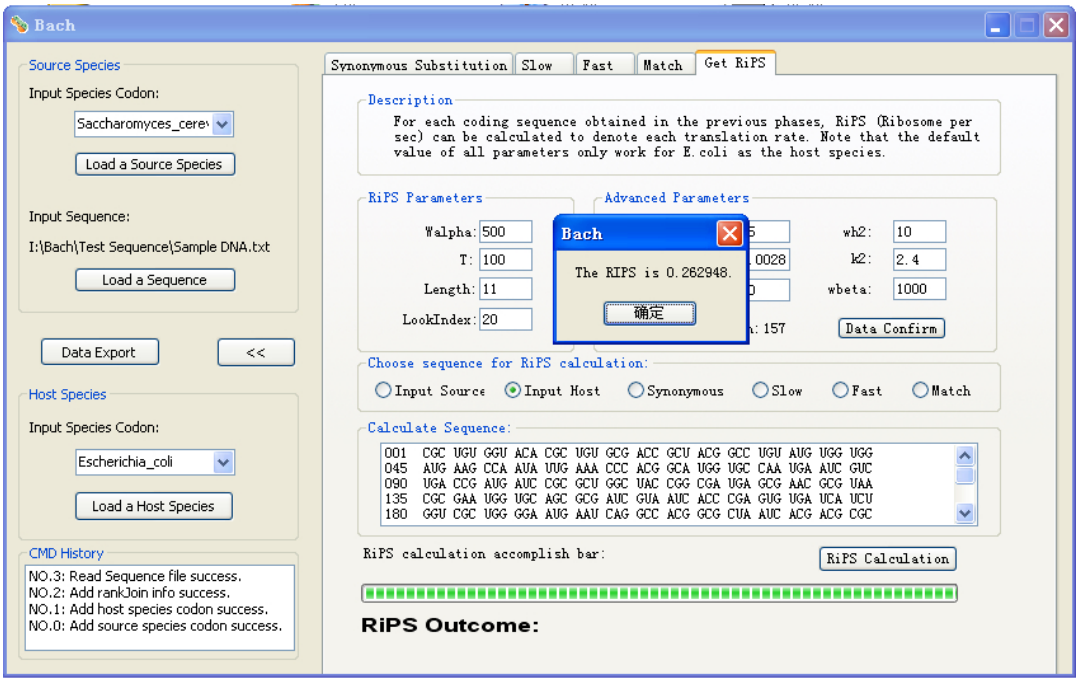
alteration has to be made about the default parameters. If not, there may be a need to reevaluate these parameters before calculating RiPS. Furthermore, if you are not satisfied with the default value of parameters in either case, you can always change it based on your own research or other more reliable sources you'd like.

Advanced parameters are as follows:

- Wg: Rate constant for transition from 4-5, in which tRNA is shifted and site A is occupied by EF-G in the GTP bound form.
- Wp: Rate constant for transition from 2-1, when falsely bound aa-tRNA-EF-Tu complex dissociates from site A.
- Wh1: Rate constant for transition from 2-3, denoting the probability of the GTP part of EF-Tu hydrolized to GDP.
- Wh2: Rate constant for transition from 5-6, representing the hydrolysis of GTP to GDP and released. As transition from 6-7-1 can be negligible compared with transition from 5-6, wh2 is used in the model to indicate the transition rate from 5-1.
- K2: Rate constant for transition from 3-4, when the phosphate group, a product of the hydrolysis, leaves and releases the EF-Tu.
- Wbeta: The dissociation rate of ribosome from mRNA at the stop codon.
- Codon total length: The length of the selected sequence in terms of codon. The value of total length casts limits on values of other parameters such as Length and Look Index, thus helping users to set reasonable values for other parameters.

Click Data Confirm button to confirm the values of all the parameters you want to use for the calculation of RiPS. Then choose the sequence you wish to compute RiPS. Upon clicking on each sequence tag, the corresponding sequence will be instantly displayed in the box beneath. Click RiPS Calculation to finally compute.

A green bar of progress will appear at the bottom of the window to show how long you have to wait for the result. When the bar is completed, a small box will pop out, on which the value of RiPS is shown. Upon closing the small box, the result will also be shown under the bar.



4.5 Data Output

By clicking the Data Export button on the left, the optimization results (containing newly composed sequences from different methods) can be exported into a text file.

5. Copyright

For further queries related to Bach program, please contact us through the email of wendyhu001@gmail.com. Any reports of failure or suggestions on Bach are welcome.

6. Appendix

Species Data Compilation

If you can't find the target species in the Species Data file, you can easily make your own data file by the following procedures:

- Go to Codon Usage Database: http://www.kazusa.or.jp/codon/.
- Search for or choose the codon usage table of your target species. Here we use Saccharomyces cerevisiae as an example.
- Choose '1. Standard' from the Format list under the table.
- Click the Submit button underneath.
- The table is changed into a new format, in which there're four columns of information for each codon. Copy the table into a new text file.
- The text file format should be strictly the same with the example figure below. The number in the first line should be bigger than 1232 and should be distinct for each species. The blank space ' ' in the Latin name of the species in the second line should be replaced with '_'.