BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 20 2012, pages 2683–2684
doi:10.1093/bioinformatics/bts465

Sequence analysis

Advance Access publication July 30, 2012

EuGene: maximizing synthetic gene design for heterologous expression

Paulo Gaspar[1,*], José Luís Oliveira[1], Jörg Frommlet[2], Manuel A.S. Santos[2] and Gabriela Moura[2,*]

[1]DETI/IEETA, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal and [2]Department of Biology and CESAM, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Numerous software applications exist to deal with synthetic gene design, granting the field of heterologous expression a significant support. However, their dispersion requires the access to different tools and online services in order to complete one single project. Analyzing codon usage, calculating codon adaptation index (CAI), aligning orthologs and optimizing genes are just a few examples. A software application, EuGene, was developed for the optimization of multiple gene synthetic design algorithms. In a seamless automatic form, EuGene calculates or retrieves genome data on codon usage (relative synonymous codon usage and CAI), codon context (CPS and codon pair bias), GC content, hidden stop codons, repetitions, deleterious sites, protein primary, secondary and tertiary structures, gene orthologs, species housekeeping genes, performs alignments and identifies genes and genomes. The main function of EuGene is analyzing and redesigning gene sequences using multi-objective optimization techniques that maximize the coding features of the resulting sequence.

Availability: EuGene is freely available for non-commercial use, at http://bioinformatics.ua.pt/eugene

Contact: paulogaspar@ua.pt

Received and revised on July 10, 2012; accepted on July 18, 2012

1 INTRODUCTION

Molecular biology has witnessed an increased use of computer science tools, especially after the first genome sequencing projects. Assessing the effect and meaning of the genetic information for cells, discovering the information underlying genes and describing protein functionality are some of the most important tasks in this field. Often these require specialized computing tools.

Moreover, the developments in the field of gene decoding and protein synthesis already show a good knowledge of the factors that are involved in it. Those factors are mostly related to codon usage (Angov *et al.*, 2008), codon context (Moura *et al.*, 2007), GC content, hidden stop-codons (Seligmann and Pollock, 2004), repetitions (Brégeon *et al.*, 2001), deleterious sequences (Jin *et al.*, 2006), messenger RNA (mRNA) secondary structure (Kozak, 1986) and charged transfer RNA availability (Welch *et al.*, 2009). Controlling such factors has become imperative in the field of synthetic gene design. Therefore, numerous tools have been developed to optimize gene sequences and modulate specific characteristics that are believed to influence protein expression efficiency, such as Optimizer (Puigbò *et al.*, 2007), Synthetic Gene Designer (Wu *et al.*, 2006) and Gene Composer (Lorimer *et al.*, 2009).

Most of them offer several forms of codon usage optimization, restriction sites management and removal of codon and nucleotide repetitions. However, selection or input of additional information is usually required, such as codon usage and context tables or orthologs to perform alignments. Nonetheless, most of the required information is available online in diverse biological databases such as NCBI, EBI, PDB and KEGG, but few systems take advantage of those.

Several other tasks are involved in synthetic gene design efficiency, such as calculating protein secondary structure and visualizing the tertiary structure to aid in mapping protein motifs to gene zones or gathering and aligning orthologs to find similarities and conservation regions. Nevertheless, despite the amount and variety of available gene information and synthetic redesign approaches, integration of those factors into a single optimization tool is still very limited. Thus, in order to advance and facilitate synthetic gene design and analysis, we developed a tool that integrates several online services, offline tools and synthetic gene redesign techniques into a single application.

2 FEATURES

EuGene capabilities can be separated into two blocks, data gathering and gene optimization. In the first block, a set of features allows the retrieval of information about a gene from several known online sources, offline tools and statistical calculations on its own genome. In the second block, EuGene allows redesigning a gene according to several factors that influence the mRNA decoding process, using information obtained in the data gathering procedure.

2.1 Data gathering

A large amount of information about genes and genomes can be considered for synthetic design. For instance, it is common to evaluate the protein's secondary and tertiary structures and try to map how the gene's codon sequence influences the formation of these structures. Unraveling which codons are responsible for the correct folding of a polypeptide might be essential when redesigning genes, in order to ensure the solubility of the final protein. Another clue to identifying important regions of a gene is given by its level of conservation among orthologs. Other forms of information are also important, such as having a set of highly expressed genes in order to calculate codon adaptation index (CAI) values.

Before retrieving related information, EuGene identifies genes using two strategies. The first is processing gene annotations that

*To whom correspondence should be addressed.

are normally present in FASTA and GenBank formats, extracting any database identifiers. Those identifiers are then used to access NCBI and obtain the names of the gene and genome and resulting protein. The second strategy uses the gene codon sequence to perform an online BLAST in NCBI (Johnson *et al.*, 2008). If an identical gene is found, the same information is retrieved (gene, protein and genome names), plus a set of orthologs resulting from BLAST.

Further information regarding the gene is then obtained by contacting PDB and KEGG online servers to automatically download more orthologs, protein tertiary structures and highly expressed genes for the same genome. Orthologs are seamlessly aligned with the gene using MUSCLE (Edgar, 2004), and colored according to codon conservation, allowing a visible assessment of conserved and functional regions. The protein 3D structure is then presented, and is visually mapped to the codon sequence, easing the understanding of which codons decode to which regions of the protein. The highly expressed genes obtained permit the automatic calculation of CAI for any gene in the same genome. Moreover, the protein secondary structure is also seamlessly calculated and presented using PsiPred (McGuffin *et al.*, 2000). Other calculations to enable assessment of relative synonymous codon usage (RSCU) and codon pair bias (CPB for codon context) are automatic upon opening a genome file.

All retrieved and calculated information are always displayed. That includes, for each gene, its CAI, G+C%, N_c effective codons, mean RSCU, CPB, protein primary, secondary and tertiary structures, orthologs and corresponding names (Fig. 1).

2.2 Synthetic gene redesign

The main functionality of EuGene is optimizing gene codon sequences according to specific aspects. A set of six redesign approaches is available to customize the gene: codon usage (RSCU and CAI); codon context (CPB); G+C content level; controlling hidden stop codons; the elimination of repetitions and removal of deleterious sites (such as Shine–Dalgarno). All modifications are performed without changing the resulting native amino acid sequence. Moreover, changes are controlled by the genetic code and codon usage/context tables of the target host species and, therefore, if the host species is non-native, the protein primary structure is always maintained and all redesign considerations are made using the statistical information of the host. This allows, for instance, harmonizing the codons usage of a gene to express in a heterologous host. The target host species can be any whose genome data (FASTA or GenBank file) was uploaded into the application.

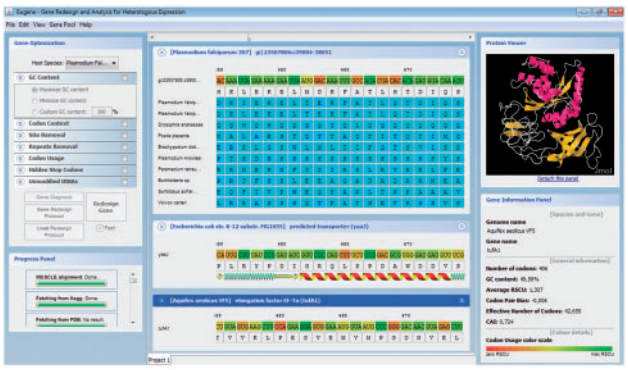

Fig. 1. EuGene workspace. In the center, several genes are being studied, and aligned orthologs are shown. On the right, information about the selected gene is shown, such as the protein tertiary structure

EuGene allows simultaneous optimization of any redesign approaches by using two multi-objective optimization techniques: a genetic algorithm and simulated annealing. Both strategies aim at finding the overall best gene candidate for the redesign parameters. However, simulated annealing is much quicker in the search for the optimal codon configuration, and is, therefore, ideal for experiments that need fast results. However, in the gene multi-goal design, there is rarely one single optimal solution due to the degeneracy of the genetic code. Thus, using a genetic algorithm in combination with a Pareto front archive (Knowles and Corne, 2000), it is possible to return a list of the best equivalent solutions to the optimization requirements, allowing selection of a solution offering the best trade-off between the selected redesign methods.

3 IMPLEMENTATION

EuGene is built in Java, using web-start technology, which enables automatic updates in order to ensure the application has the latest features and corrections. Moreover, EuGene has a modular architecture where each gene redesign procedure is independent, allowing the facilitated construction of new features. This is especially suitable to the biological field, given its constant evolution.

Funding: Partially supported by the European projects MEPHITIS and GEN2PHEN. P.G. is funded by Fundação para a Ciência e Tecnologia (FCT, SFRH/BD/71063/2010).

Conflict of Interest: none declared.

REFERENCES

Angov,E. *et al.* (2008) Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. *PLoS One*, **3**, e2189.

Brégeon,D. *et al.* (2001) Translational misreading: a tRNA modification counteracts a+ 2 ribosomal frameshift. *Genes Develop.*, **15**, 2295.

Edgar,R. (2004) Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Jin,H. *et al.* (2006) Influences on gene expression *in vivo* by a Shine–Dalgarno sequence. *Mol. Microbiol.*, **60**, 480–492.

Johnson,M. *et al.* (2008) NCBI blast: a better web interface. *Nucleic Acids Res.*, **36** (Suppl. 2), W5.

Knowles,J. and Corne,D. (2000) Approximating the nondominated front using the pareto archived evolution strategy. *Evol. Comput.*, **8**, 149–172.

Kozak,M. (1986) Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. *Proc. Natl. Acad. Sci. USA.*, **83**, 2850.

Lorimer,D. *et al.* (2009) Gene composer: database software for protein construct design, codon engineering, and gene synthesis. *BMC Biotechnol.*, **9**, 36.

McGuffin,L. *et al.* (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404.

Moura,G. *et al.* (2007) Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure. *PLoS One*, **2**, e847.

Puigbò,P. *et al.* (2007) Optimizer: a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Res.*, **35** (Suppl. 2), W126.

Seligmann,H. and Pollock,D. (2004) The ambush hypothesis: hidden stop codons prevent off-frame gene reading. *DNA Cell Biol.*, **23**, 701–705.

Welch,M. *et al.* (2009) Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS One*, **4**, e7002.

Wu,G. *et al.* (2006) The synthetic gene designer: a flexible web platform to explore sequence manipulation for heterologous expression. *Protein Expr. Purif.*, **47**, 441–445.