Bioinformatics, 31(22), 2015, 3688–3690
doi: 10.1093/bioinformatics/btv420
Advance Access Publication Date: 20 July 2015
Applications Note

OXFORD

Sequence analysis

CEMAsuite: open source degenerate PCR primer design

Courtney E. Lane*, Daniel Hulgan, Kelly O'Quinn and Michael G. Benton*

Cain Department of Chemical Engineering, Louisiana State University, Baton Rouge, LA 70803, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 9, 2015; revised on June 24, 2015; accepted on July 14, 2015

Abstract

Summary: The codon-equivalent multiple alignment suite begins conservational analysis for polymerase chain reaction primer design at the protein level, allowing the user to design consensus primers capable of detecting homologous coding sequences even when low-to-moderate sequence information is available. This package also condenses the wealth of information associated with multiple sequence alignments and presents them in an intuitive manner, allowing the user to quickly and effectively address degenerate primer design considerations.
Availability and implementation: https://sourceforge.net/projects/cemasuite/.
Contact: benton@lsu.edu or cemasuite@gmail.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In many polymerase chain reaction (PCR) applications, it is desirable to design PCR primers, which can amplify across similar targeted DNA fragments. Some examples include genetic screening to detect homologous genes across multiple organisms and working with highly variable sequences, such as in viral strains. Primers can be designed within a conserved region of the target sequences in such a way that the annealing properties of the primers remain relatively unchanged across multiple templates. Minor variations in the sequences can be accounted for by introducing degenerate positions into the primers. However, increasing the degeneracy of a primer also increases the likelihood of stable sites outside of the target region, leading to an increased risk of byproduct formation. Byproduct formation is undesirable in PCR as it can lead to false-type results (by out-competing and/or having similar product sizes as the amplicon) and/or additional purification steps for downstream cloning applications.

In this work, we present CEMAsuite (https://sourceforge.net/projects/cemasuite/), an open source Java-based application with the ability to (i) construct a codon-equivalent multiple alignment (CEMA) from a protein multiple sequence alignment (MSA) file, (ii) generate and score each consensus DNA sequence position using multiple algorithms, (iii) design single-degeneracy primer backbones using Primer3 (Untergasser *et al.*, 2012) and (iv) estimate the stability of degenerate primers on each of the coding sequences (CDSs) within the CEMA. The goal of CEMAsuite is to aid in the design of degenerate primer sets which are robust enough for the assay at hand yet retain as much specificity and product homology as possible.

2 Methods

Under the equiprobability assumption, the probability that all residues of an arbitrary column in an arbitrary MSA will match the first residue in that column due to random chance alone can be described by the equation: $P=(R+1)^{(1-n)}$, where R is the number of observed residue states (+1 for deletion) and n is the number of sequences in the MSA. This type of conservation can be considered a sort of Type I error where the conservation is present, but due to random chance alone, and not necessarily from evolutionary pressures. This conservation is undesirable in probe design since it has no influencing factors to remain conserved, meaning the use of primers based on these regions on a homolog with unknown sequence may result in a loss of primer annealing ability. For this reason,

CEMAsuite begins the conservational analysis for primer design on the protein level, where $R = 20$, as opposed to coding DNA, where $R = 4$. A CEMA is generated by obtaining the primary CDSs of proteins in an extant MSA and expanding each position within the protein MSA to the observed codon representing each amino acid.

Since many primer design applications can account for positional quality in their objective functions, each position within a consensus sequence for a CEMA is scored by one of four algorithms. These algorithms are detailed in the user manual. Briefly, the *Percent Identity* algorithm scores each position based on the normalized frequency of the consensus nucleotide. *Identity Runs* scores positions on identity, then adjusts the value based on the number of consecutive completely conserved positions in the locale. *Potential Degeneracy* scores positions on identity and then adjusts the value based on the potential degeneracy of the consensus codon positions according to one of 18 translation tables. The final algorithm, *Runs & Degeneracy*, scores values sequentially using logic from each of the three methods described above.

The location of a primer within a CEMA is considered to be the region with the least total number of mismatches throughout all sequences. For cases where there are equal numbers of mismatches, the algorithm returns the index with mismatches occurring mostly on the 5'-end. Simulated hybridization of the primers to each template is performed by iterating through all possible permutations of a degenerate primer and estimating its most stable conformation when bound to the respective location of the template. Each simulated hybridization event returns a Gibbs free energy value estimated using nearest neighbor thermodynamics which is corrected using one of two entropic correction formulas (detailed in user manual).

3 Usage

A Java Swing-based interface is deployed containing tabs that allow the user to visualize the step-wise primer design process. Before using CEMAsuite, the user must first possess a MSA file of the protein-of-interest. This file can be imported and will be displayed under the *Protein MSA* tab.

Next, the user may choose to either upload the CDSs from a file or attempt to retrieve them from the National Center for Biotechnological Information database via the Entrez Efetch utility (Geer *et al.*, 2010). Once the sequences have been successfully uploaded, they will be displayed under the CDS tab.

Next, the CEMA can be generated and displayed under the CEMA tab (Fig. 1, left). The user may select one of the four consensus scoring methods described in Section 2. The strengths and weaknesses of each algorithm are described in further detail in the user manual. At this point, the scored consensus sequence will be visible as a bar plot underneath the tabs (Fig. 1, left).

Once a scored consensus sequence has been obtained, CEMAsuite accesses a compiled Primer3 executable, which can return potential primer sets for the consensus sequence displayed under the Primer Design tab. Scored consensus information can also be readily exported for input into alternate design applications if desired.

After the primers have been designed, each primer-template hybridization combination can be calculated and displayed under the Hybridization tab (Fig. 1, right). CEMAsuite will highlight primer-template pairs, which are unlikely to successfully amplify as shown in Figure 1. Results from 94 PCR experiments were collected from literature and subjected to the hybridization algorithm to obtain the default warning threshold values; however, these values can be adjusted readily. Selective degeneracy can be added by the user to increase the stability of the primer-template pairs (Fig. 1, right), which can be viewed by reiterating the hybridization algorithm.

4 Conclusion

Because CEMAsuite begins conservational analysis at the protein level, the primers obtained through this application will have a high likelihood of successful amplification even when the template

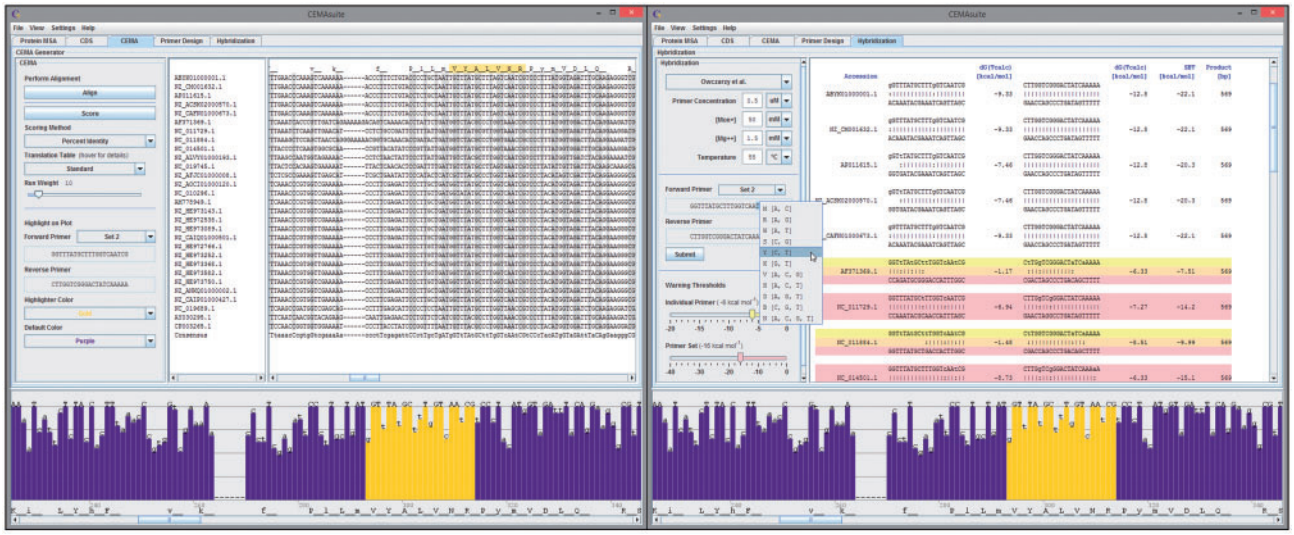

Fig. 1. *Screenshots of the CEMAsuite application. (Left)* CEMA *tab, CEMA scored using the* Percent Identity *algorithm shown in the bar plot. The forward primer region is highlighted (gold) in the plot and the first line of the CEMA alignment text. (Right)* Hybridization *tab with each calculated Gibbs free energy value at the specified conditions for each primer-template pair. Primer-template pairs that the user would like to flag as "likely to fail amplification" are highlighted if they possess an individual primer less stable than the user input* Individual Primer Warning *threshold (yellow) and/or the sum of the stability values of both primers is less stable than the* Primer Set Warning *threshold (red). Selective addition of degeneracy via the popup menu is shown in the forward primer input region* (Color version of this figure is available at *Bioinformatics* online.)

sequence information is unknown (e.g. PCR amplification of a novel environmental isolate). Additionally, because actual CDS information is used and degeneracy can be added as needed, users are free to balance the specificity and sensitivity of a PCR primer set.

It is important to note that even though CEMAsuite incorporates Primer3 functionality, the scoring algorithms and output graphics significantly speed up the process of primer design by inspection for those with moderate primer design experience.

We expect CEMAsuite to be a valuable assay design tool for applications such as genetic/environmental screening, working with highly variable sequences and DNA quality control via PCR. An active open source community dedicated to improving the degenerate PCR primer design process is the goal of this project, and the participation of those whom are interested is encouraged. This application is released under public licensing (GNU GPL v3).

Acknowledgement

We would like to thank Robert Willis of the Louisiana State University Department of Chemical Engineering for his consultation.

Funding

This work was supported by National Science Foundation through CBET 1032599.

Conflict of Interest: none declared.

References

Geer,L.Y. *et al.* (2010) The NCBI BioSystems database. *Nucleic Acids Res.*, **38**, D492–D496.

Untergasser,A. *et al.* (2012) Primer3—new capabilities and interfaces. *Nucleic Acids Res.*, **40**, e115.