Bioorganic & Medicinal Chemistry 25 (2017) 4946–4952

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Aligator: A computational tool for optimizing total chemical synthesis of large proteins ☆

Michael T. Jacobsen [a,b], Patrick W. Erickson [a,b], Michael S. Kay [a,*]

[a] Department of Biochemistry, University of Utah School of Medicine, 15 North Medical Drive East, Room 4100, Salt Lake City, UT 84112-5650, United States

ARTICLE INFO

Article history:
Received 7 April 2017
Revised 24 May 2017
Accepted 30 May 2017
Available online 3 June 2017

ABSTRACT

The scope of chemical protein synthesis (CPS) continues to expand, driven primarily by advances in chemical ligation tools (e.g., reversible solubilizing groups and novel ligation chemistries). However, the design of an optimal synthesis route can be an arduous and fickle task due to the large number of theoretically possible, and in many cases problematic, synthetic strategies. In this perspective, we highlight recent CPS tool advances and then introduce a new and easy-to-use program, Aligator (Automated Ligator), for analyzing and designing the most efficient strategies for constructing large targets using CPS. As a model set, we selected the *E. coli* ribosomal proteins and associated factors for computational analysis. Aligator systematically scores and ranks all feasible synthetic strategies for a particular CPS target. The Aligator script methodically evaluates potential peptide segments for a target using a scoring function that includes solubility, ligation site quality, segment lengths, and number of ligations to provide a ranked list of potential synthetic strategies. We demonstrate the utility of Aligator by analyzing three recent CPS projects from our lab: TNFα (157 aa), GroES (97 aa), and DapA (312 aa). As the limits of CPS are extended, we expect that computational tools will play an increasingly important role in the efficient execution of ambitious CPS projects such as production of a mirror-image ribosome.

Contents

1. Introduction

Chemical protein synthesis (CPS)[1–3] allows the precise atomic-level preparation of proteins and employs two key technologies: (1) solid-phase peptide synthesis (SPPS) to produce peptide segments[4,5] and (2) a chemoselective ligation strategy[6] to assemble peptide segments into longer synthetic products. The enabling advance in this field was the discovery of Native Chemical Ligation (NCL) in 1994[7], inspired by the pioneering selective chemical ligation concept.[8] In NCL, a peptide containing a C-terminal thioester chemoselectively reacts with another peptide containing an N-terminal cysteine (or other thiolated amino acid[9–11]) to form a native amide bond.

CPS possesses two major advantages over recombinant protein expression. First, mirror-image (D-) peptides and proteins can be

☆ A provisional patent application describing the Aligator program has been filed by the University of Utah.

* Corresponding author at: Department of Biochemistry, University of Utah School of Medicine, 15 North Medical Drive East, Room 4100, Salt Lake City, UT 84112-5650, United States.
E-mail address: kay@biochem.utah.edu (M.S. Kay).

[b] Equal contribution.

http://dx.doi.org/10.1016/j.bmc.2017.05.061
0968-0896/

directly produced. D-Peptides and proteins are attractive therapeutics due to their resistance to natural L-proteases.[12,13] Our group has used mirror-image phage display,[14] which requires total chemical synthesis of the mirror-image protein target, to develop D-peptide inhibitors of HIV[15–17] and Ebola[18] viral entry. This same approach has been used by other groups to develop mirror-image therapeutics.[19–23] Another major application of mirror-image peptides/proteins is racemic protein crystallography (reviews[24,25]), due to extended space group accessibility, extending even to quasi-racemic protein crystallography.[26] Several examples from the Kent lab have demonstrated this advantage for protein crystallization.[27–30] Besides CPS, there is currently no other method for producing D-proteins, as only a few D-residues can currently be incorporated into proteins using the ribosome.[31,32] Second, CPS offers the ability to site-specifically modify proteins for mechanistic studies. Semisynthetic proteins can be prepared by ligation of recombinantly expressed proteins with synthetic segments.[33–35] Some recent examples include ubiquitin,[36] alpha-synuclein,[37] histones,[38] and membrane proteins.[39] Additional examples include fundamental ubiquitin biology,[40] proteins with selective isotopic labeling,[41] site-specific installation of fluors (e.g., FRET pairs),[42] and interesting scaffold approaches.[43]

Using CPS methods, proteins of ~100 residues can be routinely prepared in most cases, but production of >300-residue proteins remains very difficult.[44–47] Challenges in the field include peptide thioester preparation (by Fmoc SPPS), access to reactive ligation junctions, poor SPPS synthesis quality, inefficient purification of segments and assembly intermediates, and low yield of purified product. In our hands, there are two particular challenges that hinder CPS projects: poor solubility and inefficient/suboptimal synthetic design.

The first challenge, peptide solubility, is commonly attributed to so-called "difficult" peptides (reviewed in detail[48]) that are poorly soluble even in highly denaturing buffers and/or hard-to-resolve by HPLC for analysis and purification. Several groups have addressed this challenge by designing clever chemical methods for temporarily improving handling properties. The solubility of initial peptide segments can be improved by incorporation of pH-sensitive isoacyl dipeptide building blocks[49] (at Ser/Thr residues) or application of the thioester Arg_n tag strategy.[50,51] Danishefsky's group employed custom Glu and Lys building blocks equipped with allylic ester and allylic carbamate linkers containing solubilizing guanidine groups.[52] Recently, Brik's group devised an Alloc-Phacm Cys variant for introducing poly-Arg sequences to improve peptide solubility.[53] Hojo used picolyl protection of Glu residues to improve peptide solubility and HPLC purification.[54] Photosensitive linkers have also been employed to improve segment solubility at Gln residues.[55] A very promising approach, coming from Liu's group, is termed the RBM (removable backbone modification) strategy for temporary solubilization, which was originally limited to Gly,[56] but has since been expanded to other residues.[57,58]

With the Aucagne group, we recently introduced another approach for temporarily solubilizing difficult peptides via a solubility-enhancing tag that we dubbed the "Helping Hand".[59] In this strategy, a heterobifunctional linker, Fmoc-Ddae-OH, can be used to specifically attach solubilizing sequences onto Lys side chains. Using this approach, the solubilizing sequence is easy to install and then selectively cleave using dilute aqueous hydrazine to restore the native Lys side chain. We demonstrated its use in one-pot applications following NCL and free-radical-based desulfurization.

A second major challenge to producing large proteins is the selection of the most efficient (and high-yielding) synthesis strategy, which we explicitly address in this perspective. Synthesis of large targets is laborious and may require tremendous material and human resources to identify an acceptable strategy. To illustrate and address this challenge, here we introduce a new computational tool, Aligator (Automated Ligator), which systematically scores all plausible ligation strategies to generate a ranked list of the predicted most efficient assemblies. We demonstrate the utility of our new computational tool in the context of three CPS projects originating from our lab: TNFα (157 aa), GroES (97 aa), and DapA (312 aa), followed by analysis of a ribosomal protein set that previews the challenges associated with this ambitious synthetic target. Finally, we discuss future directions for improving computational predictions of optimal ligation strategies.

2. Discussion

2.1. Selection of ribosomal protein set

As an ideal test set for our Aligator program, we selected the *E. coli* ribosomal proteins (30S and 50S subunits plus key accessory factors) for analysis (Fig. 1A). Synthesis of a mirror-image ribosome has been a longtime dream for mirror-image synthetic biology[44,60,61] and would enable production of large mirror-image proteins via *in vitro* translation. A mirror-image ribosome is also a key stepping stone towards building a fully mirror-image cell ("*D. coli*")*.*[60] The *E. coli* ribosome is ideal for this project because it has been extensively characterized (including detailed protocols for its efficient *in vitro* assembly[62]), and it is active without rRNA modifications[63] (which would be difficult to produce in mirror-image). These 65 proteins represent an ideal set with lengths from 38 to 890 residues (21 30S subunits, 33 50S subunits, and 11 key translation accessory factors) (Table S1). As shown in Fig. 1B, 57 of the proteins are within reach of current CPS techniques (<300 aa), although proteins longer than 200 aa would likely require multiple synthesis attempts with current manual synthesis designs. The remaining eight proteins would be very challenging to prepare with current CPS methods, as the largest protein synthesized to date is the 352-aa Dpo4 DNA polymerase.[44,45] These lengths, combined with the large number of subunits, illustrate the need to enhance the efficiency of current CPS strategies to achieve this ambitious goal.

Fig. 1C compares the number of Cys and Ala ligation sites available in the ribosomal data set. This analysis demonstrates the importance of including non-Cys ligation sites via the ligation-desulfurization approach[64,65] into ligation strategy prediction tools, as the ribosomal protein set is highly Cys-deficient. Here we include Ala as an alternate ligation junction since it is the most common amino acid in the test set and the most commonly used alternate ligation site.

2.2. Design of the Aligator program

To help overcome the tedious manual design of chemical protein syntheses, we developed a Python script called Aligator (Automated Ligator). This script performs two main functions: 1) generation of a list of "plausible" peptide segments with scores based on their predicted suitability for NCL and 2) systematic evaluation of all potential segment assemblies to produce a rank-ordered list of the predicted most efficient ligation strategies.

Aligator first divides the protein sequence based on the presence of Cys or Ala ligation sites to generate a list of potential peptide segments (Fig. 2A). Our initial version is designed to work with thioesters prepared using the hydrazide method,[66–69] so segments containing "incompatible" C-terminal residues are not included in the segment list (Asp, Glu, Asn, Gln, and Pro). Specifically, Asp/Glu are excluded because of their potential for thioester migration to the side chain,[70] although recent work has suggested that this is a pH-dependent reaction more prevalent in Asp thioesters.[71] Asp,