Published in final edited form as:

Bioorg Med Chem. 2017 September 15; 25(18): 4946–4952. doi:10.1016/j.bmc.2017.05.061.

Aligator: A Computational Tool for Optimizing Total Chemical Synthesis of Large Proteins

Michael T. Jacobsen^{a,b,c}, Patrick W. Erickson^{a,c}, and Michael S. Kay^{a,d}

^aDepartment of Biochemistry, University of Utah School of Medicine, 15 North Medical Drive East, Room 4100, Salt Lake City, Utah 84112-5650, United States

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: TNF α (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.

Graphical abstract

^dCorresponding author (kay@biochem.utah.edu).

^bPresent address: Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, United Kingdom

^cEqual contribution

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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

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 strategy⁶ to assemble peptide segments into longer synthetic products. The enabling advance in this field was the discovery of Native Chemical Ligation (NCL) in 1994⁷, inspired by the pioneering selective chemical ligation concept⁸. 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-) peptide and proteins can be 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 display¹⁴, which requires total chemical synthesis of the mirror-image protein target, to develop D-peptide inhibitors of HIV^{15–17} and Ebola¹⁸ 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 crystallography²⁶. 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 ubiquitin³⁶, alpha-synuclein³⁷, histones³⁸, and membrane proteins³⁹. Additional examples include fundamental ubiquitin biology⁴⁰, proteins with selective isotopic labeling⁴¹, site-specific installation of fluorophores (e.g., FRET pairs)⁴², and interesting scaffold approaches⁴³.

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/sub-optimal synthetic design.

The first challenge, peptide solubility, is commonly attributed to so-called “difficult” peptides (reviewed in detail⁴⁸) 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 blocks⁴⁹ (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 groups⁵². Recently, Brik's group devised an Alloc-Phacm Cys variant for introducing poly-Arg sequences to improve peptide solubility⁵³. Hojo used picolyl protection of Glu residues to improve peptide solubility and HPLC purification⁵⁴. Photosensitive linkers have also been employed to improve segment solubility at Gln residues⁵⁵. 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 Gly⁵⁶, 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: TNF α (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* assembly⁶²), and it is active without rRNA modifications⁶³ (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 synthetic 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 total

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 chain⁷⁰, although recent work has suggested that this is a pH-dependent reaction more prevalent in Asp thioesters⁷¹. Asp, Asn, and Gln are excluded because they cannot be directly prepared via peptide hydrazide method⁶⁷, although recent work has addressed this challenge using a two-step cleavage protocol⁷². Pro is eliminated due to its slow ligation kinetics⁷³ and propensity for diketopiperazine formation⁷⁴. The modular design of Aligator allows for changes to these restrictions as new tools are developed (or based on the ligation tools available to each group) and already includes a thioester scoring customization option (File S3), e.g., for labs using Boc SPPS. Only segments between 10 and the maximum length chosen (here we selected 80 residues) are allowed, since shorter or longer segments will likely result in inefficient strategies or unacceptable segment purity, respectively. This upper limit can be customized to reflect a lab’s preference.

Each segment is then evaluated by summing the four components of our scoring function (Fig. 2A):

1. *Thioester (ligation) kinetics*⁷³ : Segments containing a preferred C-terminal thioester receive +2 points (Ala, Arg, Cys, Gly, His, Met, Phe, Ser, Trp, or Tyr), while other acceptable thioesters (Ile, Leu, Lys⁷⁵, Thr, or Val) result in 0 points. See File S3 for details.
2. *Solubility*: In our experience^{46,59,76,77}, the handling properties of peptides under NCL and acidic RP-HPLC conditions correlate positively with the density of positively charged residues (His, Lys, Arg) and negatively with the density of negatively charged or branched hydrophobic residues (Asp, Glu, Val, Ile, and

Leu). Therefore, we assign a score of +1 for each HKR residue and -1 for each DEVIL residue. A segment's cumulative solubility score is then divided by the number of residues to produce a "per residue" solubility score. This number is then compared to the distribution of solubility scores observed for all viable segments within our entire ribosomal protein test set. Segments with average or better predicted solubility receive a score of zero, while below average segments receive a penalty score corresponding to their number of standard deviations below the mean (capped at -3.0) (File S3).

3. *Length*: In our hands, an ideal segment length is ~40 residues, as these are typically straightforward to prepare by SPPS and to purify by RP-HPLC. The length scoring function assigns a score of +2 for an ideal segment (40 residues) and subtracts 0.1 per residue deviation from this ideal value (score = $2 - |(40 - x) \times 0.1|$) for segments of length x (e.g., a 30-residue segment receives 1 point).
4. *Ligation junction*: Each Ala junction site receives a score of -2 to reflect the additional complexity (and potential yield loss) of post-NCL desulfurization.

All viable segments are then assembled into potential ligation strategies, and Aligator outputs optimal strategies in a rank-ordered list (Fig. 2B). In addition to summing the individual scores from each segment, ligation strategies receive an additional score based on the number of ligations. For traditional NCL (excluding various one-pot and solid-phase approaches, e.g.,^{78,79}), each ligation step adds an HPLC purification step that ultimately reduces yield. To discourage strategies that have more than an ideal number of ligations, this scoring function penalizes strategies having an average segment length <40 residues (-2 for each "excessive" ligation beyond this limit).

The list of potential ligation strategies is arranged by the overall Aligator scores (Fig. 2B, details in File S3). Compiling all of the segments into rank-order ligation strategies represents by far the most computationally demanding step of Aligator. As a target protein's length and number of segments increases, the number of potential strategies explodes, quickly exceeding typical memory and storage limits (e.g., the 529-aa RF3 protein has >6 million potential ligation strategies). As described above, the "ideal" segment length is set to 40 residues, so strategies are penalized for having more than $n/40$ segments (n =protein length). As a result, strategies having more than $n/35$ ligations would score very poorly in our parameters, and thus can be excluded from the analysis, greatly reducing the number of possible strategies (e.g., a 350-aa protein is limited to 10 ligations). We designed the modular Aligator script to allow users to easily adjust these cutoff variables to their preferences. For very long proteins (>400 residues or with >200 potential segments), the program enters "restriction mode" that more aggressively trims unlikely ligation strategies (see File S3 for details). Additionally, if no valid strategies are identified, the user can elect to enter "safe mode", which provides a sampling of valid ligation strategies by limiting the total number of strategies evaluated (only required for the 890-residue IF2 in the test set).

As a demonstration of how new synthetic tools can be incorporated into Aligator, we include an optional function ("HH mode") that predicts the solubilizing impact of a solubility-enhancing tag, helping hand (HH), in the strategy. This traceless linker, which is installed at

a Lys residue, can be used to increase the solubility of peptide segments⁵⁹. If the user wishes to incorporate helping hands in a synthesis strategy, then Aligator will reward segments containing at least one Lys residue by dividing its solubility score by 2, which reduces the penalty associated with negative (poor) solubility scores. Interestingly, ~93% of all valid segments in the ribosome test set are HH-compatible (contain at least one Lys), demonstrating the general applicability of this tool. The Aligator analyses described below do not use HH mode unless otherwise specified.

2.3. Aligator Case Studies

To evaluate the predictive power of our Aligator program for ranking synthesis strategies, we first applied it to three recent CPS projects originating from our lab: TNF α (157 aa)⁷⁷, GroES (97 aa)⁵⁹, and DapA (312 aa)⁴⁶.

In the first case study, we analyzed the synthesis of human Tumor Necrosis Factor alpha (TNF- α), an inflammatory cytokine overexpressed in chronic inflammatory diseases^{80,81}. This 157-residue target (residues 77 – 233, comprising the soluble domain) forms a 52 kDa trimer⁸². In our report⁷⁷, we initially confronted the synthesis using a three-segment approach centered on the Pro176-Cys177 junction. However, we were unable to obtain measurable product due to slow ligation kinetics and diketopiperazine formation at the Pro ligation site⁷⁴. We ultimately selected an alternative three-segment strategy that circumvented this ligation site, resulting in successful L- and D-syntheses of the target. Indeed, analysis of TNF α by Aligator (Fig. 3, File S4A) revealed that our published strategy (highlighted in green) scored 2nd out of 226 possible strategies with a score of +2.9 (scores ranged from +3.6 to –13.9). Fig. 3 diagrams the peptide segments and scores associated with the top five strategies. Lower scoring strategies were penalized by multiple factors: thioester NCL reactivity, segment solubility, segment lengths, and ligation-associated penalties. Compared to our published synthesis, the top-ranked strategy overcame modestly poor solubility and an extra Ala ligation site with good length and thioester scores. The 3rd-ranked strategy (+1.7) is very similar to the top-ranked one, differing only in the junction between the 1st and 2nd segments (replacing an optimal Arg thioester with a sluggish Leu, resulting in a 2-point penalty). The 4th strategy is very similar to the 1st, but was penalized by a longer 62-residue C-terminal segment. The 5th strategy had strong length and thioester scores that were offset by penalties due to an additional ligation (–2) and an additional Ala ligation junction (–2). Analyzing TNF- α with Aligator in HH mode (File S4B) predicts a 0.7-point HH benefit (modest for a protein of this size). Overall, Aligator analysis of TNF- α reinforced our synthetic experience.

In the second case study, we highlight the synthesis of *E. coli* GroES, which we made using a ligation-desulfurization approach⁵⁹. During this synthesis, we encountered significant challenges with the hydrophobic C-terminal region, which we overcame using a poly-Lys helping hand. Ultimately, we synthesized both L- and D- GroES using a two-segment approach centered on the Leu41-Ala42 junction. Analysis of GroES by Aligator (File S5A) produced ten possible strategies with scores ranging from –1.0 to –6.0. A full-length, one-segment strategy⁸³ was not considered due to elimination of peptides >80 aa. Our published strategy⁵⁹ (File S5A, green highlight) scored –1.7, fourth on the list. Examination of higher

scoring strategies suggested that increasing the length of the C-terminal segment would have produced a better strategy due to improved solubility. We had not considered this possibility, showing that Aligator can sometimes provide unexpected potential insights into synthetic strategies. Importantly, our synthesis required the assistance of a helping hand to overcome solubility problems in the C-terminal segment (solubility score of -2.0). Analyzing GroES with Aligator in HH mode (File S5B) predicts a 1-point HH benefit, which is highly significant since it arises from a single segment in a small protein.

In the final case, we analyzed our synthesis of the *E. coli* GroEL/ES chaperone-dependent protein DapA (312 residues)⁴⁶. In this work, we pursued several different synthetic strategies before initially settling on an eight-segment approach. Aligator analysis of DapA (File S6A) produced an astonishing 357,721 possible strategies (scores ranging from $+17.2$ to -25.6), which we culled to the top 1000 hits (scores ranging from $+17.2$ to $+7.8$). To our satisfaction, the top-ranked strategy ($+17.2$) produced the same eight-segment strategy that we had settled upon through experience. However, after further trial-and-error, we eventually identified a successful seven-segment strategy that combined the last two segments. This approach ranked in the top 0.02% (File S6A, green highlight, #51, $+12.5$). This dramatic score difference is almost entirely due to a nearly 5 point peptide segment length penalty (75 vs. 37 + 38 aa), reflecting the surprisingly high quality and yield of the long C-terminal segment, which is difficult to predict. Interestingly, re-analysis in HH-mode (File S6B) predicts that the helping hand will have much less impact (0.7 points) in the context of this larger protein. Thus, Aligator was effective at suggesting optimal synthesis strategies, but further tuning of the various design factors will be important in future versions of the program, based on user input.

After validating that Aligator provides an informative rank ordering of synthesis strategies for GroES, TNF α , and DapA, we next analyzed the entire ribosomal protein set, with and without HH mode (Files S7A/B). In standard mode, potential ligation strategies were found for all proteins except IF2, which was overcome using safe mode. Impressively, Aligator required only 12 min to analyze the full protein set (on a 2016 MacBook Pro). Aligator's raw output files showing all possible strategies are quite large and difficult to navigate (>3 GB for the largest proteins in the test set). Therefore, Aligator also provides a list of the top 1000 strategies in Excel format that includes all of the component scores and enables easy visualization of the differences between high-ranking strategies.

Analysis of the highest-scoring ligation strategies across this test set reveals how the tension between the different scoring functions affects overall ranking. For example, the top S12 ligation strategies take different approaches to achieve their high scores (File S7A). The winning strategy benefits from a small number of ligations and ideal segment lengths. The two similar strategies tied for 2nd-place both have one fewer ligation, but this benefit is offset by an additional Ala junction in one strategy or a less optimal thioester in the other. The 4th-place strategy makes optimal use of natural Cys junctions and minimizes ligations, but contains a suboptimal long segment (72 aa).

Perhaps the most interesting aspect of this Aligator analysis is the impact of the HH reward function on the overall strategy scores. For example, S1's top score (2.2, no-HH) is

dramatically penalized by its poor solubility (−10.7), but its score improves dramatically in HH-mode (by ~5 points). Overall, the order of the top-scoring strategies is largely unchanged because all benefit similarly from the HH solubility-enhancing reward. A striking example of the HH's potential impact is RF2, which is predicted to be among the most difficult proteins in the test set (−2.0, no-HH). Analysis in HH-mode predicts a greatly improved score (from −2.0 to 0.8), but also deposes the original top-ranked strategy (now #2). Compared to the top non-HH strategy, the best HH-assisted strategy uses better thioesters and has a better length distribution, though these benefits are partly offset by its use of one additional Ala junction. HH-mode reveals this ligation strategy by diminishing strong solubility penalties that otherwise discouraged the use of more optimal segments. As expected, proteins with good overall solubility scores are predicted to benefit much less from helping hands (e.g., L1). Experimental data from the ribosomal test set will provide valuable training to improve Aligator, particularly for balancing the relative weights of each of the scoring functions.

3. Conclusions and Outlook

In this perspective, we highlighted two of the central challenges in CPS projects: poor solubility and inefficient/suboptimal synthetic design. Chemists have been continually developing new tools to handle difficult peptides—in particular, our recently published Helping Hand solubility-enhancing tool for temporarily introducing solubilizing groups at Lys residues⁵⁹. In certain cases, there may not be an accessible Lys in a difficult segment, so it would be wise to employ the suite of solubilizing tools (e.g., Gly, Glu, or Cys-based strategies described in the introduction). In order to address strategic challenges in CPS projects, we have introduced Aligator, a new computational tool to analyze and then predict the most efficient synthesis strategies for CPS projects.

The current Aligator program (version 1.0) effectively ranked the synthesis strategies for three targets from our lab: TNF α , GroES, and DapA. We then applied it to the entire ribosomal protein set. Although Aligator already produces informative predictions, we aim to continuously revise and improve its scoring functions, as well as introduce additional functionality. For this first version of Aligator, we developed a simple and efficient solubility scoring function based on our practical experience with handling difficult peptides. However, we acknowledge that solubility prediction is a challenging problem and that a more sophisticated scoring function will be required to improve accuracy.

Future versions of Aligator can incorporate additional scoring functions to reflect other challenges in CPS. For example, aspartimide formation reduces synthetic quality and yields^{84–86}, so Aligator could reward placement of aspartimide-prone dipeptide sequences near the N-termini of segments (to minimize base-catalyzed formation). Another example would take into account proline and pseudoproline sites that introduce kinks into the peptide chain, which can help prevent on-resin aggregation during SPPS, significantly improving quality and yield^{87–90}. This type of kink analysis could account for the distribution of such sites within segments in order to predict synthetic difficulty. For mirror-image projects, this function could help predict which custom-synthesized D-pseudoproline dipeptides would have the most impact on improving overall quality.

Aligator could become a more valuable tool by considering the order of segment assembly, though such a feature will require a much more complex and computationally intensive algorithm. While one-pot and solid-phase ligation strategies can be used, convergent synthesis is generally the most efficient approach⁹¹. Consideration of the optimal ligation order will depend on the specific terminal protecting groups used, as well as the arrangement of residues requiring desulfurization. To predict optimal convergent synthesis strategies, Aligator will also need to consider the solubility of selected assembly intermediates to avoid intractable peptides that are difficult to ligate or purify.

In this work, we introduced Aligator, a new computational tool that can rapidly analyze the sequences of a CPS target and suggest the most optimal synthesis strategies based on the sequence. Aligator is centered on NCL using the Fmoc-compatible hydrazide method for generating peptide thioesters^{66–69}, but should be generally applicable to other chemoselective ligation reactions (e.g., KAHA⁹² and Ser/Thr ligation⁹³). The programs described here are freely available for non-commercial use on Github: <https://github.com/kay-lab>. Users can also share their improvements and customizations on this site.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health [Predoctoral Training Grant CA171677 to M.T.J., AI076168 and AI102347 to M.S.K].

References

1. Kent SB. Total chemical synthesis of proteins. Chemical Society reviews. 2009; 38(2):338–351. [PubMed: 19169452]
2. Dawson PE, Kent SB. Synthesis of native proteins by chemical ligation. Annu Rev Biochem. 2000; 69:923–960. [PubMed: 10966479]
3. Nilsson BL, Soellner MB, Raines RT. Chemical synthesis of proteins. Annu Rev Biophys Biomol Struct. 2005; 34:91–118. [PubMed: 15869385]
4. Merrifield B. Solid phase synthesis. Science. 1986; 232(4748):341–347. [PubMed: 3961484]
5. Merrifield RB. Solid phase peptide synthesis. I. The synthesis of a tetrapeptide. J Am Chem Soc. 1963; 85(14):2149–2154.
6. Hackenberger CP, Schwarzer D. Chemoselective ligation and modification strategies for peptides and proteins. Angew Chem Int Ed Engl. 2008; 47(52):10030–10074. [PubMed: 19072788]
7. Dawson PE, Muir TW, Clark-Lewis I, Kent SB. Synthesis of proteins by native chemical ligation. Science. 1994; 266(5186):776–779. [PubMed: 7973629]
8. Schnolzer M, Kent SB. Constructing proteins by dovetailing unprotected synthetic peptides: backbone-engineered HIV protease. Science. 1992; 256(5054):221–225. [PubMed: 1566069]
9. Malins LR, Payne RJ. Recent extensions to native chemical ligation for the chemical synthesis of peptides and proteins. Curr Opin Chem Biol. 2014; 22:70–78. [PubMed: 25285753]
10. Wong CTT, Tung CL, Li X. Synthetic cysteine surrogates used in native chemical ligation. Molecular bioSystems. 2013; 9(5):826–833. [PubMed: 23302767]
11. He Q-Q, Fang G-M, Liu L. Design of thiol-containing amino acids for native chemical ligation at non-Cys sites. Chinese Chemical Letters. 2013; 24(4):265–269.

12. Milton RC, Milton SC, Kent SB. Total chemical synthesis of a D-enzyme: the enantiomers of HIV-1 protease show reciprocal chiral substrate specificity. *Science*. 1992; 256(5062):1445–1448. [PubMed: 1604320]
13. Zawadzke LEB, J M. A racemic protein. *J Am Chem Soc*. 1992; 114(10):4002–4003.
14. Schumacher TN, Mayr LM, Minor DL Jr, Milhollen MA, Burgess MW, Kim PS. Identification of D-peptide ligands through mirror-image phage display. *Science*. 1996; 271(5257):1854–1857. [PubMed: 8596952]
15. Welch BD, VanDemark AP, Heroux A, Hill CP, Kay MS. Potent D-peptide inhibitors of HIV-1 entry. *Proc Natl Acad Sci U S A*. 2007; 104(43):16828–16833. [PubMed: 17942675]
16. Welch BD, Francis JN, Redman JS, et al. Design of a potent D-peptide HIV-1 entry inhibitor with a strong barrier to resistance. *J Virol*. 2010; 84(21):11235–11244. [PubMed: 20719956]
17. Francis JN, Redman JS, Eckert DM, Kay MS. Design of a modular tetrameric scaffold for the synthesis of membrane-localized d-peptide inhibitors of HIV-1 entry. *Bioconjug Chem*. 2012
18. Clinton TR, Weinstock MT, Jacobsen MT, et al. Design and characterization of ebolavirus GP prehairpin intermediate mimics as drug targets. *Protein Science*. 2015; 24(4):446–463. [PubMed: 25287718]
19. Liu M, Pazgier M, Li C, Yuan W, Li C, Lu W. A left-handed solution to peptide inhibition of the p53-MDM2 interaction. *Angew Chem Int Ed Engl*. 2010; 49(21):3649–3652. [PubMed: 20449836]
20. Liu M, Li C, Pazgier M, et al. D-peptide inhibitors of the p53-MDM2 interaction for targeted molecular therapy of malignant neoplasms. *Proc Natl Acad Sci U S A*. 2010; 107(32):14321–14326. [PubMed: 20660730]
21. Mandal K, Uppalapati M, Ault-Riche D, et al. Chemical synthesis and X-ray structure of a heterochiral {D-protein antagonist plus vascular endothelial growth factor} protein complex by racemic crystallography. *Proc Natl Acad Sci U S A*. 2012; 109(37):14779–14784. [PubMed: 22927390]
22. Uppalapati M, Lee DJ, Mandal K, et al. A Potent d-Protein Antagonist of VEGF-A is Nonimmunogenic, Metabolically Stable, and Longer-Circulating in Vivo. *ACS chemical biology*. 2016; 11(4):1058–1065. [PubMed: 26745345]
23. Chang HN, Liu BY, Qi YK, et al. Blocking of the PD-1/PD-L1 Interaction by a D-Peptide Antagonist for Cancer Immunotherapy. *Angew Chem Int Ed Engl*. 2015; 54(40):11760–11764. [PubMed: 26259671]
24. Yeates TO, Kent SB. Racemic protein crystallography. *Annu Rev Biophys*. 2012; 41:41–61. [PubMed: 22443988]
25. Matthews BW. Racemic crystallography—easy crystals and easy structures: what’s not to like? *Protein Science*. 2009; 18(6):1135–1138. [PubMed: 19472321]
26. Gao S, Pan M, Zheng Y, et al. Monomer/Oligomer Quasi-Racemic Protein Crystallography. *J Am Chem Soc*. 2016; 138(43):14497–14502. [PubMed: 27768314]
27. Mandal K, Pentelute BL, Tereshko V, et al. Racemic crystallography of synthetic protein enantiomers used to determine the X-ray structure of plectasin by direct methods. *Protein Sci*. 2009; 18(6):1146–1154. [PubMed: 19472324]
28. Pentelute BL, Gates ZP, Tereshko V, et al. X-ray structure of snow flea antifreeze protein determined by racemic crystallization of synthetic protein enantiomers. *J Am Chem Soc*. 2008; 130(30):9695–9701. [PubMed: 18598029]
29. Mandal K, Pentelute BL, Bang D, Gates ZP, Torbeev VY, Kent SB. Design, total chemical synthesis, and X-ray structure of a protein having a novel linear-loop polypeptide chain topology. *Angew Chem Int Ed Engl*. 2012; 51(6):1481–1486. [PubMed: 22213444]
30. Bunker RD, Mandal K, Bashiri G, et al. A functional role of Rv1738 in *Mycobacterium tuberculosis* persistence suggested by racemic protein crystallography. *Proc Natl Acad Sci U S A*. 2015; 112(14):4310–4315. [PubMed: 25831534]
31. Achenbach J, Jahnz M, Bethge L, et al. Outwitting EF-Tu and the ribosome: translation with d-amino acids. *Nucleic acids research*. 2015; 43(12):5687–5698. [PubMed: 26026160]
32. Katoh T, Tajima K, Suga H. Consecutive Elongation of D-Amino Acids in Translation. *Cell Chem Biol*. 2017; 24(1):46–54. [PubMed: 28042044]

33. Muir TW, Sondhi D, Cole PA. Expressed protein ligation: a general method for protein engineering. *Proc Natl Acad Sci U S A*. 1998; 95(12):6705–6710. [PubMed: 9618476]
34. Muralidharan V, Muir TW. Protein ligation: an enabling technology for the biophysical analysis of proteins. *Nat Methods*. 2006; 3(6):429–438. [PubMed: 16721376]
35. Muir TW. Semisynthesis of proteins by expressed protein ligation. *Annu Rev Biochem*. 2003; 72:249–289. [PubMed: 12626339]
36. Hemantha HP, Bavikar SN, Herman-Bachinsky Y, et al. Nonenzymatic polyubiquitination of expressed proteins. *J Am Chem Soc*. 2014; 136(6):2665–2673. [PubMed: 24437386]
37. Haj-Yahya M, Fauvet B, Herman-Bachinsky Y, et al. Synthetic polyubiquitinated α -Synuclein reveals important insights into the roles of the ubiquitin chain in regulating its pathophysiology. *Proc Natl Acad Sci U S A*. 2013; 110(44):17726–17731. [PubMed: 24043770]
38. Voigt P, Reinberg D. Histone tails: ideal motifs for probing epigenetics through chemical biology approaches. *Chembiochem*. 2011; 12(2):236–252. [PubMed: 21243712]
39. Focke PJ, Valiyaveetil FI. Studies of ion channels using expressed protein ligation. *Curr Opin Chem Biol*. 2010; 14(6):797–802. [PubMed: 20965773]
40. Singh SK, Sahu I, Mali SM, et al. Synthetic Uncleavable Ubiquitinated Proteins Dissect Proteasome Deubiquitination and Degradation, and Highlight Distinctive Fate of Tetraubiquitin. *J Am Chem Soc*. 2016; 138(49):16004–16015. [PubMed: 27960333]
41. Luk LY, Ruiz-Pernia JJ, Adesina AS, et al. Chemical Ligation and Isotope Labeling to Locate Dynamic Effects during Catalysis by Dihydrofolate Reductase. *Angew Chem Int Ed Engl*. 2015; 54(31):9016–9020. [PubMed: 26079622]
42. Geurink PP, van Tol BD, van Dalen D, et al. Development of Diubiquitin-Based FRET Probes To Quantify Ubiquitin Linkage Specificity of Deubiquitinating Enzymes. *Chembiochem*. 2016; 17(9):816–820. [PubMed: 26996281]
43. van de Langemheen H, Quarles van Ufford HC, Kruijtz JAW, Liskamp RMJ. Efficient Synthesis of Protein Mimics by Sequential Native Chemical Ligation. *Organic Letters*. 2014; 16(8):2138–2141. [PubMed: 24708093]
44. Xu W, Jiang W, Wang J, et al. Total chemical synthesis of a thermostable enzyme capable of polymerase chain reaction. *Cell Discov*. 2017; 3:17008. [PubMed: 28265464]
45. Pech A, Achenbach J, Jahnz M, et al. A thermostable d-polymerase for mirror-image PCR. *Nucleic acids research*. 2017
46. Weinstock MT, Jacobsen MT, Kay MS. Synthesis and folding of a mirror-image enzyme reveals ambidextrous chaperone activity. *Proc Natl Acad Sci U S A*. 2014; 111(32):11679–11684. [PubMed: 25071217]
47. Kumar KS, Bavikar SN, Spasser L, Moyal T, Ohayon S, Brik A. Total chemical synthesis of a 304 amino acid K48-linked tetraubiquitin protein. *Angew Chem Int Ed Engl*. 2011; 50(27):6137–6141. [PubMed: 21591043]
48. Paradis-Bas M, Tulla-Puche J, Albericio F. The road to the synthesis of “difficult peptides”. *Chemical Society reviews*. 2016; 45(3):631–654. [PubMed: 26612670]
49. Coin I, Dolling R, Krause E, et al. Depsipeptide methodology for solid-phase peptide synthesis: circumventing side reactions and development of an automated technique via depsidipeptide units. *J Org Chem*. 2006; 71(16):6171–6177. [PubMed: 16872202]
50. Sato T, Saito Y, Aimoto S. Synthesis of the C-terminal region of opioid receptor like 1 in an SDS micelle by the native chemical ligation: effect of thiol additive and SDS concentration on ligation efficiency. *J Pept Sci*. 2005; 11(7):410–416. [PubMed: 15635671]
51. Johnson EC, Kent SB. Towards the total chemical synthesis of integral membrane proteins: a general method for the synthesis of hydrophobic peptide-thioester building blocks. *Tetrahedron letters*. 2007; 48(10):1795–1799. [PubMed: 19177172]
52. Tan Z, Shang S, Danishefsky SJ. Rational development of a strategy for modifying the aggregability of proteins. *Proc Natl Acad Sci U S A*. 2011; 108(11):4297–4302. [PubMed: 21368182]
53. Maity SK, Mann G, Jbara M, Laps S, Kamnesky G, Brik A. Palladium-Assisted Removal of a Solubilizing Tag from a Cys Side Chain To Facilitate Peptide and Protein Synthesis. *Org Lett*. 2016; 18(12):3026–3029. [PubMed: 27268382]

54. Hojo H. A strategy for the synthesis of hydrophobic proteins and glycoproteins. *Organic & biomolecular chemistry*. 2016; 14(27):6368–6374. [PubMed: 27283517]
55. Huang YC, Li YM, Chen Y, et al. Synthesis of autophagosomal marker protein LC3-II under detergent-free conditions. *Angew Chem Int Ed Engl*. 2013; 52(18):4858–4862. [PubMed: 23554020]
56. Zheng JS, Yu M, Qi YK, et al. Expedient total synthesis of small to medium-sized membrane proteins via fmoc chemistry. *J Am Chem Soc*. 2014; 136(9):3695–3704. [PubMed: 24559202]
57. Zheng JS, He Y, Zuo C, et al. Robust Chemical Synthesis of Membrane Proteins through a General Method of Removable Backbone Modification. *J Am Chem Soc*. 2016; 138(10):3553–3561. [PubMed: 26943264]
58. Zuo C, Tang S, Si YY, Wang ZA, Tian CL, Zheng JS. Efficient synthesis of longer Abeta peptides via removable backbone modification. *Organic & biomolecular chemistry*. 2016; 14(22):5012–5018. [PubMed: 27188564]
59. Jacobsen MT, Petersen ME, Ye X, et al. A Helping Hand to Overcome Solubility Challenges in Chemical Protein Synthesis. *J Am Chem Soc*. 2016; 138(36):11775–11782. [PubMed: 27532670]
60. Weinstock MT, Francis JN, Redman JS, Kay MS. Protease-resistant peptide design-empowering nature's fragile warriors against HIV. *Biopolymers*. 2012; 98(5):431–442. [PubMed: 23203688]
61. Wang Z, Xu W, Liu L, Zhu TF. A synthetic molecular system capable of mirror-image genetic replication and transcription. *Nat Chem*. 2016; 8(7):698–704. [PubMed: 27325097]
62. Li J, Gu L, Aach J, Church GM. Improved cell-free RNA and protein synthesis system. *PLoS One*. 2014; 9(9):e106232. [PubMed: 25180701]
63. Jewett MC, Fritz BR, Timmerman LE, Church GM. In vitro integration of ribosomal RNA synthesis, ribosome assembly, and translation. *Mol Syst Biol*. 2013; 9:678. [PubMed: 23799452]
64. Yan LZ, Dawson PE. Synthesis of peptides and proteins without cysteine residues by native chemical ligation combined with desulfurization. *J Am Chem Soc*. 2001; 123(4):526–533. [PubMed: 11456564]
65. Wan Q, Danishefsky SJ. Free-radical-based, specific desulfurization of cysteine: a powerful advance in the synthesis of polypeptides and glycopolypeptides. *Angew Chem Int Ed Engl*. 2007; 46(48):9248–9252. [PubMed: 18046687]
66. Zheng JS, Tang S, Guo Y, Chang HN, Liu L. Synthesis of cyclic peptides and cyclic proteins via ligation of peptide hydrazides. *Chembiochem*. 2012; 13(4):542–546. [PubMed: 22302623]
67. Fang GM, Wang JX, Liu L. Convergent chemical synthesis of proteins by ligation of peptide hydrazides. *Angew Chem Int Ed Engl*. 2012; 51(41):10347–10350. [PubMed: 22968928]
68. Fang GM, Li YM, Shen F, et al. Protein chemical synthesis by ligation of peptide hydrazides. *Angew Chem Int Ed Engl*. 2011; 50(33):7645–7649. [PubMed: 21648030]
69. Zheng JS, Tang S, Qi YK, Wang ZP, Liu L. Chemical synthesis of proteins using peptide hydrazides as thioester surrogates. *Nat Protoc*. 2013; 8(12):2483–2495. [PubMed: 24232250]
70. Villain M, Gaertner H, Botti P. Native Chemical Ligation with Aspartic and Glutamic Acids as C-Terminal Residues: Scope and Limitations. *European Journal of Organic Chemistry*. 2003; 2003(17):3267–3272.
71. Dang B, Kubota T, Mandal K, Bezanilla F, Kent SB. Native chemical ligation at Asx-Cys, Glx-Cys: chemical synthesis and high-resolution X-ray structure of ShK toxin by racemic protein crystallography. *J Am Chem Soc*. 2013; 135(32):11911–11919. [PubMed: 23919482]
72. Tian X, Li J, Huang W. Optimal peptide hydrazide ligation with C-terminus Asp, Asn, and Gln hydrazides. *Tetrahedron letters*. 2016; 57(38):4264–4267.
73. Hackeng TM, Griffin JH, Dawson PE. Protein synthesis by native chemical ligation: expanded scope by using straightforward methodology. *Proc Natl Acad Sci U S A*. 1999; 96(18):10068–10073. [PubMed: 10468563]
74. Nakamura T, Shigenaga A, Sato K, Tsuda Y, Sakamoto K, Otaka A. Examination of native chemical ligation using peptidyl prolyl thioesters. *Chemical Communications*. 2014; 50(1):58–60. [PubMed: 24195110]
75. Siman P, Karthikeyan SV, Nikolov M, Fischle W, Brik A. Convergent chemical synthesis of histone H2B protein for the site-specific ubiquitination at Lys34. *Angew Chem Int Ed Engl*. 2013; 52(31):8059–8063. [PubMed: 23794525]

76. Clinton TR, Weinstock MT, Jacobsen MT, et al. Design and characterization of ebolavirus GP prehairpin intermediate mimics as drug targets. *Protein Sci.* 2015; 24(4):446–463. [PubMed: 25287718]
77. Petersen ME, Jacobsen MT, Kay MS. Synthesis of tumor necrosis factor alpha for use as a mirror-image phage display target. *Organic & biomolecular chemistry.* 2016; 14(23):5298–5303. [PubMed: 27211891]
78. Bang D, Kent SB. A one-pot total synthesis of crambin. *Angew Chem Int Ed Engl.* 2004; 43(19):2534–2538. [PubMed: 15127445]
79. Jbara M, Seenaiah M, Brik A. Solid phase chemical ligation employing a rink amide linker for the synthesis of histone H2B protein. *Chem Commun (Camb).* 2014; 50(83):12534–12537. [PubMed: 25196573]
80. Feldmann M, Maini RN. Lasker Clinical Medical Research Award. TNF defined as a therapeutic target for rheumatoid arthritis and other autoimmune diseases. *Nat Med.* 2003; 9(10):1245–1250. [PubMed: 14520364]
81. Carswell EA, Old LJ, Kassel RL, Green S, Fiore N, Williamson B. An endotoxin-induced serum factor that causes necrosis of tumors. *Proc Natl Acad Sci U S A.* 1975; 72(9):3666–3670. [PubMed: 1103152]
82. Tracey D, Klareskog L, Sasso EH, Salfeld JG, Tak PP. Tumor necrosis factor antagonist mechanisms of action: a comprehensive review. *Pharmacol Ther.* 2008; 117(2):244–279. [PubMed: 18155297]
83. Mascagni P, Tonolo M, Ball H, Lim M, Ellis RJ, Coates A. Chemical synthesis of 10 kDa chaperonin. Biological activity suggests chaperonins do not require other molecular chaperones. *FEBS Lett.* 1991; 286(1–2):201–203. [PubMed: 1677897]
84. Mergler M, Dick F, Sax B, Weiler P, Vorherr T. The aspartimide problem in Fmoc-based SPPS. Part I. *J Pept Sci.* 2003; 9(1):36–46. [PubMed: 12587881]
85. Mergler M, Dick F, Sax B, Stahelin C, Vorherr T. The aspartimide problem in Fmoc-based SPPS. Part II. *J Pept Sci.* 2003; 9(8):518–526. [PubMed: 12952393]
86. Lauer JLF, C G, Fields GB. Sequence dependence of aspartimide formation during 9-fluorenylmethoxycarbonyl solid-phase peptide synthesis. *Letters in Pep Sci.* 1995; 1(4):197–205.
87. Toniolo CB, G M, Mutter M, Pillai VNR. Linear oligopeptides. 77. The effect of the insertion of a proline residue on the solid-state conformation of host peptides. *Makromol Chem.* 1981; 182(7):1997–2005.
88. Haack T, Mutter M. Serine derived oxazolidines as secondary structure disrupting, solubilizing building blocks in peptide synthesis. *Tetrahedron letters.* 1992; 33(12):1589–1592.
89. Wöhr T, Wahl F, Nefzi A, et al. Pseudo-Prolines as a Solubilizing, Structure-Disrupting Protection Technique in Peptide Synthesis. *Journal of the American Chemical Society.* 1996; 118(39):9218–9227.
90. Mutter MN, Sato A, Sato T, Sun X, Wahl F, Wöhr T. Pseudo-prolines (psi Pro) for accessing “inaccessible” peptides. *Pept Res.* 1995; 8(3):145–153. [PubMed: 7670229]
91. Seenaiah M, Jbara M, Mali SM, Brik A. Convergent versus sequential protein synthesis: the case of ubiquitinated and glycosylated H2B. *Angew Chem Int Ed Engl.* 2015; 54(42):12374–12378. [PubMed: 26079184]
92. Harmand TJ, Murar CE, Bode JW. Protein chemical synthesis by alpha-ketoacid-hydroxylamine ligation. *Nat Protoc.* 2016; 11(6):1130–1147. [PubMed: 27227514]
93. Zhang Y, Xu C, Lam HY, Lee CL, Li X. Protein chemical synthesis by serine and threonine ligation. *Proc Natl Acad Sci U S A.* 2013; 110(17):6657–6662. [PubMed: 23569249]

Figure 1. Ribosomal protein test set

A) All proteins within the 30S and 50S *E. coli* ribosome, as well as important accessory factors, were compiled in our test data set (structures shown from the following PDB codes: 4V6D, 4V9O, 1EFC, 2B3T, 1EK8). See Table S1 for more information on these proteins. **B)** Histogram showing the protein length distribution for the ribosomal test set. **C)** Histogram displaying the number of acceptable Cys and Ala NCL sites available in the ribosomal test set. The data for B and C (shown in File S1) were generated using our “Protein amino acid composition analysis” (Paacman) script (described in File S2).

Figure 2. Determining and ranking all viable ligation strategies using Aligator program TNF- α is used here as an example, and “forbidden” thioesters and the maximum segment length cutoffs were selected as described in the main text (both can be easily modified by users). **A)** Aligator first divides the protein sequence at Cys and Ala ligation sites. It then generates all viable 10–80 aa segments that have acceptable thioesters (see main text for list). These segments are then scored by summing the four scoring functions shown (dotted box), as described in the main text and File S3. **B)** After segments are scored, Aligator identifies viable ligation assemblies by looping through the entire list of segments to find those that can be connected to create the entire protein. Each box corresponds to segments of different lengths within the viable segment list, and the branches after the hydrazides indicate the available number of segments for the next ligation. A branch terminates if: 1) The number of ligations exceeds $n/35$ (not shown); 2) No valid segments can be ligated to the C-terminus (shown as red X); or 3) a viable ligation assembly is completed. All complete ligation strategies are scored by summing the scores of their segments plus a penalty for “excessive” ligations (described in main text and File S3). Finally, Aligator sorts the final viable ligation strategy list by overall score. Our lab’s published TNF- α ligation strategy is highlighted.

	Scores						Full-length TNF α					
	Total	Thioester	Solubility	Length	Ala Junction	Ligation #	Segment Strategies					
1	3.6	6.0	-2.3	5.9	-4.0	-2.0	R A G C G A					
2	2.9	4.0	-1.4	2.3	-2.0	0.0	G C G A					
3	1.7	4.0	-2.2	5.9	-4.0	-2.0	L A G C G A					
4	1.2	6.0	-2.1	3.3	-4.0	-2.0	R A G C S A					
5	0.6	8.0	-3.1	5.7	-6.0	-4.0	R A G C S A S A					

Figure 3. Top five TNF- α ligation strategies calculated by Aligator

Displayed results are in a similar format to Aligator's Excel output (File S4A), in which the total and component scores are shown before the segments in the predicted strategy. In the case of TNF- α , the 2nd-ranked assembly reflects the actual strategy we used to prepare TNF α .