W402–W405 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* *Published online 24 April 2009*
doi:10.1093/nar/gkp256

ProteinCCD: enabling the design of protein truncation constructs for expression and crystallization experiments

Wijnand T. M. Mooij[1], Eirini Mitsiki[1] and Anastassis Perrakis[1,*]

[1]Department of Biochemistry, NKI, Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands

Received December 15, 2008; Revised April 2, 2009; Accepted April 6, 2009

ABSTRACT

ProteinCCD (CCD for Crystallographic Construct Design) aims to facilitate a common practice in structural biology, namely the design of several truncation constructs of the protein under investigation, based on experimental data or on sequence analysis tools. ProteinCCD functions as a meta-server, available online at http://xtal.nki.nl/ccd, that collects information from prediction servers concerning secondary structure, disorder, coiled coils, transmembrane segments, domains and domain linkers. It then displays a condensed view of all results against the protein sequence. The user can study the output and choose interactively possible starts and ends for suitable protein constructs. Since the required input to ProteinCCD is the DNA and not the protein sequence, once the starts and ends of constructs are chosen, the software can automatically design the oligonucleotides needed for PCR amplification of all constructs. ProteinCCD outputs a comprehensive view of all constructs and all oligos needed for bookkeeping or for direct copy-paste ordering of the designed oligonucleotides.

INTRODUCTION

The production of soluble proteins in amounts suitable for structural studies has been a common bottleneck in structural biology and structural genomics alike (1,2). For X-ray crystallographic studies an additional goal is to obtain not only soluble protein, but also a specific construct with a high propensity to crystallize. Similarly, for NMR studies the soluble protein domain does not only need to be relatively small, but also highly soluble in relatively high concentration to deliver clear spectra. The advent of cloning techniques that are high-throughput, inexpensive and compatible with robotic implementations (3,4) allows parallel construction of tens of expression constructs for each protein under study; a standard practice in many labs.

Expression constructs can be designed based on experimental information, typically limited proteolysis experiments followed by mass spectrometry based identification of the proteolytic fragments (4). Computational design based on sequence analysis is another method of choice. The use of multiple sequence alignments is wide spread and there are a variety of specific tools, e.g. T-Coffee (5) or MUSCLE (6). Based on similarities and differences among family members, the researcher decides what are the likely domain boundaries that will yield soluble, well-behaved proteins. Comparative modeling (7) can also be used if a structure of a homologous protein is known. Finally, a variety of sequence analysis methods aim to deliver structural information from the sequence alone. Although significant progress has been made in sequence analysis, there is no definitive method of choice. Typically, most researchers use a 'personal' collection of web-based tools for sequence analysis, based on prior experience. A clear bottleneck arises, as it is cumbersome to display the results of different tools in a condensed form. At present, this requires submitting many queries to different servers, and subsequent copying-pasting to compare the results of different methods.

After a concise and condensed representation of all analysis results is obtained, the researcher typically decides what are promising domain boundaries for the protein in hand. The next step is to design oligonucleotides to be used for PCR-based amplification of all these fragments. At this stage a trivial but time consuming additional bottleneck is encountered: the protein-based analysis has to be transformed back to the DNA sequence. Although the task is by all means trivial, it is time consuming and error prone, since the direct mapping between protein and DNA sequence is lost in the analysis step.

METHODS

We have developed a web-based tool to address these bottlenecks, which provides a simple and practical

*To whom correspondence should be addressed. Tel: +31205121951; Fax: +31205121954; Email: a.perrakis@nki.nl

solution for daily use in the research laboratory. The tool we developed, ProteinCCD, addresses these bottlenecks by:

(i) Acting as a meta-server that combines several sequence analysis methods that are available as web services and commonly used for expression construct design, and displaying all results in a condensed and concise manner.
(ii) Requiring as user input the DNA rather than the protein sequence and thus enabling the 'single-click' design of all oligonucleotides needed for PCR amplification of the user-designed constructs.

We chose methods from four groups of sequence analysis tools.

The first group concerns secondary structure prediction servers, aiming to predict stretches of sequences that are likely to be either helices or strands in the three dimensional structure (8) and should not be disrupted. Different algorithms give slightly different results especially at the secondary structure element boundaries. Therefore we have included a few methods from this group, and use the collection available at the Network Protein Sequence @nalysis (9) (NPS@).

The second group of sequence analysis methods concerns algorithms that aim to predict disordered regions in protein sequences. We chose four methods: IUPred (10) which uses the estimated pairwise energy content; RONN (11) which is based on a Bio-Basis Function Neural Network (BBFNN) to predict intrinsically disordered regions in proteins; DisEMBL (12) which uses empirical definitions for disorder predictions; and GlobPlot (13) which uses predictions for globularity against disorder to better identify disordered regions in proteins.

The third group of methods involves specialized servers for specific features of protein sequence. Currently we check for coiled-coil regions (14), and trans-membrane topology prediction combined with signal peptide prediction, as available from the Phobius webserver (15).

The fourth and final group of methods includes two of the many new algorithms that look for domains. The Simple Modular Architecture Research Tool (16) (SMART) aims to identify and annotate genetically mobile domains and to analyze domain architecture. The Domain Linker Predictor (17) in contrast attempts to flag the regions between likely domains.

A user needs to submit a cDNA sequence to ProteinCCD. This entry is first checked for validity, translated to amino-acid sequence, displayed in the output window and sequentially submitted to the selected servers. As results are returned in real time, they are displayed below the query sequence in a multiple-alignment manner allowing the direct comparison of results between different methods. The user can utilize the condensed comparative output, to scroll along the sequence and choose domain boundaries, with simple mouse-clicks for choosing N- and C-terminal boundaries. This step is entirely up to the user and no automated method is provided.

As soon as the user has selected domain boundaries, ProteinCCD outputs a list of the resulting protein sequences and all the oligonucleotides needed for the PCR amplification of these sequences. Oligonucleotides are chosen based either on the simple rule that N nucleotides are needed for annealing (default $N = 20$) or by choosing enough nucleotides to reach a user defined annealing temperature T_m (default $T_m = 65$) based on the formula:

$$T_m = 64.9 + \frac{41(GCnr)}{N} \qquad \mathbf{1}$$

where $GCnr$ is the number of guanine and cytosine nucleotides in the oligonucleotide sequence. User-defined overhangs are appended both on the 5′ and the 3′ oligonucleotide, to facilitate cloning and quick cut-and-paste ordering of the final oligonucleotides. ProteinCCD does not check oligonucleotides for secondary structure or for false-annealing sites.

EXAMPLE

We will now consider a very simple example describing the steps to analyze a new sequence and design five simple truncation constructs. For the example we will use the protein Dug2 from yeast (18). In Figure 1 there is an overview of the ProteinCCD web server after the submitted job was finished. The only input was the DNA sequence of Dug2. After pasting the sequence to the top field and pressing the 'Submit' button the servers selected by default returned the analysis results for the protein sequence. The translated protein sequence and the 'aligned' predictions appear in the panel below the input. Subsequently, we selected with the mouse three N-termini: The N-terminal residues 1, 214 and 510; and three C-termini: residues 106, 458 and the terminal residue 878. In Figure 2A, we display the predictions in the area of all the six termini. In brief,

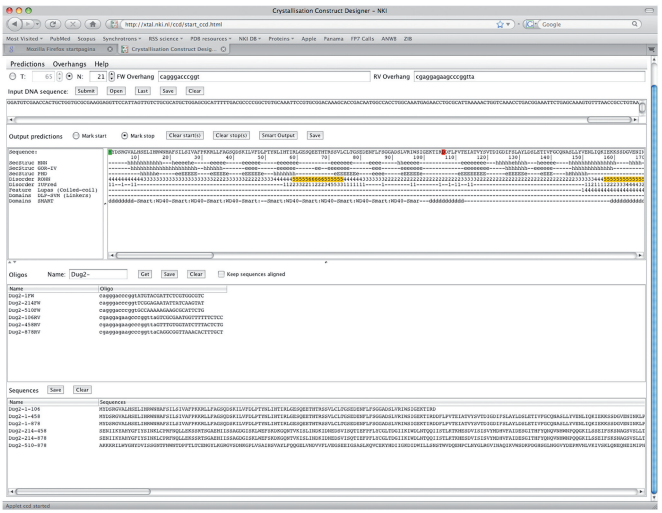


Figure 1. An overview of the ProteinCCD server after all predictions have been collected, user choices have been made, and the oligonucleotides have been suggested, for the example discussed in this paper.

SMART (16) shows that Dug2 contains four WD40 repeats, spanning the N-terminal half and a peptidase domain spanning the C-terminal half. Here we discuss how we made the exact choices for the N- and C-termini of all truncation constructs.

- N_{term} -1: The first WD40 repeat starts essentially at the very start of the protein, thus the natural N-terminus was considered as a good selection for an expression construct.
- C_{term}-106: The first pair of WD40 repeats ends at residue 98. Since a consensus strand prediction runs to residue 99, we chose to include a few more additional residues to be sure that the C-terminal predicted strand interactions are maintained.
- N_{term}-214: This residue is right at the start of the helix of the third predicted WD40 domain.
- C_{term}-458: The fourth predicted WD40 domain predicted by SMART ends at residue 396. However, the three secondary structure predictions contradict each other: they indicate a short strand followed by helix, long but weak strand prediction, short but strong strand prediction. On top, this residue is within a region predicted by RONN (11) to be disordered and—contradicting that prediction—just before a region predicted to be a domain linker (17). Finally, we noticed that two of the secondary structure prediction programs predict two helices for the next ~50 residues. Thus, we decided in this case to use as a C-terminus residue 458 and include these two secondary structure elements.
- N_{term}-510: The peptidase SMART domain is predicted to start at residue 516. However, a consensus strand is predicted to start residue 513. To be sure we don't interrupt the secondary structure elements, which might be important for domain folding, we chose to start that construct at residue 510.
- C_{term}-878: The peptidase SMART domain ends up right at the natural C-terminus and choosing exactly that was an easy choice.

After these 'starts' and 'stops' are selected by the corresponding buttons in the web server application, the 'Submit' button is pressed. This automatically calculates the needed oligonucleotides to design all possible constructs between these termini. In this particular case we only wanted constructs Dug2$^{1–878}$, Dug2$^{1–458}$, Dug2$^{214–878}$, Dug2$^{510–878}$, Dug2$^{1–106}$. The oligonucleotides that were designed to amplify these constructs and clone them to the NKI-LIC-His-3C vector, are presented in Figure 2B. In Figure 2C we show the amplified DNA under standard PCR conditions, and in Figure 2D the expression experiments, that resulted in all constructs being soluble. In conclusion, in this case study, educated 'guesses' for construct design were made very easy by the comprehensive view offered by ProteinCCD, and the

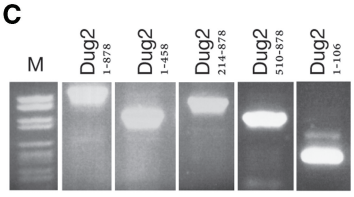
A

```
Sequence:          MYDSRGVALH...IWSIGEKTIRDDFLPVTEIA...AAIIEVPSENIIKY...KDTTNTLSLR...GKVSQVKGAKKKRIL...LTEILSKVFNRL
                         10|... 100|      110|     ...210|      220|...   460|    ...     510|       ...870|
SecStruc HNN       ------hhhh...eee-------------eeee...-eeee---hhhhhh...------hhhh...--heee-----eeee...hhhhhhhhhh--
SecStruc GOR       -------hhh...eeee-------------eee...-eeee-----hhhh...--------hh...--hhhhhh---eeee...eeeeeeeeeee-
SecStruc PHD       ----------...EEEe-----------eEEEE...-eEEEE--hhhhhh...----hhhHHH...-----------eEEE...HHHHHHHHHHh-
Disorder RONN      4444444444...2222222222222223333...44333333333332...5554444444...333333333333333...333333322222
Disorder IUPred    11--1--11-...11-------1--1-11--11...1111111-------...222222221-...11111-12111---1...11-1-11-1--1
Feature  Lupas     ----------...--------------------...--------------...----------...---------------...------------
Domains  DLP-SVM   ----------...--------------------...--------------...----------...---------------...------------
Domains  SMART     ddddddddd-S...mar---ddddddddddd---...------Smart:WD...----------...--------------P...idase_M20-P-
```

B

```
Dug2-1FW       cagggacccggtATGTACGATTCTCGTGGCGTC
Dug2-214FW     cagggacccggtTCGGAGAATATTATCAAGTAT
Dug2-510FW     cagggacccggtGCCAAAAAGAAGCGCATTCTG
Dug2-106RV     cgaggagaagcccggttaGTCGCGAATGGTTTTTTCTCC
Dug2-458RV     cgaggagaagcccggttaGTTTGTGGTATCTTTACTCTG
Dug2-878RV     cgaggagaagcccggttaCAGGCGGTTAAACACTTTGCT
```

C



D

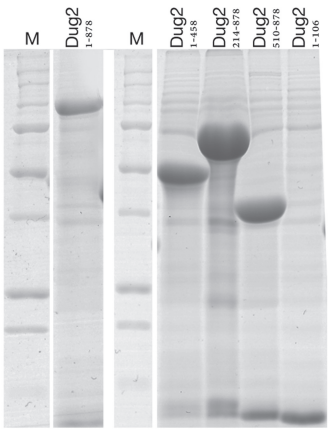


Figure 2. (**A**) Regions of interest from the ProteinCCD output where starts (green letters) and stops (red letters) are marked by the user. Discontinuities in the sequence have been marked with (...) (**B**) The oligonucleotide sequences for PCR amplifications of the regions of interest as suggested by ProteinCCD. (**C**) An ethidium bromide stained agarose gel showing the PCR products for the five selected truncation constructs of interest obtained with the designed oligonucleotides under standard experimental conditions (**D**) A polyacrylamide SDS gel stained with Coomassie blue showing the expressed and IMAC purified truncation constructs designed and cloned above. The (*) denotes the protein of interest in each lane.

experiments designed with the aid of the web server resulted in soluble proteins suitable for structural and functional studies.

CONCLUSION AND PERSPECTIVE

ProteinCCD provides a user-friendly tool to facilitate protein construct design and oligonucleotide ordering. By consolidating tools that are common in structural biology in a single platform ProteinCCD enables comparative analysis of the sequence, and by keeping track of both the protein and DNA sequence it allows the straightforward design of oligonucleotides for PCR amplification of the protein constructs.

The choice of protein constructs for structural studies is not a straightforward task. Although some notable automation attempts exist, there is no widely accepted method. Thus, at this stage we chose to only implement a tool that collects the results of a variety of popular web servers. This enables informed decisions by the user, but at present not further automation is provided, and we are unable to quantify or benchmark the server, since it is clearly dependent on user choices. It provides an enabling technology and not an automated tool.

Among the plans for future work however, is to allow users to submit their constructs choices to a connected database. This could allow for the future development of supervised learning methods to imitate user choices. This would still not allow the development of algorithms that actually predict which constructs end up being soluble (or even crystallizable), since that information is not readily available and is very difficult to collect outside the context of well-managed high-throughput projects. It can be argued however, that a learning system that mimics common user choices for a variety of targets could indeed be a useful direction for future research.

ACKNOWLEDGEMENTS

We thank all member of the Perrakis and Sixma labs at the NKI for their input, and in particular Dene Littler, Rick Hibbert, and Evangelos Christodoulou for their ideas, beta-testing the software and their comments for this manuscript.

FUNDING

ProteinCCD has been both inspired and tested as part of our participation in the EU FP6 programs 3D-Repertoire (LSHG-CT-2005-512028) and SPINE2-Complexes (LSH-2004-1.1.2-1). Funding for open access charge: Netherlands Cancer Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Graslund,S., Nordlund,P., Weigelt,J., Hallberg,B.M., Bray,J., Gileadi,O., Knapp,S., Oppermann,U., Arrowsmith,C., Hui,R. *et al.* (2008) Protein production and purification. *Nat. Methods*, **5**, 135–146.
2. Lesley,S.A. (2001) High-throughput proteomics: protein expression and purification in the postgenomic world. *Protein Expr. Purif.*, **22**, 159–164.
3. Alzari,P.M., Berglund,H., Berrow,N.S., Blagova,E., Busso,D., Cambillau,C., Campanacci,V., Christodoulou,E., Eiler,S., Fogg,M.J. *et al.* (2006) Implementation of semi-automated cloning and prokaryotic expression screening: the impact of SPINE. *Acta Crystallogr. D Biol. Crystallogr.*, **62**, 1103–1113.
4. Banci,L., Bertini,I., Cusack,S., de Jong,R.N., Heinemann,U., Jones,E.Y., Kozielski,F., Maskos,K., Messerschmidt,A., Owens,R. *et al.* (2006) First steps towards effective methods in exploiting high-throughput technologies for the determination of human protein structures of high biomedical value. *Acta Crystallogr. D Biol. Crystallogr.*, **62**, 1208–1217.
5. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
6. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
7. Ginalski,K. (2006) Comparative modeling for protein structure prediction. *Curr. Opin. Struct. Biol.*, **16**, 172–177.
8. Dunbrack,R.L. Jr. (2006) Sequence comparison and protein structure prediction. *Curr. Opin. Struct. Biol.*, **16**, 374–384.
9. Combet,C., Blanchet,C., Geourjon,C. and Deleage,G. (2000) NPS@: network protein sequence analysis. *Trends Biochem. Sci.*, **25**, 147–150.
10. Dosztanyi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.
11. Yang,Z.R., Thomson,R., McNeil,P. and Esnouf,R.M. (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, **21**, 3369–3376.
12. Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. and Russell,R.B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.
13. Linding,R., Russell,R.B., Neduva,V. and Gibson,T.J. (2003) GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.
14. Lupas,A. (1997) Predicting coiled-coil regions in proteins. *Curr. Opin. Struct. Biol.*, **7**, 388–393.
15. Kall,L., Krogh,A. and Sonnhammer,E.L. (2007) Advantages of combined transmembrane topology and signal peptide prediction – the Phobius web server. *Nucleic Acids Res.*, **35**, W429–W432.
16. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.
17. Miyazaki,S., Kuroda,Y. and Yokoyama,S. (2002) Characterization and prediction of linker sequences of multi-domain proteins by a neural network. *J. Struct. Funct. Genomics.*, **2**, 37–51.
18. Holmstrom,K., Brandt,T. and Kallesoe,T. (1994) The sequence of a 32,420 bp segment located on the right arm of chromosome II from Saccharomyces cerevisiae. *Yeast*, **10 (Suppl. A)**, S47–S62.