DOI 10.1186/s40064-015-1609-z

SpringerPlus
a SpringerOpen Journal

SOFTWARE

Open Access

Structural homology guided alignment of cysteine rich proteins

Thomas M. A. Shafee[1*], Andrew J. Robinson[2,3], Nicole van der Weerden[1] and Marilyn A. Anderson[1]

Abstract

Background: Cysteine rich protein families are notoriously difficult to align due to low sequence identity and frequent insertions and deletions.

Results: Here we present an alignment method that ensures homologous cysteines align by assigning a unique 10 amino acid barcode to those identified as structurally homologous by the DALI webserver. The free inter-cysteine regions of the barcoded sequences can then be aligned using any standard algorithm. Finally the barcodes are replaced with the original columns to yield an alignment which requires the minimum of manual refinement.

Conclusions: Using structural homology information to constrain sequence alignments allows the alignment of highly divergent, repetitive sequences that are poorly dealt with by existing algorithms. Tools are provided to perform this method online using the CysBar web-tool (http://CysBar.science.latrobe.edu.au) and offline (python script available from http://github.com/ts404/CysBar).

Keywords: Alignment, Barcode, Cysteine-rich proteins, Defensin

Background

Cysteine Rich Proteins (CRPs) are found in all kingdoms of life and are involved in diverse functions—from innate immunity to signalling to neurotoxicity (Craik et al. 2001; Koppers et al. 2011; Van der Weerden and Anderson 2013). Their properties are markedly different from globular proteins since their stability stems from the covalent disulphide bonds that cross-link their sequence, rather than relying on a hydrophobic core (Colgrave and Craik 2004; Fass 2012). This robustness allows families to evolve high sequence diversity in the inter-cysteine loops as well as making them interesting scaffolds for protein engineering (Gracy and Chiche 2011; Northfield et al. 2014). In this work we use the example of defensins, a family of small, cationic CRPs found in plants in invertebrates which perform key roles in defence against pathogens.

Protein sequence alignment is the first step in many bioinformatic analyses necessary to understand sequence-function relationships. Errors in alignment may lead to erroneous conclusions being drawn and so having accurate alignments is very important. As well as increasingly accurate multiple sequence alignment tools being developed, some software tools take existing alignments and refine them. For example the RASCAL and Refiner software (Thompson et al. 2003; Chakrabarti et al. 2006), however CRPs still pose problems for most alignment algorithms, leading to a high reliance on extensive manual alignment (Dassanayake et al. 2007; Whittington et al. 2008). This is due to two factors. Firstly, the only conserved part of the sequences is often the cysteines comprising the disulphide bridges and sequence identity within a family is frequently below 15 % (Van der Weerden and Anderson 2013). Secondly, inter-cysteine loops typically have a high rate of insertion and deletion leading to large gaps in the alignment. Together these factors lead to misalignment of sequences such that alignment columns frequently contain structurally non-homologous cysteines (Russell and Ponting 1998; Liu et al. 2009; Dwivedi and Gadagkar 2009).

Here we solve these problems with a method for aligning divergent, disulphide-rich protein sequences by identifying structurally homologous cysteines using the DALI

*Correspondence: T.Shafee@latrobe.edu.au
[1] Department of Biochemistry and Genetics, La Trobe Institute for Molecular Sciences, La Trobe University, Melbourne 3086, Australia
Full list of author information is available at the end of the article

© 2016 Shafee et al. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

server (Holm and Rosenström 2010) and replacing them with 10 amino acid barcodes (Fig. 1). Tertiary structure is widely considered a good indicator of homology since structure is more conserved than protein or DNA sequence (Grishin 2001; Hasegawa and Holm 2009). Barcoding homologous cysteines by CysBar links the known structural information to the sequences.

The barcoded sequences can then be aligned by existing algorithms with the barcoded columns effectively forced to align. Finally, the barcode sequences are removed from the alignment and the original columns restored by CysBar to yield a final alignment. This process reduces the need for manual manipulation leading to a more impartial alignment. Barcoding and reconstruction steps can be performed online with the new CysBar web-tool (http://CysBar.science.latrobe.edu.au) or with the offline python script (*cysbar.py*) and results analysed with the *loopproperties.xlsx* spreadsheet.

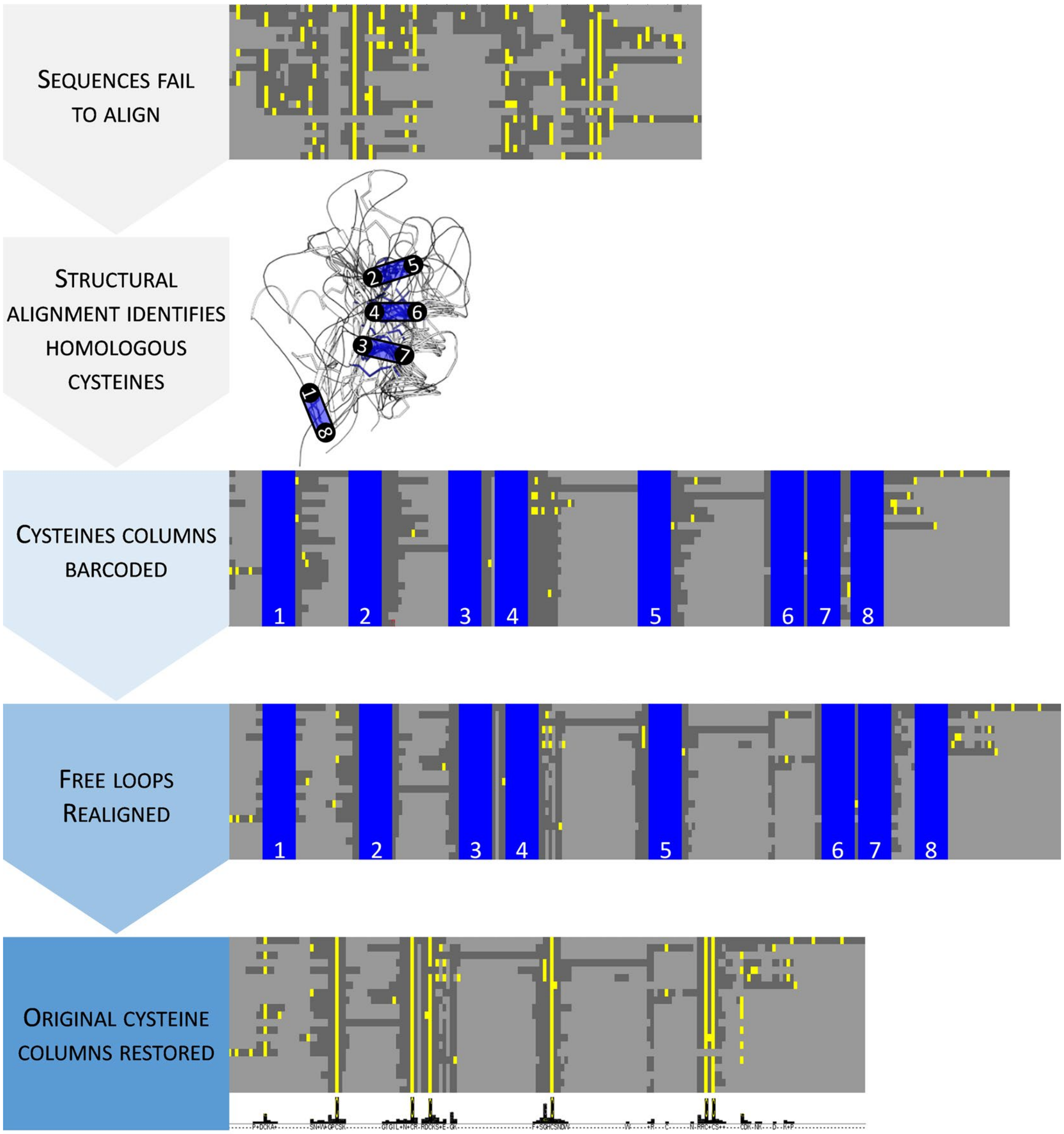
Sequences fail to align
Structural alignment identifies homologous cysteines
2
5
4
6
3
7
1
8
Cysteines columns barcoded
1
2
3
4
5
6
7
8
Free loops realigned
1
2
3
4
5
6
7
8
Original cysteine columns restored

Fig. 1 Overview of barcode alignment method. Lack of sequence conservation and abundance of cysteines prevents automatic alignment by standard methods. Homologous cysteines identified from structural alignment are replaced with 10aa barcodes to pin them in place. Standard algorithms are used to realign free loops between the *barcoded columns*. *Barcodes* are exchanged for the original columns for the final alignment and phylogeny calculation. Sequences are *coloured* with cysteines in *yellow*, any other residue in *grey*, gaps in *light grey*, and barcode sequences in *blue*

Implementation and results

Repeat cysteines and divergent loop sequences cause cysteine misalignment

CRP superfamilies are typically highly sequence diverse, possibly due to the stability of the small, disulphide-constrained fold. Indeed, the cysteines are often the only conserved residue. In the absence of similar sequence, alignment algorithms are typically heavily biased by attempting to align cysteines. Additionally, the high insertion and deletion rate and variations in disulphide pattern causes frequent cysteine misalignment (Additional file 1: Figure S1). Misalignment of 15–25 % of the cysteines (Additional file 1: Figure S2a) causes further misalignment between non-homologous inter-cysteine residues. For example, the frequent insertions and deletions within known secondary structure (Additional file 1: Figure S2b) conflicts with known trends in structure evolution (Pascarella and Argos 1992; Zhang et al. 2012).

Structure used to align homologous cysteines

To address the deficiencies of standard alignment protocols, it is necessary to use structural homology to identify the homologous cysteines that should be barcoded. The DALI (Holm and Rosenström 2010) structural homology webserver can be queried using a PDB structure and return both a list of homologous structures, and a sequence alignment based on backbone positions of the overlaid structures (Fig. 2). This alignment is used to identify homologous cysteines. In the sample dataset of defensins, cysteines are counted as homologous if they are either in the same column, or a neighbouring column (having shifted <3Å along the structure). Checking which columns contain a set of homologous cysteines is the only step that requires manual decision making.

Sequences without known structure are assigned to the closest relative with known structure by querying BLAST (Altschul et al. 1990) with the sequences of known

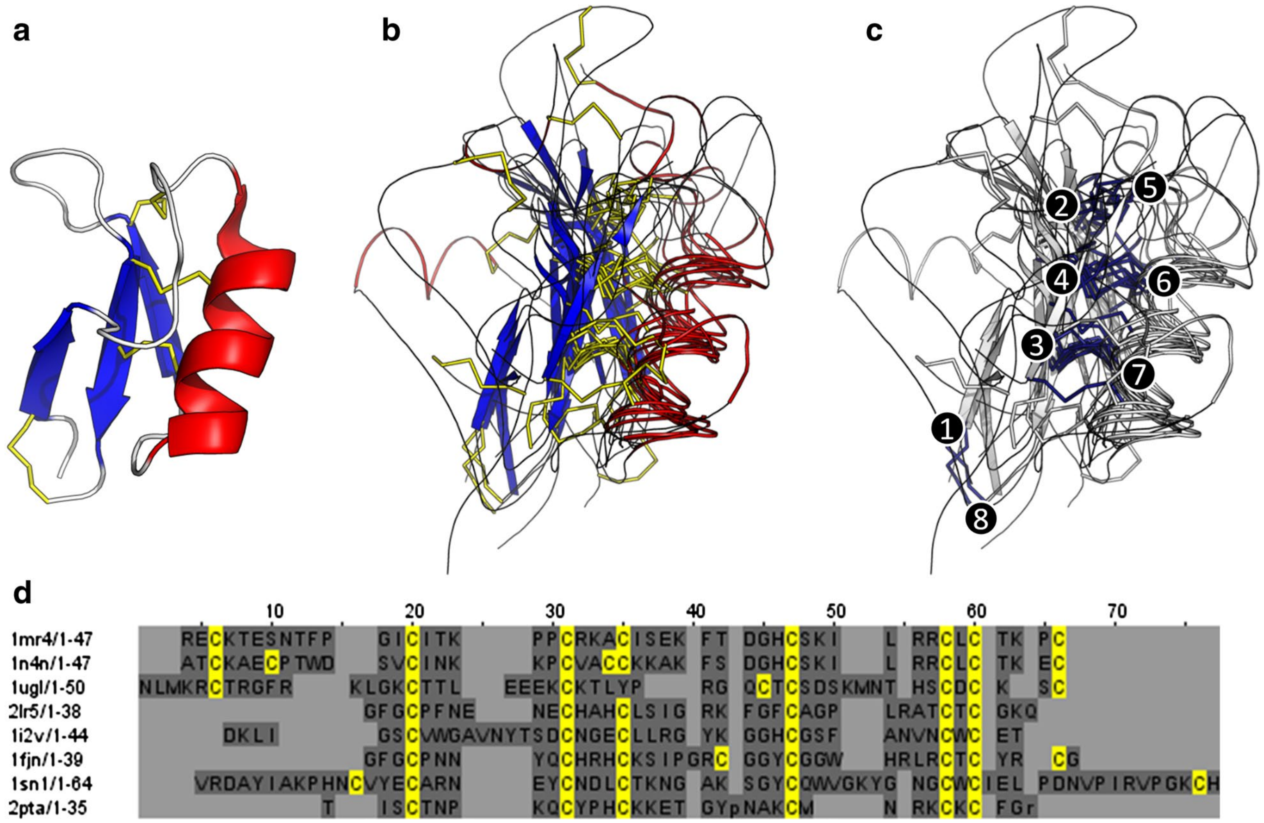
a
b
c
1
2
3
4
5
6
7
8
d
10
20
30
40
50
60
70
1mr4/1-47 RECKTESNTFP GICITK PPCRKACISEK FT DGHCSKI L RRCLC TK PC
1n4n/1-47 ATCKAECPTWD SVCINK KPCVACCKKAK FS DGHCSKI L RRCLC TK EC
1ugl/1-50 NLMKRCTRGFR KLGKCTTL EEEKCKTLYP RG QCTCSDSKMNT HSCDC K SC
2lr5/1-38 GFGCPFNE NECHAHCLSIG RK FGFCAGP LRATCTC GKQ
1i2v/1-44 DKLI GSCWWGAVNYTSDCNGECLLRG YK GGHCGSF ANVNCWC ET
1fjn/1-39 GFGCPNN YQCHRHCKSIPGRC GGYCGGW HRLRCTC YR CG
1sn1/1-64 VRDAYIAKPHNCVYECARN EYCNDLCTKNG AK SGYCQWVGKYG NGCWCIEL PDNVPIRVPGKCH
2pta/1-35 T ISCTNP KQCYPHCKKET GYpNAKCM N RKCKC FGr

Fig. 2 Identifying homologous cysteines by structural alignment. **a** The starting query structure (1MR4). **b** Overlay of aligned structures identified by DALI (Holm and Rosenström 2010). **c** The cysteine pairs indicated by DALI to be homologous in the structures. **d** Alignment of sequence based on structure by DALI. *PDB accession numbers: 1MR4, 1N4N, 1UGL, 2LR5, 1I2V, 1FJN, 1SN1, 2PTA*

structure. This process generates sub-group alignments containing at least one sequence of known structure (Additional file 1: Figure S3). These initial sub-groups only need to accurately align cysteines. In the absence of any sequence similarity to a known structure, homologous cysteines cannot be reliably identified, although a number of prediction programs exist (Fariselli and Casadio 2001; Vullo and Frasconi 2004; Ceroni et al. 2006).

Barcoding of cysteines with CysBar maintains the structural alignment of sequences

In order to merge these sub-group alignments, the non-homologous cysteines need to be distinguished from one another by retaining structural information in the sequences. To do this, alignment columns known to contain structurally homologous residues are replaced with 10 amino acid barcodes (Additional file 1: Figure S4a). This is done with either the user-friendly CysBar online web-tool or running the barcoder tool (*cysbar.py*) locally.

Two user inputs are required for each sub-group: firstly an alignment fasta file, and secondly the alignment positions to be barcoded. Each barcode is composed of the four least common amino acids (Wilkins et al. 1999) to prevent accidental mis-alignment to non-barcode sequence (Table 1).

The default set of barcode sequences minimise the likelihood of misalignment with the input sequences or with each other, based on principles developed for multiplex sequencing (Bystrykh 2012; Faircloth and Glenn 2012). Each sequence consists of the 4 least common amino acids in high-complexity sequences (Shannon entropy $\geq$1.8). Chance identity to random sequence is extremely low, with P(match) $\approx 10^{-10}$ for an alignment of 1000 sequences of length 1000. Additionally, to minimise the probability that barcodes erroneously align with one another, each differs from the others by at least 4 substitutions or indels (Hamming^{+} distance $\geq$4), with the first 8 barcodes differing from any others by 7 changes. The default set of 949 barcodes is ordered such that the sequences with higher robustness are used first. Equally robust barcodes are then ranked from high to low complexity. A full list of 948 default sequences, their Shannon entropy, and Hamming^{+} distance is contained in supplementary file *default_barcode_info.csv*.

Table 1 First 8 default barcode sequences used to replace homologous cysteine columns

ID	Sequence
bc.001	WWYHWYYHMM
bc.002	WHWMMHYHYY
bc.003	WWHHMWMMYW
bc.004	WHYYMMWMWM
bc.005	HWWMYHHMHW
bc.006	HMHYYWHHYM
bc.007	MMYMWMWHHW
bc.008	MYYHHMYWYY

Finally, the barcodes to be used are checked automatically for matches to the input sequences and the user notified that the next suitable barcode has been used instead. Custom barcodes of any length or sequence can be entered if required. The output fasta file is compatible with any standard alignment program and the residues in loops between the barcoded columns are free to realign (Additional file 1: Figure S4b). The identities of residues that were replaced with barcodes are stored in the fasta sequence ID for later reinsertion into the final alignment.

The barcoded sequences can then be re-aligned with any standard alignment algorithm, which allows the inter-cysteine loops to optimally align with the cysteine columns constrained by the barcode sequences. In this example Clustal Omega was used to align the barcoded defensin sequences since it is both accurate and scales well to large alignments (Sievers et al. 2011). The simultaneous alignment of all inter-cysteine regions, rather than aligning each block in isolation, allows the entire sequence to inform optimal alignment.

Once the loops have been realigned, the second page of the CysBar web-tool or reconstructer function of the python tool will return the original columns in place of the barcodes to generate a final sequence alignment (Fig. 3). The only user input required for this is the realigned fasta file.

The final alignment by this method is superior to the initial, naïve alignments. Homologous cysteine alignment is retained at 100 %, rather than the 75–85 % by standard methods. Indels are also primarily predicted in loop regions, in agreement with the known evolutionary trends (Additional file 1: Figure S2). The sequence of the default barcodes do not affect the final alignment. When the method was repeated 10 times on the example dataset (using the first 80 default barcodes), no misalignment of barcodes to target sequences was observed. These alignments differed by <2 % (the same margin for repeating identical alignments with different random seeds). Repeating the method on a larger set of 965 sequences (Additional file 1: Figure S5), finds that the misalignment of cysteines and secondary structure by standard algorithms is slightly poorer than for the smaller example set (Additional file 1: Figure S6).

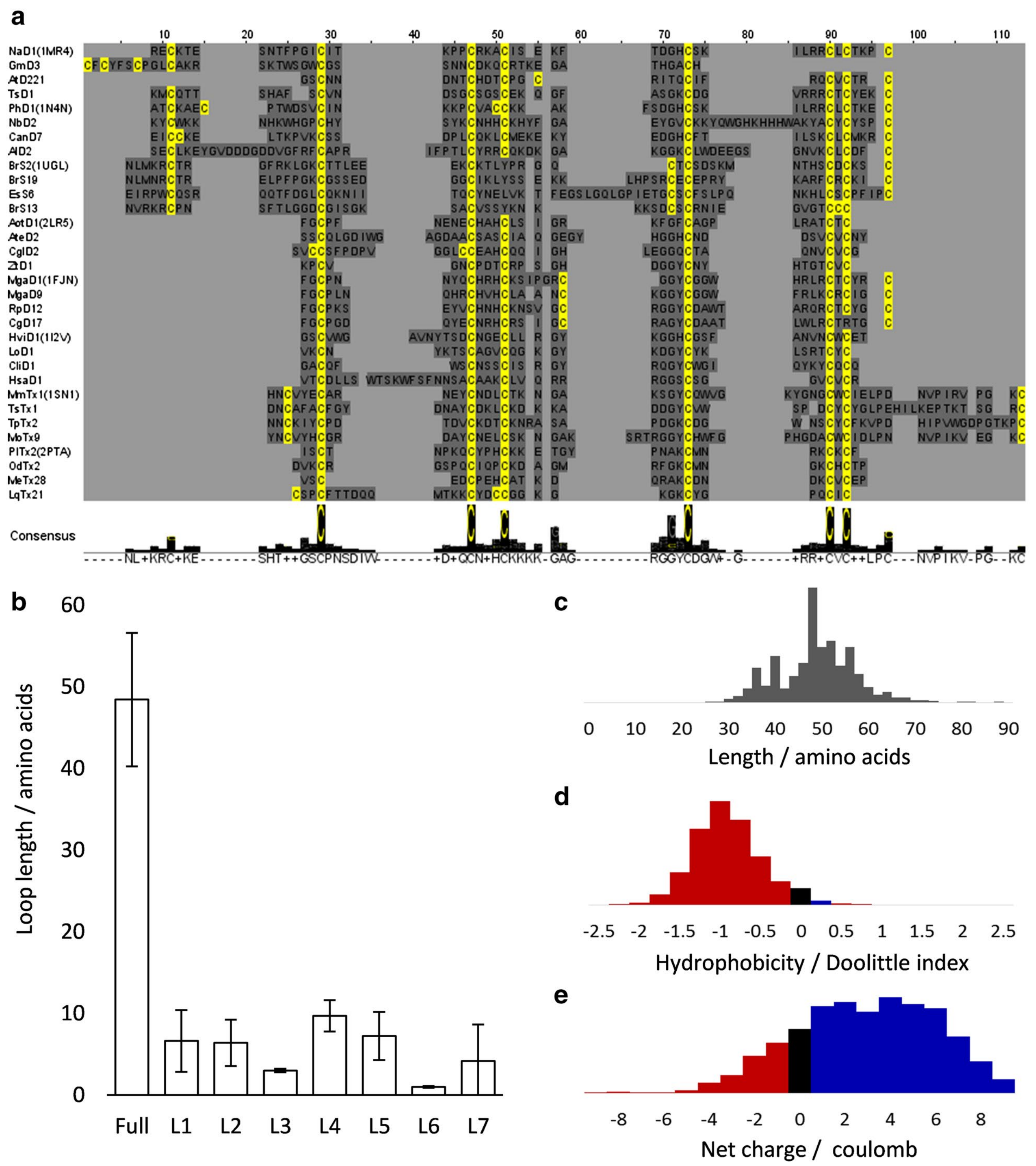

Fig. 3 Final alignment. **a** Alignment of the sequences after barcodes have been replaced with original sequence columns by CysBar-r. Sequences *coloured* with cysteines in *yellow*, any other residue in *grey*, gaps in *light grey*. **b** Distribution of lengths of inter-cysteine loops. **c**–**e** Distribution of sequence length, hydrophobicity and net charge. Data from *oop_statistics.csv* processed by *loopproperties.xlsx*

In addition to the final alignment, a.csv file is produced containing the biophysical properties for each loop of each sequence: length, hydrophobicity and net charge (Kyte and Doolittle 1982). These variables can also be user-defined if an alternative set of values is required. This file can be pasted into *loopproperties*.

xlsx spreadsheet to display and graphically summarise the loop property results in Microsoft Excel (Additional file 1: Figure S7). This allows trends and similarities to be identified in the properties of sequences that are too diverse for easy comparison.

Conclusion

Using the CysBar webtool and offline tools allows the alignment of previously unalignable protein sequences. This enables more robust bioinformatics on divergent, cysteine-rich sequences that previously had to be aligned manually. Most importantly, cysteine misalignment is abolished. A secondary effect of the improved cysteine alignment is that the inter-cysteine loop alignment is not disturbed by non-homologous loops. Finally, minimising manual alignment reduces user bias. Although extreme sequence diversity means that there is never enough information encoded in sequence or structure for perfect alignments, this method represents a significant improvement on previous protocols.

The CysBar web-tool provides a simple graphical user interface for performing the barcoding and reconstructing steps of the method. The biophysical properties are summarised by the *loopproperties.xlsx* spreadsheet interface and can be used to categorise sequences by their general properties even when their sequences are highly diverse. The annotated python script *cysbar.py* is also included in the supplementary materials along with a detailed readme containing step-by-step instructions and example data sets.

Defensin sequences have been used here as examples, however this method is equally applicable to other CRP superfamilies for which protein structures are available. Finally, the method is also applicable to sequences containing residues for which homology can be unambiguously assigned based on structural alignment (for example key catalytic residues).

Availability and requirements

Project name: CysBar.

Project home page: http://CysBar.science.latrobe.edu.au/.

Scripts repository: http://github.com/TS404/CysBar/.

Operating system(s): The web-tool can be accessed through any web browsers. The python script runs on any Linux-like platform, such as the terminal on Linux/Unix/MacOS, or runs on Microsoft Windows with Python installed. The excel spreadsheet requires Microsoft Excel 2007 or newer.

Programming language: Python and PHP.

License: Academic Free License 3.0.

Any restrictions to use by non-academics: None beyond the general restriction against redistribution in the license.

Additional files

Additional file 1:Figure S1. Misalignment by standard algorithms. **Figure S2.** Misalignment quantification. **Figure S3.** Alignments of subgroups to closest sequence of known structure. **Figure S4.** Barcoded sequences. **Figure S5.** Misalignment by standard algorithms (larger data set). **Figure S6.** Misalignment quantification (larger data set). **Figure S7.** Summary of biophysical attributes by *loopproperties.xls*.

Authors' contributions

TS conceived and developed the method, wrote excel spreadsheet. AR wrote the python script and PHP scripts for online implementation. All authors read and approved the final manuscript.

Author details

[1] Department of Biochemistry, La Trobe Institute of Molecular Sciences, La Trobe University, Melbourne 3086, Australia. [2] College of Science, Health and Engineering, La Trobe University, Melbourne 3086, Australia. [3] Life Sciences Computation Centre, Victorian Life Sciences Computation Initiative, Melbourne 3053, Australia.

Competing interests

The authors declare that there are no competing interests.

Received: 2 June 2015 Accepted: 13 December 2015
Published online: 12 January 2016

References

Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. doi:10.1016/S0022-2836(05)80360-2

Bystrykh LV (2012) Generalized DNA barcode design based on Hamming codes. PLoS One. doi:10.1371/journal.pone.0036852

Ceroni A, Passerini A, Vullo A, Frasconi P (2006) DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. Nucleic Acids Res 34:W177–W181. doi:10.1093/nar/gkl266

Chakrabarti S, Lanczycki CJ, Panchenko AR et al (2006) Refining multiple sequence alignments with conserved core regions. Nucleic Acids Res 34:2598–2606. doi:10.1093/nar/gkl274

Colgrave ML, Craik DJ (2004) Thermal, chemical, and enzymatic stability of the cyclotide kalata B1: the importance of the cyclic cystine knot. Biochemistry 43:5965–5975. doi:10.1021/bi049711q

Craik DJ, Daly NL, Waine C (2001) The cystine knot motif in toxins and implications for drug design. Toxicon 39:43–60. doi:10.1016/S0041-0101(00)00160-4

Dassanayake RS, Silva Gunawardene YIN, Tobe SS (2007) Evolutionary selective trends of insect/mosquito antimicrobial defensin peptides containing cysteine-stabilized alpha/beta motifs. Peptides 28:62–75. doi:10.1016/j.peptides.2006.09.022

Dwivedi B, Gadagkar SR (2009) Phylogenetic inference under varying proportions of indel-induced alignment gaps. BMC Evol Biol 9:211. doi:10.1186/1471-2148-9-211

Faircloth BC, Glenn TC (2012) Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. PLoS One. doi:10.1371/journal.pone.0042543

Fariselli P, Casadio R (2001) Prediction of disulfide connectivity in proteins. Bioinformatics 17:957–964

Fass D (2012) Disulfide bonding in protein biophysics. Annu Rev Biophys 41:63–79. doi:10.1146/annurev-biophys-050511-102321

Gracy J, Chiche L (2011) Structure and modeling of Knottins, a promising molecular scaffold for drug discovery. Curr Pharm Des 17:4337–4350

Grishin NV (2001) Fold change in evolution of protein structures. J Struct Biol 134:167–185. doi:10.1006/jsbi.2001.4335

Hasegawa H, Holm L (2009) Advances and pitfalls of protein structural alignment. Curr Opin Struct Biol 19:341–348. doi:10.1016/j.sbi.2009.04.003

Holm L, Rosenström P (2010) Dali server: conservation mapping in 3D. Nucleic Acids Res 38:W545–W549. doi:10.1093/nar/gkq366

Koppers AJ, Reddy T, O'Bryan MK (2011) The role of cysteine-rich secretory proteins in male fertility. Asian J Androl 13:111–117. doi:10.1038/aja.2010.77

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132

Liu K, Nelesen S, Raghavan S et al (2009) Barking up the wrong treelength: the impact of gap penalty on alignment and tree accuracy. IEEE ACM Trans Comput Biol Bioinform 6:7–21. doi:10.1109/TCBB.2008.63

Northfield SE, Wang CK, Schroeder CI et al (2014) Disulfide-rich macrocyclic peptides as templates in drug design. Eur J Med Chem 77:248–257. doi:10.1016/j.ejmech.2014.03.011

Pascarella S, Argos P (1992) Analysis of insertions/deletions in protein structures. J Mol Biol 224:461–471. doi:10.1016/0022-2836(92)91008-D

Russell RB, Ponting CP (1998) Protein fold irregularities that hinder sequence analysis. Curr Opin Struct Biol 8:364–371

Sievers F, Wilm A, Dineen D et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. doi:10.1038/msb.2011.75

Thompson JD, Thierry JC, Poch O (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. Bioinformatics 19:1155–1161. doi:10.1093/bioinformatics/btg133

Van der Weerden NL, Anderson MA (2013) Plant defensins: common fold, multiple functions. Fungal Biol Rev 26:121–131. doi:10.1016/j.fbr.2012.08.004

Vullo A, Frasconi P (2004) Disulfide connectivity prediction using recursive neural networks and evolutionary information. Bioinformatics 20:653–659. doi:10.1093/bioinformatics/btg463

Whittington CM, Papenfuss AT, Bansal P et al (2008) Defensins and the convergent evolution of platypus and reptile venom genes. Genome Res 18:986–994. doi:10.1101/gr.7149808

Wilkins MR, Gasteiger E, Bairoch A et al (1999) Protein identification and analysis tools in the ExPASy server. Methods Mol Biol 112:531–552

Zhang Z, Xing C, Wang L et al (2012) IndelFR: a database of indels in protein structures and their flanking regions. Nucleic Acids Res 40:512–518. doi:10.1093/nar/gkr1107

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Immediate publication on acceptance
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com