Homology Modeling in the Twilight Zone: Improved Accuracy by Sequence Space Analysis

Rym Ben Boubaker, Asma Tiss, Daniel Henrion, Marie Chabbert

Homology modeling in the twilight zone: Improved accuracy by sequence space analysis

Rym Ben Boubaker, Asma Tiss, Daniel Henrion and Marie Chabbert*

UMR CNRS 6015 – INSERM 1083, Laboratoire MITOVASC, Université d'Angers, Angers, France

* To whom correspondence should be addressed

Key Words: Molecular modeling, threading, twilight zone, profile-profile mining, cytokine, plethodontid receptivity factor

Abstract

The analysis of the relationship between sequence and structure similarities during the evolution of a protein family has revealed a limit of sequence divergence for which structural conservation can be confidently assumed and homology modeling is reliable. Below this limit, the twilight zone corresponds to sequence divergence for which homology modeling becomes increasingly difficult and requires specific methods. Either with conventional "threading" methods or with recent deep-learning methods, such as AlphaFold, the challenge relies on the identification of a template that shares not only a common ancestor (homology) but also a conserved structure with the query. As both homology and structural conservation are transitive properties, mining of sequence databases followed by multidimensional scaling (MDS) of the query sequence space can reveal intermediary sequences to infer homology and structural conservation between the query and the template. Here, as a case study, we studied the plethodontid receptivity factor isoform 1 (PRF1) from *Plethodon jordani*, a member of a pheromone protein family present only in lungless salamanders and weakly related to cytokines of the IL6 family. A variety of conventional threading methods led to the cytokine CNTF as a template. Sequence mining, followed by phylogenetic and MDS analysis, provides missing links between PRF1 and CNTF and allows reliable homology modeling. In addition, we compare automated models obtained from web servers to a customized model to show how modeling can be improved by expert information.

1 Introduction

Since the resolution of the myoglobin structure in 1958 [1], the number of protein structures deposited in the Protein Data Bank [2] has increased exponentially to reach more than 160 000 structures in 2020. These structures led to a better understanding of protein functions and mechanisms of action. They have paved the way to computational approaches for rational drug design, search of targetable allosteric sites, better understanding of structure-function relationships, and so on. However, in spite of the huge advances in the field, the sequence space increases much more than structural space and computational approaches towards many proteins still rely on molecular modeling.

Presently, based on available structural information and deposited structures, proteins (or protein regions) can be classified into four categories: (1) proteins with resolved structures, (2) proteins with

closely related structurally resolved homologs that can be straightforwardly modeled by homology, (3) proteins within the twilight zone, for which structural information can be reached in spite of the absence of close structurally resolved homologs, and (4) proteins within the dark zone that lack similarity to any known structure and are inaccessible to homology modeling [3]. A recent analysis of the human genome, carried out with the deep-learning based AlphaFold program, revealed that on a residue basis, 58% of total residues have a resolved structure or are modelled with high confidence, whereas about 20% are dark and the remaining ones are in the twilight zone [4].

The initial studies from the 1990s on the relationship between sequence and structural evolution remain valid today. In a hallmark paper, Sander and Schneider [5] determined a curve describing the limit of confidence between evolution of the sequence and conservation of the structure. This study was extended by Rost [6] who corroborated the main result: a length dependent cutoff line separates close homologs with structural conservation from remote homologs with unknown structural similarity. The length of the aligned sequence is a key factor for confidence level in structural conservation. A cutoff of about 20% for a sequence of 200 amino acids or longer separates the safe zone for homology modeling from the twilight zone (Fig. 1). This cutoff does not mean that homology modeling is not possible in the twilight zone, but it points to the additional difficulties that arise in the twilight zone.

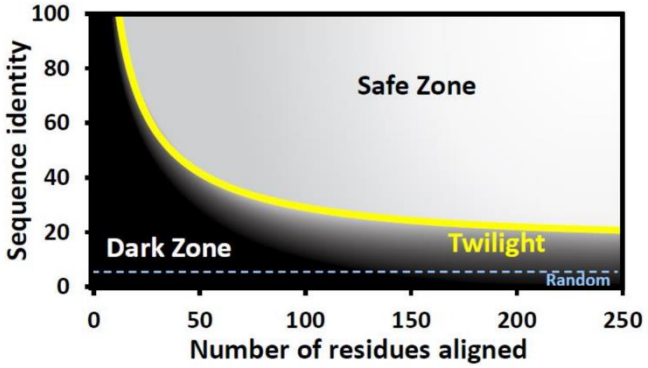


Fig. 1: Schematic representation of the dark, twilight and safe zones for molecular modeling of a protein as a function of the sequence identity and aligned length between the query and the template. Above the yellow line (drawn from [5]), the light grey zone indicates the safe zone of homology modeling. Below the yellow zone, the dark grey zone indicates the twilight zone for which molecular modeling becomes increasingly difficult because sequence identity does not infer structural conservation. In the twilight zone, templates cannot be found with BLAST and threading methods must be used. When they fail, the dark zone is reached. Note that any two proteins have at least 5% sequence identity (dashed blue line).

Homology modeling is based on evolution. Proteins do not arise from scratch and can be classified into families. Homologous proteins within a family evolved from a common ancestor and share sequence, structural, and, to a lesser extent, functional similarities [7]. The structural conservation within a family is the keystone of molecular modeling. Using known structures of homologous proteins (the templates) and a multiple sequence alignment between the query and the templates, homology modeling programs such as MODELLER [8] build restraints and optimize the query structure from the template structures. In the safe zone, the sequence identity between the query and putative templates is high enough to infer that these proteins belong to the same family and share a common structure. By contrast, in the twilight zone, the similarities may arise from chance, convergence or common ancestry, which raises several issues:

(1) Finding templates by straightforward sequence-sequence comparison methods is not possible. This point has led to the development of "threading" methods based either on compatibility of sequence and fold or on profile-based search (see below).
(2) Additional information may be necessary to find evidence of common ancestry (homology) between template and query and infer the conservation of the structure. Indeed, very low identity rates may correspond to divergent or convergent evolution. In addition, random alignment leads to 5% sequence identity whereas, in highly divergent families, sequence identities can be as low as 8% (see an example in Fig. 2).
(3) Alignment between query and template(s) may be difficult. Indeed, the cutoff also separates easy from tricky alignments which may alter the quality of the homology modeling. Alignments are greatly improved by the use of multiple sequence alignment methods [9,10] but their accuracy remains challenging at low sequence identities.

Thus, the challenge of molecular modeling in the twilight zone relies on the recognition of correct templates and generation of accurate sequence-template alignments [6]. In this chapter, for clarity purpose, we will use, as a case study, the plethodontid receptivity factor isoform 1 (PRF1) from *Plethodon jordani*. This protein, for which structural and evolutionary data are missing, was discovered in 1999. It is a member of a pheromone protein family present only in lungless salamanders, with weak similarity with cytokines of the IL6 family [11]. We will show how sequence analysis methods, in particular multidimensional scaling (MDS), support the homology and structural conservation between PRF1 and cytokines of the IL6 family, by revealing intermediary sequences. We will also show how modeling of PRF1 can be improved by a variety of techniques.

2 Materials

2.1 Databases

1. UniProt (https://www.uniprot.org/) is a comprehensive resource of protein sequences and functional information [12]. It is composed of the manually curated SwissProt and of the automatically annotated trEMBL repositories. It contains not only protein sequences but also additional information including related 3D structures or models, and identifiers of the protein family in different family databases such as Pfam [13] and InterPro [14].
2. The Protein Data Bank (PDB, accessible at https://www.rcsb.org/) is the repository of biological macromolecular structures [2,15].
3. SCOP (Structural classification of proteins) [16] is a repository of protein folds, based on an initial classification into five structural classes: all alpha, all beta, alpha/beta, alpha+beta and small proteins.

2.2 Sequence analysis

1. NRDB90.pl [17] is a perl script aimed at clusterizing sequences based on sequence identity to build non-redundant sets. It can be downloaded from ftp://biodisk.org .
2. Different programs such as CLUSTAL [18], MUSCLE [19,20] and T-COFFEE [21] can be used to perform multiple sequence alignments (MSA).
3. The EXPRESSO program from the T-COFFEE suite [22] provides a multiple sequence alignment based on structural alignments, which may be a useful initial step for aligning proteins with low identities (http://tcoffee.crg.cat/apps/tcoffee/do:expresso).

4. MSA can be manually edited using Genedoc [23], a program aimed at editing and analyzing MSA through a graphical interface and available at https://genedoc.software.informer.com/ . Subsequent phylogenetic analysis can be performed by the user-friendly MEGA software (Molecular Evolutionary Genetics Analysis) [24] available at https://www.megasoftware.net/ .
5. The R package Bios2mds [25] is aimed at analyzing MSAs by multidimensional scaling and provides tools for user-friendly visualization of the sequence space (see Note 1).

2.3 Template mining

A variety of programs can be used to mine homologs in sequence databases. The choice of the program depends on the putative sequence identities between the query and the hits. Sequence - sequence comparison programs are adequate in the safe zone whereas sophisticated profile - profile searches are adapted to the twilight zone. Initially, "threading" referred to the search of a template by analyzing the compatibility of a sequence with a protein fold. Presently, "threading" refers to any method searching a template by sequence-profile or profile-profile comparison.

Here is a non-exhaustive list of sequence database mining programs:

1. BLAST (Basic Local Alignment Search Tool) [26], based on local sequence similarity, allows fast sequence-sequence comparison.
2. PSI-BLAST (Position-Specific Interactive BLAST) [27] is based on sequence-profile comparison. It derives a position-specific scoring matrix (PSSM) from the multiple sequence alignment of sequences detected above a given score threshold using protein–protein BLAST.
3. HMMER [28] is a sequence-profile comparison method based on profile hidden Markov models (HMMs) (https://toolkit.tuebingen.mpg.de/tools/hmmer).
4. Phyre2 (Protein Homology/analogY Recognition Engine V 2.0) [29] performs its searches by mining a database of profile HMMs, one for each known 3D structure (http://www.sbg.bio.ic.ac.uk/~phyre2/).
5. HHPred [30-32] performs HMM-HMM profile searches in sequence databases to find homologs (https://toolkit.tuebingen.mpg.de/tools/hhpred).
6. LOMETS (local meta-threading-server) [33] performs template searches using eleven different threading methods (see Note 2). Starting from a query sequence, LOMETS works in three steps: (1) Building a sensitive (or deep) MSA, (2) Threading the deep MSA by individual programs, and (3) Ranking templates with a specific scoring function which takes into account normalized Z-scores and sequence identities. For each method, the normalized Z-scores differentiate good/bad templates (threshold of 1) (https://zhanglab.ccmb.med.umich.edu/LOMETS/).
7. SUPERFAMILY finds a protein fold based on a collection of hidden Markov models, which represent structural protein domains at the SCOP superfamily level [34,35]. SUPERFAMILY is a "true" threading program, aimed at finding a protein fold (https://supfam.org/SUPERFAMILY/).

2.4 Secondary structure prediction

When searching information for a protein without close structurally resolved homologs, prediction of secondary structure (SS) may yield useful information. Best performances for SS prediction are obtained with programs based on multiple sequence alignment profiles and neural networks, such as PSIPRED [36,37] (http://bioinf.cs.ucl.ac.uk/psipred/) and SPIDER3 [38] (https://sparks-lab.org/server/spider3/).

2.5 Automated structure prediction

Several web servers can perform automated structural prediction from a query sequence. They are based on threading methods to find templates, and then they use different methods for the subsequent modeling steps. Here, we list three automated 3D prediction servers compared in this Chapter:

1. Phyre2 [29] : After template detection by mining a HMM database, the subsequent molecular modeling step is carried out with MODELLER based on the resulting sequence alignment between query and templates (http://www.sbg.bio.ic.ac.uk/~phyre2/).
2. I-TASSER (Iterative Threading ASSEmbly Refinement) [39-41] is a hierarchical approach to protein structure prediction. It first identifies structural templates by comparing the best hits from the "best" ten out of fourteen threading methods with LOMETS, and then it builds full-length atomic models by iterative template-based fragment assembly simulations (https://zhanglab.ccmb.med.umich.edu/I-TASSER/).
3. ROBETTA [42] is a protein structure prediction service. In the option for Rosetta Comparative Modeling (RosettaCM) [42], four independent methods (see Note 3) are used to detect templates and generate sequence alignments, and then models are built from template hybridization (https://robetta.bakerlab.org).

2.6 Customized structure prediction

For users who wish to build their own models, the MODELLER program [8] builds molecular models of the query from the template structure(s) by minimizing structural, stereochemical and user-defined restraints. The structural restraints are based on the structure of the template(s) and the alignment between query and template(s). To customize models in order to match structural and functional requirements, expert information can be introduced by (1) adding user-defined restraints such as distance between two residues or secondary structure elements in the modeling procedure and (2) by combining user-selected templates and template fragments.

2.7 Model validation

With threading methods, the metrics to compare models and templates must be more sensitive to the global fold similarity than to local structural variation. This is not the case of the traditional root-mean-square deviation (RMSD). The TM-score (see Note 4) has been specifically designed to solve this problem [43]. A threshold of 0.5 differentiates proteins with similar fold from proteins with different fold [44]. TM-scores can be calculated after structural alignment with TM-align [45] at the Zhang lab server (https://zhanglab.ccmb.med.umich.edu/TM-align/).

2.8 Graphical analysis

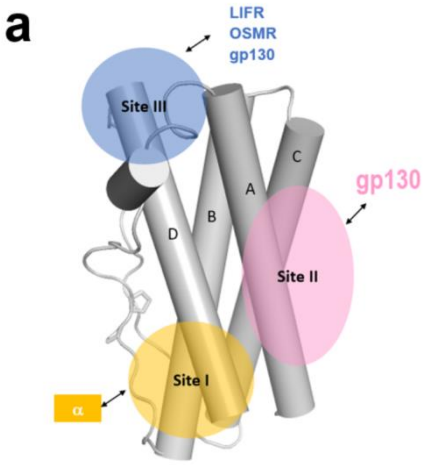
Graphical analysis of templates and models can be carried out by a variety of molecular visualization programs, such as Chimera [46] or Pymol (https://pymol.org/). Note that the low identity rates between template and query sequences in the twilight zone prevent the use of sequence-based structural superposition functions, such as the align function in Pymol.

3 Methods

3.1 Our case study

As a case study, we chose a small protein, the plethodontid receptivity factor isoform 1 (PRF1) from *Plethodon jordani* (UniProt entry: Q9PUJ2_PLEJO) [11]. This 215 amino acid protein (including a 23 amino acid peptide signal) is a courtship pheromone produced by males to increase female receptivity (see Note 5). When PRF1 was discovered in 1999, it was acknowledged that its sequence was weakly related (around 16% sequence identity) to the ciliary neurotrophic factor (CNTF) and cardiotrophin-1 (CTF1), two cytokines of the interleukin-6 (IL6) family [11]. Since then, two additional cytokines of the IL6 family have been discovered: cardiotrophin-2 (CTF2, absent in humans) [47] and the cardiotrophin-like cytokine factor 1 (CLCF1) [48]. CTF2 and CLCF1 have, respectively, 26% and 20% sequence identity with PRF1.

In mammals, the IL6 family of cytokines includes IL6, interleukin-11 (IL11), leukemia Inhibitory factor (LIF), oncostatin M (ONCM), CNTF, CTF1, CTF2 and CLCF1. Albeit the sequence identities can be as low as 8%, these cytokines share a common four-helix bundle fold with an up-up-down-down topology (Fig. 2). In addition, they signal through the gp130 receptor subunit and share similar binding sites with cognate receptors (see Note 6) [49-52]. Crystal structures have been resolved for five cytokines from the IL6 family: IL6 (1ALU [53], 5FUC [54]), IL11 (4MHL [55]), CNTF (1CNT [56]) , LIF (1EMR, 1LKI [57], 2Q7N [58], and ONCM (1EVS [59]). No crystal or NMR structure has been reported to date for PRF1 or cardiotrophin-like cytokines.



b

	PRF_PLEJO	K7FE02_PELSI	M7BAK3_CHEMY	A0A3Q0FUK7_ALLSI	H3AGE1_LATCH	CTF2_MOUSE	CLCF1_HUMAN	CTF1_HUMAN	CNTF_CHICK	CNTF_HUMAN	LIF_HUMAN	ONCM_HUMAN	IL11_HUMAN
K7FE02_PELSI	28												
M7BAK3_CHEMY	28	85											
A0A3Q0FUK7_ALLSI	27	56	51										
H3AGE1_LATCH	31	46	47	36									
CTF2_MOUSE	26	53	49	46	42								
CLCF1_HUMAN	20	27	28	28	28	28							
CTF1_HUMAN	17	29	28	35	24	29	21						
CNTF_CHICK	20	24	23	26	22	23	18	22					
CNTF_HUMAN	16	21	22	18	19	19	17	19	44				
LIF_HUMAN	16	17	17	14	18	16	14	18	12	12			
ONCM_HUMAN	14	14	13	13	10	14	13	12	14	12	15		
IL11_HUMAN	8	16	16	15	13	20	17	12	13	10	8	13	
IL6_HUMAN	11	9	8	8	7	10	12	8	10	10	8	12	13

Fig. 2: Cytokines of the IL6 family. (a) General up-up-down-down topology of the four-helix bundle fold of this cytokine family. The helices are numbered from A to D. The positions of the three conserved sites of interaction with the cognate receptors are indicated. Site II interacts with gp130, site III interacts with either gp130, LIFR or OSMR, while site I can interact with a third, specific, "α" receptor. See Note 6 for details; (b) Sequence identities between PRF1, its closest homologs and the cytokines of the IL6 family. The color code indicates the reliability of the sequence identity (green: safe zone, yellow: transition zone, red: twilight zone).

3.2Template search by PDB mining

In InterPro, PRF1 is described as belonging to the *4_helix_cytokine-like_core* superfamily (IPR009079) and to the *PRF/Cardiotrophin-like* family (IPR010681). Additional information from the PRF1 sequence was searched for using the SUPERFAMILY assignment server [34]. SUPERFAMILY predicts that PRF1 is in the class of *All alpha proteins*, and belongs to the fold/superfamily of *4-helical*

cytokines with an E-value of 3 10^{-55} (see Note 7). It also suggests an "uncertain" classification for the family level as *Long-chain cytokine* with an E-value of only 0.005.

The next step to find a template was the mining of the PDB in search of homologs. Using the mature sequence of PRF1 (residues 24-215) as a query, we performed different searches (Table 1):

1. Straightforward mining of the PDB with BLAST: this search led to no hit
2. Sequence - profile search: A PSI-BLAST search seeded with the PRF1 sequence, followed by selection of hits on most query sequence (>60%), led to CNTF as a hit with an unreliable E-value of 5.9
3. Profile HMM search: Profile search using the HMMER program [28] was carried out on the HHpred server. The search led to two hits, CNTF and LIF, as putative templates with very significant E-values of 10^{-10} or lower
4. Profile HMM - profile HMM search: the HHpred algorithm led to six hits with E-values lower than 0.1: CNTF, LIF, ONCM, IL11, GCSF (granulocyte colony-stimulating factor) and IL6. Among them, GCSF (E-value of 8 10^{-13}) is a 4-helical cytokine that does not belong to the IL6 family, but shares the same up-up-down-down four-helix bundle fold.

TABLE 1

PDB mining using the PRF1 sequence as a query

Search method	Program	Hits[1]	E-value
Sequence based	BLAST	No hit	
Sequence profile based	PSI-BLAST	CNTF	5.9
Profile HMM based	HMMER	CNTF	7 10^{-20}
		LIF	4 10^{-10}
Profile HMM – Profile HMM based	HHpred	CNTF	7 10^{-32}
		LIF	3 10^{-31}
		IL11	5 10^{-28}
		ONCM	8 10^{-26}
		GCSK	8 10^{-13}
		IL6	2 10^{-7}

[1] For clarity purpose, only the proteins (and not the PDB numbers) are indicated. Italics fonts indicate a growth factor with the same four-helix bundle fold as the IL6 family.

3.3 Template search with LOMETS

Finally, a comparison of 11 threading methods was carried out with the LOMETS server [33] (Table 2). All the methods, but one, classified CNTF or LIF as best hits. Only CEthreader, which is a contact-based method, privileged prolactin (PDB 1RW5). This growth factor shares the four-helix bundle fold of the IL6 cytokines. Most additional hits include cytokines of the IL6 family (ONCM, IL11, IL6) or cytokines/growth factors with same four-helix bundle fold (prolactin, lactogen, IL23). However, several methods also found IL1Ra, the interleukin-1 receptor antagonist (PDB 1ILR) [60] that has a beta barrel fold (see Note 8). This finding serves as a reminder of how cautious users need to be when analyzing threading results.

TABLE 2

Comparison of the LOMETS threading programs using the PRF1 sequence as a query

Program[1]	Method[2]	Top hit[3]	Additional hits with $Z_n > 1$[3,4]
HHpred	HMM	CNTF	LIF, IL11, ONCM
CEthreader	Contact	*Prolactin*	IL11, CNTF, *lactogen*, LIF, **IL1Ra**
SparksX	Profile	CNTF	LIF, IL11, ONCM, **IL1Ra**, IL6, *prolactin*
FFAS3D	Profile	CNTF	LIF, ONCM, IL11, *GCSF,* IL6, **IL1Ra***, prolactin, lactogen*
MUSTER	Profile	LIF	CNTF, ONCM, IL11
Neff-MUSTER	Profile	LIF	CNTF
HHsearch	HMM	LIF	CNTF, ONCM, IL11, *GCSF*, IL6, **IL1Ra**
SP3	Profile	LIF	CNTF, ONCM, IL11, IL6, *GCSF*, **IL1Ra**, *prolactin*
PPAS	Profile	CNTF	LIF, ONCM, IL11, *GCSF*, IL6, **IL1Ra**, *IL23*
PROSPECTOR2	Profile	CNTF	LIF, ONCM, IL11
PRC	HMM	LIF	CNTF, ONCM, IL11

[1] The programs are ranked as determined by LOMETS. See Note 2 for references.

[2] HMM corresponds to profile HHM – profile HMM based searches; profile corresponds to sequence profile – sequence profile based searches.

[3] For clarity purpose, only the proteins (and not the PDB numbers) are indicated.

[4] Normalized Z-scores (Z_n) indicate the quality of the hits. They are considered "good" above the threshold of 1. The hits are sorted by decreasing Z_n. When several hits correspond to the same protein (different origins, conditions or methods), only the first hit is indicated. Italics fonts indicate four-helix bundle cytokines/growth factors that do not belong to the IL6 family but share the same fold. The bold fonts for IL1Ra highlight a hit with a beta barrel fold.

3.4 Sequence space investigation

As exemplified by IL1Ra in LOMETS results (Table 2), finding a template by threading does not prove that there is homology, i.e. a common ancestor, between the template and the query, nor that the structure is conserved. The identity of 16% between the PRF1 query and the CNTF or LIF templates is positioned in the twilight zone (Fig. 1). However, homology and structural conservation are transitive properties. Investigation of the sequence space of the query analogs may reveal sequences homologous to both query and template(s) in the safe zone and consequently validate the template. Indeed, finding intermediates is an efficient strategy to reduce false positives [6,30].

To investigate the query sequence space, several steps must be carried out:

1. Blast search of the query homologs in sequence databases. Here, using PRF1 as a query in UniProt vertebrate sequences, we obtained 602 hits with E-value lower than 10. Among them, 190 sequences corresponded to salamander receptivity factors and shares with PRF1 sequence identities larger than 60%. Among very significant hits (E-value < 10^{-10}), several sequences are known cytokines: human and mouse CLCF1, human and mouse CTF1, mouse CTF2. Interestingly, the sequence of chicken CNTF was identified with an E-value of 5 and a sequence identity of 20%.
2. Building of a non-redundant (NR) set of the hits. This step considerably reduces the number of sequences investigated by suppressing non informative, highly similar sequences. Here, we built a NR set of PRF1 hits (id <90%), first by selecting sequences with length between 150 and 360 aa, and then, by clustering these sequences using the NRDB90.pl script [17]. The MSA of the NR set

was carried with Clustal [18] and manually edited with Genedoc [23]. At this stage, a few additional truncated sequences were removed. This led to a non-redundant set of 125 aligned sequence sequences to which we added the human CNTF sequence. Computation of sequence identities on this alignment revealed that the closest neighbors to PRF1 are proteins from coelacanth (H3AGE1_LATCH, 31% id), turtle (K7FE02_PELSI, M7BAK3_CHEMY, 28%) and alligator (A0A3Q0FUK7_ALLSI, 27%).

3. Building a phylogenetic tree of the NR set. Using the alignment of the NR set, a Neighbor Joining tree was built with the MEGA software [24] (Fig. 3a). This tree visualizes the PRF, CTF1, CTF2, CLCF1 and CNTF sub-families along with the closest neighbors of PRF1.
4. Visualizing the sequence space of the non-redundant set. Using the bios2mds R package [25], we could visualize the sequence space of the PRF1 homologs by multidimensional scaling (Fig. 3b). The first two dimensions are driven by the differences between CTF1, CTF2 and CLCF1 cytokines. CNTFs and PRFs are projected towards the center of the first two dimensions. To better visualize the relationship between these proteins, we projected the sequences onto the third and fourth dimensions. In this case, we observed the closest neighbors of PRF1 at intermediary positions between PRFs and CNTFs. Indeed, these sequences are above the twilight threshold for human and/or chicken CNTF, strongly supporting the assumption of homology between PRF1 and CNTF (Fig. 2b).

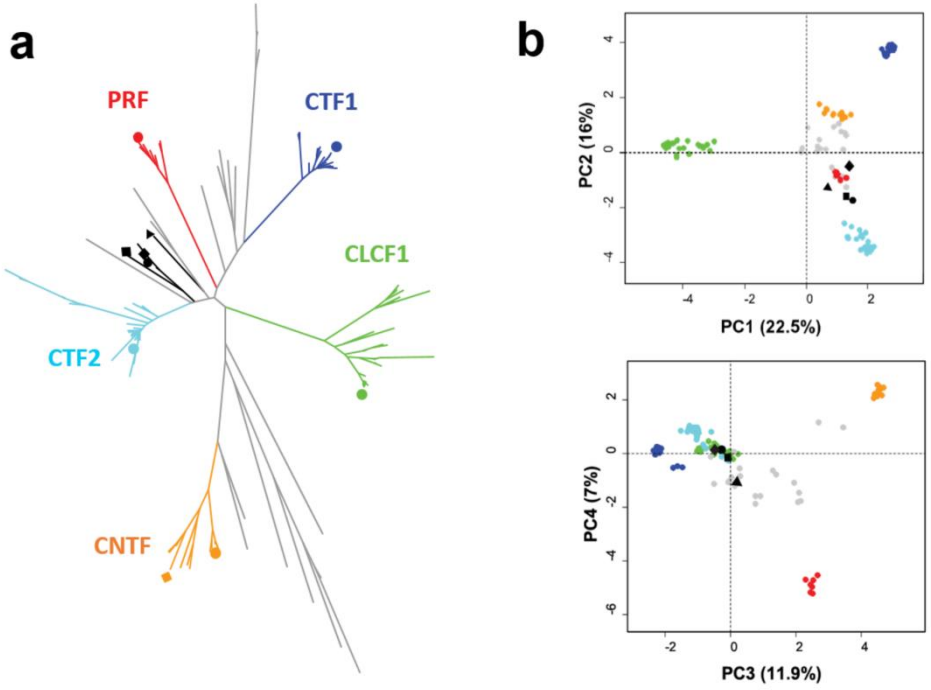


Fig. 3: Evolutionary information on PRF1. (a) NJ tree of PRF1 homologs (500 bootstraps); (b) Sequence space of the PRF1 homologs with projection of the sequences onto the first and second dimensions (top) and onto the third and fourth dimensions (bottom) of the MDS analysis. The PRF1 homologs were obtained by mining UniProt with BLAST, followed by clustering to obtain a non-redundant set. The MDS analysis was carried out with the bios2mds package. In (a) and (b), the color code is as follows: PRF proteins: red, CNTF: orange, CTF1: dark blue, CTF2: magenta, CLCF1: green, the closest four homologs: black, others: grey. In (a), labels indicate PRF1 (red circle), human CTF1 (blue circle), murine CTF2 (cyan circle), human CLCF1 (green circle), human CNTF (orange diamond), chicken CNTF (orange circle), K7FE02_PELSI (black circle), M7BAK3_CHEMY (black square), H3AGE1 LATCH (black triangle) and A0A3Q0FUQ7_ALLSI (black diamond). In (b), only black labels for closest PRF1 homologs are shown.

3.5 Template analysis

As several structures of IL6 cytokines are available, careful graphical analysis may give information on conserved and variable elements that should be taken into account for modeling. Models are less reliable in the variable parts and more reliable in variable parts.

Fig. 4a displays the crystal structure of human CNTF superimposed with structure of LIF, IL6, ONCM and IL11. The superposition was carried out with the super function of Pymol, based on structures because the low sequence identities between these cytokines prevented the use of the align function based on sequence identities.

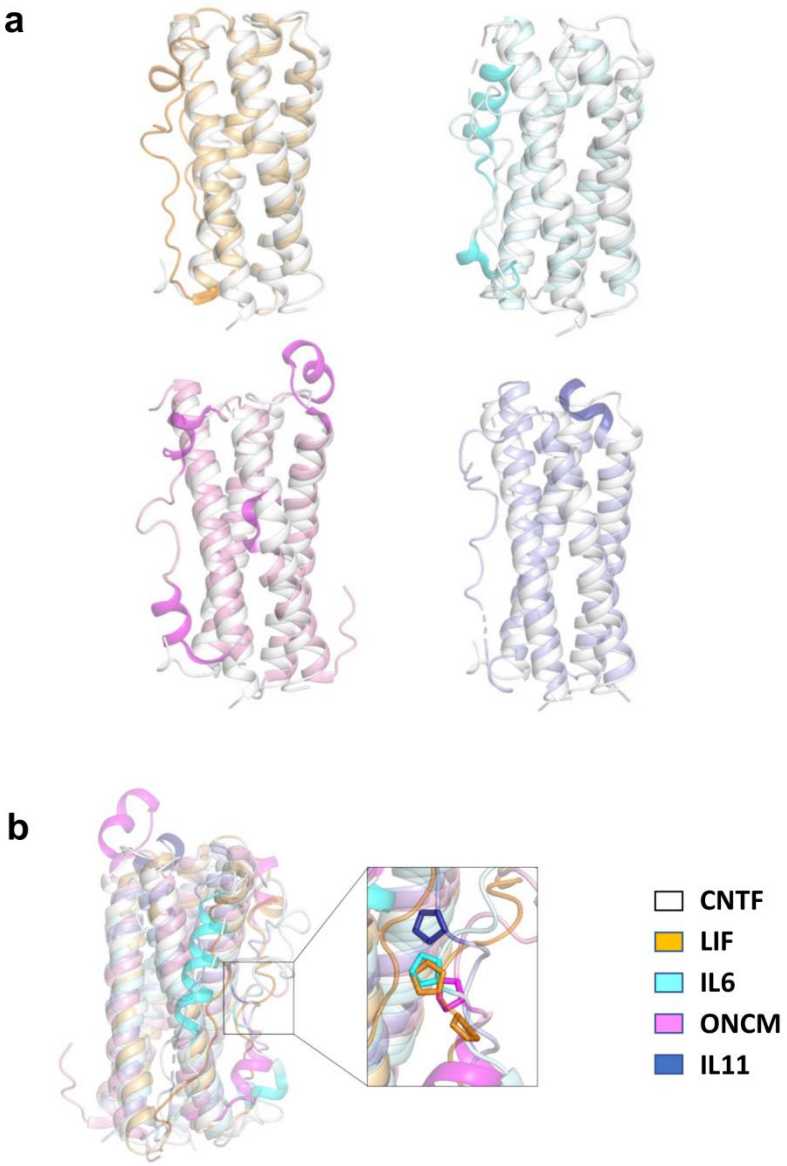


Fig. 4: Comparison of different templates. In (a), a ribbon representation of CNTF (white) is superimposed with LIF (orange), IL6 (cyan), ONCM (magenta) and IL11 (blue). Regions structurally different as compared to CNTF are indicated by a darker color for clarity purpose. In (b), zooming on the AB loop reveals the conserved positioning of a proline residue (shown as stick) in the cleft between helices B and D in LIF, IL6, ONCM and IL11. Note that the loop AB is absent in the CNTF structure. The PDB files are 1CNT (CNTF), 1EMR (LIF), 5FUC (IL6), 1EVS (ONCM) and 4MHL (IL11).

The general topology of the four-helix bundle is conserved but specific properties are observed. LIF has shorter helices A and D on the "bottom" side but otherwise does not present marked differences. Shorter helices A and D are also observed for IL6. In addition, IL6 has shorter helices B and C on the "top" side and a 12 residue long α-helix in loop CD. ONCM has a strong distortion in helix A, an insertion with two helical turns in loop BC and two helical elements in loop AB. A helical element is also observed in loop BC of IL11 and distorts the C-terminus of helix B.

3.6 Sequence alignment

MODELLER uses spatial restraints based on the alignment between query and template(s) to build a model. The alignment step is thus crucial for the quality of the model. In the twilight zone, straightforward alignment of the query and the template should be avoided as they do not lead to reliable alignment. It was early acknowledged that MSA methods lead to improved quality of the alignment [9]. Here are some tips:

1. First, use large homolog sets to build an MSA. Then, manually inspect the MSA, compare it to structural alignment of the templates and correct it if necessary with Genedoc [23]. This should limit template/query alignment errors.
2. You may use EXPRESSO [22] to obtain an initial sequence alignment of all the putative templates based on their structural alignment. Compare the sequence alignment to the superimposed structures and correct if necessary.
3. Subsequently, you may use this template profile as a seed to align the query and its homologs.
4. Visual inspection of MSA is mandatory to insure that the alignment is correct.
5. In any case, never forget that sequence alignment may be the limiting step of homology modeling.

3.7 Loop modeling

Long loops are challenging to reliably model due to combination of different factors: (1) they have high sequence and structural variability, (2) they are frequently missing in templates (e.g. the CNTF template in which loops AB and CD are missing and partly missing, respectively), and (3) the functions developed for loop modeling, such as model_loop in MODELLER [61], are suited to short loops but not long loops. Two options are possible to improve modeling of long loops: (1) SS prediction programs may suggest SS elements in these long loops that can be introduced in the modeling procedure; and (2) loops with the same (or, if not possible, similar) size in related proteins may be used as templates.

In our case study, the long loops joining helices A and B, and helices C and D are highly variable and they are missing or partially missing in all the templates except LIF (Fig. 5). However, we note that the loops AB of LIF, ONCM and IL6 present helical elements. In loops CD, a long helix is observed only for IL6. To gain information on putative secondary elements in the long loops of PRF1, we used the SS prediction programs PSIPRED [36] and SPIDER3 [38]. Both programs predicted a helical element at the C-terminus of loop AB that should be taken into account for modeling (Fig. 5). In addition, careful scrutiny of the templates indicates the strong conservation of a proline residue from loop AB at the cleft between helices B and D (Fig. 4b). This position might tether the loop AB at the cytokine surface and should be an element to take into account for PRF1 modeling.

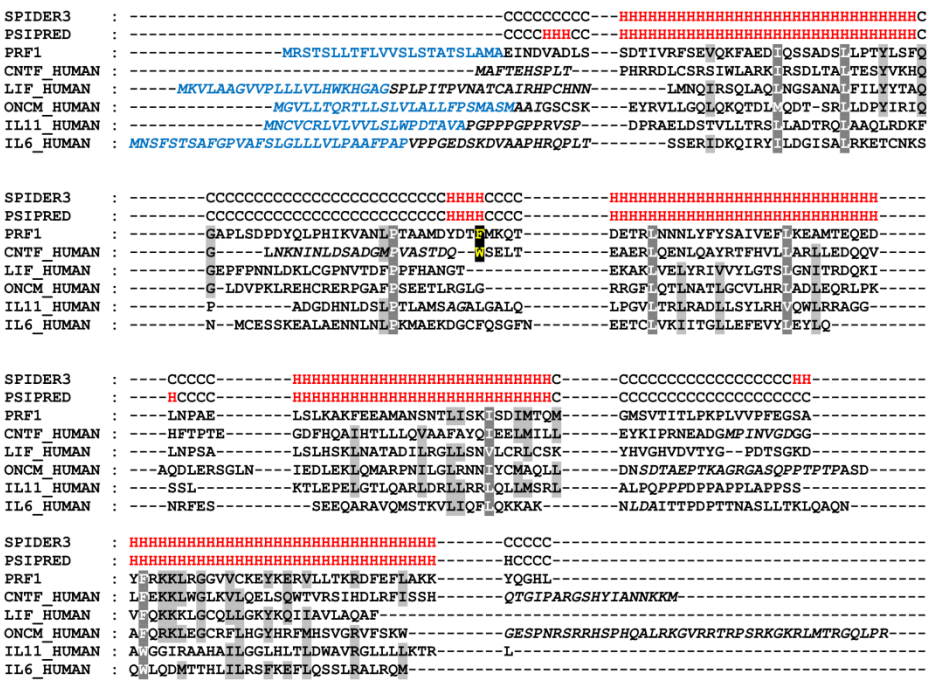


Fig. 5: Alignment of PRF1 versus a structural alignment of the templates. The initial alignment based on structure superposition was obtained with EXPRESSO and manually refined. The SS predictions were obtained by SPIDER3 and PSIPRED from the sequence of PRF1. Residues in blue indicate the sequences of the peptide signal. Residues in italic are missing in the PDB files. The homologous CNTF Trp-64 and PRF1 Trp-70 residues are highlighted in yellow on black background because this position may be involved in the binding of a specific "α" receptor.

3.8 Model building

In this last step, we compare the models of PRF1 obtained with three automated modeling procedures through web servers (Phyre2, RosettaCM and I-TASSER) to a customized model built with MODELLER. Concerning the automated models, the four-helix bundle fold was obviously recognized in the three cases, although RosettaCM did not provide information on the templates used. I-TASSER uses the first hits from the best ten out of fourteen threading programs: CNTF (best hit for four programs), LIF (best hit for five programs) and ONCM (best hit for one program). In Phyre2, CNTF, LIF, IL11 and ONCM were found with a confidence level ≥ 99.8%.

The customized PRF1 model was built with MODELLER with the following parameters:

1. LIF (PDB 1EMR) and CNTF (PDB 1CNT) were used as templates after structural comparison (Fig. 4).
2. Loop CD was modeled by homology with LIF, with the two residue insertion positioned as indicated in Fig. 5.
3. Loop AB of PRF1 contains a six residue insertion compared to LIF. Thus, the N-terminus of loop AB (up to the anchor proline) was built from the LIF template and the insertion was positioned at the C-terminus with a helical restraint corresponding to the helical prediction.
4. PRF1 Phe-70 is homologous to CNTF Trp-64 which is crucial for receptor alpha binding [62-64]. We did not constraint this residue in the modeling procedure but checked its orientation in the models as a criterion for structural validation.

Fig. 6 displays the three automated models superimposed on the customized MODELLER model. No significant difference was observed for the four-helix bundle core, but the loops AB were markedly different. Among automated models, only the Phyre2 model predicted a position of Pro-61 similar to the customized model. By contrast, only RosettaCM predicted a helical element at the C-terminus of loop AB. However, the exposed orientation of Phe-70, similar to the orientation of Trp-64 in CNTF, was observed only in the customized model, supporting the positioning of the helical element predicted by SPIDER3 and PSIPRED.

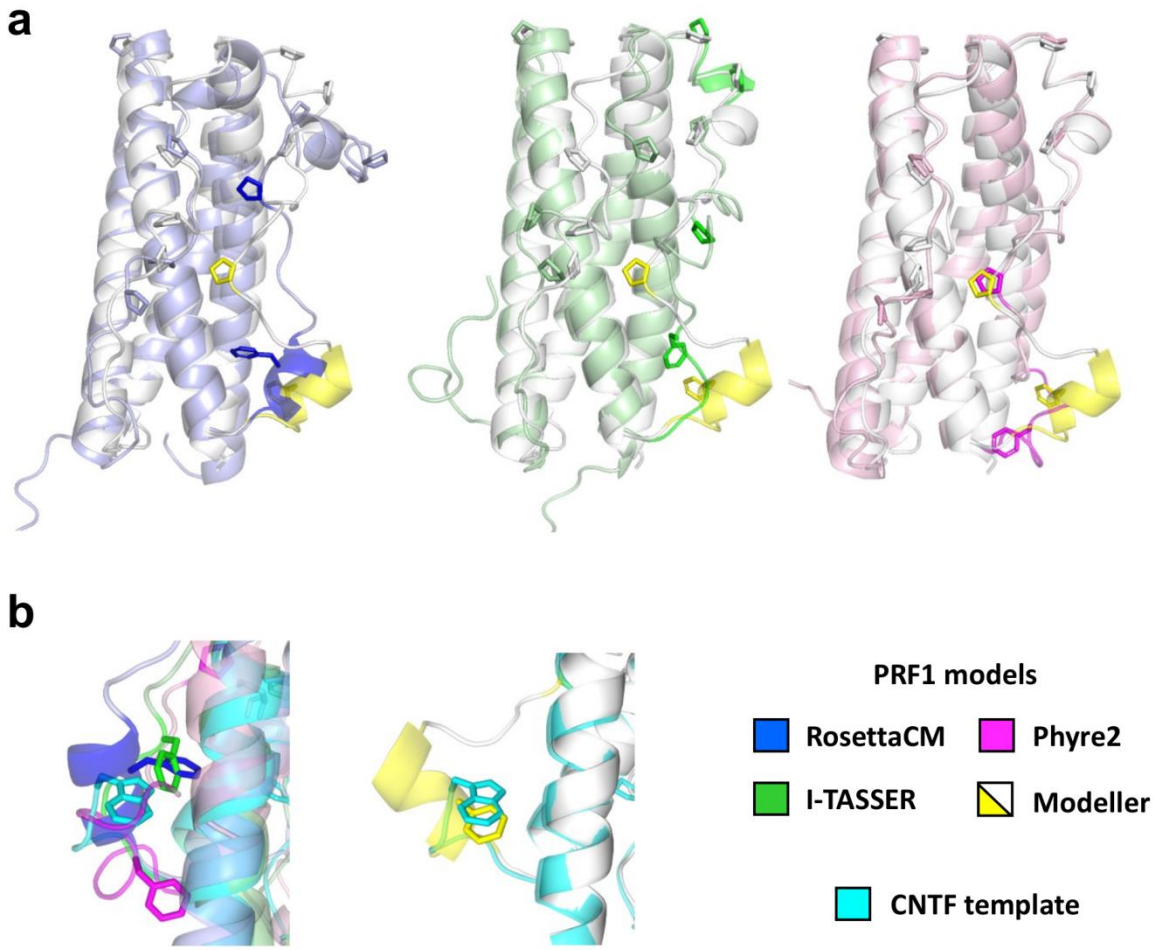


Fig. 6: Comparison of the PRF1 models. In (a), the customized MODELLER model (white and yellow) is superimposed with the best models obtained with RosettaCM (blue), I-TASSER (green) and Phyre2 (pink). Pro-61 and Phe-70 are indicated by sticks. The yellow ribbon in the MODELLER model indicates the user-added helical restrains. In (b), the zooming on the C-terminal part of loop AB reveals that the superposition of the phenyl ring of CNTF Trp-64 (cyan) with PRF1 Phe-70 (yellow) in the customized MODELLER model is not observed in the other three models.

The TM-scores between models and templates (Table 3) can assess the quality of the models. The higher scores obtained by the customized and the Phyre2 models with the CNTF and LIF templates, respectively, reflect the templates favored by MODELLER in the modelling procedure, while the I-TASSER and RosettaCM work by template based fragment assembly or model hybridization, and do not match a template as closely as the customized and Phyre2 models. In any case, the high TM-scores are consistent with similar folds and corroborate the “threadability” of PRF1 [65].

TABLE 3

TM-scores[1] of the PRF1 models

Method	CNTF	LIF
I-TASSER	0.71	0.76
RosettaCM	0.69	0.69
Phyre2	0.68	0.88
Customized	0.78	0.85

[1] The TM-scores between the PRF1 models and the CNTF and LIF templates were computed using the length of PRF1 as reference.

3.9 Concluding remarks

This chapter provides an example of homology modeling in the twilight zone, using conventional methods. It is important to note that all the threading methods used to find templates led to similar results (Tables 1 and 2). This is a general observation that threading methods succeed or fail together for molecular modeling of proteins in the twilight zone [65]. Indeed, "best" templates of any method, including CEthreader, share the up-up-down-down four-helix bundle fold of the IL6 family. Nevertheless, several methods found a hit with a beta barrel fold. Caution is thus always recommended when threading methods must be used. Here, the TM-scores obtained for the PRF1 models (Table 3) indicate high "threadability", which is consistent with the four-helix bundle fold recognized as "best" hit by the different threading methods. The sequence space analysis corroborates the homology between the query and the templates by finding intermediate sequences linking PRF1 and CNTF. The most difficult task remains the modeling of the long loops. In this case, additional information provided by the user or by deep learning, based on structural or functional criteria, may improve the quality of the model. While this chapter was in press, the PRF1 model obtained with AlphaFold was released (AF-Q9PUJ2-F1, available on Uniprot). Notably, this model predicted the same alpha-helical structure at the C-terminus of the loop AB (rmsd of 1.7 Å), with the same *trans* orientation of Phe-70 towards helix D, as our customized model (Supplementary Fig. S1). Thus, AlphaFold was able to automatically mine and integrate additional information (here secondary structure predictions) in its modeling procedure, as we have done manually. Finally, "unthreadability" appears as a property inherent to protein fold [65]. Exploration of the dark zone will require the development of *ab initio* approaches to deduce the three dimensional structure of a protein from physics and physico-chemical principles alone.

4 Notes

1. Multidimensional scaling is a multivariate analysis method that transforms a distance matrix into points in a low dimensional space. The distances between points in the resulting space are as close as possible to the distances in the original matrix [66]. When applied to an MSA, this method allows the 2D or 3D visualization of the sequences (the sequence space) based on the identity or on the similarity matrix of the sequences in the MSA.
2. LOMETS [33] compares eleven threading methods. Four of these methods (CEthreader, MUSTER, Neff-MUSTER and PPAS) have been developed by Zhang and coworkers [33,67,68]. The methods are based on contact searches (CEthreader), sequence profile – sequence profile searches

(MUSTER, Neff-MUSTER, FFAS-3D [69], PPAS, PROSPECT2 [70], SP3 [71], SparksX [72]) and profile HMM – profile HHM searches (HHsearch [31], HHpred [30], and PRC [73]).

3. In RosettaCM, the four independent threading methods used to detect templates are RaptorX [74], HHpred [30], Sparks-X [72], and Map align [75].
4. TM-score is a metric for assessing the topological similarity of protein structures. It gives more weight to smaller deviations than to larger deviations and is thus more sensitive to the global fold than traditional root mean square deviations. In addition, it is scaled by sequence length. TM-scores have values in [0,1] range, with 1 indicating a perfect match between two structures. Scores are higher than 0.5 for structures with similar folds and lower than 0.17 for randomly chosen unrelated proteins [41,43]. Interestingly, a TM-score threshold of 0.4 can also be used to differentiate "threadable" from "non threadable" proteins. The "threadability" of a protein appears related to inherent properties of the protein fold [42].
5. The *Plethodontidae* or lungless salamanders have complex mating behaviors and courtship rituals. PRF1 is produced in a gland on the male's chin and delivered to the female during courtship by scratching or slapping. A male using the scratching mode of delivery administers the pheromone by wiping his mental gland across the female's dorsum while scraping her skin to deliver the pheromone directly into the female's circulatory system. In the slapping delivery mode, the male slaps its gland directly across the female's nares to deliver the pheromone to chemoreceptors in the vomeronasal organ [76].
6. The cytokines of the IL6 family [49-51] are long chain four-helix bundle proteins. In spite of low sequence identities (8-20%), these cytokines share a common fold with an up-up-down-down topology for the four A to D helices. In mammals, the family includes IL6, IL11, LIF, ONCM, CNTF, CTF1, CTF2 (absent in humans) and CLCF1. These cytokines recruit two receptors, including the common gp130 receptor and a second variable receptor (gp130, LIFR or OSMR). The two receptors transactivate each other to trigger a cellular response through JAK/STAT pathways. In addition, a third receptor (the α receptor) without intracellular part may be recruited to increase the affinity. These receptors bind three conserved sites of the cytokines (Fig. 2). The site I located on helix D and loop AB binds the α receptor, site II located on helices A and C binds gp130 and site III at the top of helix D binds either gp130, LIFR or OSMR. The signature of LIFR binding is an FxxK motif at the N-terminus of helix D.
7. Expect-value or E-value describes the number of hits "expected" by chance when searching a database of a particular size. It depends on both the matching score and the length of the query. The closer to 0 the E-value is, the more significant the hit is.
8. Six threading programs (CEthreader, SparksX, FFAS3D, HHsearch, SP3, PPAS), based on three different search methods, find the interleukin-1 receptor antagonist (IL1Ra) as a putative "good" hit. This protein (PDB: 1ILR) has a very different beta trefoil structure [60]. The beta trefoil fold consists of six beta hairpins, each formed by two beta strands. Together they form a beta barrel with a triangular cap and an approximate three-fold symmetry.

Acknowledgments: This study was supported by institutional grants from INSERM, CNRS and University of Angers. MC is supported by CNRS. RB is supported by a fellowship from the University of Angers (France). AT is supported by a fellowship from the University of Carthage (Tunisia).

References

1. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 181:662-666. doi:10.1038/181662a0
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235-242. doi:10.1093/nar/28.1.235
3. (!!! INVALID CITATION !!!)
4. Perdigao N, Rosa A (2019) Dark Proteome Database: Studies on Dark Proteins. High Throughput 8 doi:10.3390/ht8020008
5. Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56-68. doi:10.1002/prot.340090107
6. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12:85-94. doi:10.1093/protein/12.2.85
7. Devos D, Valencia A (2000) Practical limits of function prediction. Proteins 41:98-107.
8. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779-815. doi:10.1006/jmbi.1993.1626
9. Wallace IM, Blackshields G, Higgins DG (2005) Multiple sequence alignments. Curr Opin Struct Biol 15:261-266. doi:10.1016/j.sbi.2005.04.002
10. Kemena C, Notredame C (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. Bioinformatics 25:2455-2465. doi:10.1093/bioinformatics/btp452
11. Rollmann SM, Houck LD, Feldhoff RC (1999) Proteinaceous pheromone affecting female receptivity in a terrestrial salamander. Science 285:1907-1909. doi:10.1126/science.285.5435.1907
12. UniProt C (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47:D506-D515. doi:10.1093/nar/gky1049
13. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD (2019) The Pfam protein families database in 2019. Nucleic Acids Res 47:D427-D432. doi:10.1093/nar/gky995
14. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang HY, El-Gebali S, Fraser MI, Gough J, Haft DR, Huang H, Letunic I, Lopez R, Luciani A, Madeira F, Marchler-Bauer A, Mi H, Natale DA, Necci M, Nuka G, Orengo C, Pandurangan AP, Paysan-Lafosse T, Pesseat S, Potter SC, Qureshi MA, Rawlings ND, Redaschi N, Richardson LJ, Rivoire C, Salazar GA, Sangrador-Vegas A, Sigrist CJA, Sillitoe I, Sutton GG, Thanki N, Thomas PD, Tosatto SCE, Yong SY, Finn RD (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res 47:D351-D360. doi:10.1093/nar/gky1100
15. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, Christie C, Dalenberg K, Duarte JM, Dutta S, Feng Z, Ghosh S, Goodsell DS, Green RK, Guranovic V, Guzenko D, Hudson BP, Kalro T, Liang Y, Lowe R, Namkoong H, Peisach E, Periskova I, Prlic A, Randle C, Rose A, Rose P, Sala R, Sekharan M, Shao C, Tan L, Tao YP, Valasatava Y, Voigt M, Westbrook J, Woo J, Yang H, Young J, Zhuravleva M, Zardecki C (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res 47:D464-D474. doi:10.1093/nar/gky1004
16. Andreeva A, Kulesha E, Gough J, Murzin AG (2020) The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res 48:D376-D382. doi:10.1093/nar/gkz1064
17. Holm L, Sander C (1998) Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 14:423-429. doi:10.1093/bioinformatics/14.5.423
18. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947-2948. doi:10.1093/bioinformatics/btm404

19. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 47:W636-W641. doi:10.1093/nar/gkz268
20. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113. doi:10.1186/1471-2105-5-113
21. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302:205-217. doi:10.1006/jmbi.2000.4042
22. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V, Notredame C (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res 34:W604-608. doi:10.1093/nar/gkl092
23. Nicholas KB, Jr NHB, Deerfield DWI (1999) GeneDoc: Analysis and Visualization of Genetic Variation. EMBNEW.NEWS 4:14.
24. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-2739. doi:10.1093/molbev/msr121
25. Pele J, Becu JM, Abdi H, Chabbert M (2012) Bios2mds: an R package for comparing orthologous protein families by metric multidimensional scaling. BMC Bioinformatics 13:133. doi:10.1186/1471-2105-13-133
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-410. doi:10.1016/S0022-2836(05)80360-2
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402. doi:10.1093/nar/25.17.3389
28. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755-763. doi:10.1093/bioinformatics/14.9.755
29. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10:845-858. doi:10.1038/nprot.2015.053
30. Soding J, Remmert M (2011) Protein sequence comparison and fold recognition: progress and good-practice benchmarking. Curr Opin Struct Biol 21:404-411. doi:10.1016/j.sbi.2011.03.005
31. Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951-960. doi:10.1093/bioinformatics/bti125
32. Zimmermann L, Stephens A, Nam SZ, Rau D, Kubler J, Lozajic M, Gabler F, Soding J, Lupas AN, Alva V (2018) A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol 430:2237-2243. doi:10.1016/j.jmb.2017.12.007
33. Zheng W, Zhang C, Wuyun Q, Pearce R, Li Y, Zhang Y (2019) LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. Nucleic Acids Res 47:W429-W436. doi:10.1093/nar/gkz384
34. Pandurangan AP, Stahlhacke J, Oates ME, Smithers B, Gough J (2019) The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. Nucleic Acids Res 47:D490-D494. doi:10.1093/nar/gky1130
35. Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313:903-919. doi:10.1006/jmbi.2001.5080
36. Buchan DWA, Jones DT (2019) The PSIPRED Protein Analysis Workbench: 20 years on. Nucleic Acids Res 47:W402-W407. doi:10.1093/nar/gkz297
37. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195-202. doi:10.1006/jmbi.1999.3091
38. Heffernan R, Yang Y, Paliwal K, Zhou Y (2017) Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Bioinformatics 33:2842-2849. doi:10.1093/bioinformatics/btx218
39. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725-738. doi:10.1038/nprot.2010.5
40. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER Suite: protein structure and function prediction. Nat Methods 12:7-8. doi:10.1038/nmeth.3213
41. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40. doi:10.1186/1471-2105-9-40

42. Song Y, DiMaio F, Wang RY, Kim D, Miles C, Brunette T, Thompson J, Baker D (2013) High-resolution comparative modeling with RosettaCM. Structure 21:1735-1742. doi:10.1016/j.str.2013.08.005
43. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57:702-710. doi:10.1002/prot.20264
44. Xu J, Zhang Y (2010) How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 26:889-895. doi:10.1093/bioinformatics/btq066
45. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33:2302-2309. doi:10.1093/nar/gki524
46. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 25:1605-1612. doi:10.1002/jcc.20084
47. Derouet D, Rousseau F, Alfonsi F, Froger J, Hermann J, Barbier F, Perret D, Diveu C, Guillet C, Preisser L, Dumont A, Barbado M, Morel A, deLapeyriere O, Gascan H, Chevalier S (2004) Neuropoietin, a new IL-6-related cytokine signaling through the ciliary neurotrophic factor receptor. Proc Natl Acad Sci U S A 101:4827-4832. doi:10.1073/pnas.0306178101
48. Senaldi G, Varnum BC, Sarmiento U, Starnes C, Lile J, Scully S, Guo J, Elliott G, McNinch J, Shaklee CL, Freeman D, Manu F, Simonet WS, Boone T, Chang MS (1999) Novel neurotrophin-1/B cell-stimulating factor-3: a cytokine of the IL-6 family. Proc Natl Acad Sci U S A 96:11458-11463. doi:10.1073/pnas.96.20.11458
49. Heinrich PC, Behrmann I, Haan S, Hermanns HM, Muller-Newen G, Schaper F (2003) Principles of interleukin (IL)-6-type cytokine signalling and its regulation. Biochem J 374:1-20. doi:10.1042/BJ20030407
50. Huising MO, Kruiswijk CP, Flik G (2006) Phylogeny and evolution of class-I helical cytokines. J Endocrinol 189:1-25. doi:10.1677/joe.1.06591
51. Rose-John S (2018) Interleukin-6 Family Cytokines. Cold Spring Harb Perspect Biol 10 doi:10.1101/cshperspect.a028415
52. Sims NA (2015) Cardiotrophin-like cytokine factor 1 (CLCF1) and neuropoietin (NP) signalling and their roles in development, adulthood, cancer and degenerative disorders. Cytokine Growth Factor Rev 26:517-522. doi:10.1016/j.cytogfr.2015.07.014
53. Somers W, Stahl M, Seehra JS (1997) 1.9 A crystal structure of interleukin 6: implications for a novel mode of receptor dimerization and signaling. EMBO J 16:989-997. doi:10.1093/emboj/16.5.989
54. Adams R, Burnley RJ, Valenzano CR, Qureshi O, Doyle C, Lumb S, Del Carmen Lopez M, Griffin R, McMillan D, Taylor RD, Meier C, Mori P, Griffin LM, Wernery U, Kinne J, Rapecki S, Baker TS, Lawson AD, Wright M, Ettorre A (2017) Discovery of a junctional epitope antibody that stabilizes IL-6 and gp80 protein:protein interaction and modulates its downstream signaling. Sci Rep 7:37716. doi:10.1038/srep37716
55. Putoczki TL, Dobson RC, Griffin MD (2014) The structure of human interleukin-11 reveals receptor-binding site features and structural differences from interleukin-6. Acta Crystallogr D Biol Crystallogr 70:2277-2285. doi:10.1107/S1399004714012267
56. McDonald NQ, Panayotatos N, Hendrickson WA (1995) Crystal structure of dimeric human ciliary neurotrophic factor determined by MAD phasing. EMBO J 14:2689-2699.
57. Robinson RC, Grey LM, Staunton D, Vankelecom H, Vernallis AB, Moreau JF, Stuart DI, Heath JK, Jones EY (1994) The crystal structure and biological function of leukemia inhibitory factor: implications for receptor binding. Cell 77:1101-1116. doi:10.1016/0092-8674(94)90449-9
58. Huyton T, Zhang JG, Luo CS, Lou MZ, Hilton DJ, Nicola NA, Garrett TP (2007) An unusual cytokine:Ig-domain interaction revealed in the crystal structure of leukemia inhibitory factor (LIF) in complex with the LIF receptor. Proc Natl Acad Sci U S A 104:12737-12742. doi:10.1073/pnas.0705577104
59. Deller MC, Hudson KR, Ikemizu S, Bravo J, Jones EY, Heath JK (2000) Crystal structure and functional dissection of the cytostatic cytokine oncostatin M. Structure 8:863-874. doi:10.1016/s0969-2126(00)00176-3
60. Schreuder HA, Rondeau JM, Tardif C, Soffientini A, Sarubbi E, Akeson A, Bowlin TL, Yanofsky S, Barrett RW (1995) Refined crystal structure of the interleukin-1 receptor antagonist. Presence of a disulfide link and a cis-proline. Eur J Biochem 227:838-847. doi:10.1111/j.1432-1033.1995.tb20209.x
61. Fiser A, Sali A (2003) Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol 374:461-491. doi:10.1016/S0076-6879(03)74020-8
62. Panayotatos N, Radziejewska E, Acheson A, Somogyi R, Thadani A, Hendrickson WA, McDonald NQ (1995) Localization of functional receptor epitopes on the structure of ciliary neurotrophic factor indicates a

conserved, function-related epitope topography among helical cytokines. J Biol Chem 270:14007-14014. doi:10.1074/jbc.270.23.14007
63. Perret D, Guillet C, Elson G, Froger J, Plun-Favreau H, Rousseau F, Chabbert M, Gauchat JF, Gascan H (2004) Two different contact sites are recruited by cardiotrophin-like cytokine (CLC) to generate the CLC/CLF and CLC/sCNTFRalpha composite cytokines. J Biol Chem 279:43961-43970. doi:10.1074/jbc.M407686200
64. Plun-Favreau H, Elson G, Chabbert M, Froger J, deLapeyriere O, Lelievre E, Guillet C, Hermann J, Gauchat JF, Gascan H, Chevalier S (2001) The ciliary neurotrophic factor receptor alpha component induces the secretion of and is required for functional responses to cardiotrophin-like cytokine. EMBO J 20:1692-1703. doi:10.1093/emboj/20.7.1692
65. Skolnick J, Zhou H (2017) Why Is There a Glass Ceiling for Threading Based Protein Structure Prediction Methods? J Phys Chem B 121:3546-3554. doi:10.1021/acs.jpcb.6b09517
66. Abdi H (2007) Metric multidimensional scaling. In: *Encyclopedia of Measurement and Statistics*. Salkind NJ, editor. Sage. Thousand Oaks (CA). pp. 598-605.
67. Wu S, Zhang Y (2008) MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 72:547-556. doi:10.1002/prot.21945
68. Wu S, Zhang Y (2007) LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res 35:3375-3382. doi:10.1093/nar/gkm251
69. Xu D, Jaroszewski L, Li Z, Godzik A (2014) FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking. Bioinformatics 30:660-667. doi:10.1093/bioinformatics/btt578
70. Xu Y, Xu D (2000) Protein threading using PROSPECT: design and evaluation. Proteins 40:343-354.
71. Zhou H, Zhou Y (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 58:321-328. doi:10.1002/prot.20308
72. Yang Y, Faraggi E, Zhao H, Zhou Y (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 27:2076-2082. doi:10.1093/bioinformatics/btr350
73. Madera M (2008) Profile Comparer: a program for scoring and aligning profile hidden Markov models. Bioinformatics 24:2630-2631. doi:10.1093/bioinformatics/btn504
74. Kallberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J (2012) Template-based protein structure modeling using the RaptorX web server. Nat Protoc 7:1511-1522. doi:10.1038/nprot.2012.085
75. Ovchinnikov S, Park H, Varghese N, Huang PS, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D (2017) Protein structure determination using metagenome sequence data. Science 355:294-298. doi:10.1126/science.aah4043
76. Palmer CA, Watts RA, Gregg RG, McCall MA, Houck LD, Highton R, Arnold SJ (2005) Lineage-specific differences in evolutionary mode in a salamander courtship pheromone. Mol Biol Evol 22:2243-2256. doi:10.1093/molbev/msi219

Supplementary Figure 1

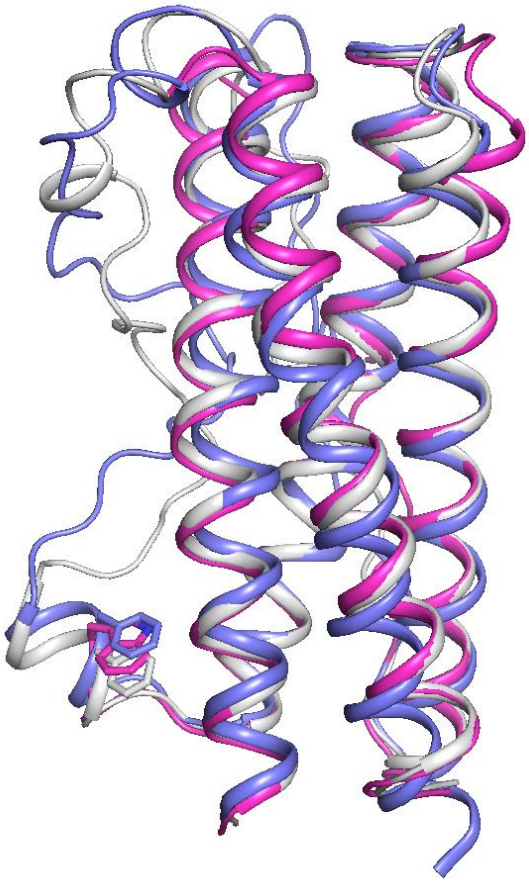


Fig S1: Comparison of the customized and AlphaFold models. The customized MODELLER model (residues 10-187 of mature protein, white ribbon) is superimposed with the AlphaFold model deposited in UniProt (AF_Q9PUJ2-F1, slate) and with the CNTF template (PDB access number: 1CNT, magenta). Phe-70 of PRF1 and Trp-64 of CNTF are shown as sticks.