Origin of a folded repeat protein from an intrinsically disordered ancestor

Hongbo Zhu, Edgardo Sepulveda, Marcus D. Hartmann, Manjunatha Kogenaru†, Astrid Ursinus, Eva Sulz, Reinhard Albrecht, Murray Coles, Jörg Martin, Andrei N. Lupas§

Department of Protein Evolution,
Max Planck Institute for Developmental Biology,
Spemannstr. 35, D-72076 Tübingen, Germany

hongbo.zhu@tuebingen.mpg.de
† present address: Department of Life Sciences, Imperial College London, London SW7 2AZ, United Kingdom. E-mail: m.kogenaru@imperial.ac.uk
§ andrei.lupas@tuebingen.mpg.de; Corresponding author

Abstract

Repetitive proteins are thought to have arisen through the amplification of subdomain-sized peptides. Many of these originated in a non-repetitive context as cofactors of RNA-based replication and catalysis, and required the RNA to assume their active conformation. In search of the origins of one of the most widespread repeat protein families, the tetratricopeptide repeat (TPR), we identified several potential homologs of its repeated helical hairpin in non-repetitive proteins, including the putatively ancient ribosomal protein S20 (RPS20), which only becomes structured in the context of the ribosome. We evaluated the ability of the RPS20 hairpin to form a TPR fold by amplification and obtained structures identical to natural TPRs for variants with 2-5 point mutations per repeat. The mutations were neutral in the parent organism, suggesting that they could have been sampled in the course of evolution. TPRs could thus have plausibly arisen by amplification from an ancestral helical hairpin.

1 Introduction

Most present-day proteins arose through the combinatorial shuffling and differentiation of a set of domain prototypes. In many cases, these prototypes can be traced back to the root of cellular life and have since acted as the primary unit of protein evolution (Anantharaman et al., 2001; Apic et al., 2001; Koonin, 2003; Kyrpides et al., 1999; Orengo and Thornton, 2005; Ponting and Russell, 2002; Ranea et al., 2006). The mechanisms by which they themselves arose are however still poorly understood. We have proposed that the first folded domains emerged through the repetition, fusion, recombination, and accretion of an ancestral set of peptides, which supported RNA-based replication and catalysis (the RNA world (Bernhardt, 2012; Gilbert, 1986)) (Alva et al., 2015; Lupas et al., 2001; Söding and Lupas, 2003). Repetition would have been a particularly prominent mechanism by which these peptides yielded folds; six of the ten most populated folds in the Structural Classification of Proteins (SCOP) (Murzin et al., 1995) – including the five most frequent ones – have repetitive structures. In all cases, their amplification from subdomain-sized fragments can also be retraced at the sequence level in at least some of their members.

One of these highly populated repetitive folds is the $\alpha\alpha$-solenoid (SCOP a.118), whose most widespread superfamily is the tetratricopeptide repeat (TPR; a.118.8). This was originally identified as a repeating 34 amino-acid motif in Cdc23p of *Saccharomyces cerevisiae* (Sikorski et al., 1990) – hence its name. Since then, TPR-containing proteins have been discovered in all kingdoms of life, where they mediate protein-protein interactions in a broad range of biological processes, such as cell cycle control, transcription, protein translocation, protein folding, signal transduction and innate immunity (Cortajarena and Regan, 2006; Dunin-Horkawicz et al., 2014; Katibah et al., 2014; Keiski et al., 2010; Kyrpides and Woese, 1998; Lamb et al., 1995; Sikorski et al., 1990). The first crystal structure of a TPR domain (Das et al., 1998) showed that the repeat units are helical hairpins, stacked into a continuous, right-handed superhelical architecture with an inner groove that mediates the interaction with target proteins (Forrer et al., 2004). The hairpins interact via a specific geometry

58 involving knobs-into-holes packing (Crick, 1953) and burying about 40% of their surface between
59 repeat units. This tightly packed, superhelical arrangement of a repeating structural unit is typical of
60 all αα-solenoid proteins (Di Domenico et al., 2014; Kajava, 2012; Kobe and Kajava, 2000).

61 Comparison of TPRs from a variety of proteins reveals a high degree of sequence diversity, with
62 conservation observed mainly in the size of the repeating unit and the hydrophobicity of a few key
63 residues (D'Andrea and Regan, 2003; Magliery and Regan, 2004). Nevertheless, almost all known
64 TPR-containing proteins can be detected using a single sequence profile (Karpenahalli et al., 2007),
65 underscoring their homologous origin. As their name implies, TPR proteins generally contain at least
66 two unit hairpins in a repeated fashion. The few that have only one hairpin, notably the
67 mitochondrial import protein Tom20 (Abe et al., 2000), are clearly not ancestral based on their
68 phylogenetic distribution and functionality, implying that the ancestor of the superfamily already had
69 a repeated structure. In searching for the origin of TPRs, we hypothesized that the hairpin at the root
70 of the fold might either have been part of a different, non-repetitive fold or have given rise to both
71 repetitive and non-repetitive folds at the origin of folded domains. Either way we hoped that we
72 might find α-hairpins in non-repetitive proteins that are similar in both sequence and structure to the
73 TPR unit, suggesting a common origin. Here we show that such hairpins are detectable and that one
74 of them, from the ribosomal protein RPS20 (Schluenzen et al., 2000), can be customized to yield a
75 TPR fold by repetition, with only a small number of point mutations that are neutral for the parent
76 organism. Ribosomal proteins most likely constitute some of the oldest proteins observable today
77 and are still intimately involved in an RNA-driven process: translation (Fox, 2010; Hsiao et al., 2009).
78 They are mostly incapable of assuming their folds outside the ribosomal context (Peng et al., 2014)
79 and thus belong to a class of intrinsically disordered proteins that become structured upon binding to
80 a macromolecular scaffold (Dyson and Wright, 2005; Habchi et al., 2014; Oldfield and Dunker, 2014;
81 Peng et al., 2014; Varadi et al., 2014). This hairpin therefore plausibly retains today many of the
82 properties likely to have been present in the ancestral peptide that gave rise to the TPR fold.

2 Results and Discussion

2.1 Recently amplified TPR arrays in present-day proteins

Repetitive folds with variable numbers of repeats, such as HEAT, LRR, TPR or β-propellers, usually have some members with a high level of sequence identity between their repeat units (Dunin-Horkawicz et al., 2014). In these proteins, the units are more similar to each other than to any other unit in the protein sequence database, showing that they were recently amplified. In a detailed study of β-propellers (Chaudhuri et al., 2008), we found that this process of amplification and differentiation has been ongoing since the origin of the fold. TPR proteins show a similar evolutionary history. In some proteins, most of the repeats can be seen to have been amplified separately and to a different extent in each ortholog, pointing to their recent origin (Figure 1a); in others, the amplification must have occurred much earlier, as their ancestor already had fully differentiated repeats (Figure 1b). In recently amplified proteins, such as the ones shown in Figure 1a, within which repeats frequently have >80% pairwise sequence identity, tracking the probable α-hairpin at the root of the amplification is a fairly straightforward proposition. We wondered, however, whether it might be possible to go much further back in time and track the original α-hairpin from which the first TPR protein was amplified. We therefore searched for TPR-like α-hairpins in non-repetitive proteins as present-day descendants of the original hairpin.

2.2 Identification of helical hairpins resembling the TPR unit

We had previously developed a profile-based method, named TPRpred, specially designed for the detection of TPRs and related repeat proteins with high sensitivity from sequence data (Karpenahalli et al., 2007). Here, in a first step, we used TPRpred to scan protein sequences in the Protein Data Bank (PDB) (Berman et al., 2000) for peptides that share statistically significant similarity to the TPR sequence profile and yet have not been annotated as TPR in Pfam (Finn et al., 2014); we used a *P*-value cutoff=1.0e−4, which leads to an estimated false discovery rate of 1.0%, see Materials and Methods. We ignored tandem repeats in the hit list and focused only on the singleton cases.

Subsequently, we compared the structures of these helical hairpin singletons to the average TPR hairpin and removed non-hairpin-like structures. This yielded 31 helical hairpins that are similar to the TPR unit with respect to both sequence and structure. Among them, 21 are part of solenoid-like structures and were discarded. The remaining nine hits belong to three families: (I) mitochondrial import receptor subunit Tom20; (II) microtubule interacting and transport (MIT) domain including katanin (Iwaya et al., 2010); and (III) 30S ribosomal protein S20 (RPS20) (Figure 2).

The similarity of Tom20 and MIT domains to TPR proteins has been noted before (Abe et al., 2000; Iwaya et al., 2010; Scott et al., 2005), but the similarity of RPS20 was surprising and drew our attention particularly due to the ancestrality attributed to ribosomal proteins. To further explore the similarity between the helical hairpin in RPS20 (in short, RPS20-hh) and TPR, we used TPRpred to rank the RPS20 sequences in Pfam (Finn et al., 2014). The top-scoring hit was RPS20-hh from *Thermus aquaticus* (NCBI accession number=WP_003044315.1, UniProt id=B7A5L8_THEAQ), which matches the TPR unit sequence profile at a *P*-value of 5.4e−07, almost an order of magnitude better than the second hit (see supplementary file 1D). Furthermore, we examined the surface residues of RPS20-hh fragments to assess their suitability to occur in a tandem repeat mode, as in TPRs. To this end, we first defined five interface positions on the TPR helical hairpin and transferred the definition to RPS20-hh according to their structure alignment (positions 3, 7, 10, 21 and 28 using TPR unit numbering). Then, we searched for RPS20-hhs with as many hydrophobic residues as possible at these interface positions. We found 42 RPS20-hhs that contain at least three hydrophobic residues out of the five interface positions. Among them, the only RPS20-hh predicted to match the TPR unit profile above a *P*-value of 1.0e−4 was again the RPS20 from *T. aquaticus*, in which three of the five interface residues are hydrophobic (L10, I21 and V28). We therefore chose this helical hairpin (RPS20-hhta) to construct a TPR-like solenoid by amplification (Figure 3).

2.3 Design of a TPR array from a RPS20

We focused on the construction of three-repeat TPRs, which represent the most common form of this fold (D'Andrea and Regan, 2003; Sawyer et al., 2013). For instance, 18 of the 54 non-identical

TPR domains in the extended Structural Classification of Proteins database (SCOPe v2.05) (Fox et al., 2014) have three repeats. A previously designed three-repeat TPR protein, CTPR3, was also demonstrated to be highly stable, even more so than natural three-repeat TPR proteins (Main et al., 2003b). We concatenated three copies of RPS20-hhta as an initial construct, connected by the TPR consensus loop sequence (DPNN). We annotate the two helices in each repeat unit as helix *Ai* and *Bi*, where *i* is the index of the repeat unit (*i=1, 2 or 3*) (Figure 3). Under the hypothesis of common descent between TPR and RPS20 from the same ancestral peptide and retention of ancestral features in RPS20, this basic construct would fold as a TPR solenoid with a minimal number of mutations, ideally none.

When we experimentally made a construct containing no mutations (M0, Table 1), it was soluble but remained unfolded under all conditions tested (see Section 2.4). We therefore introduced point mutations into the sequence of RPS20-hhta, aimed at favoring the target structure. Here, we followed the principle of consensus design (Forrer et al., 2004; Main et al., 2003a), which requires the mutation positions to be occupied by the most commonly observed residues in homologous proteins (Forrer et al., 2004). Consensus design methods have been successful in engineering several different repeat proteins with solenoid folds, including ankyrin repeats (Binz et al., 2003; Kohl et al., 2003; Mosavi et al., 2002), TPRs (Doyle et al., 2015; Kajander et al., 2007; Main et al., 2003b), pentatricopeptide repeats (PPRs) (Coquille et al., 2014; Shen et al., 2016) and leucine rich repeats (Rämisch et al., 2014; Stumpp et al., 2003). Following these principles, four different sites of mutation (L4W, K7L/R, V9N, I23D/Y, see Figure 4) were considered to improve interface hydrophobicity or preserve coevolved positions observed in TPRs (Sawyer et al., 2013) (see Materials and Methods). Furthermore, as natural TPR proteins tend to exhibit zero net charge (Magliery and Regan, 2004), four positively charged residues were also targeted (K2E, K6N, K22E, R25Q/E, see Figure 4). This resulted in a set of eight candidate mutation sites. In order to preserve the character of the RPS20-hhta sequence, we restricted the number of mutations in any repeat unit to be at most five.

In most TPR proteins, there is an α-helix at the C-terminus, which interacts with the last TPR unit by covering the hydrophobic surface. This so-called C-terminal "stop helix" had been observed in all known TPR structures and was considered essential for the solubility of natural TPR proteins (D'Andrea and Regan, 2003; Das et al., 1998; Main et al., 2003b). Most other designed TPRs employ purpose-designed stop helix sequences. Here, we chose to use the RPS20 C-terminal helix to become a natural stop helix, since it is already known to interact favorably with RPS20-hhta (Figure 3). Further, we inserted two residues (Asn-Ser) before the first TPR unit as an N-terminal cap to the first helix (Aurora and Rose, 1998; Kumar and Bansal, 1998), in analogy to a previously designed idealized TPR protein, CTPR3 (Main et al., 2003b).

To model the structure of the designed proteins *in silico*, we fused two structures to create a hybrid template: We used CTPR3 (PDB id: 1na0 chain A) as the structural template for the three RPS20-hhta fragments, and the best-resolved RPS20 structure (PDB id: 2vqe chain T; 2.5 Å) for helix B3 and the stop helix. We built structural models on this hybrid template and tested a variety of mutants using the Rosetta programs *fixbb* and *relax*, which perform fixed-backbone design and structural refinement (Das and Baker, 2008; Doyle et al., 2015; Park et al., 2015; Parmeggiani et al., 2015). The Rosetta energy score of the models calculated for all mutants is depicted in a boxplot (Figure 4—figure supplement 2). Among them, five were selected for further testing *in vitro* (see Materials and Methods). These five tested mutants are termed M2, M4E, M4N, M4RD and M5. Their primary structures are listed in Table 1.

2.4 Biophysical characterization of designed TPRs and RPS20

We cloned the five TPR designs plus the unmutated construct M0 into pET vectors for expression in *Escherichia coli*. Three proteins (M0, M4RD and M5) could be purified from soluble extracts; the other constructs were insoluble and were refolded from inclusion bodies. In far UV circular dichroism (CD) spectra, all proteins displayed a strong alpha-helical pattern, except M0 and M4RD, which appeared to be unfolded, but not prone to aggregation and precipitation, even at high concentrations. When we studied the melting curves, M4N showed cooperative unfolding with a T_m

of 77°C (supplementary file 1F), while the unfolding of M2, M4E and M5 did not conform to a classical two-state transition, consistent with an unstable molten globule-like state. On the other hand, non-cooperative unfolding processes have been demonstrated for perfectly stable TPR repeats and suggested to be common for various types of repeat proteins (Cortajarena and Regan, 2006; Kajander et al., 2007; Stumpp et al., 2003). To clarify this point, urea-induced unfolding transitions were monitored by CD. Like M4N, the three variants M2, M4E and M5 yielded typical cooperative denaturation curves, indicative of folded polypeptides (Figure 5—figure supplement 2). The $\Delta G_{U\text{-}F}{}^{H2O}$ values agree well with those reported for other designed TPRs (supplementary file 1F) (Main et al., 2005). In line with these findings, M5, the only protein containing tryptophan residues, had a λ_{max} of 336 nm in fluorescence emission spectra, as expected for partially shielded aromatic residues. We conclude that four of the five designed TPR variants, M2, M4E, M4N and M5, result in well-folded repeat proteins. To determine the oligomeric state of our folded proteins, we performed static light scattering experiments. Surprisingly, all four constructs were exclusively dimers (supplementary file 1F).

We also examined the ribosomal parent protein RPS20. Within the ribosome, RPS20 is partially embedded in the 16S rRNA, making many nucleic acid contacts. Like many other ribosomal proteins, it is not expected to adopt a stable structure in isolation. Indeed, it has a biased amino acid composition and is predicted to be largely unstructured by many prediction programs (Figure 4—figure supplement 1, see also Supplementary file 1J). It had been shown previously that isolated RPS20 exhibits only one third helical content by CD (Paterakis et al., 1983). For *Thermus* RPS20 specifically, simulations predict a flexible conformation in solution (Burton et al., 2012). We cloned RPS20 from *T. aquaticus* and its close relative *T. thermophilus*. Upon expression, both proteins were insoluble and had to be refolded. In static light scattering measurements, both proteins behaved as monomers (supplementary file 1F). Based on CD spectra, which showed a high proportion of random structure, and the absence of defined melting and urea-denaturation curves (supplementary file 1F), we conclude that RPS20 indeed exhibits considerable conformational variation in solution.

2.5 Structure of a designed TPR

To obtain high-resolution structural information on our designed proteins, we set up crystallization trials for all four folded constructs. We obtained crystals and solved the structure of M4N to a resolution of 2.2 Å (Figure 5a). The asymmetric unit (ASU) contains three polypeptide chains of almost identical structure (all pairwise C_α RMSD values below 1.4 Å). Notably, all three chains exhibit the desired TPR architecture with three repetitive hairpins, which interact via knobs-into-holes packing between helices Ai and B(i-1), as is characteristic of TPR hairpins. A superposition to the CTPR3 structure yields C_α RMSD values below 2.6 Å (supplementary file 1I). An unexpected difference to canonical TPR structure is that the stop helix of M4N is not resolved in any of the three chains. However, this missing helix is compensated for by a specific dimerization mode of two M4N protomers. Therein, the C-terminal TPR units of the two protomers form a tight interface, in which the B3 helix of each chain substitutes for the stop helix of the other, mimicking the capping effect of the stop helix (Figure 6). A superposition of this mimicry to the last TPR unit and stop helix of CTPR3 yields C_α RMSD values as low as 1.2 Å over 44 residues. The third chain of the ASU, however, was found as a monomer, capping its C-terminal TPR unit in a more unspecific manner by packing it orthogonally against the two A1 helices of the dimer (Figure 5a).

Analysis by mass spectrometry revealed that the M4N stop helix had been partially proteolyzed upon expression of the protein (Figure 5—figure supplement 3). Although we did not observe proteolysis in the other folded constructs (M2, M4E and M5), which were also all dimeric, we analyzed whether proteolysis might have favored the dimerization of M4N. Extending the stop helix with a C-terminal His_6-tag prevented proteolysis, but did not affect stability or dimerization (M4N-His; supplementary file 1F). We conclude that in the amplified constructs, the observed interactions are more favorable than the interaction with the native stop helix, releasing it and rendering it prone to degradation. This led us to ask whether this helix is in fact dispensable. Indeed, an M4NΔC construct, which terminates with the B3 helix, showed the same stability and dimerization as M4N. We obtained two structures for M4NΔC from different crystal forms at 2.0 Å and 1.7 Å resolution, respectively, the first

(CF I) with two dimers in the ASU and the second (CF II) with a single chain in the ASU, for which we constructed the dimer by crystallographic symmetry. All three dimers superimpose to the M4N dimer with C_α RMSD below 1.9 Å (Figure 7, supplementary file 1I). We conclude that the stop helix is dispensable for folding, dimerization and stability of our designed constructs.

The geometry of dimerization in M4N has not been observed in TPR structures before. Although there have been reports on the self-association of TPR-containing proteins involved in various regulatory biological processes (Bansal et al., 2009a; Bansal et al., 2009b; Ramarao et al., 2001; Serasinghe and Yoon, 2008), only a small number of oligomeric TPR structures have been determined to date (Krachler et al., 2010; Lunelli et al., 2009; Zeytuni et al., 2012; Zeytuni et al., 2015; Zhang et al., 2010). None of these resemble the ring-shaped dimer of M4N.

2.6 Mutations introduced into RPS20-hhta are neutral to Thermus

The results shown above suggest that the mutations we made to RPS20-hhta were crucial for obtaining the TPR fold. If RPS20 and TPR proteins indeed share a common ancestor, such mutations may have been sampled in the course of evolution. Since we cannot reconstruct the ancestor and do not know what its function was beyond a general expectation of RNA binding, we decided to test whether the mutations we introduced impaired the interaction between RPS20 and its cognate RNA, as an indication of their compatibility with RNA interaction. Each mutation in M2 and M4N occurs in natural RPS20 sequences (see supplementary file 1A), but no RPS20 sequence has all four mutations simultaneously and we therefore tested if they can be tolerated *in vivo*. As genetic engineering in *T. aquaticus* turned out to be unfeasible, we performed these tests in *T. thermophilus* HB8, which is a well-established model organism. The RPS20 helical hairpins in *T. aquaticus* and *T. thermophilus* differ only at four positions, of which two are highly conservative substitutions (Figure 8a).

We first attempted to substitute the chromosomal RPS20-encoding gene, *rpsT*, with a kanamycin resistance cassette, to obtain *T. thermophilus* strain KM4 (Figure 8b). For complementation we introduced plasmids bearing wild type *rpsT* from *T. thermophilus* (TT) or *T. aquaticus* (TA), *rpsT* from

T. aquaticus carrying the mutations from M2 (TA2) or M4N (TA4), or merely empty plasmids as negative control (E). We monitored the substitution of *rpsT* by a PCR screening protocol, which will amplify a 1500 bp region if WT *rpsT* is substituted and an 800 bp region otherwise (Figure 8b). Under selection pressure from kanamycin, only the 1500 bp product was obtained in all cases where plasmid-borne *rpsT* was introduced, whether in wild-type or mutated form (Figure 8c panels 1 and 2, lanes TT, TA, TA2 and TA4), showing that the chromosomal gene had been fully substituted. In contrast, PCR screening of strain KM4 complemented with an empty plasmid produced both 800 bp and 1500 bp fragments (Figure 8c panels 1 and 2, lane E). Since *T. thermophilus* HB8 is a polyploid organism (minimally tetraploid (Ohtani et al., 2010)), this result shows that *rpsT* can be reduced in copy number, but not fully eliminated, suggesting that the gene is essential.

To assess the level of substitution achieved with the various plasmids, we designed a second PCR screening protocol to specifically detect chromosomal *rpsT* via a 300 bp product. At low kanamycin concentrations this protocol always generated a product (Figure 8d panel 1), but at increased kanamycin concentration we did not obtain product for any *rpsT* allele (Figure 8d panel 2, lanes TT, TA, TA2 and TA4). This demonstrates that plasmid-borne *rpsT* and its mutants were able to complement the chromosomal *rpsT* and that the latter was displaced from the population to a level that left it undetectable by PCR. In contrast, we could never completely suppress chromosomal *rpsT* in strain KM4 complemented with an empty plasmid, even under high kanamycin conditions (120 µg/ml).

In *E. coli* and *Salmonella enterica*, *rpsT* has been reported to be non-essential, but its deletion significantly lowers growth rates (Bubunenko et al., 2007; Tobin et al., 2010). We found that *rpsT* is essential in *T. thermophilus*, but that its loss could be complemented by wild-type and mutant *T. aquaticus rpsT*, and that this complementation restored wild-type levels of growth (Figure 8e). Moreover, when the selection pressure from kanamycin was removed, no reversal in the PCR products was detected for any strain (Figure 8c and d, panel 3), which confirms that chromosomal *rpsT* was substantially displaced during kanamycin treatment. We conclude that *rpsT* from *T.*

289 *aquaticus* and its two mutated alleles are neutral with respect to survival and growth for *T.*
290 *thermophilus*. This demonstrates that the mutations we introduced do not affect negatively the
291 interaction between RPS20 and its cognate RNA, and that therefore such mutations could have been
292 sampled multiply and in a cumulative fashion by neutral drift in the course of evolution.

293 2.7 Implications for the emergence of folded proteins

294 Proteins are the most complex macromolecules synthesized in nature and by and large need to
295 assume defined structures for their activity. This folding process is complicated and easily disrupted,
296 witness the elaborate systems for protein folding, quality control and degradation universal to all
297 living beings. Despite widespread problems to reach and maintain the folded state, natural proteins
298 nevertheless form a best-case group, since the overwhelming majority of random polypeptides do
299 not appear to have a folded structure (Keefe and Szostak, 2001; Wei et al., 2003). It thus seems
300 impossible that, at the origin of life, the prototypes for the folded proteins we see today could have
301 arisen by random concatenation of amino acids. We have proposed that folding resulted from the
302 increasing complexity of peptides that supported RNA replication and catalysis, and that these
303 peptides assumed their structure through the interaction with the RNA scaffold (Lupas et al., 2001;
304 Söding and Lupas, 2003). In this view, protein folding was an emergent property of RNA-peptide
305 coevolution. We have recently described 40 such peptides whose conservation in diverse folds
306 suggests that they predated folded proteins (Alva et al., 2015). These peptides are enriched for
307 nucleic-acid binders, particularly in the context of the ribosome.

308 Due to its extremely slow rate of change, the ribosome essentially represents a living fossil, providing
309 the possibility to study the chronology of ancient events in molecular evolution (Hsiao et al., 2009).
310 Thus, core ribosomal proteins offer a window into the time when proteins were acquiring the ability
311 to fold. Those close to the catalytic center almost entirely lack secondary structure. Further away
312 from the center, their secondary structure content gradually increases and at the periphery, these
313 secondary structure elements become arranged into topologies that parallel those seen in cytosolic
314 proteins (Hsiao et al., 2009). Collectively, the structures of ribosomal proteins chart a path of

progressive emancipation from the RNA scaffold. Even the peripheral proteins, however, still mostly assume their structure only in the context of the ribosomal RNA, as exemplified by RPS20 in our study (supplementary file 1F, see also (Paterakis et al., 1983)).

The simplest mechanism to achieve an increase in complexity is the repetition of building blocks and nature provides many examples for this, at all levels of organization. The dominant role of repetition in the genesis of protein folds has been documented in many publications since the 1960s (Alva et al., 2007; Blundell et al., 1979; Broom et al., 2012; Eck and Dayhoff, 1966; Kopec and Lupas, 2013; Lee and Blaber, 2011; McLachlan, 1972; McLachlan, 1987; Remmert et al., 2010; Söding et al., 2006). As a test of this mechanism, we explored whether a peptide originating from a ribosomal protein that is disordered outside the context of the ribosome, could form a folded protein through an increase in complexity afforded by repetition. For this, we chose a present-day representative of one of the 40 fragments we reconstructed (Alva et al., 2015); this fragment is naturally found in single copy in several different folds, including that of ribosomal protein RPS20, and repetitively in one fold, TPR. Simple repetition was not sufficient in our case, but the repeat protein was so close to a folded structure that only two point mutations per repeat were necessary to allow it to fold reliably. The mutations needed for this transition did not appear to affect negatively the interaction with the RNA scaffold, raising the possibility that they could have been among the variants sampled multiply in the course of evolution.

Our experiments recapitulate a scenario for the emergence of a protein fold by a widespread and well-documented mechanism, and show that this could have proceeded in a straightforward way. These experiments represent a proof of concept, starting with a modern peptide likely to still retain many features of an ancestral αα-hairpin that gave rise to a number of folds, including TPR. Rather than proposing proto-RPS20 as the parent of TPR domains, we see it as one of many proteins emerging from this ancestral hairpin. Given the ease with which repetition of the RPS20 hairpin yielded a TPR-like fold, we consider it likely that the hairpins belonging to the ancestral group were amplified many times during the emergence of folded proteins to yield a range of TPR-like offspring,

of which only one may have survived to this day (but see also the figure legend to Figure 1). The reason for this limited survival may lie in the fact that structure is a prerequisite for protein function, but it is the function that is under biological selection. It could be that the newly emerged TPR-like folds required many additional changes to achieve a useful activity and that therefore only very few – possibly just one – survived. We consider a different scenario more probable, however. All present-day TPR domains whose function has been characterized mediate protein-protein interactions by binding to linear sequence motifs in unstructured polypeptide segments (D'Andrea and Regan, 2003; Zeytuni and Zarivach, 2012). This activity would have been particularly relevant at a time of transition from peptides dependent on RNA scaffolds for their structure, to autonomously folded polypeptides. Functions relevant in this context would have been to prevent aggregation and increase the solubility of newly emerging (poly)peptides, to promote autonomous folding, to serve as assembling factors for RNA-protein and protein-protein complexes, and to recognize targeting sequences in the emerging cellular networks. It therefore seems likely to us that many of the newly evolved TPR-like folds became established in one or the other of these activities, only to be subsequently displaced by folding becoming a general property of cellular polypeptides and by more advanced, energy-dependent folding factors, which offered much better regulation. Exploring the extent to which our new TPR protein could fulfill such functions represents the next frontier in our studies.

3 Materials and Methods

3.1 Phylogeny for recently amplified TPR arrays

All sequence similarity searches in this work were performed using the Web BLAST (RRID:SCR_004870) from the National Institute for Biotechnology Information (NCBI; http://blast.ncbi.nlm.nih.gov; (Boratyn et al., 2013)) and in the MPI Bioinformatics Toolkit (RRID:SCR_010277, https://toolkit.tuebingen.mpg.de/; (Alva et al., 2016)). Examples of recently amplified repeat units in TPR were taken from a previous investigation (Dunin-Horkawicz et al., 2014). The TPR domain in serine/threonine-protein phosphatase 5 was chosen as a representative

three-repeat TPR, the most common TPR form in natural proteins (D'Andrea and Regan, 2003; Sawyer et al., 2013), to study divergent evolution of TPR. We ran BLAST on the non-redundant protein sequence database (nr) with an E-value threshold of 0.05 using the TPR domain of serine/threonine-protein phosphatase 5 from *Homo sapiens* as query (Das et al., 1998). From the results, we chose seven taxa to cover a diverse range of life.

TPRpred program (Karpenahalli et al., 2007) was used to help identify tandem repeats of TPR units. The construction of multiple sequence alignments (MSAs) for TPR units was straightforward as all TPR units are of the same size (34 aa) and no indels were allowed in the MSAs. We used Clustal X 2.1 (Larkin et al., 2007) to build phylogenetic trees using the neighbor-joining clustering algorithm and 1000 bootstrap trials (Bootstrap N-J Tree). SplitTree4 (Huson and Bryant, 2006) was used to render the phylogenetic trees.

3.2 Identification of helical hairpins resembling the TPR unit

To find proteins homologous to the TPR unit, we first employed the TPRpred program (Karpenahalli et al., 2007) to identify proteins that share significant sequence similarity to the TPR sequence profile, then filtered them by comparing to the TPR structures.

First, TPRpred program with TPR profile tpr2.8 was used to identify TPR unit like sequences from all protein sequences of known structures in the Protein Data Bank (PDB, RRID:SCR_012820) (Berman et al., 2000). Protein sequences from the SEQRES record in PDB files were downloaded from the PDB. We only considered sequences with at least 34 residues, which is the length of the TPR unit. Redundancy was removed by keeping only non-identical sequences. In total, 68,197 sequences were scanned by using TPRpred with default parameters. Only fragments predicted to be TPR with a *P*-value lower than 1.0e−4 were retained (646 hits). We estimated the false discovery rate (FDR) (Noble, 2009) associated with this *P*-value cutoff using a simulated sequence dataset generated by using the amino-acid composition derived from the PDB sequences. The dataset contains the same number of sequences of the same length distribution as the PDB sequences. The FDR was estimated

391 to be the ratio of the number of hits in the simulated dataset to the number of detected hits in the
392 PDB sequences (Noble, 2009). We repeated the simulation 100 times and the FDR was estimated to
393 be 1.0 ± 0.4%.

394 Within the 646 hits, we kept only TPR unit singletons, which are TPR units that have no other TPR
395 units close to them within a distance of 10 residues in the same sequence. TPR units of identical
396 sequences are considered only once. Subsequently, these TPR unit singletons were filtered by
397 removing those annotated to belong to clan *TPR* (CL0020) in Pfam 27.0 (RRID:SCR_004726).

398 The structures of the predicted TPR units obtained from the previous step were then compared to an
399 average TPR unit structure. A predicted TPR unit was discarded if the C_α RMSD of the 34 residues is
400 greater than 2.0 Å after superposition. The average TPR unit structure was generated by considering
401 all proteins belonging to family tetratricopeptide repeat (TPR) (a.118.8.1) in SCOP 1.75
402 (RRID:SCR_007039) (Murzin et al., 1995). TPR repeats in these proteins were again detected using
403 TPRpred and a per-repeat *P*-value cutoff of 1.0e−4 was used. In total, 50 non-redundant TPR repeat
404 fragments were identified and superposed using a multiple structure alignment tool MultiProt
405 (Shatsky et al., 2004). The average C_α positions were calculated from the 50 structures after
406 superposition. We obtained 31 fragments after the structure filtering step (supplementary file 1C).
407 We then inspected the protein structures using PyMOL (RRID:SCR_000305) (Schrödinger, 2010).
408 Among them, 22 were observed to be involved in the formation of solenoid or tandem repeat
409 structures and were thus not further considered.

410 3.3 Identification of TPR homologs in RPS20

411 We applied TPRpred to scan all RPS20 sequences belonging to Pfam 27.0 family *Ribosomal S20p*
412 (PF01649), including sequences from both datasets “full” and “ncbi”. There are 4,402 and 2,284
413 sequences in the two sets. We merged the two sets and removed identical sequences to create a
414 dataset of 3,742 RPS20 sequences. TPRpred was used to detect TPR unit homologs in them. We

obtained 24 hits in these RPS20 sequences predicted by TPRpred to match TPR unit profile with a *P*-value smaller than 1.0e−4 (see supplementary file 1D).

We defined “interface positions” in the TPR unit and then transferred the definition to RPS20-hh according to their structure superposition. We considered the residues on the outer side of the two helices facing neighboring TPR units. Both helix A and helix B in the TPR unit are α-helices, which have on average 3.6 residues per turn. Thus, every third or fourth residue always appears on the same side of the helix. They are positions 3, 7 and 10 in helix A and positions 17, 21, 24 and 28 in helix B. According to the TPR sequence profile compiled by Main et al. (Main et al., 2003b), the most common residues at these positions are hydrophobic except for positions 17 and 24, where the most common residues are both Tyr (see also Figure 4a). Therefore, positions 17 and 24 were not included in the definition of interface positions. Furthermore, the residue at position equivalent to position 24 in RPS20 structure faces its C-terminal helix and is already an interface residue (Figure 4c). Thus, it was not considered as an interface position to be checked in the study. In the end, only positions 3, 7, 10, 21 and 28 in RPS20-hh were defined to be interface positions to be examined, because they are exposed to the solvent or interact with the RNA molecules in the ribosome, but would interact with neighboring repeats in the TPR fold.

We searched all RPS20 sequences in Pfam 27.0 family *Ribosomal_S20p* (PF01649), including both datasets “full” and “ncbi”, for candidates in which the interface positions are occupied by as many hydrophobic residues as possible. In the MSA provided by Pfam, we extracted the 34 columns that correspond to the sequence fragment of RPS20-hh from *Thermus aquaticus*, which was found by TPRpred to be the hit with the best *P*-value and was thus used as the reference RPS20-hh. We obtained 1,370 sequence fragments that do not contain any indels, in which the interface positions were examined for hydrophobicity. Here, Ala, Ile, Leu, Met, Phe, Val were considered as hydrophobic residues. Trp was not included as its side chain may be too large to be accommodated at the interface.

We employed several low-complexity / intrinsically disordered region prediction methods (SEG (Wootton, 1994), PONDR (Romero et al., 2001), DisEMBL (Linding et al., 2003), IUPred (Dosztányi et al., 2005a; Dosztányi et al., 2005b)) to investigate putative intrinsically disordered regions in the RPS20 of *Thermus aquaticus*. We ran SEG with three sets of recommended parameters (Wootton and Federhen, 1996) and the other approaches with default parameters.

3.4 Optimization of RPS20-hh in the designed TPRs

We considered eight positions (2, 4, 6, 7, 9, 22, 23 and 25) in RPS20-hhta for optimization apart from the four residues at the C-terminus.

Main et al. (Main et al., 2003b) discovered a set of eight "TPR signature residues" in the consensus design: W4, L7, G8, Y11, A20, Y24, A27 and P32. Six of them are missing in RPS20-hhta except A20 and A27. Following the principle of consensus design, we introduced L4W and K7L into RPS20-hhta. K7 is also one of the interface positions that ought to be mutated to hydrophobic residue for better packing at interfaces. A8 and L11 were not optimized because they are the second and third most common residues at positions 8 and 11 in the TPR profile, respectively. M24 was also retained because it seems long hydrophobic side chains are favored at position 24 though Met is not one of the three most common residues (YFL). P32 was introduced to replace D32 in RPS20-hhta as part of the C-terminal consensus loop (DPNN) between repeats.

Co-evolution is commonly observed between physically interacting residues (de Juan et al., 2013). We investigated if any positions we optimized are involved in co-evolution relationship so that we can preserve such correlations. We performed direct coupling analysis (Morcos et al., 2011) and computed the mutual information using MatrixPlot (Gorodkin et al., 1999) between all positions in TPR repeat sequences. The results of both approaches revealed that the highest correlation occurs between positions 7 and 23 (Figure 4—figure supplement 1), with the most commonly observed combinations being R7-D23 and L7-Y23. Therefore, we always mutated I23 to the most commonly observed residue tyrosine (I23Y) in the TPR consensus sequence together with aforementioned

mutation K7L. In addition, we considered combination K7R and I23D together. Combination K7-I23D was also tested because of highly similar physicochemical properties between Lys and Arg side chains.

The hydrophobic side chain of valine at position 9 in RPS20-hhta is buried between helices in RPS20, but would be exposed on the surface of the designed protein except in the last repeat, in which V9 interacts with the stop helix. Therefore, it is considered to be mutated to the most common residue asparagine (V9N) in the TPR repeat consensus except in the last repeat (Figure 4c).

RPS20-hhta sequence and surface is enriched with positively charged residues (Figure 4b). This would lead to the exceedingly high theoretical iso-electric point (pI) of the designed proteins. Natural TPR proteins tend to exhibit zero net charge (Magliery and Regan, 2004). Hence, we decided to randomly mutate the positively charged residues (Lys and Arg) in the two helices of RPS20-hhta to the corresponding most common residues in TPR sequence profile (K2E, K6N, K22E, R25Q/E). K26 was not mutated as Lys is already the most common residue in the TPR profile.

At the C-terminus of the designed TPR, the last four residue of RPS20-hhta (IDKA) were replaced with the TPR consensus loop sequence (DPNN) between repeat units. The reason is as follows. The secondary structure of the TPR unit is helix (13 aa) – loop (3aa) – helix (14 aa) – loop (4aa), while the secondary structure of the RPS20-hhta identified to be homologous to TPR unit is helix (13 aa) – loop (3 aa) – helix (18 aa) (Figure 2 and 4). The last four residues may have been included in the prediction by TPRpred merely to fulfill the size requirement of TPR repeat (34 aa). Indeed, when we scanned RPS20-hhta sequence using the hidden Markov model constructed for Pfam family *TPR_1*, only positions 2-28 were found to be similar to the *TPR_1* profile using HMMER 3.0 (RRID:SCR_005305) (Eddy, 2009), even if all filters were switched off. So the four very C-terminal residues in RPS20-hhta were not used in the designed TPR between repeat units. They were not replaced in the last repeat unit (Figure 3).

3.5 Structure modeling and refinement *in silico*

CTPR3 structure of an idealized TPR repeat (Main et al., 2003b) (PDB id: 1na0, chain A) was taken as the main template to build an initial TPR structure model using RPS20-hhta. Helix B3 and the stop helix in our designed protein are different from natural TPRs, but more similar to natural RPS20s. So we also used a RPS20 protein as the structure template for the last repeat and the stop helix. The structure of RPS20 from *Thermus thermophilus* HB8 (PDB id: 2vqe, chain T) was used because it was the structure with the best resolution (2.5 Å). The C-terminal loop in 2vqeT was discarded. The two structures 1na0A and 2vqeT were merged into a hybrid template based on the superposition of their homologous helical hairpins: the third TPR unit in 1na0A and the RPS20-hh in 2vqeT (the very C-terminal four residues were not used). We then modeled the designed TPR sequences using RPS20-hhta onto the hybrid structure template using Rosetta programs *fixbb* and *relax* (Das and Baker, 2008). The Rosetta fixed backbone design application *fixbb* was used to make the initial model. Subsequently, these models were relaxed using the Rosetta structure refinement application *relax*. The two steps were iterated three times. See the supplementary file 1E for the command lines. Rosetta 3.4 was used in the work.

We selected five constructs for further testing *in vitro* (Table 1). They are among the best-scoring constructs according to the *in silico* simulation (Figure 4—figure supplement 2). If two constructs have comparable scores (they are adjacent in the score ranking), the one with fewer mutations was preferred. The selected constructs all differ at least at two positions in their sequences. When we searched these optimized RPS20-hhta fragments in the NCBI *nr* database using BLAST (Camacho et al., 2009), the top hits were still RPS20s.

3.6 Cloning, protein expression and purification

DNA sequences coding for the designed TPR repeats were gene-synthesized in codon-optimized form (Eurofins) and cloned into vector pET-28b (Novagen) using NcoI/HindIII restriction sites, and into pETHis_1a to generate proteins with an N-terminal cleavable His_6-tag. RPS20 *T. aquaticus* and *T.*

thermophilus genes were amplified from genomic DNA and cloned likewise. Recombinant plasmids were transformed into *E. coli* strain BL21-Gold (DE3) grown on LB agar plates containing 50 µg/ml kanamycin. For expression, cells were cultured at 25°C and induced with 1 mM isopropyl-D-thiogalactopyranoside (IPTG) at an OD_{600} of 0.6 for continued growth overnight.

Bacterial cell pellets were resuspended in buffer A (50 mM Tris pH 8, 150 mM NaCl), supplemented with 5 mM $MgCl_2$, DNaseI (Applichem) and protease inhibitor cocktail (cOmplete, Roche). After breaking the cells in a French Press, the suspension was centrifuged twice at 37,000 g. Soluble His_6-tagged proteins were purified by binding proteins to Ni-NTA columns (GE Healthcare) in buffer A (50 mM Tris pH 8.0, 300 mM NaCl) and elution with increasing concentrations of imidazole up to 0.6 M. Eluted proteins were dialyzed against buffer A for cleavage by His_6-TEV-protease (50 U/mg protein). Cleavage leaves two additional residues (Gly-Ala) as N-terminal extension to all proteins produced in this manner. After incubation overnight, cleaved proteins were re-run on Ni-NTA columns and collected in the flow-through. They were finally purified by gel size exclusion chromatography (Superdex G75, GE Healthcare) in buffer A containing 0.5 mM EDTA. Insoluble proteins were dissolved in 6 M guanidinium chloride and refolded by dialysis overnight against buffer A. Refolded proteins were further purified by sequential anion-exchange (Q Sepharose HP) and cation-exchange (SP Sepharose HP) chromatography using 0-500 mM NaCl salt gradients in buffer D (20 mM Tris pH 8, 1 mM EDTA), and by gel size exclusion chromatography (Superdex G75) in buffer A.

3.7 Biophysical characterization

To determine the native molecular mass of designed TPR repeats, static light scattering experiment were performed by applying samples onto a superdex S200 gel size exclusion column to which a miniDAWN Tristar Laser photometer (Wyatt) and an RI 2031 differential refractometer (JASCO) were coupled. Runs were performed in buffer A. Data analysis and molecular mass calculations were carried out with ASTRA V software (Wyatt). Tryptophan fluorescence spectra were recorded on a Jasco FP-6500 spectrofluorometer at 23°C; excitation was at 280 nm, emission spectra were collected from 300-400 nm. Circular dichroism (CD) spectra from 200-250 nm were recorded with a Jasco J-810

spectropolarimeter at 23°C in buffer E (30 mM MOPS pH 7.2, 150 mM NaCl). Cuvettes of 1 mm path length were used in all measurements. For melting curves and determination of Tm, CD measurements were recorded at 222 nm from 20-95°C, the temperature change was set to 1°C per minute, using a Peltier-controlled sample holder unit. For equilibrium-unfolding experiments performed at 23°C, native protein was mixed with different concentrations of urea in buffer A. After equilibration, circular dichroism was monitored at 222 nm. The fraction of unfolded protein f_U was determined based on $f_u = (y_F - y)/(y_F - y_U)$, where y_F and y_U are the values of y typical of the folded and unfolded states. Data were fitted to a two-state model with the software ProFit (6.1) as described (Grimsley et al., 2013), assuming a linear urea *[D]* dependence of ΔG: $\Delta G_{U\text{-}F}^{D} = \Delta G_{U\text{-}F}^{H2O} - m[D]$, where $\Delta G_{U\text{-}F}^{D}$ is the free energy change at a given denaturant concentration, $\Delta G_{U\text{-}F}^{H2O}$ the free energy change in the absence of denaturant, and m the sensitivity of the transition to denaturant. Fragment sizes of M4N were determined by ESI-micrOTOF mass spectrometry (Bruker Daltonics, Max Planck Institute core facility Martinsried), followed by bioinformatic analysis using the Find-Pept tool (ExPASy).

3.8 Crystallization, structure solution and refinement

For crystallization, the M4N and M4NΔC protein solutions were concentrated to 70 and 30 mg/ml, respectively, in buffer A. The buffer for M4NΔC additionally contained 0.5 mM EDTA. Crystallization trials were performed at 295 K in 96-well sitting-drop vapor-diffusion plates with 50 µl of reservoir solution and drops consisting of 300 nl protein solution and 300 nl reservoir solution in the case of M4N, and 400 nl protein solution and 200 nl reservoir solution in the case of M4NΔC. Crystallization conditions for the crystals used in the diffraction experiments are listed in supplementary file 1H together with the solutions used for cryo-protection. Single crystals were transferred into a droplet of cryo-protectant before loop-mounting and flash-cooling in liquid nitrogen. For experimental phasing, crystals of M4N were soaked overnight in a droplet containing reservoir solution supplemented with 5 mM K_2PtCl_4 prior to cryo-protection and flash-cooling. All data were collected at beamline X10SA (PXII) at the Swiss Light Source (Paul Scherrer Institute, Villigen, Switzerland) at

100 K using a PILATUS 6M detector (DECTRIS) at the wavelengths indicated in supplementary file 1H. Diffraction images were processed and scaled using the XDS program suite (Kabsch, 1993). Using SHELXD (Sheldrick, 2008), three strong Pt-sites were identified in the M4N derivative dataset. After density modification with SHELXE, the resulting electron density map could be traced by Buccaneer (Cowtan, 2006) to large extents, and revealed three chains of M4N in the asymmetric unit (ASU), organized as one dimer and one monomer. Refinement was continued using the native dataset. The two different crystal forms of M4NΔC, CF I and CF II, were solved by molecular replacement on the basis of the refined M4N coordinates. Using MOLREP (Vagin and Teplyakov, 2000), two copies of the dimeric assembly of the M4N structure were located in the ASU of CF I, and one monomer in the ASU of CF II. All models were completed by cyclic manual modeling with Coot (Emsley and Cowtan, 2004) and refinement with PHENIX (RRID:SCR_014224) (Adams et al., 2010). Analysis with PROCHECK (Laskowski et al., 1993) showed excellent geometries for all structures. Data collection and refinement statistics are summarized in supplementary file 1H. The three structures are deposited in the PDB (Berman et al., 2000) with accession codes: 5FZQ (M4N), 5FZR (M4NΔC CF I), 5FZS (M4NΔC CF II).

3.9 Testing mutations in *T. thermophilus*

T. thermophilus HB8 and *T. aquaticus* YT-1 were obtained from the German Collection of Microorganisms and Cell Cultures (DSMZ). Growth in liquid medium was performed under mild stirring (160 rpm) in long necked flasks at 68°C with DSMZ Medium 74 for *T. thermophilus* and DSMZ Medium 878 for *T. aquaticus*. Agar (1.6% w/v) was added to the medium for growth on plates. When required, kanamycin (30 µg/ml) and bleocin (10 µg/ml) were added to the media. For purification experiments 25 ml cultures were grown to an optic density of 0.7 OD_{600} (~12 hours) and then re-inoculated in the same volume to an optical density of 0.035 OD_{600}. The process was repeated serially three times and two 5 ml samples were taken in each step for glycerol stocks and DNA purification. Transformation of *T. thermophilus* was performed as described previously (Nguyen and Silberg,

2010). Genomic and plasmid DNA from Thermus were purified from 5 ml cultures using the QIAamp DNA Mini Kit and the QIAprep Spin Miniprep Kit, respectively.

T. thermophilus KM4 strain was generated by gene replacement as follows: two PCR products comprising each one 1 Kb of DNA upstream and downstream of *rpsT* were amplified from *T. thermophilus* HB8 genomic DNA and then fused by overlapping PCR. The resulting fragment, in which *rpsT* is substituted by a PstI site, was cloned in the KpnI/XbaI sites of plasmid pBlueScript II SK (+). Next, a fragment from plasmid pKT1 (Biotools, Spain), which contains the thermostable kanamycin resistance *Kat* gene under the control of the constitutive PslpA promoter, was inserted into the new PstI site. Direction of the *Kat* cassette insertion was selected, so transcription from the PslpA promoter continues through *thx*, a gene that is located downstream and is predicted to form an operon with *rpsT*. The 3 Kb final construct cloned in pBluescript was subsequently amplified by PCR and the linear product was purified and transformed by electroporation in *T. thermophilus* HB8. Integration of the *Kat* cassette was selected by growth in kanamycin.

For the complementation in trans of *rpsT* from *T. thermophilus*, a PCR product of *rpsT* was amplified from genomic DNA and cloned in the SpeI/PstI sites of plasmid pJJSpro (Nguyen and Silberg, 2010) generating plasmid pJJSprorps20Tt. The same approach was followed for *rpsT* in *T. aquaticus* (pJJSpro-rpsTTa) and in *T. aquaticus rpsT* alleles with two (pJJSpro-rpsTTaM2) and four (pJJSpro-rpsTTaM4N) amino-acid substitutions. The PCR product for the two later constructs was amplified using the plasmids in which the synthesized genes were delivered as a template.

4 Acknowledgments

We thank Elisabeth Weyher from the Core Facility of the MPI for Biochemistry, Martinsried, for analyzing proteins by mass spectrometry. We are grateful to the staff of beamline PXII/Swiss Light Source for their technical support. We also thank Birte Höcker, Vikram Alva and Sergey Samsonov for many helpful comments and discussions. This work was supported by institutional funds of the Max Planck Society.

5 References

Abe Y, Shodai T, Muto T, Mihara K, Torii H, Nishikawa S, Endo T, Kohda D. 2000. Structural basis of presequence recognition by the mitochondrial protein import receptor Tom20. *Cell* **100**(5): 551-560.

Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung L-W, Kapral GJ, Grosse-Kunstleve RW, others. 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr Sect D Biol Crystallogr* **66**(2): 213-221.

Alva V, Ammelburg M, Söding J, Lupas AN. 2007. On the origin of the histone fold. *BMC Struct Biol* **7**(17): 1-10.

Alva V, Söding J, Lupas AN. 2015. A vocabulary of ancient peptides at the origin of folded proteins. *Elife* **4**(pii:e09410). doi: 10.7554/eLife.09410.

Alva V, Nam S-Z, Söding J, Lupas AN. 2016. The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Res* **44**(W1): W410-W415.

Anantharaman V, Koonin EV, Aravind L. 2001. TRAM, a predicted RNA-binding domain, common to tRNA uracil methylation and adenine thiolation enzymes. *FEMS Microbiol Lett* **197**(2): 215-221.

Apic G, Gough J, Teichmann SA. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* **310**(2): 311-325.

Aurora R, Rose GD. 1998. Helix capping. *Protein Sci* **7**(1): 21-38.

Bansal PK, Mishra A, High AA, Abdulle R, Kitagawa K. 2009a. Sgt1 dimerization is negatively regulated by protein kinase CK2-mediated phosphorylation at Ser361. *J Biol Chem* **284**(28): 18692-18698.

Bansal PK, Nourse A, Abdulle R, Kitagawa K. 2009b. Sgt1 dimerization is required for yeast kinetochore assembly. *J Biol Chem* **284**(6): 3586-3592.

641 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000.
642 The Protein Data Bank. *Nucleic Acids Res* **28**(1): 235-242.
643 Bernhardt HS. 2012. The RNA world hypothesis: the worst theory of the early evolution of life (except
644 for all the others). *Biol Direct* **7**(23): 1-10.
645 Biegert A, Mayer C, Remmert M, Söding J, Lupas AN. 2006. The MPI Bioinformatics Toolkit for protein
646 sequence analysis. *Nucleic Acids Res* **34**(Web Server issue): W335-W339.
647 Binz HK, Stumpp MT, Forrer P, Amstutz P, Plückthun A. 2003. Designing repeat proteins: well-
648 expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin
649 repeat proteins. *J Mol Biol* **332**(2): 489-503.
650 Blundell TL, Sewell BT, McLachlan AD. 1979. Four-fold structural repeat in the acid proteases.
651 *Biochim Biophys Acta* **580**(1): 24-31.
652 Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis
653 SD, Merezhuk Y, Raytselis Y, Sayers EW, Tao T, Ye J, Zaretskaya I. 2013. BLAST: a more
654 efficient report with usability improvements. *Nucleic Acids Res* **41**(Web Server issue): W29-
655 W33.
656 Broom A, Doxey AC, Lobsanov YD, Berthin LG, Rose DR, Howell PL, McConkey BJ, Meiering EM. 2012.
657 Modular evolution and the origins of symmetry: reconstruction of a three-fold symmetric
658 globular protein. *Structure* **20**(1): 161-171.
659 Bubunenko M, Baker T, Court DL. 2007. Essentiality of ribosomal and transcription antitermination
660 proteins analyzed by systematic gene replacement in Escherichia coli. *J Bacteriol* **189**(7):
661 2844-2853.
662 Burton B, Zimmermann MT, Jernigan RL, Wang Y. 2012. A computational investigation on the
663 connection between dynamics properties of ribosomal proteins and ribosome assembly.
664 *PLoS Comput Biol* **8**(5): e1002530.
665 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+:
666 architecture and applications. *BMC Bioinformatics* **10**(421): 1-9.

667 Chaudhuri I, Söding J, Lupas AN. 2008. Evolution of the beta-propeller fold. *Proteins* **71**(2): 795-803.
668 Chen C, Natale DA, Finn RD, Huang H, Zhang J, Wu CH, Mazumder R. 2011. Representative
669 proteomes: a stable, scalable and unbiased proteome set for sequence analysis and
670 functional annotation. *PLoS One* **6**(4): e18910.
671 Coquille S, Filipovska A, Chia T, Rajappa L, Lingford JP, Razif MFM, Thore S, Rackham O. 2014. An
672 artificial PPR scaffold for programmable RNA recognition. *Nat Commun* **5**: 5729.
673 Cortajarena AL, Regan L. 2006. Ligand binding by TPR domains. *Protein Sci* **15**(5): 1193-1198.
674 Cowtan K. 2006. The Buccaneer software for automated model building. 1. Tracing protein chains.
675 *Acta Crystallogr Sect D Biol Crystallogr* **62**(9): 1002-1011.
676 Crick FHC. 1953. The packing of α-helices: simple coiled-coils. *Acta Crystallogr* **6**(8-9): 689-697.
677 Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome*
678 *Res* **14**(6): 1188-1190.
679 D'Andrea LD, Regan L. 2003. TPR proteins: the versatile helix. *Trends Biochem Sci* **28**(12): 655-662.
680 Das AK, Cohen PW, Barford D. 1998. The structure of the tetratricopeptide repeats of protein
681 phosphatase 5: implications for TPR-mediated protein-protein interactions. *EMBO J* **17**(5):
682 1192-1199.
683 Das R, Baker D. 2008. Macromolecular modeling with rosetta. *Annu Rev Biochem* **77**: 363-382.
684 de Juan D, Pazos F, Valencia A. 2013. Emerging methods in protein co-evolution. *Nat Rev Genet* **14**(4):
685 249-261.
686 Di Domenico T, Potenza E, Walsh I, Parra RG, Giollo M, Minervini G, Piovesan D, Ihsan A, Ferrari C,
687 Kajava AV, Tosatto SCE. 2014. RepeatsDB: a database of tandem repeat protein structures.
688 *Nucleic Acids Res* **42**(Database issue): D352-D357.
689 Dosztányi Z, Csizmók V, Tompa P, Simon I. 2005a. The pairwise energy content estimated from amino
690 acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol*
691 *Biol* **347**(4): 827-839.

692 Dosztányi Z, Csizmók V, Tompa P, Simon I. 2005b. IUPred: web server for the prediction of
693 intrinsically unstructured regions of proteins based on estimated energy content.
694 *Bioinformatics* **21**(16): 3433-3434.
695 Doyle L, Hallinan J, Bolduc J, Parmeggiani F, Baker D, Stoddard BL, Bradley P. 2015. Rational design of
696 α-helical tandem repeat proteins with closed architectures. *Nature* **528**(7583): 585-588.
697 Dunin-Horkawicz S, Kopec KO, Lupas AN. 2014. Prokaryotic ancestry of eukaryotic protein networks
698 mediating innate immunity and apoptosis. *J Mol Biol* **426**(7): 1568-1582.
699 Dyson HJ, Wright PE. 2005. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell*
700 *Biol* **6**(3): 197-208.
701 Eck RV, Dayhoff MO. 1966. Evolution of the structure of ferredoxin based on living relics of primitive
702 amino acid sequences. *Science* **152**(3720): 363-366.
703 Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome*
704 *Inform* **23**(1): 205-211.
705 Emsley P, Cowtan K. 2004. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol*
706 *Crystallogr* **60**(Pt 12 Pt 1): 2126-2132.
707 Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L,
708 Mistry J, Sonnhammer ELL, Tate J, Punta M. 2014. Pfam: the protein families database.
709 *Nucleic Acids Res* **42**(Database issue): D222-D230.
710 Forrer P, Binz HK, Stumpp MT, Plückthun A. 2004. Consensus design of repeat proteins.
711 *Chembiochem* **5**(2): 183-189.
712 Fox GE. 2010. Origin and evolution of the ribosome. *Cold Spring Harb Perspect Biol* **2**(9): a003483.
713 Fox NK, Brenner SE, Chandonia J-M. 2014. SCOPe: Structural Classification of Proteins--extended,
714 integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res*
715 **42**(Database issue): D304-D309.
716 Gilbert W. 1986. Origin of life: The RNA world. *Nature* **319**: 618.

717 Gorodkin J, Staerfeldt HH, Lund O, Brunak S. 1999. MatrixPlot: visualizing sequence constraints.
718 *Bioinformatics* **15**(9): 769-770.
719 Grimsley GR, Trevino SR, Thurlkill RL, Scholtz JM. 2013. Determining the conformational stability of a
720 protein from urea and thermal unfolding curves. *Curr Protoc Protein Sci* **Chapter**
721 **28**(Unit28.4): 28.24.21-28.24.14.
722 Habchi J, Tompa P, Longhi S, Uversky VN. 2014. Introducing protein intrinsic disorder. *Chem Rev*
723 **114**(13): 6561-6588.
724 Hsiao C, Mohan S, Kalahar BK, Williams LD. 2009. Peeling the onion: ribosomes are ancient molecular
725 fossils. *Mol Biol Evol* **26**(11): 2415-2425.
726 Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol*
727 *Evol* **23**(2): 254-267.
728 Iwaya N, Kuwahara Y, Fujiwara Y, Goda N, Tenno T, Akiyama K, Mase S, Tochio H, Ikegami T,
729 Shirakawa M, Hiroaki H. 2010. A common substrate recognition mode conserved between
730 katanin p60 and VPS4 governs microtubule severing and membrane skeleton reorganization.
731 *J Biol Chem* **285**(22): 16822-16829.
732 Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of
733 hydrogen-bonded and geometrical features. *Biopolymers* **22**(12): 2577-2637.
734 Kabsch W. 1993. Automatic processing of rotation diffraction data from crystals of initially unknown
735 symmetry and cell constants. *J Appl Crystallogr* **26**(6): 795-800.
736 Kajander T, Cortajarena AL, Mochrie S, Regan L. 2007. Structure and stability of designed TPR protein
737 superhelices: unusual crystal packing and implications for natural TPR proteins. *Acta*
738 *Crystallogr D Biol Crystallogr* **63**(Pt 7): 800-811.
739 Kajava AV. 2012. Tandem repeats in proteins: from sequence to structure. *J Struct Biol* **179**(3): 279-
740 288.
741 Karpenahalli MR, Lupas AN, Söding J. 2007. TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like
742 repeats from protein sequences. *BMC Bioinformatics* **8**(2): 1-8.

743 Katibah GE, Qin Y, Sidote DJ, Yao J, Lambowitz AM, Collins K. 2014. Broad and adaptable RNA
744 structure recognition by the human interferon-induced tetratricopeptide repeat protein
745 IFIT5. *Proc Natl Acad Sci U S A* **111**(33): 12025-12030.
746 Keefe AD, Szostak JW. 2001. Functional proteins from a random-sequence library. *Nature* **410**(6829):
747 715-718.
748 Keiski C-L, Harwich M, Jain S, Neculai AM, Yip P, Robinson H, Whitney JC, Riley L, Burrows LL, Ohman
749 DE, Howell PL. 2010. AlgK is a TPR-containing protein and the periplasmic component of a
750 novel exopolysaccharide secretin. *Structure* **18**(2): 265-273.
751 Kobe B, Kajava AV. 2000. When protein folding is simplified to protein coiling: the continuum of
752 solenoid protein structures. *Trends Biochem Sci* **25**(10): 509-515.
753 Kohl A, Binz HK, Forrer P, Stumpp MT, Plückthun A, Grütter MG. 2003. Designed to be stable: crystal
754 structure of a consensus ankyrin repeat protein. *Proc Natl Acad Sci U S A* **100**(4): 1700-1705.
755 Koonin EV. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor.
756 *Nat Rev Microbiol* **1**(2): 127-136.
757 Kopec KO, Lupas AN. 2013. β-Propeller blades as ancestral peptides in protein evolution. *PLoS One*
758 **8**(10): e77074.
759 Krachler AM, Sharma A, Kleanthous C. 2010. Self-association of TPR domains: Lessons learned from a
760 designed, consensus-based TPR oligomer. *Proteins* **78**(9): 2131-2143.
761 Kumar S, Bansal M. 1998. Dissecting alpha-helices: position-specific analysis of alpha-helices in
762 globular proteins. *Proteins* **31**(4): 460-476.
763 Kyrpides N, Overbeek R, Ouzounis C. 1999. Universal protein families and the functional content of
764 the last universal common ancestor. *J Mol Evol* **49**(4): 413-423.
765 Kyrpides NC, Woese CR. 1998. Tetratrico-peptide-repeat proteins in the archaeon Methanococcus
766 jannaschii. *Trends Biochem Sci* **23**(7): 245-247.
767 Lamb JR, Tugendreich S, Hieter P. 1995. Tetratrico peptide repeat interactions: to TPR or not to TPR?
768 *Trends Biochem Sci* **20**(7): 257-259.

769 Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace
770 IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X
771 version 2.0. *Bioinformatics* **23**(21): 2947-2948.
772 Laskowski RA, MacArthur MW, Moss DS, Thornton JM. 1993. PROCHECK: a program to check the
773 stereochemical quality of protein structures. *J Appl Crystallogr* **26**(2): 283-291.
774 Lee J, Blaber M. 2011. Experimental support for the evolution of symmetric protein architecture from
775 a simple peptide motif. *Proc Natl Acad Sci U S A* **108**(1): 126-130.
776 Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. 2003. Protein disorder prediction:
777 implications for structural proteomics. *Structure* **11**(11): 1453-1459.
778 Lunelli M, Lokareddy RK, Zychlinsky A, Kolbe M. 2009. IpaB-IpgC interaction defines binding motif for
779 type III secretion translocator. *Proc Natl Acad Sci U S A* **106**(24): 9661-9666.
780 Lupas AN, Ponting CP, Russell RB. 2001. On the evolution of protein folds: are similar motifs in
781 different protein folds the result of convergence, insertion, or relics of an ancient peptide
782 world? *J Struct Biol* **134**(2-3): 191-203.
783 Magliery TJ, Regan L. 2004. Beyond consensus: statistical free energies reveal hidden interactions in
784 the design of a TPR motif. *J Mol Biol* **343**(3): 731-745.
785 Main ERG, Jackson SE, Regan L. 2003a. The folding and design of repeat proteins: reaching a
786 consensus. *Curr Opin Struct Biol* **13**(4): 482-489.
787 Main ERG, Xiong Y, Cocco MJ, D'Andrea L, Regan L. 2003b. Design of stable alpha-helical arrays from
788 an idealized TPR motif. *Structure* **11**(5): 497-508.
789 Main ERG, Lowe AR, Mochrie SGJ, Jackson SE, Regan L. 2005. A recurring theme in protein
790 engineering: the design, stability and folding of repeat proteins. *Curr Opin Struct Biol* **15**(4):
791 464-471.
792 McLachlan AD. 1972. Repeating sequences and gene duplication in proteins. *J Mol Biol* **64**(2): 417-
793 437.

McLachlan AD. 1987. Gene duplication and the origin of repetitive protein structures. *Cold Spring Harb Symp Quant Biol* **52**: 411-420.

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* **108**(49): E1293-E1301.

Mosavi LK, Minor J, Daniel L, Peng Z-Y. 2002. Consensus-derived structural determinants of the ankyrin repeat motif. *Proc Natl Acad Sci U S A* **99**(25): 16029-16034.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**(4): 536-540.

Nguyen PQ, Silberg JJ. 2010. A selection that reports on protein-protein interactions within a thermophilic bacterium. *Protein Eng Des Sel* **23**(7): 529-536.

Noble WS. 2009. How does multiple testing correction work? *Nat Biotechnol* **27**(12): 1135-1137.

Ohtani N, Tomita M, Itaya M. 2010. An extreme thermophile, Thermus thermophilus, is a polyploid bacterium. *J Bacteriol* **192**(20): 5499-5505.

Oldfield CJ, Dunker AK. 2014. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem* **83**: 553-584.

Orengo CA, Thornton JM. 2005. Protein families and their evolution-a structural perspective. *Annu Rev Biochem* **74**: 867-900.

Park K, Shen BW, Parmeggiani F, Huang P-S, Stoddard BL, Baker D. 2015. Control of repeat-protein curvature by computational protein design. *Nat Struct Mol Biol* **22**(2): 167-174.

Parmeggiani F, Huang P-S, Vorobiev S, Xiao R, Park K, Caprari S, Su M, Seetharaman J, Mao L, Janjua H, Montelione GT, Hunt J, Baker D. 2015. A general computational approach for repeat protein design. *J Mol Biol* **427**(2): 563-575.

Paterakis K, Littlechild J, Woolley P. 1983. Structural and functional studies on protein S20 from the 30-S subunit of the Escherichia coli ribosome. *Eur J Biochem* **129**(3): 543-548.

819 Peng Z, Oldfield CJ, Xue B, Mizianty MJ, Dunker AK, Kurgan L, Uversky VN. 2014. A creature with a
820 hundred waggly tails: intrinsically disordered proteins in the ribosome. *Cell Mol Life Sci* **71**(8):
821 1477-1504.
822 Ponting CP, Russell RR. 2002. The natural history of protein domains. *Annu Rev Biophys Biomol Struct*
823 **31**: 45-71.
824 Ramarao MK, Bianchetta MJ, Lanken J, Cohen JB. 2001. Role of rapsyn tetratricopeptide repeat and
825 coiled-coil domains in self-association and nicotinic acetylcholine receptor clustering. *J Biol*
826 *Chem* **276**(10): 7475-7483.
827 Rämisch S, Weininger U, Martinsson J, Akke M, André I. 2014. Computational design of a leucine-rich
828 repeat protein with a predefined geometry. *Proc Natl Acad Sci U S A* **111**(50): 17875-17880.
829 Ranea JAG, Sillero A, Thornton JM, Orengo CA. 2006. Protein superfamily evolution and the last
830 universal common ancestor (LUCA). *J Mol Evol* **63**(4): 513-525.
831 Remmert M, Biegert A, Linke D, Lupas AN, Söding J. 2010. Evolution of outer membrane beta-barrels
832 from an ancestral beta beta hairpin. *Mol Biol Evol* **27**(6): 1348-1358.
833 Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. 2001. Sequence complexity of
834 disordered protein. *Proteins* **42**(1): 38-48.
835 Sawyer N, Chen J, Regan L. 2013. All repeats are not equal: a module-based approach to guide repeat
836 protein design. *J Mol Biol* **425**(10): 1826-1838.
837 Schluenzen F, Tocilj A, Zarivach R, Harms J, Gluehmann M, Janell D, Bashan A, Bartels H, Agmon I,
838 Franceschi F, Yonath A. 2000. Structure of functionally activated small ribosomal subunit at
839 3.3 angstroms resolution. *Cell* **102**(5): 615-623.
840 Schrödinger L, 2010. The PyMOL Molecular Graphics System, Version 1.3r1.
841 Scott A, Gaspar J, Stuchell-Brereton MD, Alam SL, Skalicky JJ, Sundquist WI. 2005. Structure and
842 ESCRT-III protein interactions of the MIT domain of human VPS4A. *Proc Natl Acad Sci U S A*
843 **102**(39): 13813-13818.

844 Serasinghe MN, Yoon Y. 2008. The mitochondrial outer membrane protein hFis1 regulates
845 mitochondrial morphology and fission through self-interaction. *Exp Cell Res* **314**(19): 3494-
846 3507.
847 Shatsky M, Nussinov R, Wolfson HJ. 2004. A method for simultaneous alignment of multiple protein
848 structures. *Proteins* **56**(1): 143-156.
849 Sheldrick GM. 2008. A short history of SHELX. *Acta Crystallogr Sect A Found Crystallogr* **64**(1): 112-
850 122.
851 Shen C, Zhang D, Guan Z, Liu Y, Yang Z, Yang Y, Wang X, Wang Q, Zhang Q, Fan S, Zou T, Yin P. 2016.
852 Structural basis for specific single-stranded RNA recognition by designer pentatricopeptide
853 repeat proteins. *Nat Commun* **7**: 11285.
854 Sikorski RS, Boguski MS, Goebl M, Hieter P. 1990. A repeating amino acid motif in CDC23 defines a
855 family of proteins and a new relationship among genes required for mitosis and RNA
856 synthesis. *Cell* **60**(2): 307-317.
857 Söding J, Lupas AN. 2003. More than the sum of their parts: on the evolution of proteins from
858 peptides. *Bioessays* **25**(9): 837-846.
859 Söding J, Remmert M, Biegert A. 2006. HHrep: de novo protein repeat detection and the origin of
860 TIM barrels. *Nucleic Acids Res* **34**(Web Server issue): W137-W142.
861 Stumpp MT, Forrer P, Binz HK, Plückthun A. 2003. Designing repeat proteins: modular leucine-rich
862 repeat protein libraries based on the mammalian ribonuclease inhibitor family. *J Mol Biol*
863 **332**(2): 471-487.
864 Tobin C, Mandava CS, Ehrenberg M, Andersson DI, Sanyal S. 2010. Ribosomes lacking protein S20 are
865 defective in mRNA binding and subunit association. *J Mol Biol* **397**(3): 767-776.
866 Vagin A, Teplyakov A. 2000. An approach to multi-copy search in molecular replacement. *Acta*
867 *Crystallogr D Biol Crystallogr* **56**(Pt 12): 1622-1624.
868 Varadi M, Kosol S, Lebrun P, Valentini E, Blackledge M, Dunker AK, Felli IC, Forman-Kay JD, Kriwacki
869 RW, Pierattelli R, Sussman J, Svergun DI, Uversky VN, Vendruscolo M, Wishart D, Wright PE,

870 Tompa P. 2014. pE-DB: a database of structural ensembles of intrinsically disordered and of
871 unfolded proteins. *Nucleic Acids Res* **42**(Database issue): D326-D335.
872 Wei Y, Kim S, Fela D, Baum J, Hecht MH. 2003. Solution structure of a de novo protein from a
873 designed combinatorial library. *Proc Natl Acad Sci U S A* **100**(23): 13270-13273.
874 Wootton JC. 1994. Non-globular domains in protein sequences: automated segmentation using
875 complexity measures. *Comput Chem* **18**(3): 269-285.
876 Wootton JC, Federhen S. 1996. Analysis of compositionally biased regions in sequence databases.
877 *Methods Enzymol* **266**: 554-571.
878 Zeytuni N, Baran D, Davidov G, Zarivach R. 2012. Inter-phylum structural conservation of the
879 magnetosome-associated TPR-containing protein, MamA. *J Struct Biol* **180**(3): 479-487.
880 Zeytuni N, Zarivach R. 2012. Structural and functional discussion of the tetra-trico-peptide repeat, a
881 protein interaction module. *Structure* **20**(3): 397-405.
882 Zeytuni N, Cronin S, Lefèvre CT, Arnoux P, Baran D, Shtein Z, Davidov G, Zarivach R. 2015. MamA as a
883 Model Protein for Structure-Based Insight into the Evolutionary Origins of Magnetotactic
884 Bacteria. *PLoS One* **10**(6): e0130394.
885 Zhang Z, Kulkarni K, Hanrahan SJ, Thompson AJ, Barford D. 2010. The APC/C subunit Cdc16/Cut9 is a
886 contiguous tetratricopeptide repeat superhelix with a homo-dimer interface similar to Cdc27.
887 *EMBO J* **29**(21): 3733-3744.

888

889 **Figure legends**
890

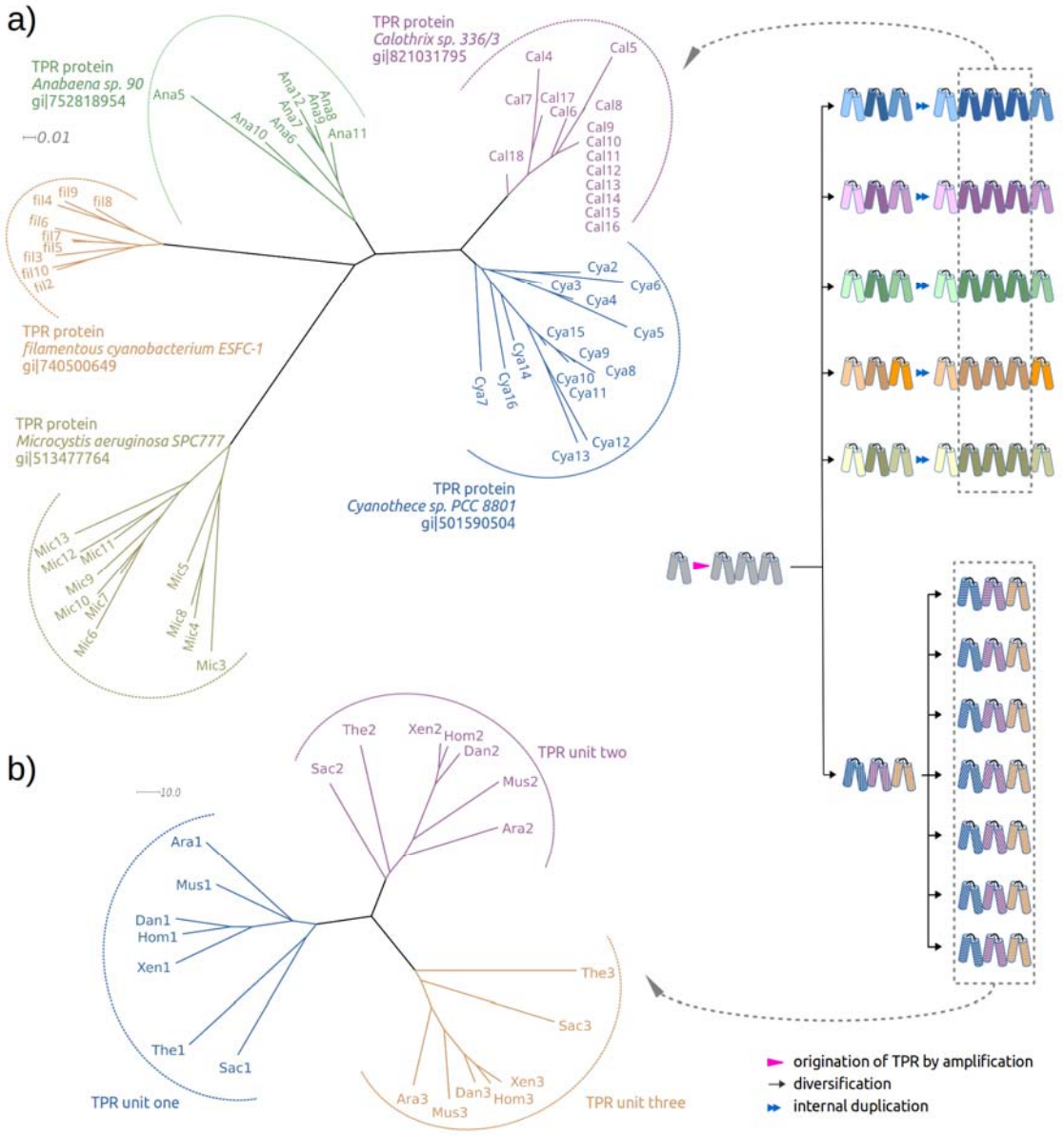


Figure 1. Two evolutionary scenarios for TPRs, illustrated by neighbor-joining phylogenetic trees. **a**) Amplification from single helical hairpin, as seen in TPR proteins from Cyanobacteria. **b**) Divergent evolution of a TPR with multiple repeat units, as seen in the TPR domains of Serine/threonine-protein phosphatase 5 (Ara: *Arabidopsis thaliana*, Dan: *Danio rerio*, Hom: *Homo sapiens*, Mus: *Musca domestica*, Sac: *Saccharomyces cerevisiae*, The: *Theileria annulata*, Xen: *Xenopus (Silurana) tropicalis*). Since evolutionary reconstructions are subject to Occam's razor and reflect the hypothesis with the fewest assumptions, we have postulated here one amplification event from one precursor hairpin. Our findings would however also be fully compatible with the precursor hairpin yielding a population of homologous variants, some of which were independently amplified to TPR-like folds;

one or more survivors among these would have become the ancestor(s) of today's TPR proteins. In this more complex scenario, the homology of TPR proteins, which we trace through the comparison of individual hairpins, is still given, but the TPR fold could have arisen from several independent amplifications, and not just a single one.

a)

```
Ana1   VQQLFKQGETAESVGDNSQAETIWRKVLQVEPNN
Ana2   GKAYNNLGNALRRQGKLPEALTAHQKALQLNPND
Ana3   AEAYVGIGNVLNAQGKPDEAVAAYRKAIEFDPKY
Ana4   AKAYNSLGNALYDQEKLKEAVAAYRKAIEFDHKY
Ana5   AAAYYNLGNVLYEQKELDEAVAAYRKAIELNPKY
Ana6   ATAYNNLGNALSDQKKLDEAVAAYQEAIKLNPKD
Ana7   ATAYNNLGIALSDQKKLDEAVAAYQKAIELDPKY
Ana8   ATAYYNLGNALSDQKKLDEAVAAYQKAIELDPKY
Ana9   ATAYYNLGNALSDQKKLDEAVAAYQKAIELDPKY
Ana10  ATAYYNLGNALRGQKKLDEAVAAYQKAIELNPKY
Ana11  ATAYNNLGIALSDQKKLDEAVAAYQKAIELNPKD
Ana12  ATAYYNLGIALSDQKKLDEAVAAYQKAIELDPKD
Ana13  AAVYNNLGNALSDQKKLKEAISNYKTALSLPEDT
Ana14  TLANNNLGLALQDQEKFAEAIKYFDKAEELDPNF

Mic1   LEQLWQQGETAQAQKKYPEAERIWRQIIQLDPNS
Mic2   AVAFSNLCAALFRQNKLDEAPIFCQKALALDPKL
Mic3   PETYYNLGNVLYNQKKLTEAEEMYRRTIELDDKF
Mic4   VYAYNNLGLVLYDQKKLKEAEEMYRRTIELDDKL
Mic5   ALVYNNLGLVLYDQKKLKEAEEMYRRAIELDDKF
Mic6   VYAYNGLGNVLRAQNKLTEAEEMYRRALALDDQY
Mic7   VDAYNGLGNVLYDQKKLTEAEEMYRRALALDDQY
Mic8   VYAYNGLGNVLYDQKKLKEAEEMYRRTIELDDKY
Mic9   VYAYKGLGNVLYNQKKLKEAEEMYRRAIALDDQY
Mic10  VHAYNSLGNVLYNQKKLTEAEEMYRRALALDDQY
Mic11  VHAYNNLGNVLYDQKKLTEAEEMYRRALALDDKY
Mic12  VPAYHNLGNVLYNQKKLTEAEEMYRRALALDDKF
Mic13  VYAYNNLGNVLYDQKKLTEAEEMYRRALDLPDDT
Mic14  TLAHNNLGLLLQEQGKLEAAIAEFEKATKIDPQY

fil1   INQLFEQGNTAQQEGRYAEAEAIWRQILEANPDN
fil2   AGAYNNLGVALYNLGQLPEAVSAYQQAIALDPDY
fil3   AIAYNNLGIALRNLGQLPEAVEAYQQAIALDPDF
fil4   AIAYYNLGIALFDLGQLPEAVSAYQQAIALDPDD
fil5   AIAYNNLGNALSNLGQLPEAVSAYQQAIALDPDD
fil6   AIAYYNLGNALSNLGQLPEAVSAYQQAITLDPDY
fil7   AIAYNNLGNALSNLGQLPEAVEAYQQAIALDPDD
fil8   ADAYYNLGNALRDLGQLPEAVSAYQQAIALDPDF
fil9   ADAYYNLGIALRDLGQLPEAVEAYQQAIALDPDD
fil10  AIAYNNLGVALYNLGQLPEAVSAYQQAIALDPDN
fil11  AFAYNNLGYAYQQQGNLEAAITEYKKAIALAPNY

Cal1   TEQLFKQGEAAESVGNNSQAETIWRQVLQLEPSN
Cal2   GKAYNNLGNALRRQGKLDEALAAHQKALQLNLND
Cal3   AEAYVGIGNVLNAQGKPEEGIAQHKKALQINPNL
Cal4   AAAYNGLGNALYDQKKLEPAVAAYQKAIQLDPNY
Cal5   AAAYYNLGNALRDQKKLEPAIAAFQKAIQLNPNF
Cal6   AAAYNGLGNALYDQKKLEPAVAAFQKAIQLNPNF
Cal7   AAAYNGLGNALYDQKKLEPAVAAYQKAIQLDPNF
Cal8   AFAYYNLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal9   AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal10  AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal11  AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal12  AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal13  AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal14  AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal15  AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal16  AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal17  AFAYNGLGNALYDQKKLEPAIAAYQKAIQLDPND
Cal18  ANAYNNLGNALYDQKKLEPAIAAYQKAIQLNPNF
Cal19  AAAYNNLGVALSDQKKLEPAIAAYQKVLTLPEDT
Cal20  TAANNGLGLVFQEQGKLKQAIDYFDKSEALDPDY

Cya1   IDQLFQQGRTAGKMGKYTEAEAIFRRVIELDPNL
Cya2   ADAYNNLGNALYYQGKLDEAIAAYQKAIQLNPND
Cya3   ADAYNNLGNALSDQGKLEEAIAAYQKAIQLNPNY
Cya4   ADAYYNLGIALSDQGKLEEAIAAYQKAIQLNPNF
Cya5   TQAYYNLGIALSDQGKLEEAIAAYQKAIQLNPNY
Cya6   ADAYYNLGNALFDQGKLDEAIAAYQKAIQLDPND
Cya7   ANAYNNLGAALYKQGKLEEAIAAYQKAIQLNPNL
Cya8   AEAYNNLGVALSDQGKRDEAIAAYQKAIQLNPNL
Cya9   AEAYNNLGVALSDQGKRDEAIAAYQKAIQLNPNF
Cya10  ALAYNNLGVALSDQGKRDEAIAAYQKAIQLNPNF
Cya11  ALAYNNLGVALSDQGKRDEAIAAYQKAIQLNPNF
Cya12  ALAYNNLGVALRNQGKRDEAIAAYQKAIQLDPND
Cya13  ANAYNNLGLALRNQGKRDEAITAYQKAIQLNPNF
Cya14  ALAYNNLGNALYSQGKREEAIAAYQKAIQLNPNF
Cya15  ALAYNNLGNALSDQGKRDEAIAAYQKAIQLNPNF
Cya16  ALAYNNLGNALSDQGKLNEAIATYQKAIQLNPNF
Cya17  ALAYNNLGNALKDQGKLNEAIAAYQKALSLPEDT
Cya18  TLAHNNLGLVYQPQGKLEEALREYEAALKIDPKF
```

b)

```
>gb|CP011382.1|:c3792015-3790486 Calothrix sp. 336/3, complete genome
GCTGCTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCAGTTGCCGCCTACCAAAAAGCCATTCAACTCGACCCTAACTAT
GCCGCTGCTTACTACAATCTCGGCAATGCCCTGAGAGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTGCTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCAGTTGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTGCTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCAGTTGCCGCCTACCAAAAAGCCATTCAACTCGACCCTAACTTT
GCTTTTGCTTACTACAATCTCGGCAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTACCAAAAAGCCATACAGCTCGACCCTAACGAT
GCCAACGCTTACAACAATCTCGGCAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTACCAAAAAGCCATTCAACTCAACCCTAACTTT
```

Figure 1—figure supplement 1. Multiple sequence alignments of recently amplified TPR repeat units. a) Alignments of the TPR units used for the phylogeny in Figure 1a. Residues different from the most common one in each column are shown in bold face and highlighted in yellow. Abbreviations: Ana: Anabaena sp. 90 (gi: 752818954, accession: WP_041458168.1); Cal: Calothrix sp. 336/3 (gi:

821031795, accession: WP_046815017.1); Cya: Cyanothece sp. PCC 8801 (gi: 501590504, accession: WP_012594639.1); fil: filamentous cyanobacterium ESFC-1 (gi: 740500649, accession: WP_038331513.1); Mic: *Microcystis aeruginosa* SPC777 (gi: 513477764, accession: EPF24195.1). b) The corresponding alignment of the DNA sequences for the most recently amplified TPR units, Cal4-Cal18, of which the central repeats, Cal9-Cal16, are fully identical. Synonymous mutations (highlighted in gray) are found at less than 1% of the nucleotides, illustrating the recent time point of the amplification. Non-synonymous mutations (highlighted in yellow) are about 2.5 times as frequent as synonymous ones.

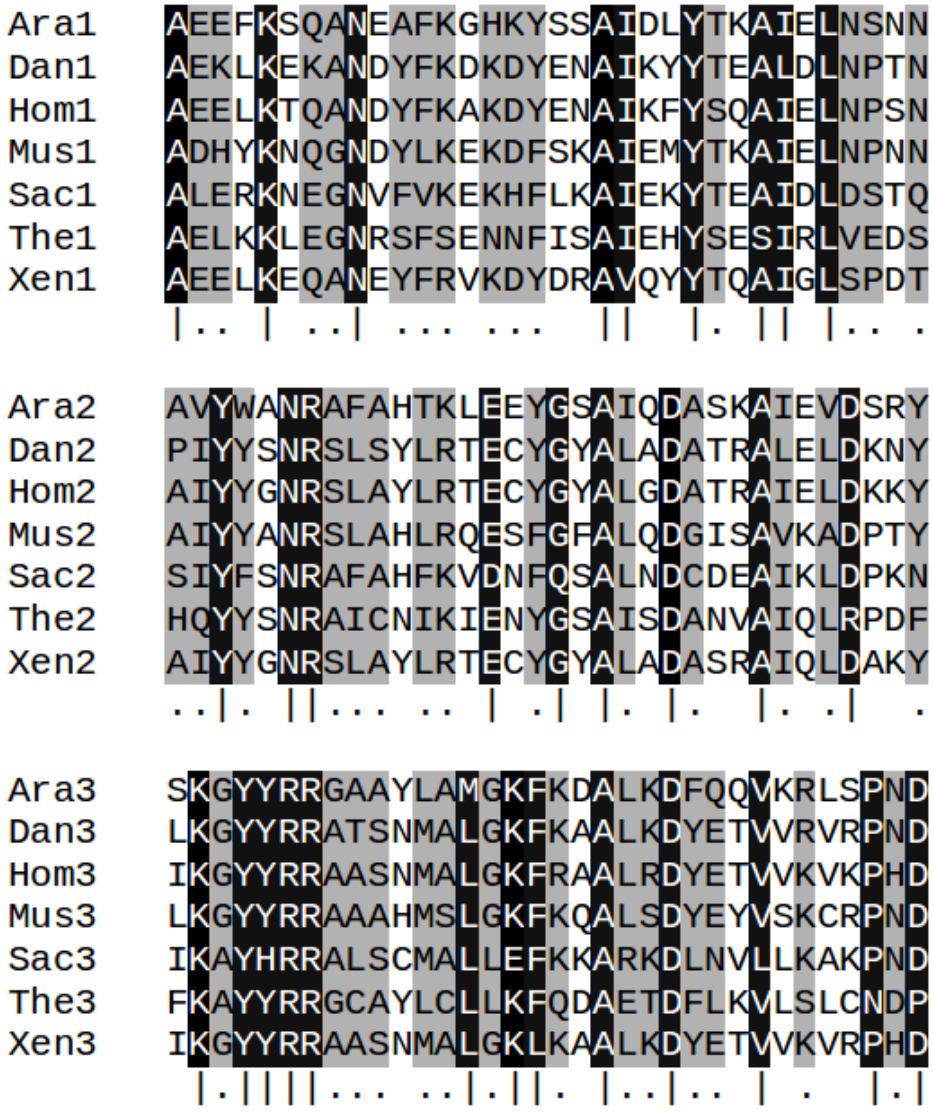
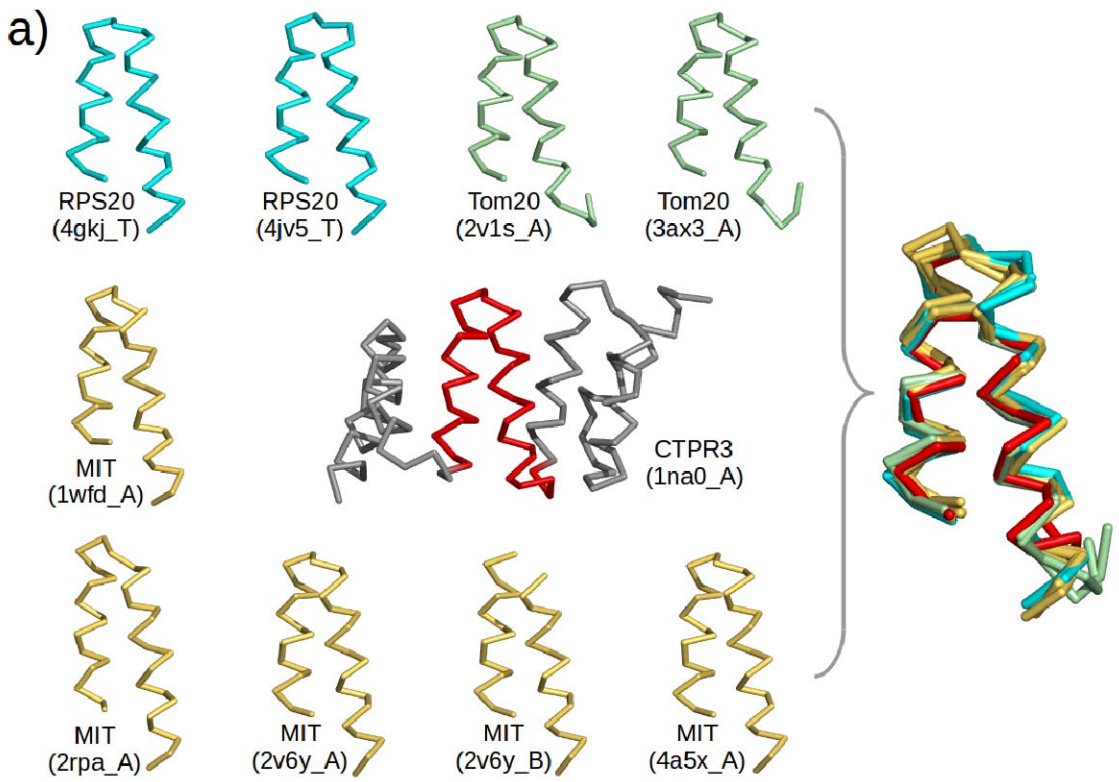
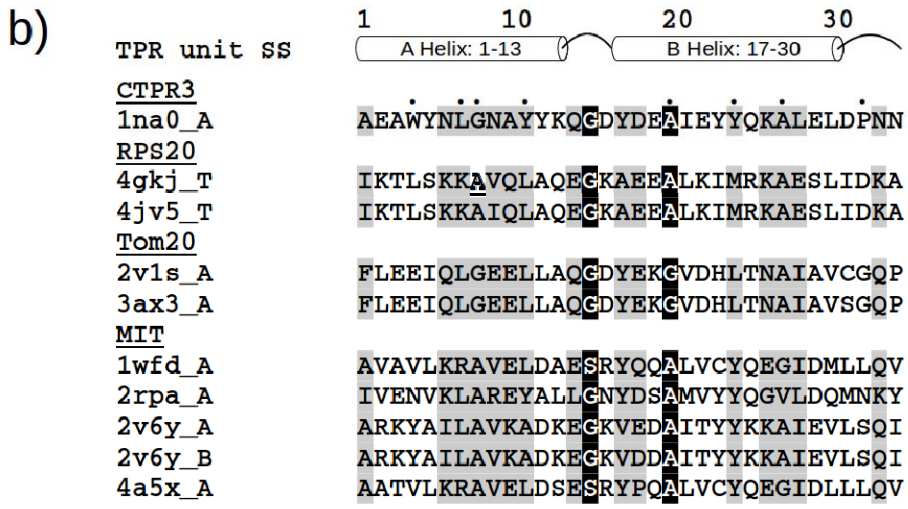


Figure 1—figure supplement 2. Multiple sequence alignment of the three TPR repeat units in serine/threonine-protein phosphatase 5 from seven taxa. Columns with identify ≥80% are highlighted in black and marked by vertical bars (|); column with identify <80% but ≥50% are highlighted in gray and marked by dots (.). Abbreviations: Ara: *Arabidopsis thaliana* (gi: 18406066, accession: NP_565985.1); Dan: *Danio rerio* (gi: 126158897, accession: NP_001014372.2); Hom: *Homo sapiens* (gi: 5453958, accession: NP_006238.1) ; Mus: *Musca domestica* (gi: 557765703, accession: XP_005182549.1); Sac: *Saccharomyces cerevisiae* S288c (gi: 398365781, accession: NP_011639.3); The: *Theileria annulata* strain Ankara (gi: 84994100, accession: XP_951772.1); Xen: *Xenopus tropicalis* (gi: 56118654, accession: NP_001007891.1).





930
931 **Figure 2. a)** Structure gallery of non-repetitive helical hairpins in the PDB that share both sequence
932 and structure similarity to TPR unit hairpin. Only the 34 amino-acid helical hairpins are shown. The
933 helical hairpins in 30S ribosomal protein s20 (RPS20), mitochondrial import receptor subunit
934 (Tom20), and microtubule interacting and transport domain (MIT) are depicted in cyan, green, and
935 yellow, respectively. The structure of a TPR with a consensus sequence, CTPR3, is shown in the
936 center with the middle TPR unit highlighted in red. PDB IDs and chain names of the proteins are given

937 in parentheses. In the superposition, all helical hairpins are superimposed onto the middle TPR unit
938 of CTPR3. **b)** Multiple sequence alignment of the helical hairpin sequences listed in a). The eight TPR
939 signature positions are marked by dots in CTPR3. Columns with sequence identity ≥ 80% are in black,
940 and columns with sequence identity ≥ 50% are in gray.

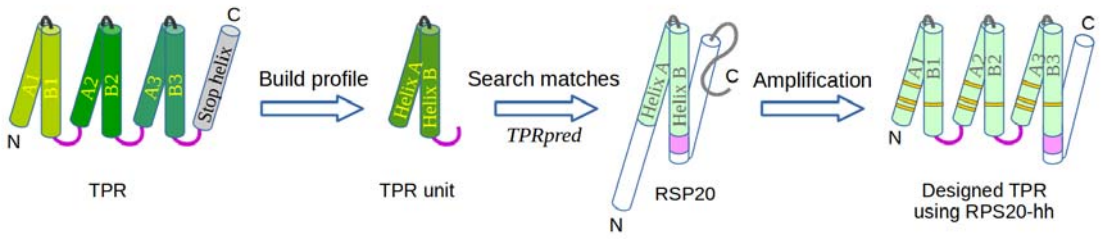
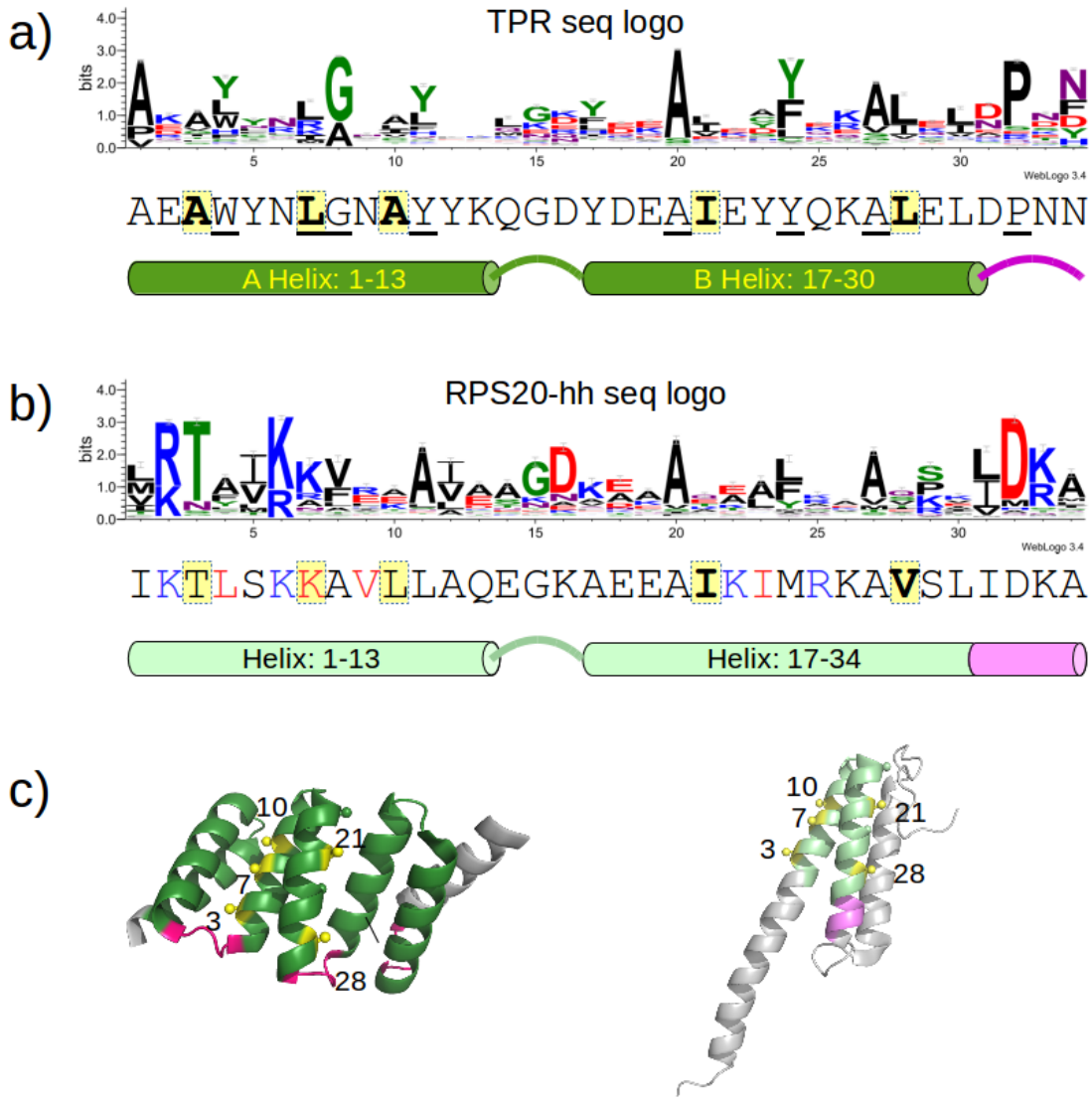


Figure 3. The design of TPR using RPS20. RPS20-hh is identified by TPRpred to match the sequence profile of TPR units. Their structures are also very similar (helices are shown as cylinders), except for the last four residues (colored in light and dark magenta). We designed a TPR protein using a RPS20-hh with up to five mutations (yellow strips) in each repeat unit. The C-terminal loop in the TPR unit (dark magenta loop) is used to replace the corresponding C-terminus (light magenta cylinder) of RPS20-hh to connect adjacent repeats. The C-terminal helix in RPS20 (white cylinder) was used as the stop helix in the design.



950
951 **Figure 4. a**) Sequence logo of the TPR motif. A TPR consensus sequence (Main et al., 2003b) (PDB:
952 1na0, chain A) and its secondary structure determined by DSSP (Kabsch and Sander, 1983) are
953 aligned below the sequence logo. The eight TPR signature positions are underscored in the consensus
954 sequence. The five interface positions are highlighted in yellow. **b**) Sequence logo of RPS20-hh. The
955 RPS20-hhta sequence and its predicted secondary structure using Quick2D (Biegert et al., 2006) is
956 aligned below the sequence logo. The derived interface positions are highlighted in yellow. The four
957 residues subjected to mutations are colored in red. The four positively charged residues selected for
958 mutation to lower the surface charge are in blue. **c)** The locations of the interface positions displayed
959 on a TPR (left) and a RPS20 structure (right). In both structures, the interface positions are labeled
960 and highlighted as yellow spheres. The TPR structure is CTPR3 (PDB: 1na0, chain A), which is shown

as a cartoon and is colored using the same scheme as the secondary structure representation in a). The stop helix is in gray. The RPS20 structure is from *T. thermophilus* (PDB: 4gkj, chain T), in which the RPS20-hh fragment is colored using the same scheme as the secondary structure representation in b).

The sequence logos were generated using WebLogo (Crooks et al., 2004). Sequences from representative proteome 75% (Chen et al., 2011) downloaded from Pfam families *TPR_1* and *Ribosomal_S20p* were used as input to WebLogo (9338 and 972 sequences, respectively). The structures were rendered using PyMOL (Schrödinger, 2010).

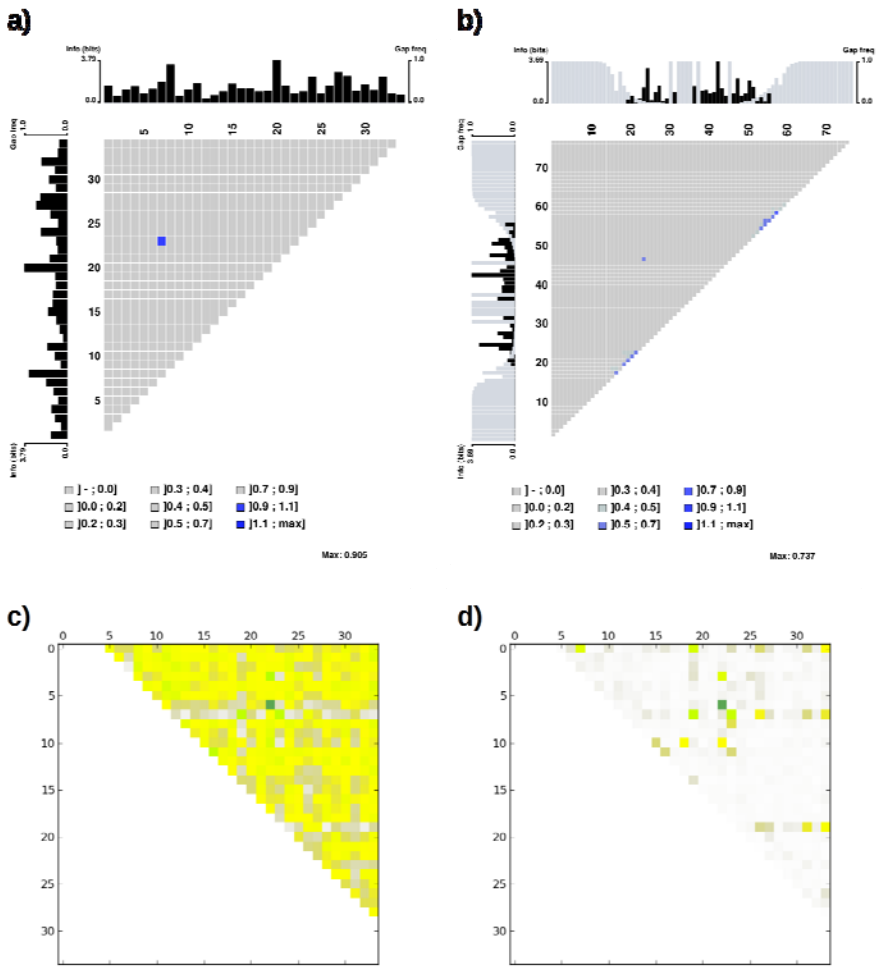
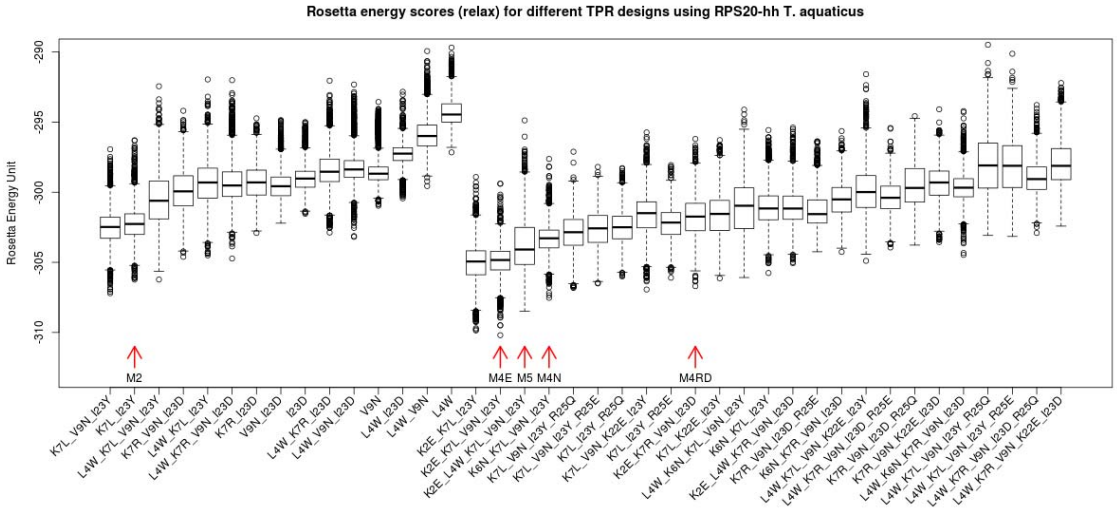


Figure 4—figure supplement 1. Mutual information plot (a and b) and direct coupling analysis plot (c and d) for TPR repeat sequences. The subfigures a) and c) were generated using the seed alignment sequences from Pfam family *TPR_1* (558 sequences. Sequence Q29585_PIG/28-61 was removed as it contains unknown residue X). The largest mutual information value is observed between position 7 and 23. The subfigures b) and d) were generated using the multiple alignment of representative proteomes rp75 sequences from Pfam family *TPR_1* (9338 sequences). The largest non-local mutual information value was observed between position 24 and 47, corresponding to position 7 and 23 using TPR repeat numbering. Alignments were taken from Pfam 27.0. Subfigures a) and b) were generated using MatrixPlot. Subfigures c) and d) were generated using DCA Workbench (http://dca.rice.edu/portal/dca/workbench).

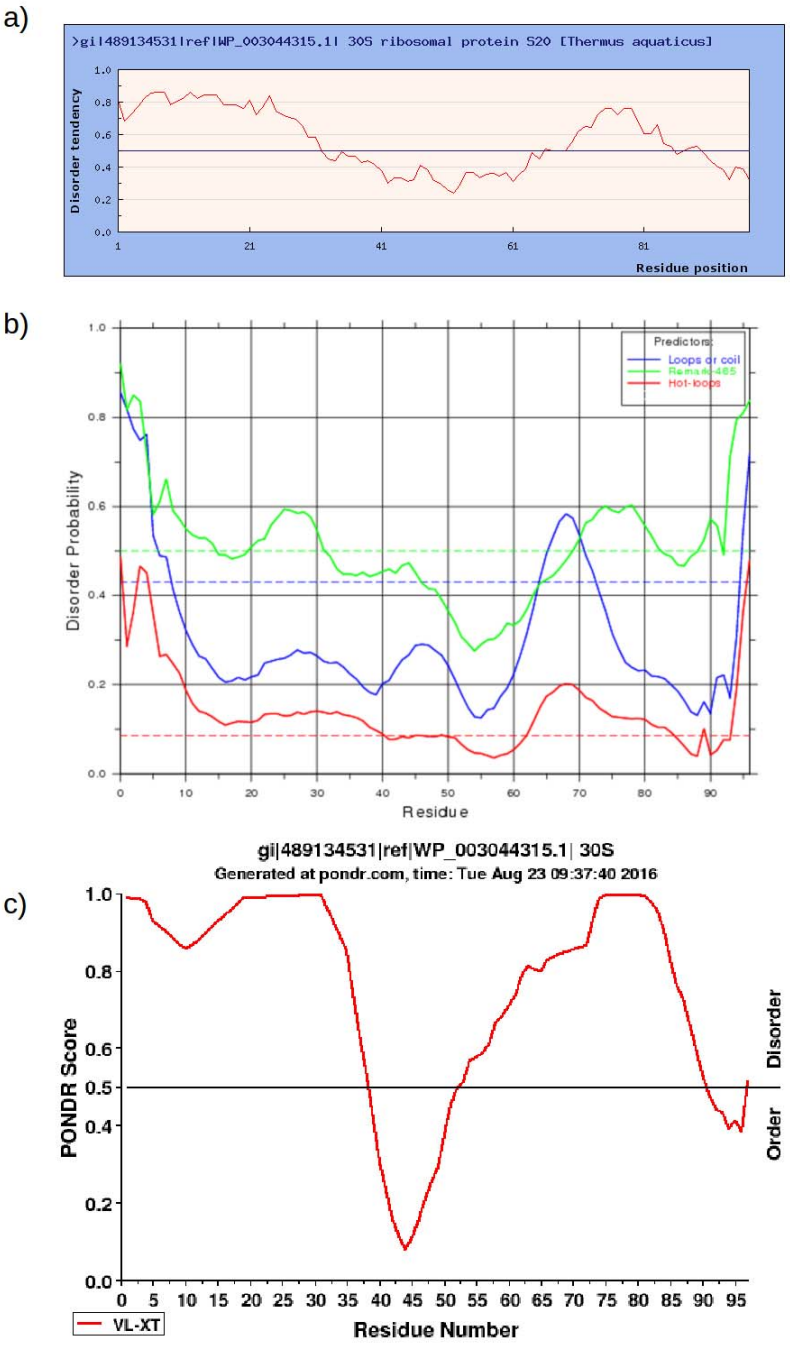


982

983 **Figure 4—figure supplement 2.** Rosetta energy scores (*fixbb+relax*) for TPR designs based on RPS20-
984 hhta sequence and various sets of mutations. The scores for the designs are shown in two groups:
985 the group to the left are combinations involving only primary mutations (see supplementary file 1E).
986 The group to the right are designs involving both primary and secondary mutations (supplementary
987 file 1E). The design variants are sorted by the average of the lowest 10% scores. The designs tested in
988 the lab are marked by red arrows (M2, M4E, M5, M4N, M4RD). The *in silico* simulation was
989 performed using Rosetta 3.4.

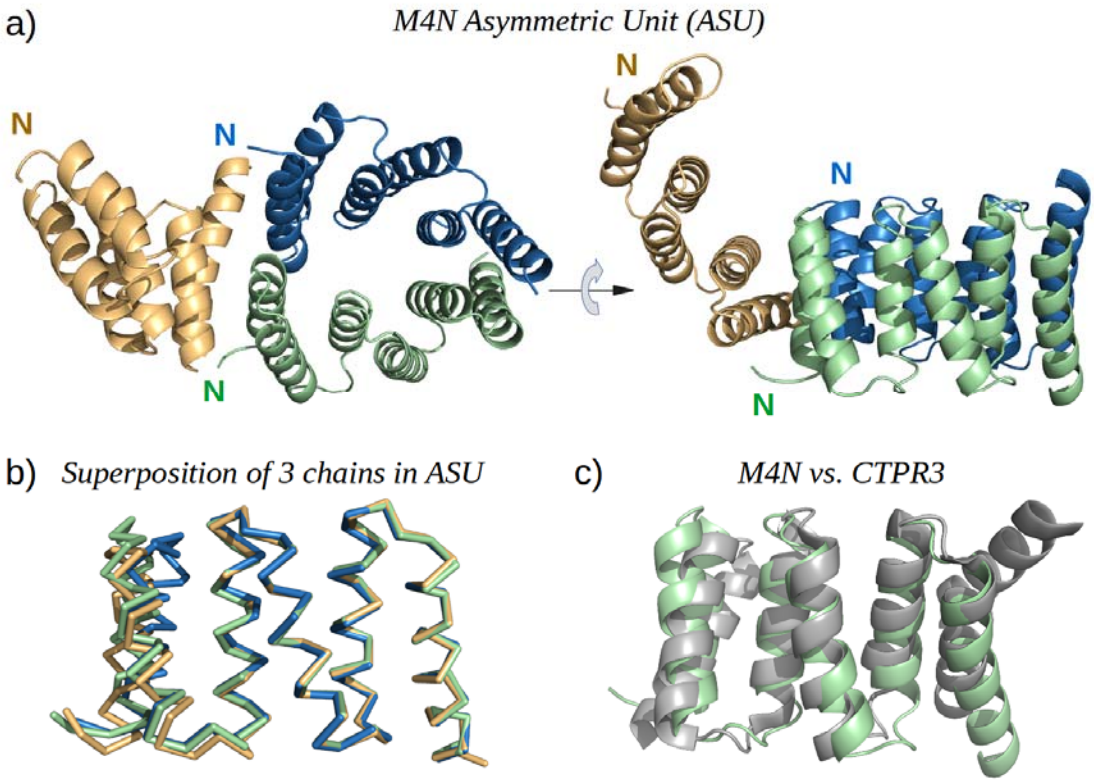
990

991



992

993 **Figure 4—figure supplement 3.** Prediction of intrinsically disordered regions in RPS20 of *Thermus*
994 *aquaticus* (NCBI gi: 489134531, accession: WP_003044315.1) using a) IUPred
995 (http://iupred.enzim.hu/) ; b) DisEMBL (http://dis.embl.de/) and c) PONDR
996 (http://www.pondr.com/).



997
998 **Figure 5. a**) The X-ray structure of M4N. The three chains A, B and C in the asymmetric unit are
999 colored green, blue and yellow, respectively. Chains A and B form a dimer. **b**) Superposition of the
1000 three chains. Only Cα traces are shown for clarity. **c**) Superposition of M4N (chain A, green) and the
1001 designed consensus TPR CTPR3 (PDB: 1na0, chain A, gray).

1002

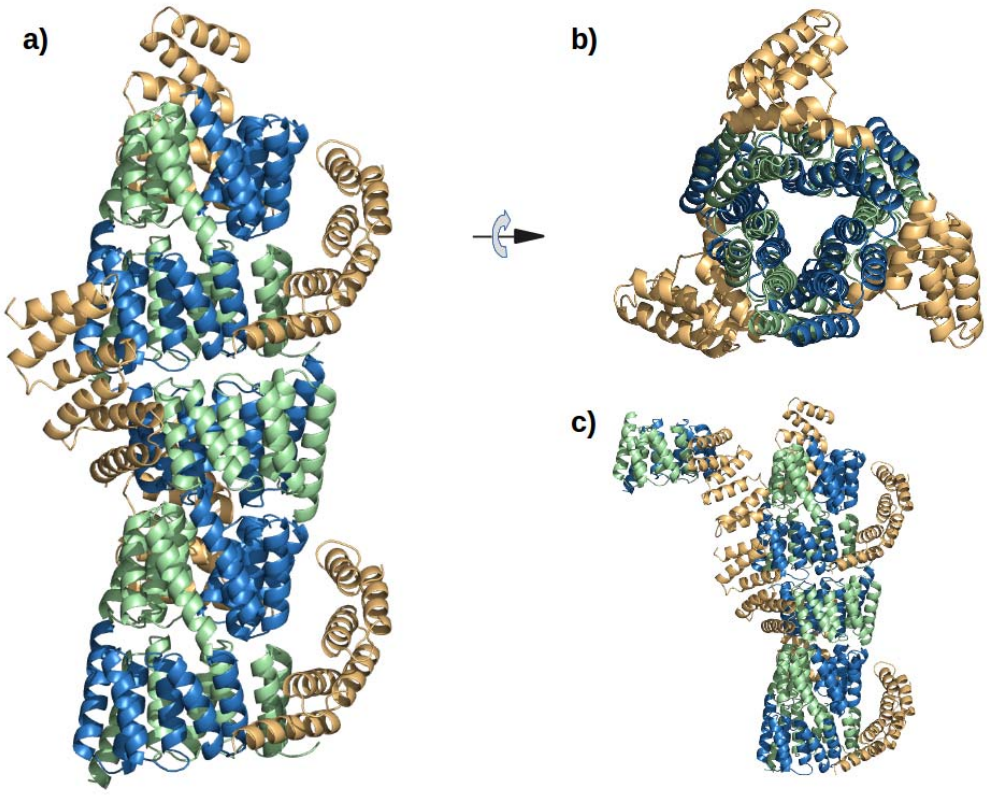
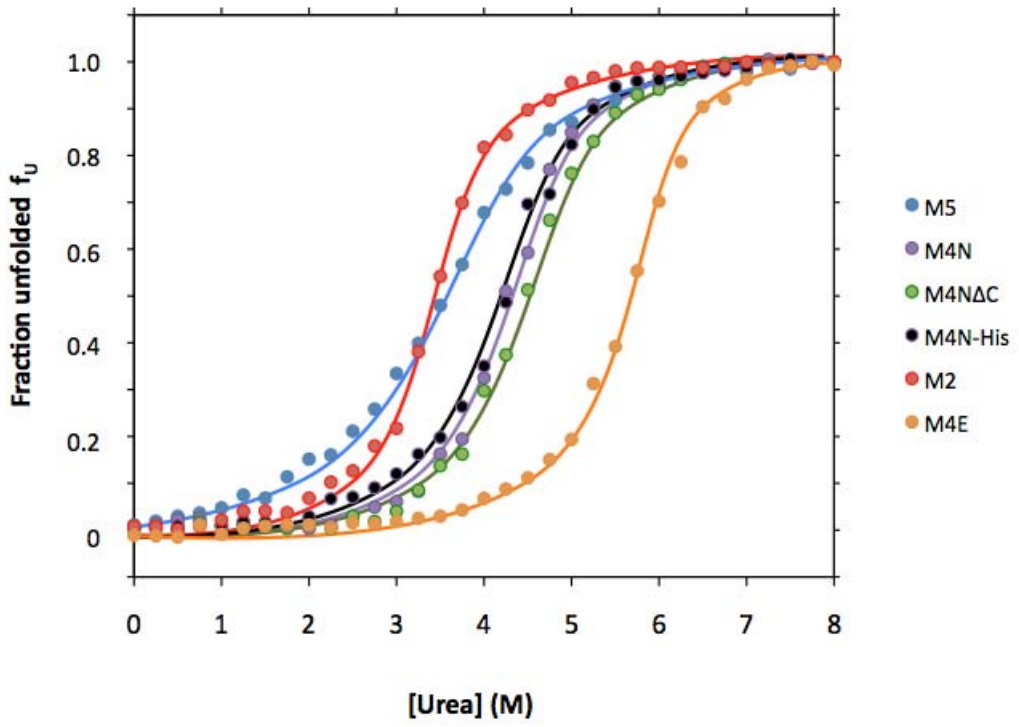


Figure 5—figure supplement 1. The interaction of M4N molecules in the crystal. **a**) Five adjacent ASUs are depicted. Chain A (green) and B (blue) form a dimer, while chain C (yellow) packs its C-terminus to the N-termini of chains A and B. **b**) Top view. **c**) An additional ASU (top-left) is shown to demonstrate the packing of N-termini of chains C.



1009

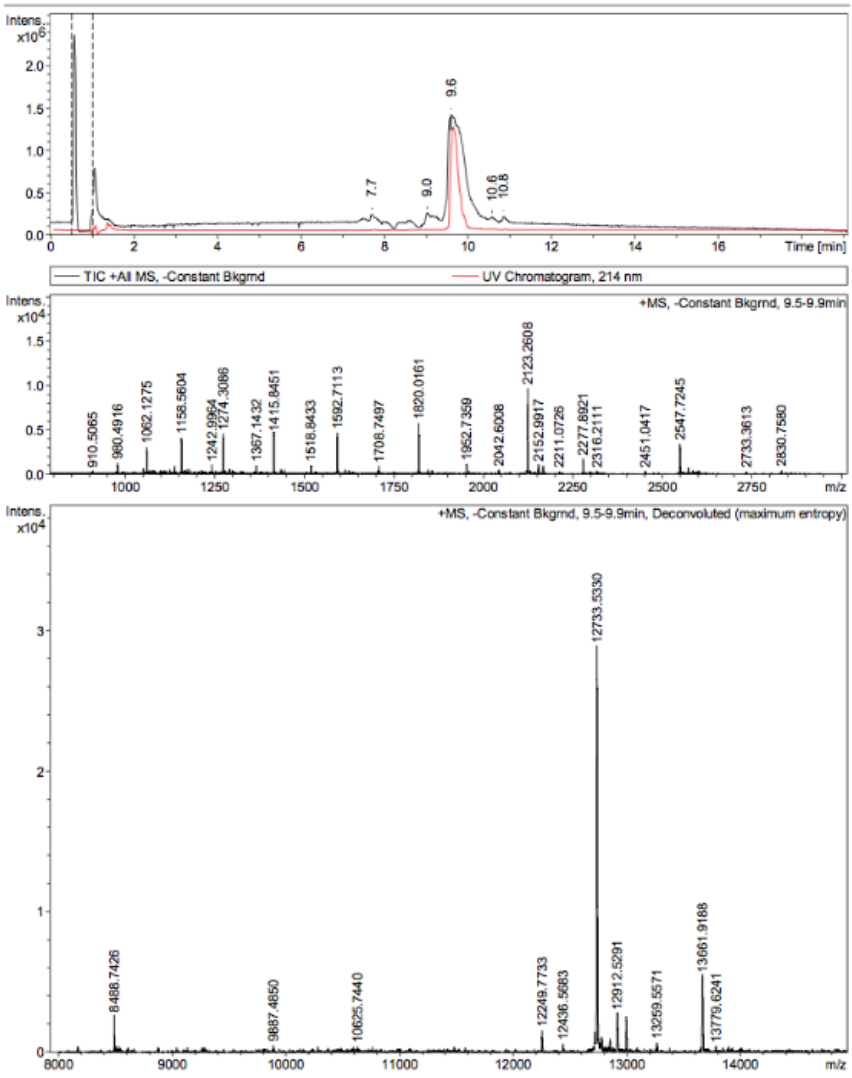
1010 **Figure 5—figure supplement 2.** Urea denaturation of designed TPR repeats. Urea-induced

1011 equilibrium unfolding at 23°C was monitored by circular dichroism at 222 nm. Data were converted

1012 to the fraction of unfolded protein f_U and fitted to a two-state model. The protein concentration was

1013 15 µM. See supplement file 1F for obtained parameters.

1014



>M4N
MGNS
IKTLSNLANLLAQEGKAEEAIKYMRKAVSLDPNN
IKTLSNLANLLAQEGKAEEAIKYMRKAVSLDPNN
IKTLSNLAVLLAQEGKAEEAIKYMRKAVSLIDKA
AKGSTLHKNAAARRKSRLMRKVQKL

1015

1016 **Figure 5—figure supplement 3.** Mass spectrometry (MS) analysis of M4N. The M4N fragment with a

1017 mass of 12733.533 Da in MS is underlined and highlighted in blue (theoretical mass 12733.77 Da).

1018 The C-terminus of M4N as observed in the crystal structure is marked by a red arrow.

1019

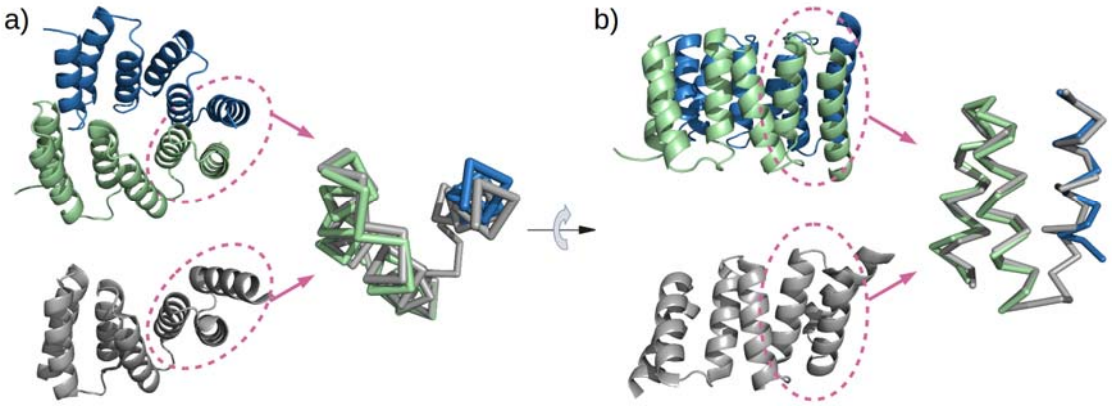


Figure 6. Mimicry of the stop helix in the M4N dimer. The C-terminal TPR unit in chain A (green) and the C-terminal helix B3 in chain B (blue) are superposed to the last TPR unit plus the stop helix in CTPR3 (gray).

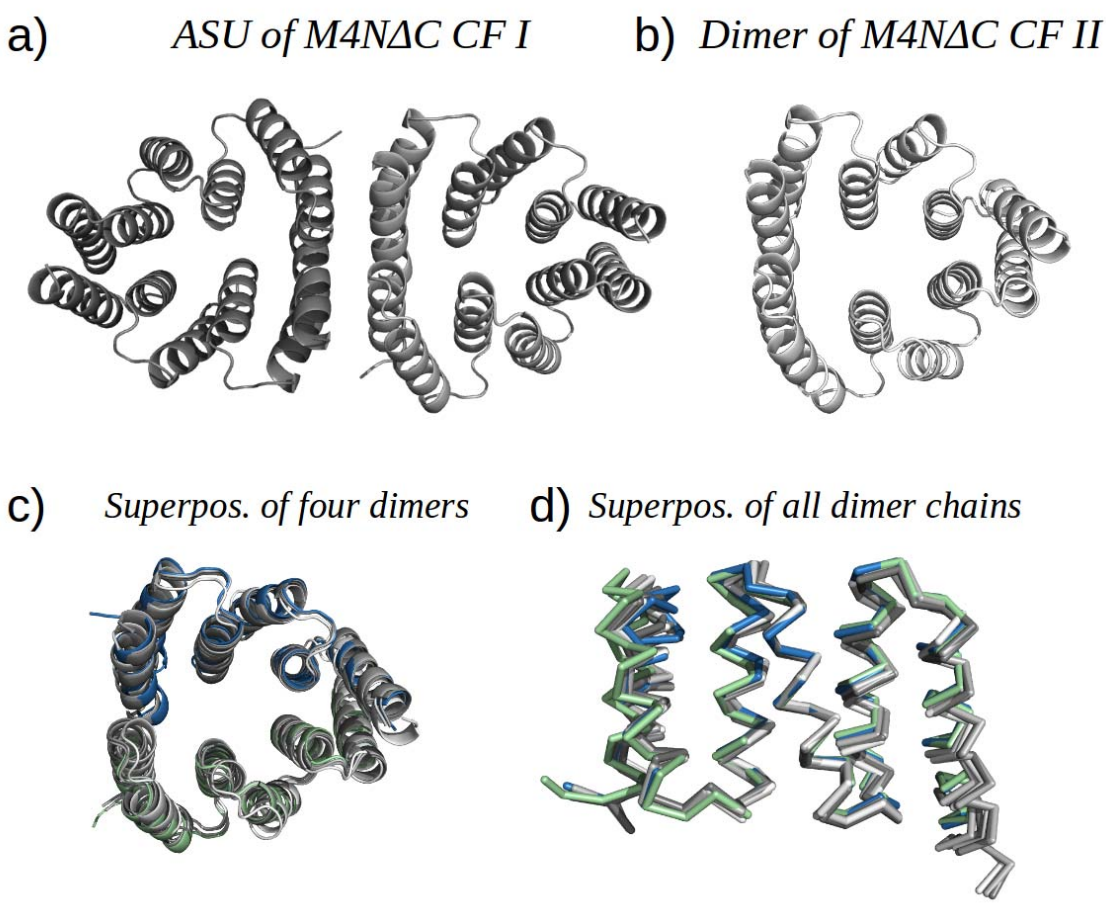


Figure 7. M4NΔC structures of two different crystal forms and their comparison to the M4N dimer. **a**) Two dimers in the ASU of M4NΔC CF I. **b**) Dimer constructed by applying the crystallographic symmetry to the single chain in the ASU of M4NΔC CF II. **c**) Superposition of all the four M4N and M4NΔC dimers. The M4N dimer is in green and blue. The three M4NΔC dimers are in different shades of gray as in a) and b). **d**) Superposition of all the chains in the M4N and M4NΔC dimers (eight chains in total). Only Cα traces of proteins are shown for clarity.

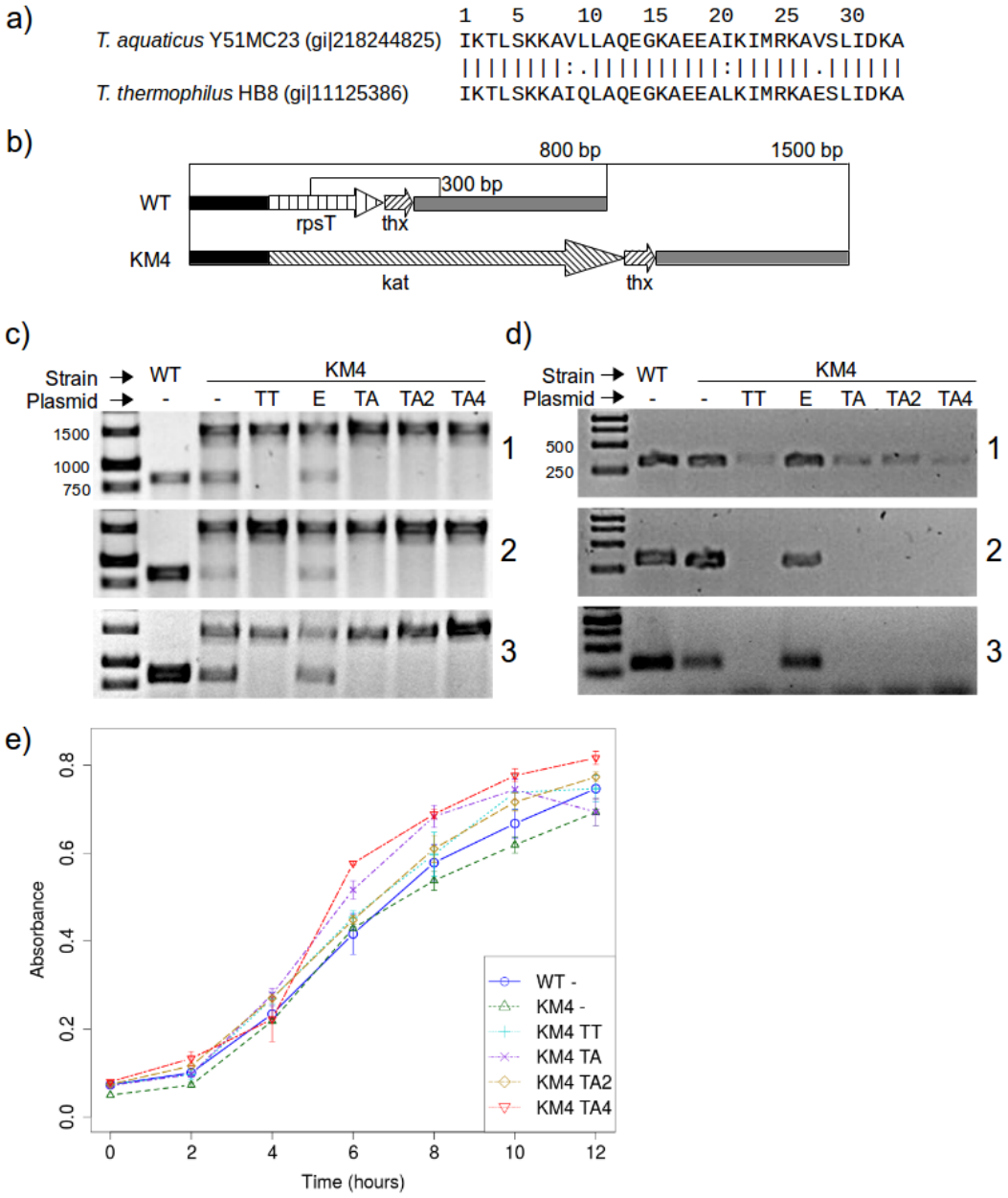


Figure 8. RPS20 variants M2 and M4N are functional proteins. **a**) The 34 amino-acid long RPS20-hh fragments in *T. aquaticus* and *T. thermophilus* differ only at four positions, including two conservative mutations (V9I and I21L). **b**) Scheme of the *rpsT* region before (upper) and after (lower) substitution of *rpsT* with the kanamycin resistance cassette (*kat*). Base pair (bp) values indicate the PCR products that can be amplified. Regions depicted with the same pattern are identical. Regions in solid black and gray also contain genes which are not marked for clarity. **c**) PCR to detect substitution of rps20 by the *kat* gene and **d**) PCR to detect the presence of chromosomal *rpsT* in *T. thermophilus* strains (WT: *T. thermophilus* HB8; KM4:*T. thermophilus* KM4) carrying various plasmids (TT: pJJSpro-rps20Tt;

E: pJJSpro; TA: pJJSpro-rpsTTa; TA2: pJJSpro-rpsTTaM2; TA4: pJJSpro-rpsTTaM4N; -: No plasmid) after sequential grow under different selective pressures (1: 30 µg/ml kanamycin; 2: 120 µg/ml kanamycin; 3: 0 µg/ml kanamycin). **e**) Corresponding growth curves of the host bacteria with various substitutions and plasmids.

Supplementary file 1: Further supporting computational and experimental results.

Section A: Sequence variation in RPS20-hh at positions 6, 7, 9 and 23 (TPR unit numbering) observed in RPS20 sequences.

Section B: Most commonly observed amino acids in RPS20-hh.

Section C: List of putative TPR homologs identified in the PDB by sequence and structure analysis.

Section D: RPS20-hh sequences that resemble a TPR profile according to TPRpred.

Section E: Mutations tested in silico on RPS20-hh for TPR design.

Section F: Biophysical parameters of designed TPRs.

Section G: Primary structures of M4N molecules observed in the crystal structures.

Section H: Crystallization conditions, and data collection/refinement statistics.

Section I: Detailed structure comparison results of different chains in M4N structures, and of M4N to CTPR3.

Section J: SEG prediction of low-complexity regions in RPS20-hhta.

Name	Mutations	Sequence
M0	-	NS IKTLSKKAVLLAQEGKAEEAIKIMRKAVSLDPNN IKTLSKKAVLLAQEGKAEEAIKIMRKAVSLDPNN IKTLSKKAVLLAQEGKAEEAIKIMRKAVSLIDKA *AKGSTLHKNAAARRKSRLMRKVQKL*
M2	K7L, I23Y	NS IKTLSK**L**AVLLAQEGKAEEAIK**Y**MRKAVSLDPNN IKTLSK**L**AVLLAQEGKAEEAIK**Y**MRKAVSLDPNN IKTLSK**L**AVLLAQEGKAEEAIK**Y**MRKAVSLIDKA *AKGSTLHKNAAARRKSRLMRKVQKL*
M4E	K2E, K7L, V9N, I23Y	NS I**E**TLSK**L**A**N**LLAQEGKAEEAIK**Y**MRKAVSLDPNN I**E**TLSK**L**A**N**LLAQEGKAEEAIK**Y**MRKAVSLDPNN I**E**TLSK**L**AVLLAQEGKAEEAIK**Y**MRKAVSLIDKA *AKGSTLHKNAAARRKSRLMRKVQKL*
M4N	K6N, K7L, V9N, I23Y	NS IKTLS**NL**A**N**LLAQEGKAEEAIK**Y**MRKAVSLDPNN IKTLS**NL**A**N**LLAQEGKAEEAIK**Y**MRKAVSLDPNN IKTLS**NL**AVLLAQEGKAEEAIK**Y**MRKAVSLIDKA *AKGSTLHKNAAARRKSRLMRKVQKL*
M4RD	K2E, K7R, V9N, I23D	NS I**E**TLSK**R**A**N**LLAQEGKAEEAIK**D**MRKAVSLDPNN I**E**TLSK**R**A**N**LLAQEGKAEEAIK**D**MRKAVSLDPNN I**E**TLSK**R**AVLLAQEGKAEEAIK**D**MRKAVSLIDKA *AKGSTLHKNAAARRKSRLMRKVQKL*
M5	K2E, L4W, K7L, V9N, I23Y	NS I**E**TLSK**L**A**N**LLAQEGKAEEAIK**Y**MRKAVSLDPNN I**E**T**W**SK**L**A**N**LLAQEGKAEEAIK**Y**MRKAVSLDPNN I**E**T**W**SK**L**AVLLAQEGKAEEAIK**Y**MRKAVSLIDKA *AKGSTLHKNAAARRKSRLMRKVQKL*
M4NΔC	K6N, K7L, V9N, I23Y	NS IKTLS**NL**A**N**LLAQEGKAEEAIK**Y**MRKAVSLDPNN IKTLS**NL**A**N**LLAQEGKAEEAIK**Y**MRKAVSLDPNN IKTLS**NL**AVLLAQEGKAEEAIK**Y**MRKAVSLIDKA *AK*

Table 1. The primary structures of the six designed proteins using RPS20-hhta tested *in vitro*. Point mutations introduced into RPS20-hhta are shown in bold and underlined. The C-terminal four residues in RPS20-hhta were replaced by the consensus loop sequence DPNN in TPRs (underlined). The stop helix is in gray (italic). M4NΔC is M4N without stop helix.

a)

TPR protein
Anabaena sp. 90
gi|752818954

0.01

Ana5
Ana10
Ana6
Ana7
Ana12
Ana9
Ana8
Ana11

TPR protein
Calothrix sp. 336/3
gi|821031795

Cal4
Cal5
Cal7
Cal17
Cal6
Cal8
Cal9
Cal10
Cal11
Cal12
Cal13
Cal14
Cal15
Cal16
Cal18

TPR protein
filamentous cyanobacterium ESFC-1
gi|740500649

fil4
fil9
fil8
fil6
fil7
fil5
fil3
fil10
fil2

TPR protein
Cyanothece sp. PCC 8801
gi|501590504

Cya2
Cya3
Cya4
Cya6
Cya5
Cya15
Cya9
Cya8
Cya10
Cya11
Cya14
Cya7
Cya16
Cya12
Cya13

TPR protein
Microcystis aeruginosa SPC777
gi|513477764

Mic13
Mic12
Mic11
Mic9
Mic10
Mic7
Mic6
Mic5
Mic8
Mic4
Mic3

b)

10.0

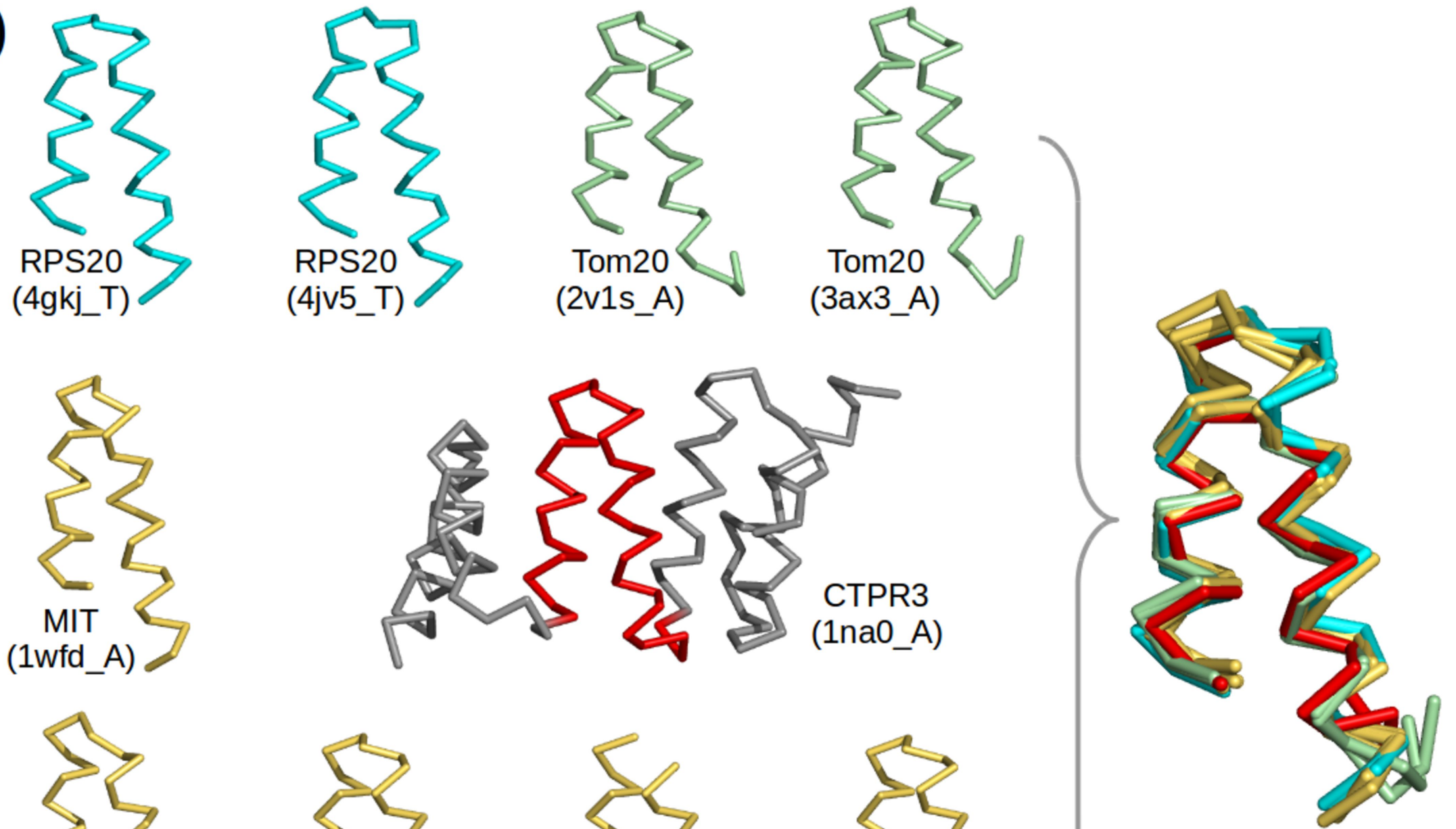
TPR unit two
The2
Xen2
Hom2
Dan2
Sac2
Mus2
Ara2

TPR unit one
Ara1
Mus1
Dan1
Hom1
Xen1
The1
Sac1

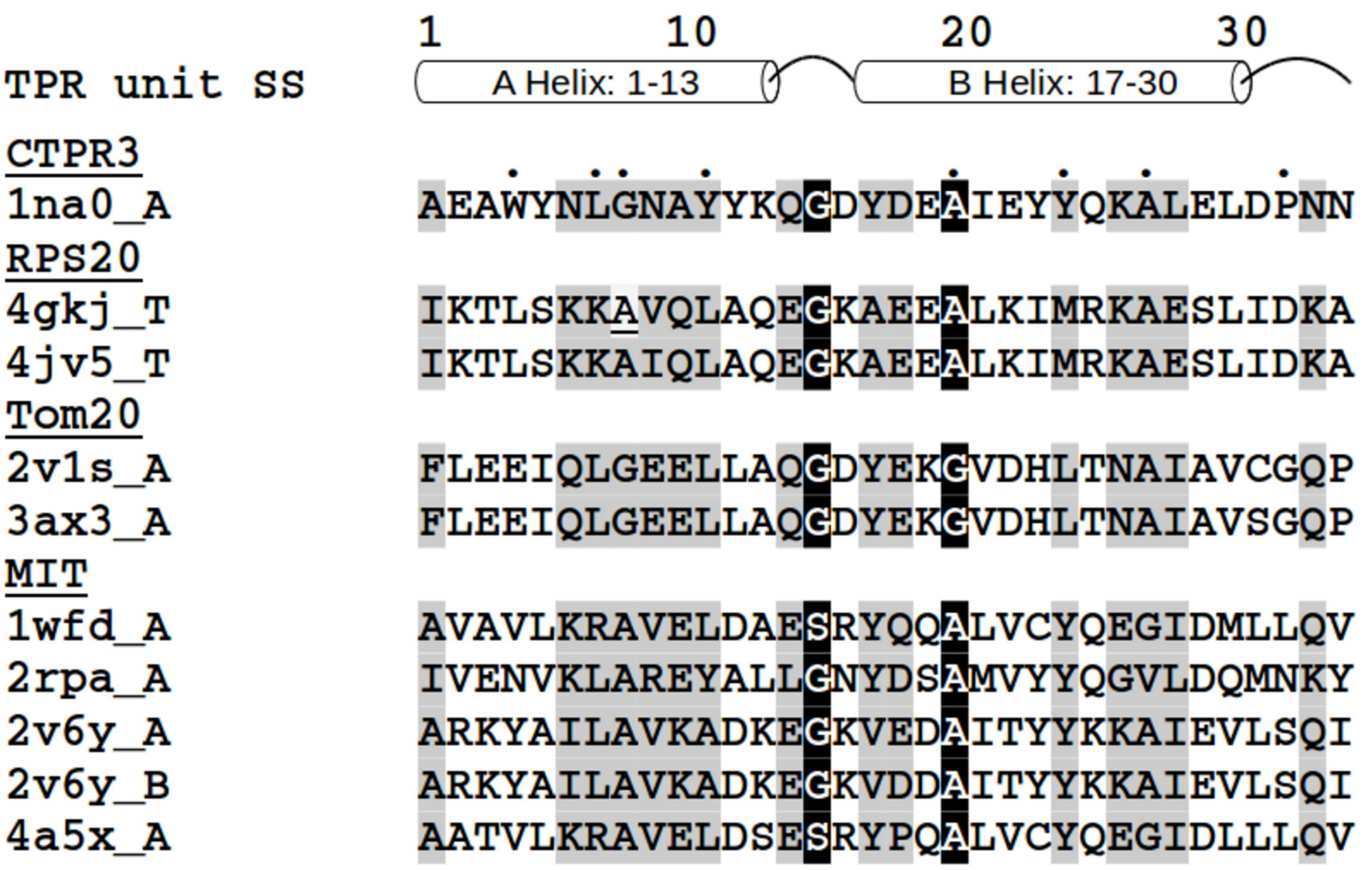
TPR unit three
The3
Sac3
Xen3
Dan3
Hom3
Ara3
Mus3

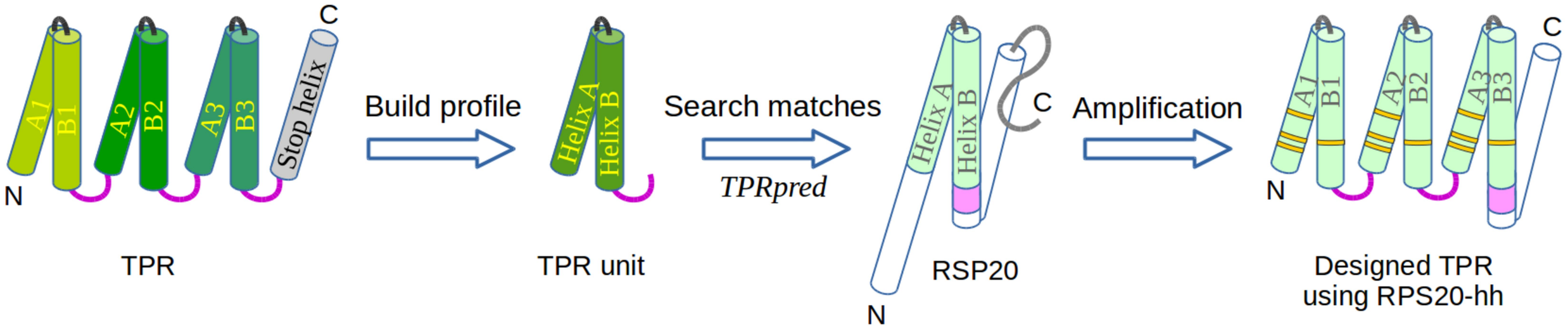
origination of TPR by amplification
diversification
internal duplication

a)

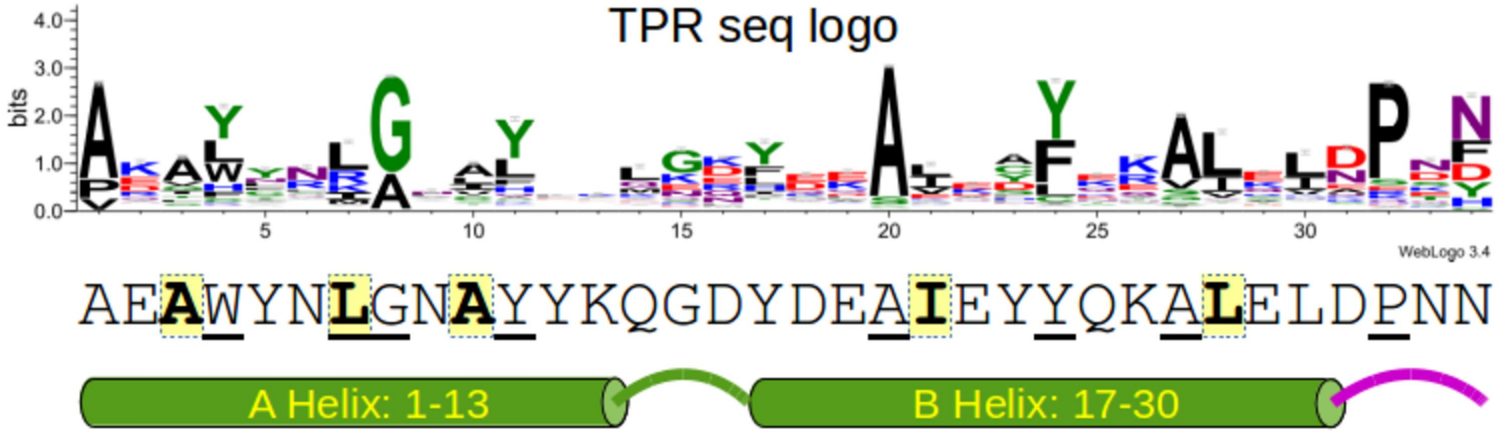

RPS20
(4gkj_T)
RPS20
(4jv5_T)
Tom20
(2v1s_A)
Tom20
(3ax3_A)
MIT
(1wfd_A)
CTPR3
(1na0_A)
MIT
(2rpa_A)
MIT
(2v6y_A)
MIT
(2v6y_B)
MIT
(4a5x_A)

b)

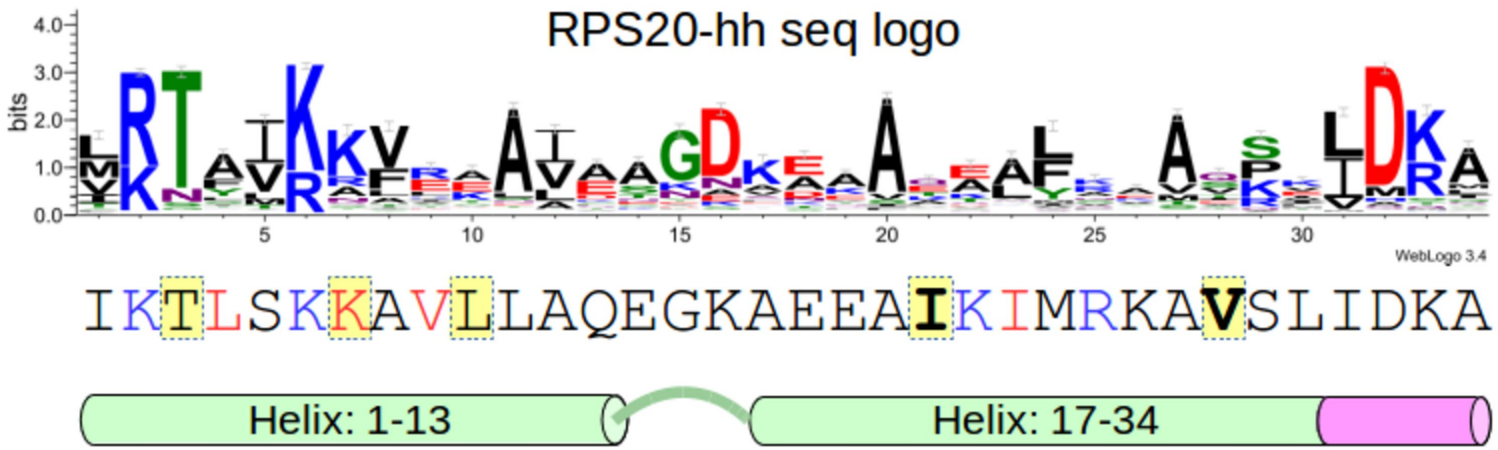

1 10 20 30
TPR unit SS
A Helix: 1-13
B Helix: 17-30
CTPR3
1na0_A AEAWYNLGNAYYKQGDYDEAIEYYQKALELDPNN
RPS20
4gkj_T IKTLSKKAVQLAQEGKAEEALKIMRKAESLIDKA
4jv5_T IKTLSKKAIQLAQEGKAEEALKIMRKAESLIDKA
Tom20
2v1s_A FLEEIQLGEELLAQGDYEKGVDHLTNAIAVCGQP
3ax3_A FLEEIQLGEELLAQGDYEKGVDHLTNAIAVSGQP
MIT
1wfd_A AVAVLKRAVELDAESRYQQALVCYQEGIDMLLQV
2rpa_A IVENVKLAREYALLGNYDSAMVYYQGVLDQMNKY
2v6y_A ARKYAILAVKADKEGKVEDAITYYKKAIEVLSQI
2v6y_B ARKYAILAVKADKEGKVDDAITYYKKAIEVLSQI
4a5x_A AATVLKRAVELDSESRYPQALVCYQEGIDLLLQV


A1
B1
A2
B2
A3
B3
Stop helix
N
C
TPR
Build profile
Helix A
Helix B
TPR unit
Search matches
TPRpred
Helix A
Helix B
C
N
RSP20
Amplification
A1
B1
A2
B2
A3
B3
C
N
Designed TPR
using RPS20-hh

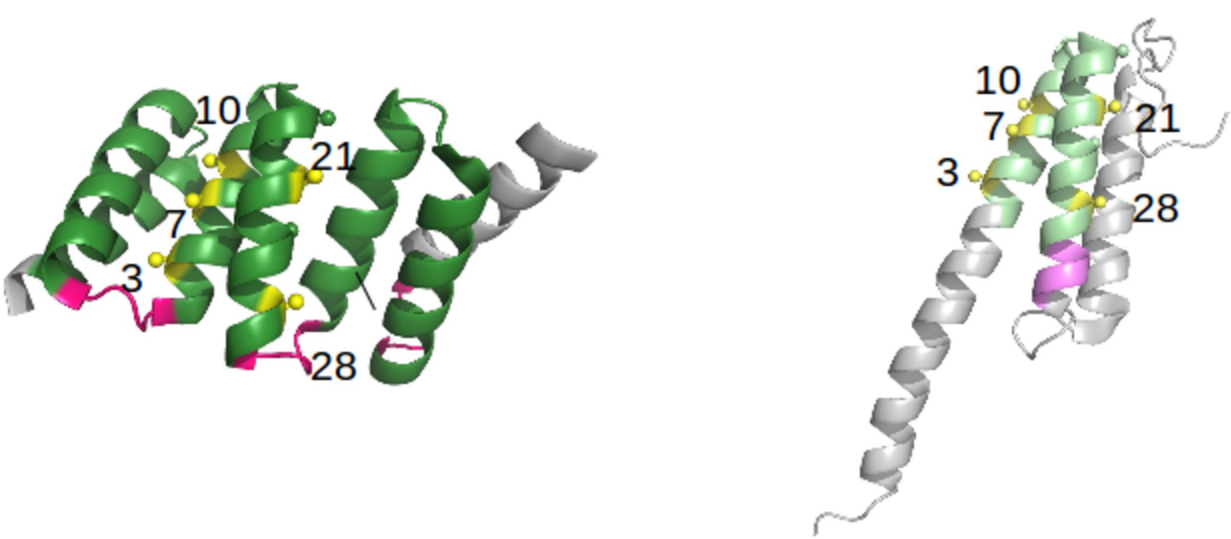
a)

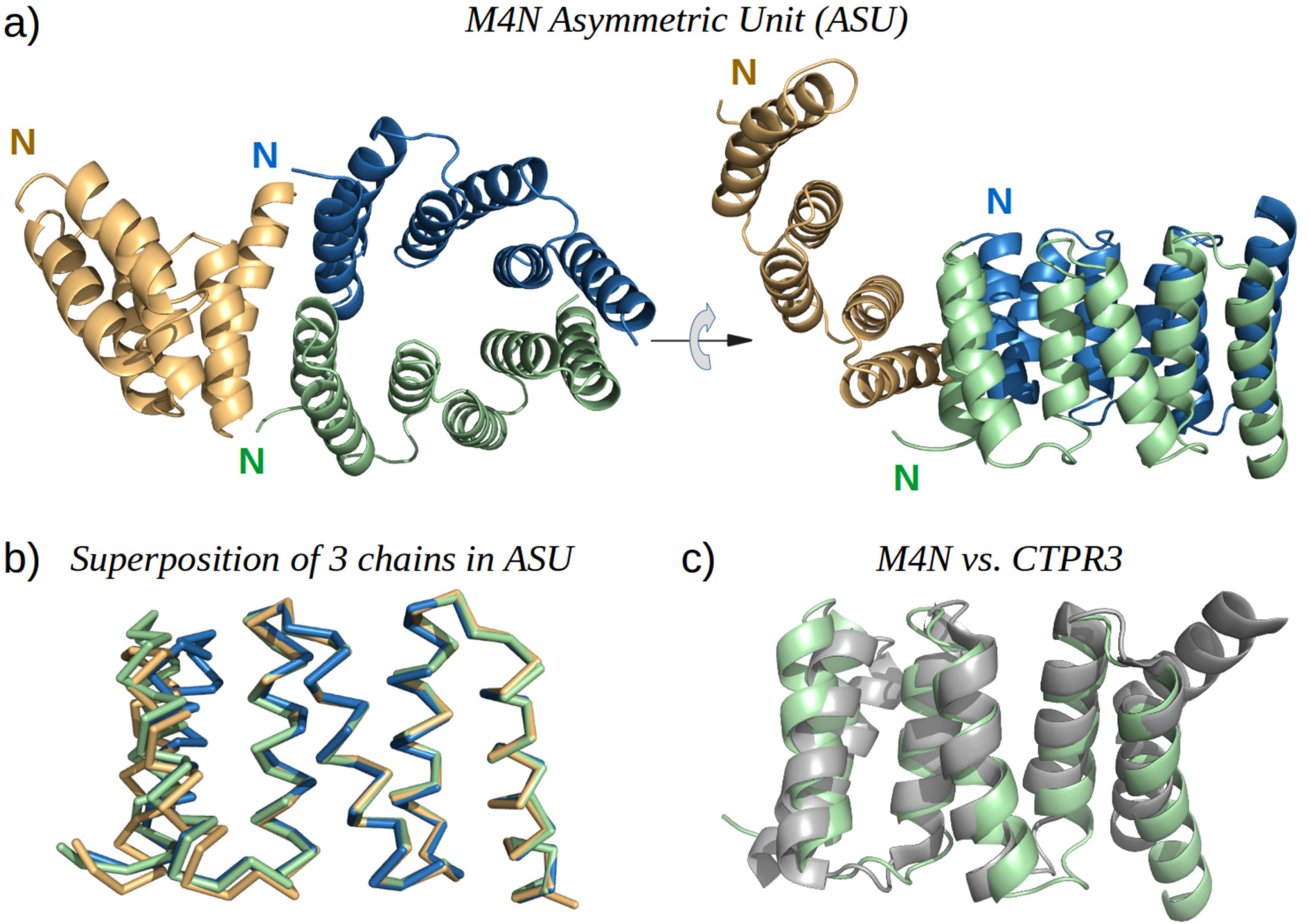

TPR seq logo
bits
WebLogo 3.4
AEAWYNLGNAYYKQGDYDEAIEYYQKALELDPNN
A Helix: 1-13
B Helix: 17-30

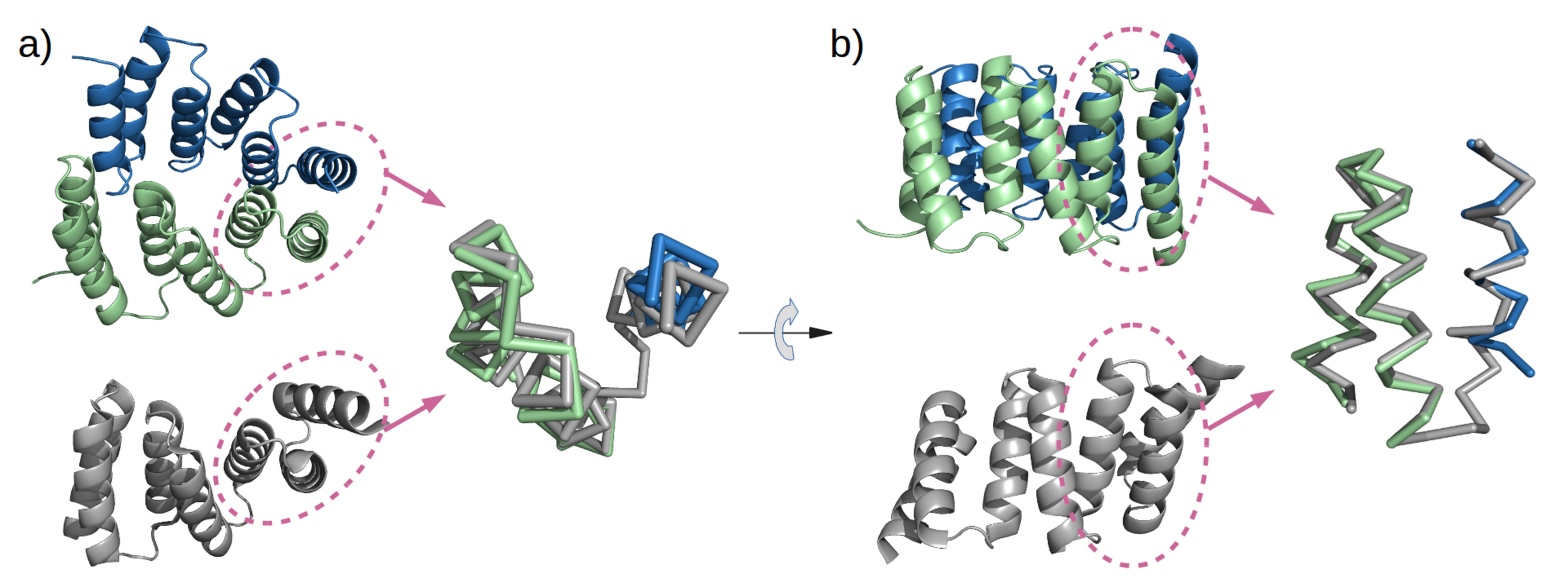
b)


RPS20-hh seq logo
bits
WebLogo 3.4
IKTLSKKAVLLAQEGKAEEAIKIMRKAVSLIDKA
Helix: 1-13
Helix: 17-34

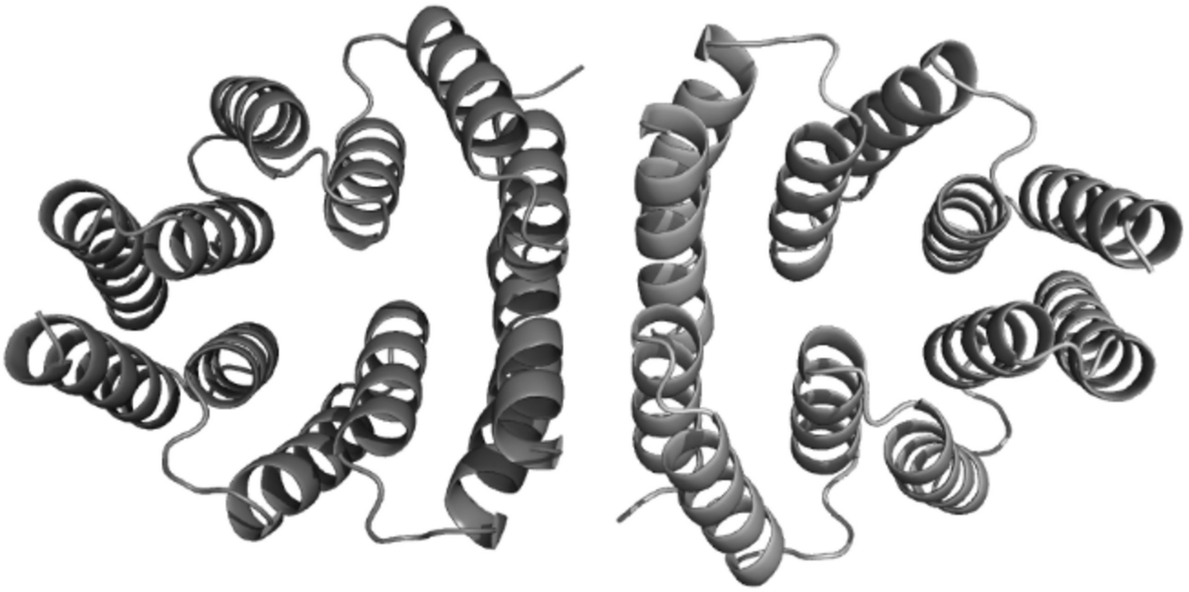
c)


10
21
7
3
28
10
7
21
3
28

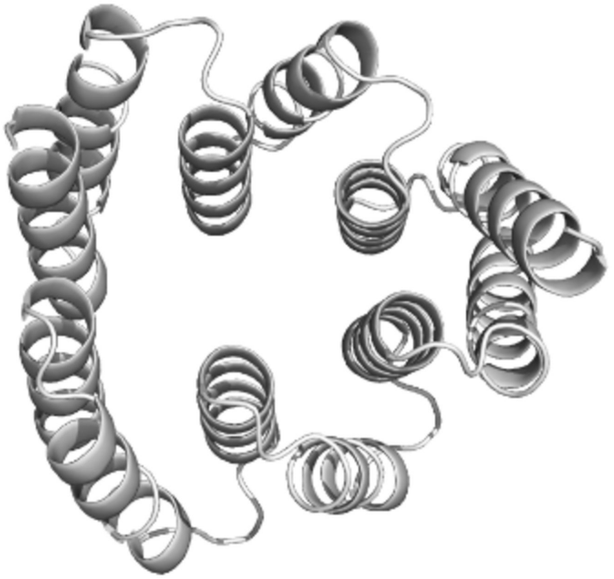

a)
M4N Asymmetric Unit (ASU)
N
N
N
N
N
b)
Superposition of 3 chains in ASU
c)
M4N vs. CTPR3


a)
b)

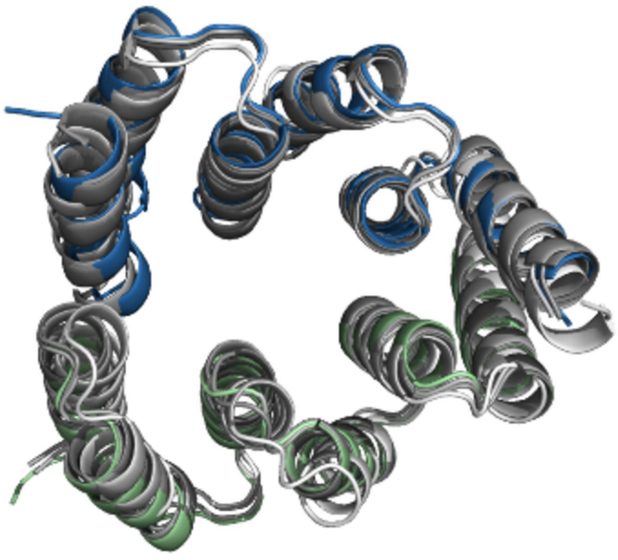
a) *ASU of M4NΔC CF I*



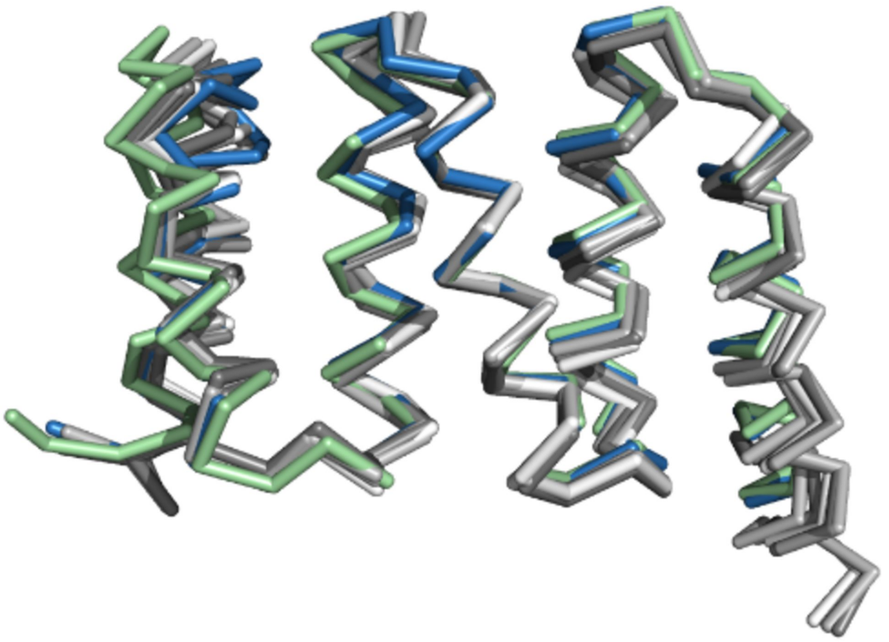
b) *Dimer of M4NΔC CF II*

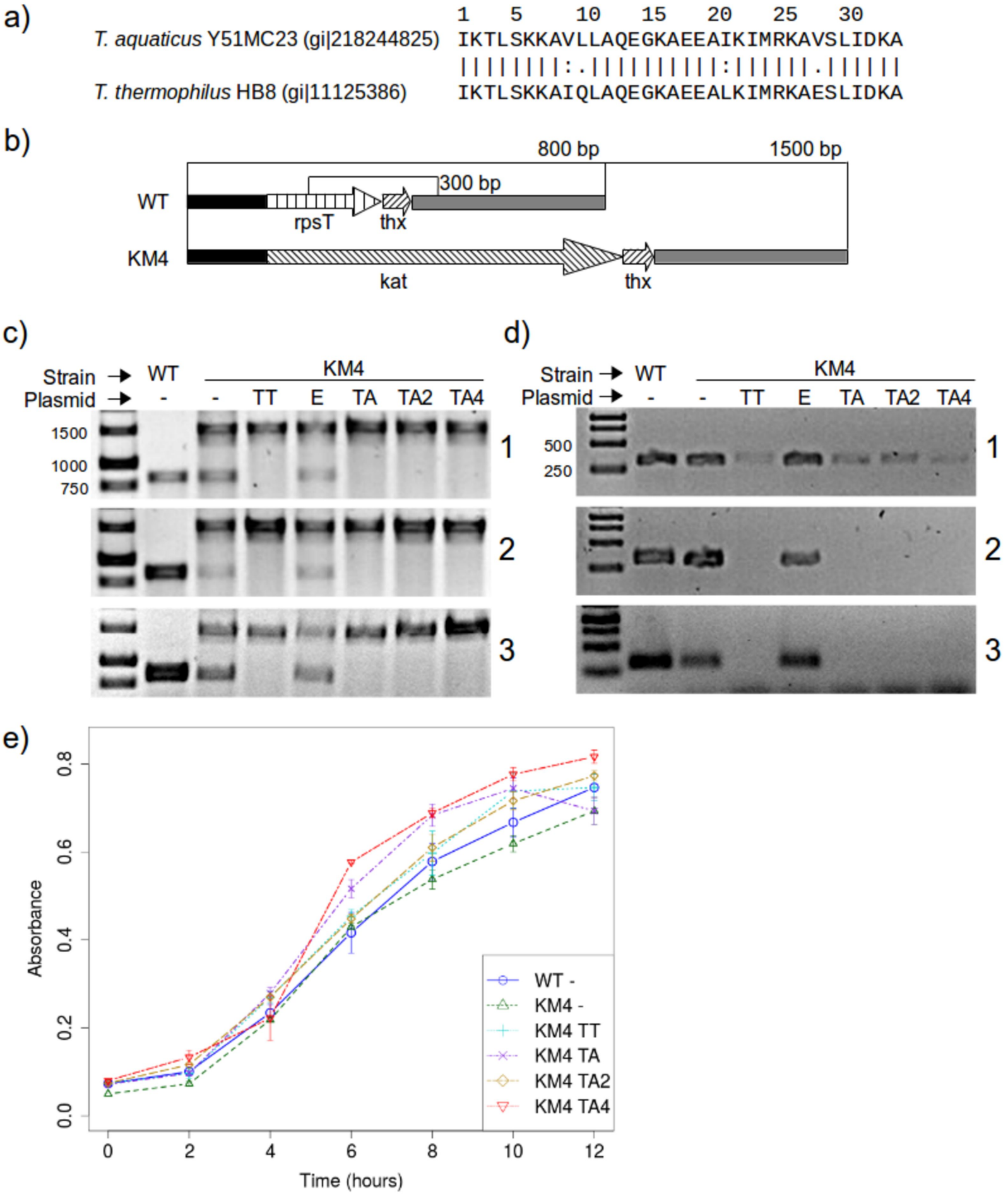


c) *Superpos. of four dimers*



d) *Superpos. of all dimer chains*




a)
1 5 10 15 20 25 30
T. aquaticus Y51MC23 (gi|218244825) IKTLSKKAVLLAQEGKAEEAIKIMRKAVSLIDKA
T. thermophilus HB8 (gi|11125386) IKTLSKKAIQLAQEGKAEEALKIMRKAESLIDKA
b)
800 bp
1500 bp
300 bp
WT
rpsT
thx
KM4
kat
thx
c)
Strain
Plasmid
WT
KM4
- - TT E TA TA2 TA4
1500
1000
750
1
2
3
d)
Strain
Plasmid
WT
KM4
- - TT E TA TA2 TA4
500
250
1
2
3
e)
Absorbance
0.0 0.2 0.4 0.6 0.8
0 2 4 6 8 10 12
Time (hours)
WT -
KM4 -
KM4 TT
KM4 TA
KM4 TA2
KM4 TA4

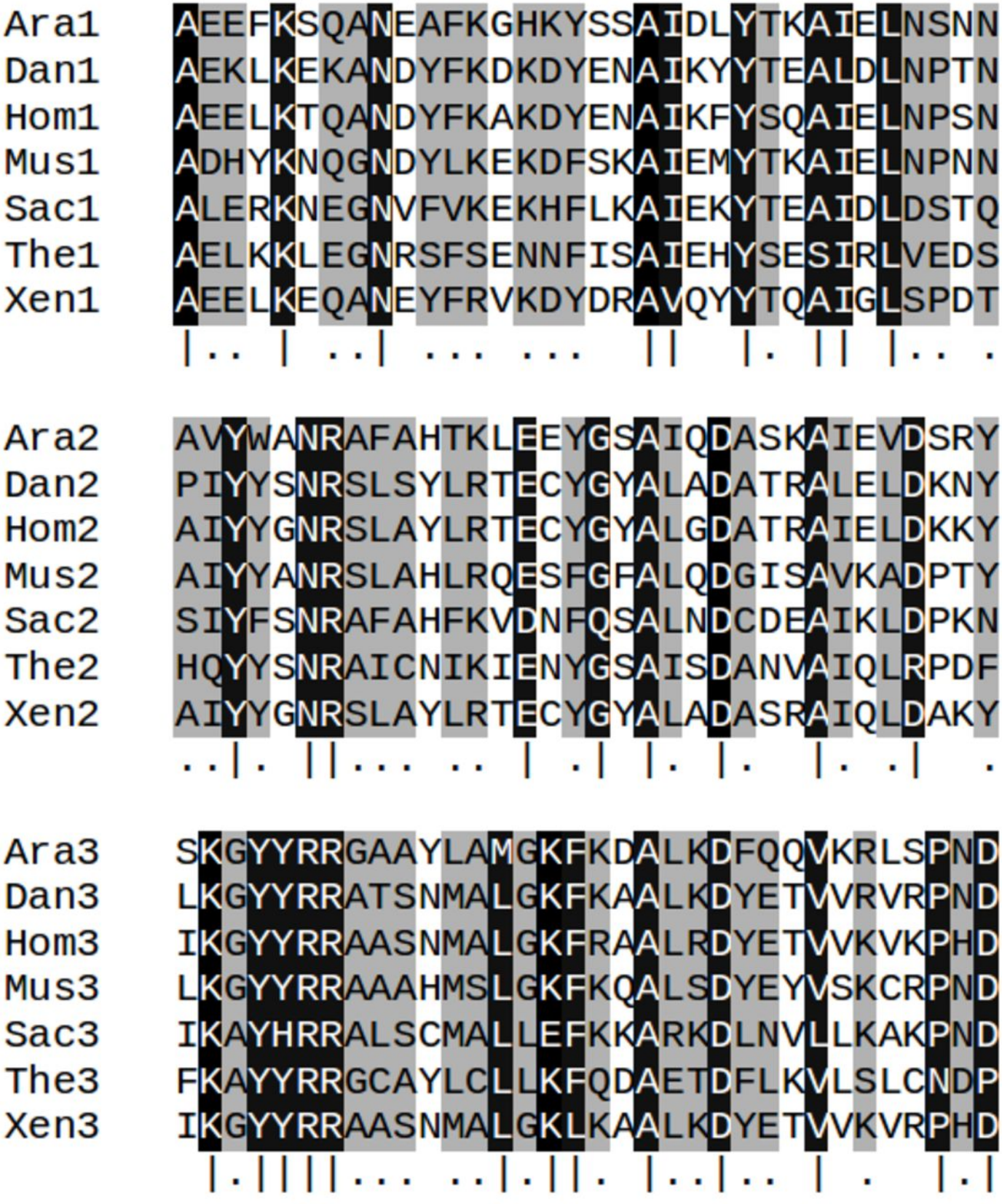
a)

```
Ana1    VQQLFKQGETAESVGDNSQAETIWRKVLQVEPNN
Ana2    GKAYNNLGNALRRQGKLPEALTAHQKALQLNPND
Ana3    AEAYVGIGNVLNAQGKPDEAVAAYRKAIEFDPKY
Ana4    AKAYNSLGNALYDQEKLKEAVAAYRKAIEFDHKY
Ana5    AAAYYNLGNVLYEQKELDEAVAAYRKAIELNPKY
Ana6    ATAYNNLGNALSDQKKLDEAVAAYQEAIKLNPKD
Ana7    ATAYNNLGIALSDQKKLDEAVAAYQKAIELDPKY
Ana8    ATAYYNLGNALSDQKKLDEAVAAYQKAIELDPKY
Ana9    ATAYYNLGNALSDQKKLDEAVAAYQKAIELDPKY
Ana10   ATAYYNLGNALRGQKKLDEAVAAYQKAIELNPKY
Ana11   ATAYNNLGIALSDQKKLDEAVAAYQKAIELNPKD
Ana12   ATAYYNLGIALSDQKKLDEAVAAYQKAIELDPKD
Ana13   AAVYNNLGNALSDQKKLKEAISNYKTALSLPEDT
Ana14   TLANNNLGLALQDQEKFAEAIKYFDKAEELDPNF

Mic1    LEQLWQQGETAQAQKKYPEAERIWRQIIQLDPNS
Mic2    AVAFSNLCAALFRQNKLDEAPIFCQKALALDPKL
Mic3    PETYYNLGNVLYNQKKLTEAEEMYRRTIELDDKF
Mic4    VYAYNNLGLVLYDQKKLKEAEEMYRRTIELDDKL
Mic5    ALVYNNLGLVLYDQKKLKEAEEMYRRAIELDDKF
Mic6    VYAYNGLGNVLRAQNKLTEAEEMYRRALALDDQY
Mic7    VDAYNGLGNVLYDQKKLTEAEEMYRRALALDDQY
Mic8    VYAYNGLGNVLYDQKKLKEAEEMYRRTIELDDKY
Mic9    VYAYKGLGNVLYNQKKLKEAEEMYRRAIALDDQY
Mic10   VHAYNSLGNVLYNQKKLTEAEEMYRRALALDDQY
Mic11   VHAYNNLGNVLYDQKKLTEAEEMYRRALALDDKY
Mic12   VPAYHNLGNVLYNQKKLTEAEEMYRRALALDDKF
Mic13   VYAYNNLGNVLYDQKKLTEAEEMYRRALDLPDDT
Mic14   TLAHNNLGLLLQEQGKLEAAIAEFEKATKIDPQY

fil1    INQLFEQGNTAQQEGRYAEAEAIWRQILEANPDN
fil2    AGAYNNLGVALYNLGQLPEAVSAYQQAIALDPDY
fil3    AIAYNNLGIALRNLGQLPEAVEAYQQAIALDPDF
fil4    AIAYYNLGIALFDLGQLPEAVSAYQQAIALDPDD
fil5    AIAYNNLGNALSNLGQLPEAVSAYQQAIALDPDD
fil6    AIAYYNLGNALSNLGQLPEAVSAYQQAITLDPDY
fil7    AIAYNNLGNALSNLGQLPEAVEAYQQAIALDPDD
fil8    ADAYYNLGNALRDLGQLPEAVSAYQQAIALDPDF
fil9    ADAYYNLGIALRDLGQLPEAVEAYQQAIALDPDD
fil10   AIAYNNLGVALYNLGQLPEAVSAYQQAIALDPDN
fil11   AFAYNNLGYAYQQQGNLEAAITEYKKAIALAPNY
```

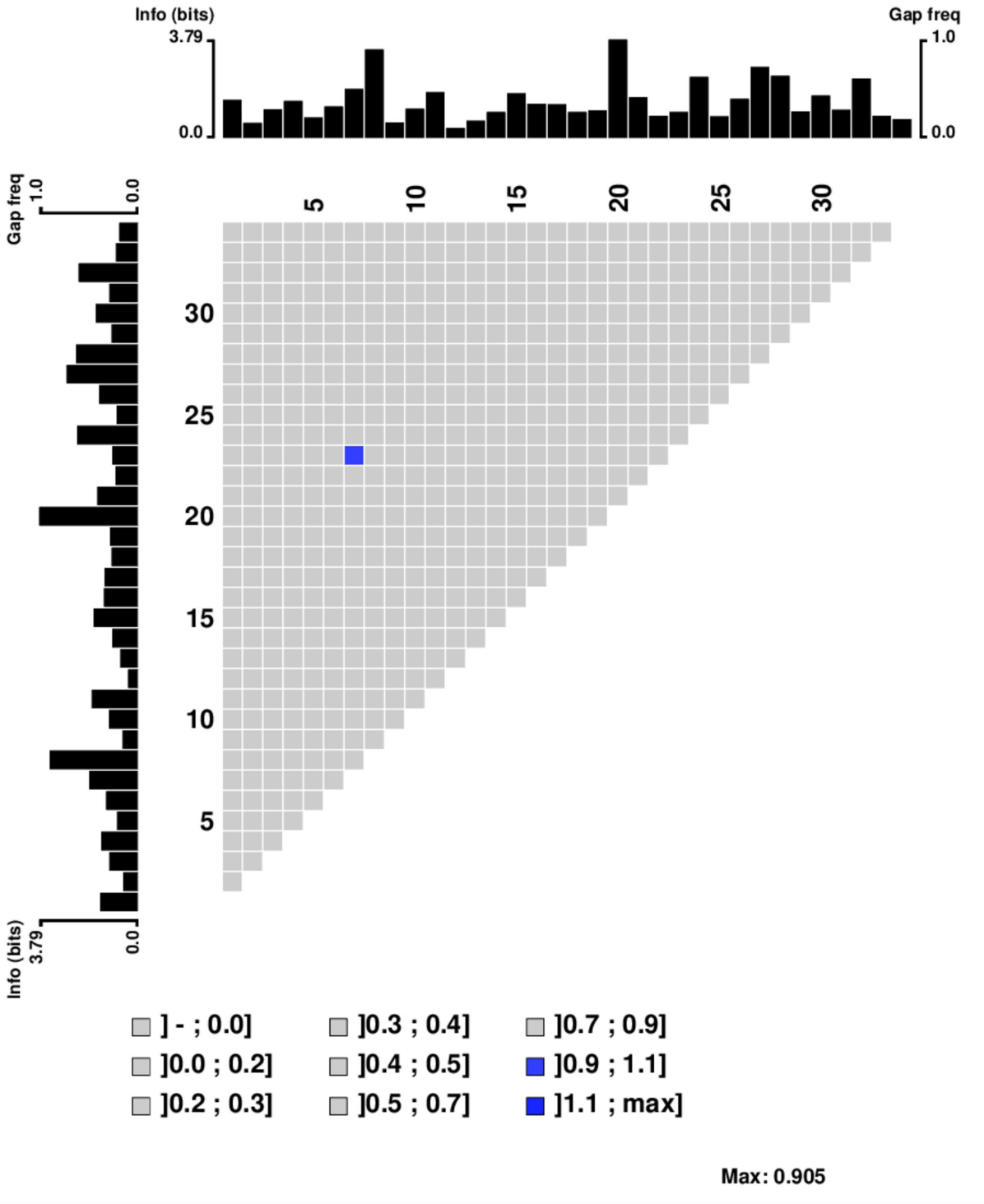
```
Cal1    TEQLFKQGEAAESVGNNSQAETIWRQVLQLEPSN
Cal2    GKAYNNLGNALRRQGKLDEALAAHQKALQLNLND
Cal3    AEAYVGIGNVLNAQGKPEEGIAQHKKALQINPNL
Cal4    AAAYNGLGNALYDQKKLEPAVAAYQKAIQLDPNY
Cal5    AAAYYNLGNALRDQKKLEPAIAAFQKAIQLNPNF
Cal6    AAAYNGLGNALYDQKKLEPAVAAFQKAIQLNPNF
Cal7    AAAYNGLGNALYDQKKLEPAVAAYQKAIQLDPNF
Cal8    AFAYYNLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal9    AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal10   AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal11   AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal12   AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal13   AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal14   AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal15   AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal16   AFAYNGLGNALYDQKKLEPAIAAFQKAIQLNPNF
Cal17   AFAYNGLGNALYDQKKLEPAIAAYQKAIQLDPND
Cal18   ANAYNNLGNALYDQKKLEPAIAAYQKAIQLNPNF
Cal19   AAAYNNLGVALSDQKKLEPAIAAYQKVLTLPEDT
Cal20   TAANNGLGLVFQEQGKLKQAIDYFDKSEALDPDY

Cya1    IDQLFQQGRTAGKMGKYTEAEAIFRRVIELDPNL
Cya2    ADAYNNLGNALYYQGKLDEAIAAYQKAIQLNPND
Cya3    ADAYNNLGNALSDQGKLEEAIAAYQKAIQLNPNY
Cya4    ADAYYNLGIALSDQGKLEEAIAAYQKAIQLNPNF
Cya5    TQAYYNLGIALSDQGKLEEAIAAYQKAIQLNPNY
Cya6    ADAYYNLGNALFDQGKLDEAIAAYQKAIQLDPND
Cya7    ANAYNNLGAALYKQGKLEEAIAAYQKAIQLNPNL
Cya8    AEAYNNLGVALSDQGKRDEAIAAYQKAIQLNPNL
Cya9    AEAYNNLGVALSDQGKRDEAIAAYQKAIQLNPNF
Cya10   ALAYNNLGVALSDQGKRDEAIAAYQKAIQLNPNF
Cya11   ALAYNNLGVALSDQGKRDEAIAAYQKAIQLNPNF
Cya12   ALAYNNLGVALRNQGKRDEAIAAYQKAIQLDPND
Cya13   ANAYNNLGLALRNQGKRDEAITAYQKAIQLNPNF
Cya14   ALAYNNLGNALYSQGKREEAIAAYQKAIQLNPNF
Cya15   ALAYNNLGNALSDQGKRDEAIAAYQKAIQLNPNF
Cya16   ALAYNNLGNALSDQGKLNEAIATYQKAIQLNPNF
Cya17   ALAYNNLGNALKDQGKLNEAIAAYQKALSLPEDT
Cya18   TLAHNNLGLVYQPQGKLEEALREYEAALKIDPKF
```

b)

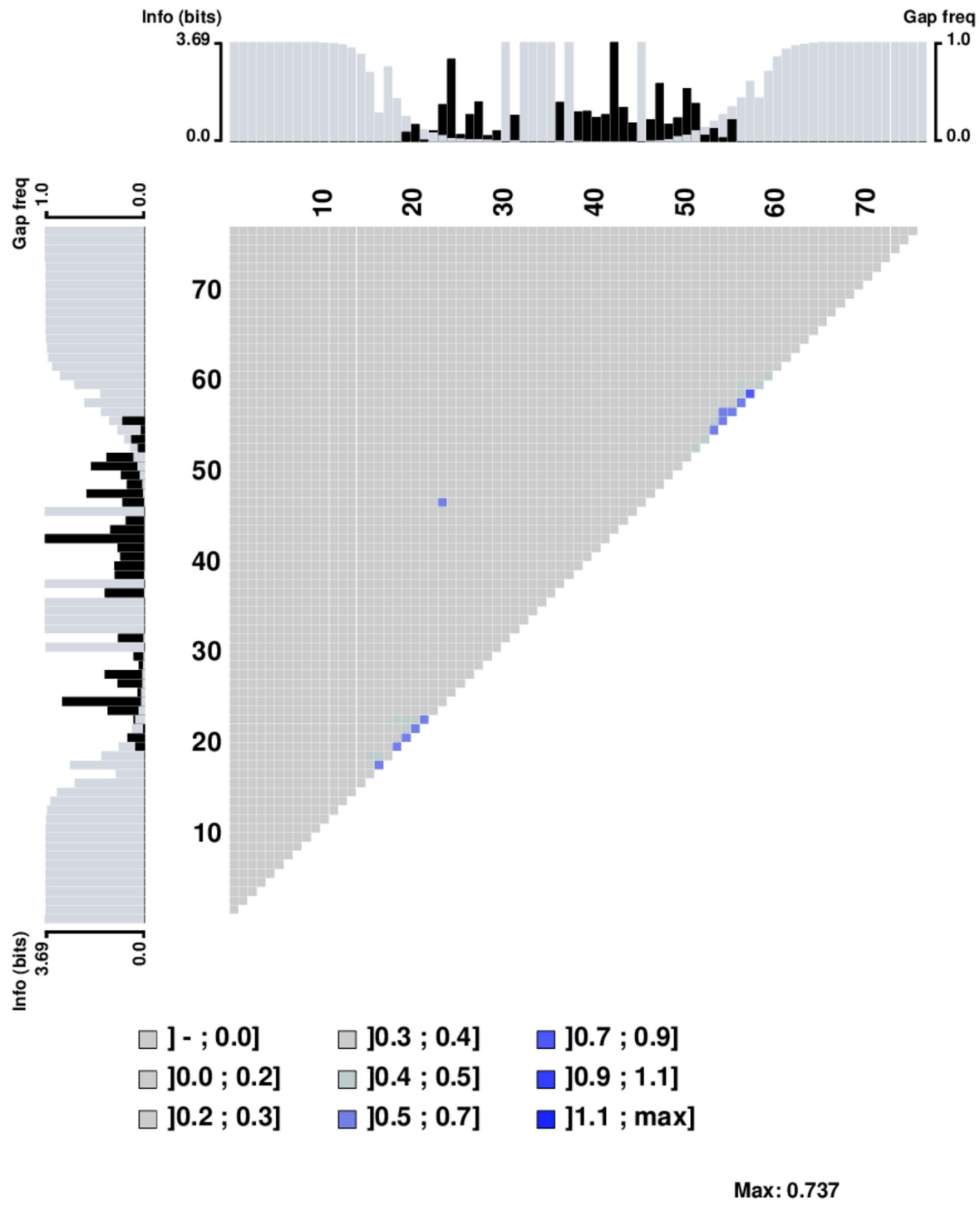
```
>gb|CP011382.1|:c3792015-3790486 Calothrix sp. 336/3, complete genome
GCTGCTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCAGTTGCCGCCTACCAAAAAGCCATTCAACTCGACCCTAACTAT
GCCGCTGCTTACTACAATCTCGGCAATGCCCTGAGAGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTGCTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCAGTTGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTGCTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCAGTTGCCGCCTACCAAAAAGCCATTCAACTCGACCCTAACTTT
GCTTTTGCTTACTACAATCTCGGCAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTTCCAAAAAGCCATACAACTCAACCCTAACTTT
GCTTTTGCTTACAATGGTCTCGGTAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTACCAAAAAGCCATACAGCTCGACCCTAACGAT
GCCAACGCTTACAACAATCTCGGCAATGCCCTATATGACCAGAAGAAACTAGAACCAGCGATCGCCGCCTACCAAAAAGCCATTCAACTCAACCCTAACTTT
```


Ara1 AEEFKSQANEAFKGHKYSSAIDLYTKAIELNSNN
Dan1 AEKLKEKANDYFKDKDYENAIKYYTEALDLNPTN
Hom1 AEELKTQANDYFKAKDYENAIKFYSQAIELNPSN
Mus1 ADHYKNQGNDYLKEKDFSKAIEMYTKAIELNPNN
Sac1 ALERKNEGNVFVKEKHFLKAIEKYTEAIDLDSTQ
The1 AELKKLEGNRSFSENNFISAIEHYSESIRLVEDS
Xen1 AEELKEQANEYFRVKDYDRAVQYYTQAIGLSPDT
Ara2 AVYWANRAFAHTKLEEYGSAIQDASKAIEVDSRY
Dan2 PIYYSNRSLSYLRTECYGYALADATRALELDKNY
Hom2 AIYYGNRSLAYLRTECYGYALGDATRAIELDKKY
Mus2 AIYYANRSLAHLRQESFGFALQDGISAVKADPTY
Sac2 SIYFSNRAFAHFKVDNFQSALNDCDEAIKLDPKN
The2 HQYYSNRAICNIKIENYGSAISDANVAIQLRPDF
Xen2 AIYYGNRSLAYLRTECYGYALADASRAIQLDAKY
Ara3 SKGYYRRGAAYLAMGKFKDALKDFQQVKRLSPND
Dan3 LKGYYRRATSNMALGKFKAALKDYETVVRVRPND
Hom3 IKGYYRRAASNMALGKFRAALRDYETVVKVKPHD
Mus3 LKGYYRRAAAHMSLGKFKQALSDYEYVSKCRPND
Sac3 IKAYHRRALSCMALLEFKKARKDLNVLLKAKPND
The3 FKAYYRRGCAYLCLLKFQDAETDFLKVLSLCNDP
Xen3 IKGYYRRAASNMALGKLKAALKDYETVVKVRPHD

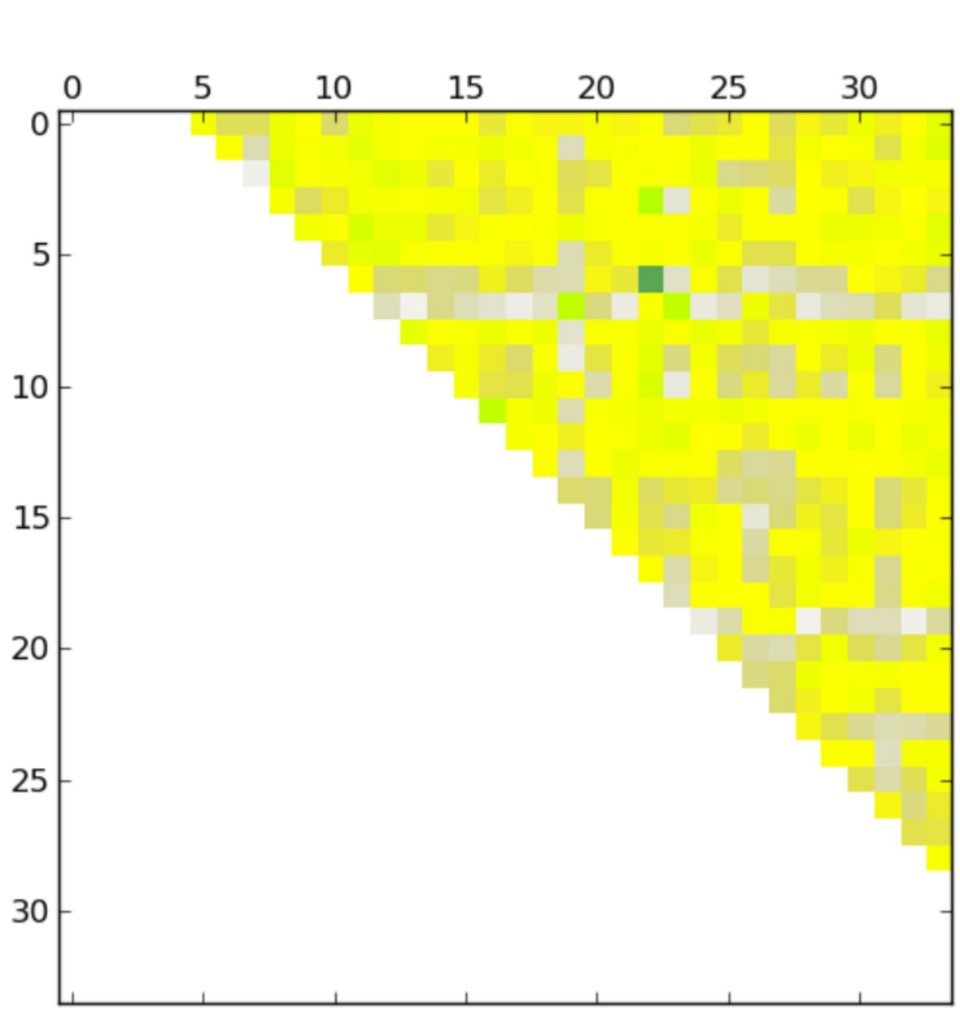
a)


Info (bits)
3.79
0.0
Gap freq
1.0
0.0
5
10
15
20
25
30
] - ; 0.0]
]0.0 ; 0.2]
]0.2 ; 0.3]
]0.3 ; 0.4]
]0.4 ; 0.5]
]0.5 ; 0.7]
]0.7 ; 0.9]
]0.9 ; 1.1]
]1.1 ; max]
Max: 0.905

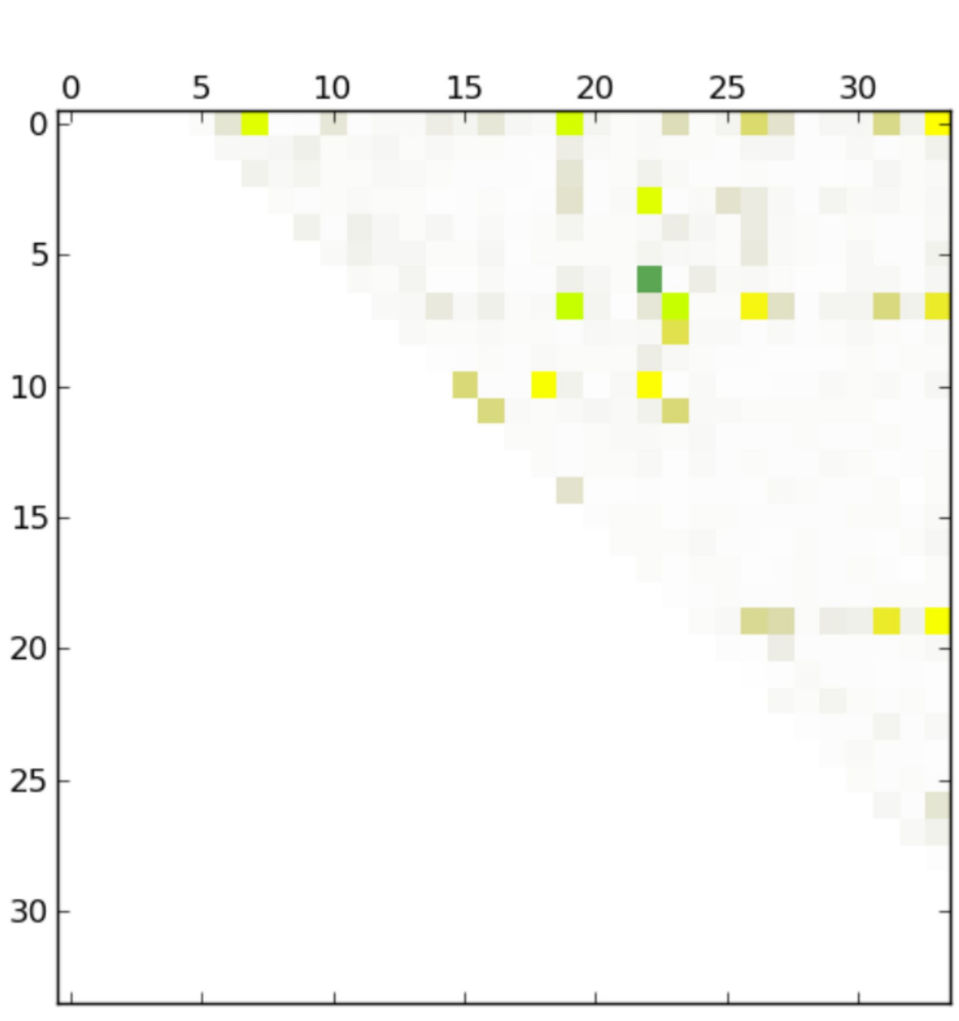
b)

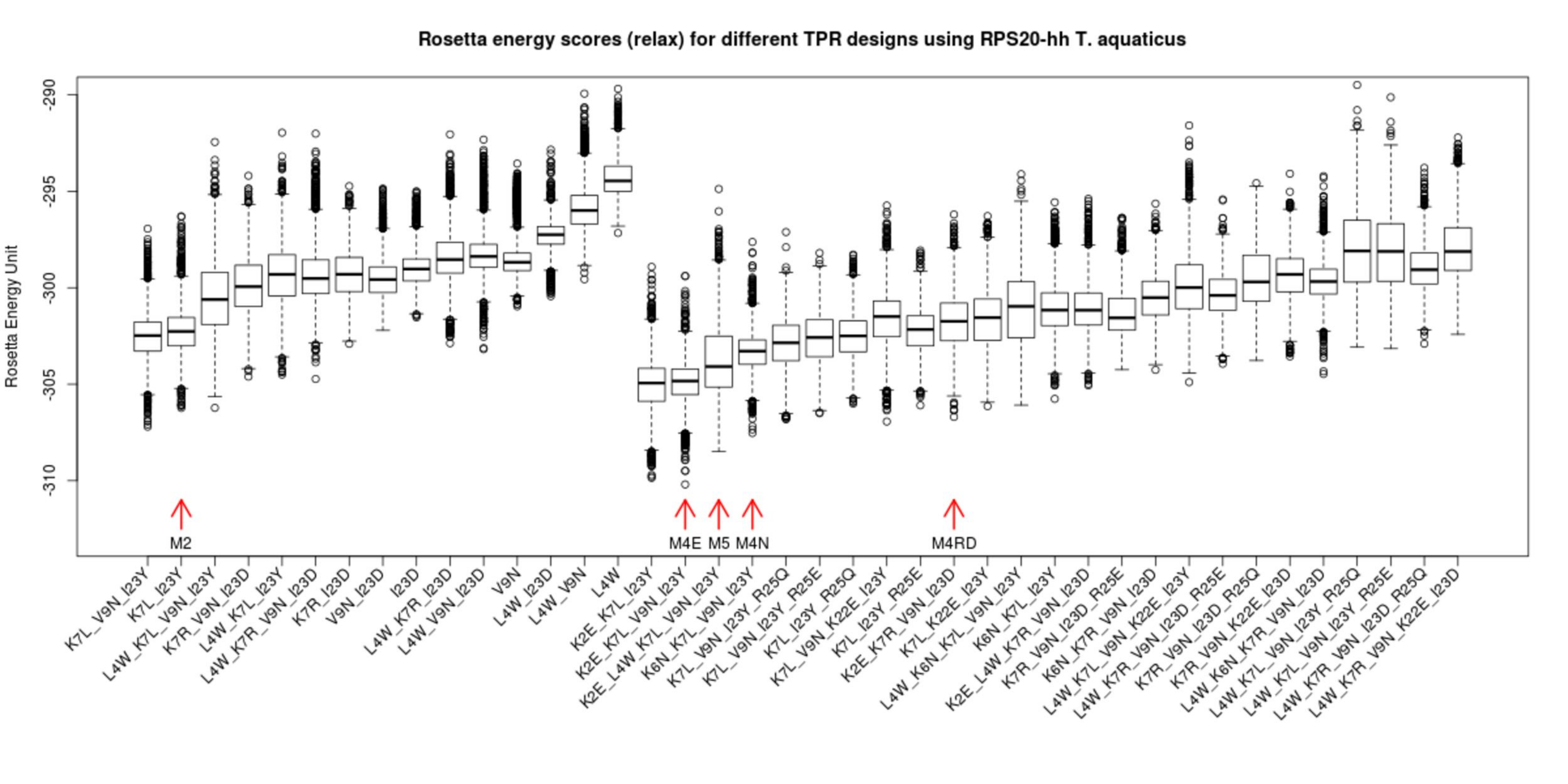

Info (bits)
3.69
0.0
Gap freq
1.0
0.0
10
20
30
40
50
60
70
] - ; 0.0]
]0.0 ; 0.2]
]0.2 ; 0.3]
]0.3 ; 0.4]
]0.4 ; 0.5]
]0.5 ; 0.7]
]0.7 ; 0.9]
]0.9 ; 1.1]
]1.1 ; max]
Max: 0.737

c)

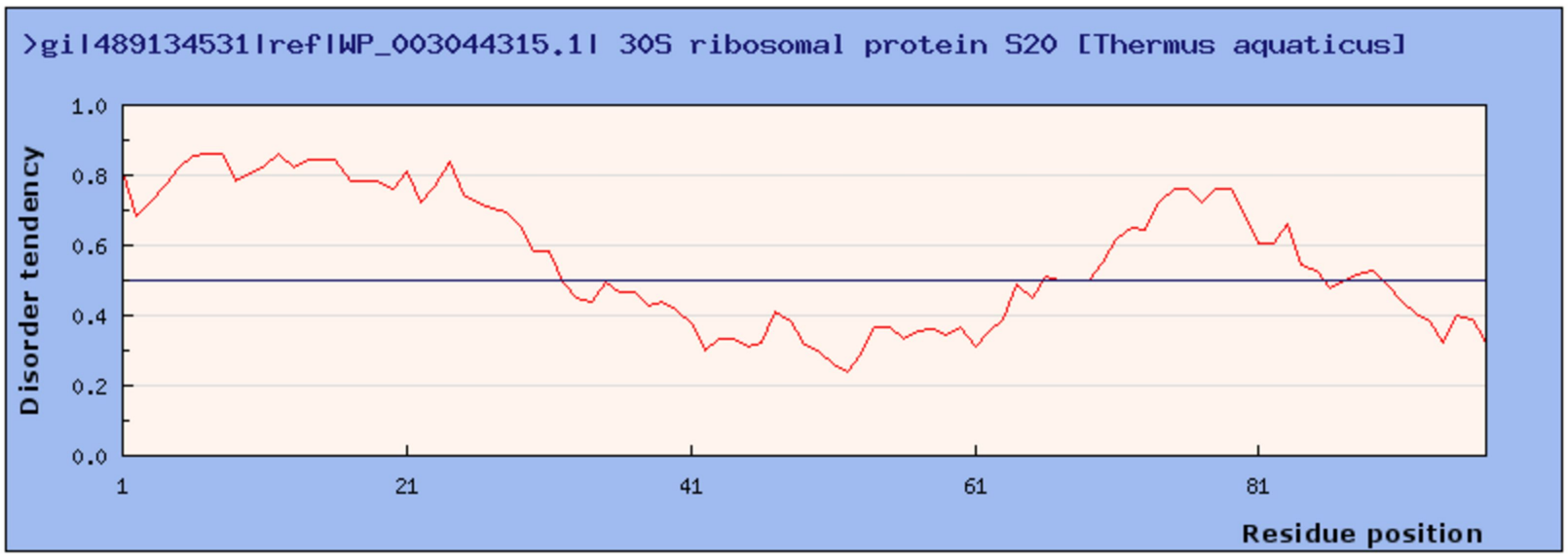

0
5
10
15
20
25
30

d)

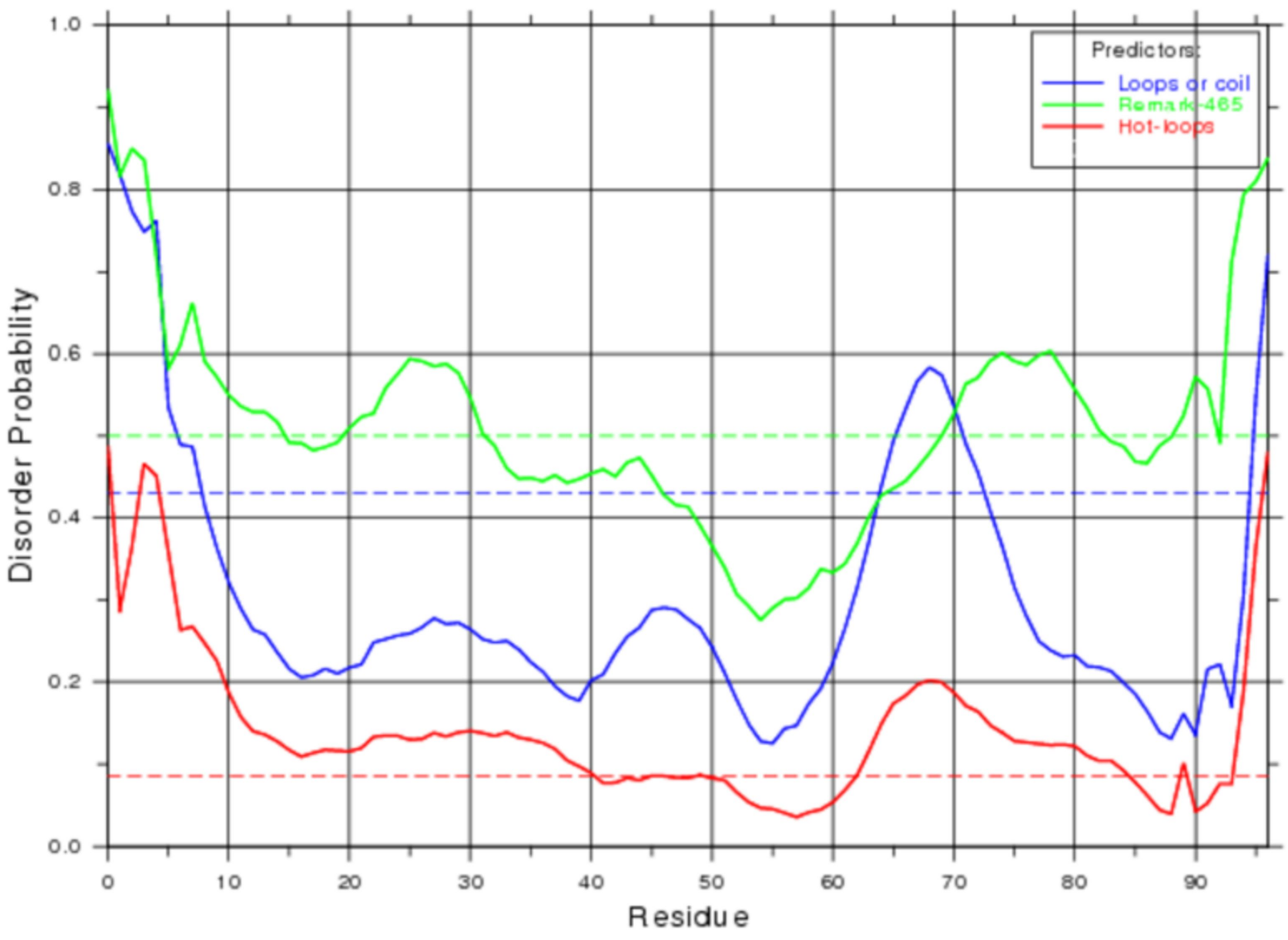

0
5
10
15
20
25
30


Rosetta energy scores (relax) for different TPR designs using RPS20-hh T. aquaticus
Rosetta Energy Unit
-290
-295
-300
-305
-310
M2
M4E
M5
M4N
M4RD
K7L_V9N_I23Y
K7L_I23Y
L4W_K7L_V9N_I23Y
K7R_V9N_I23D
L4W_K7L_I23Y
L4W_K7R_V9N_I23D
K7R_I23D
V9N_I23D
I23D
L4W_K7R_I23D
L4W_V9N_I23D
V9N
L4W_I23D
L4W_V9N
L4W
K2E_K7L_I23Y
K2E_K7L_V9N_I23Y
K2E_L4W_K7L_V9N_I23Y
K6N_K7L_V9N_I23Y
K7L_V9N_I23Y_R25Q
K7L_V9N_I23Y_R25E
K7L_I23Y_R25Q
K7L_V9N_K22E_I23Y
K7L_I23Y_R25E
K2E_K7R_V9N_I23D
K7L_K22E_I23Y
L4W_K6N_K7L_V9N_I23Y
K6N_K7L_I23Y
K2E_L4W_K7R_V9N_I23D
K7R_V9N_I23D_R25E
K6N_K7R_V9N_I23D
L4W_K7L_V9N_K22E_I23Y
L4W_K7R_V9N_I23D_R25E
K7R_V9N_I23D_R25Q
L4W_K6N_K7R_V9N_I23D
L4W_K7L_V9N_I23Y_R25Q
L4W_K7L_V9N_I23Y_R25E
L4W_K7R_V9N_I23D_R25Q
L4W_K7R_V9N_K22E_I23D

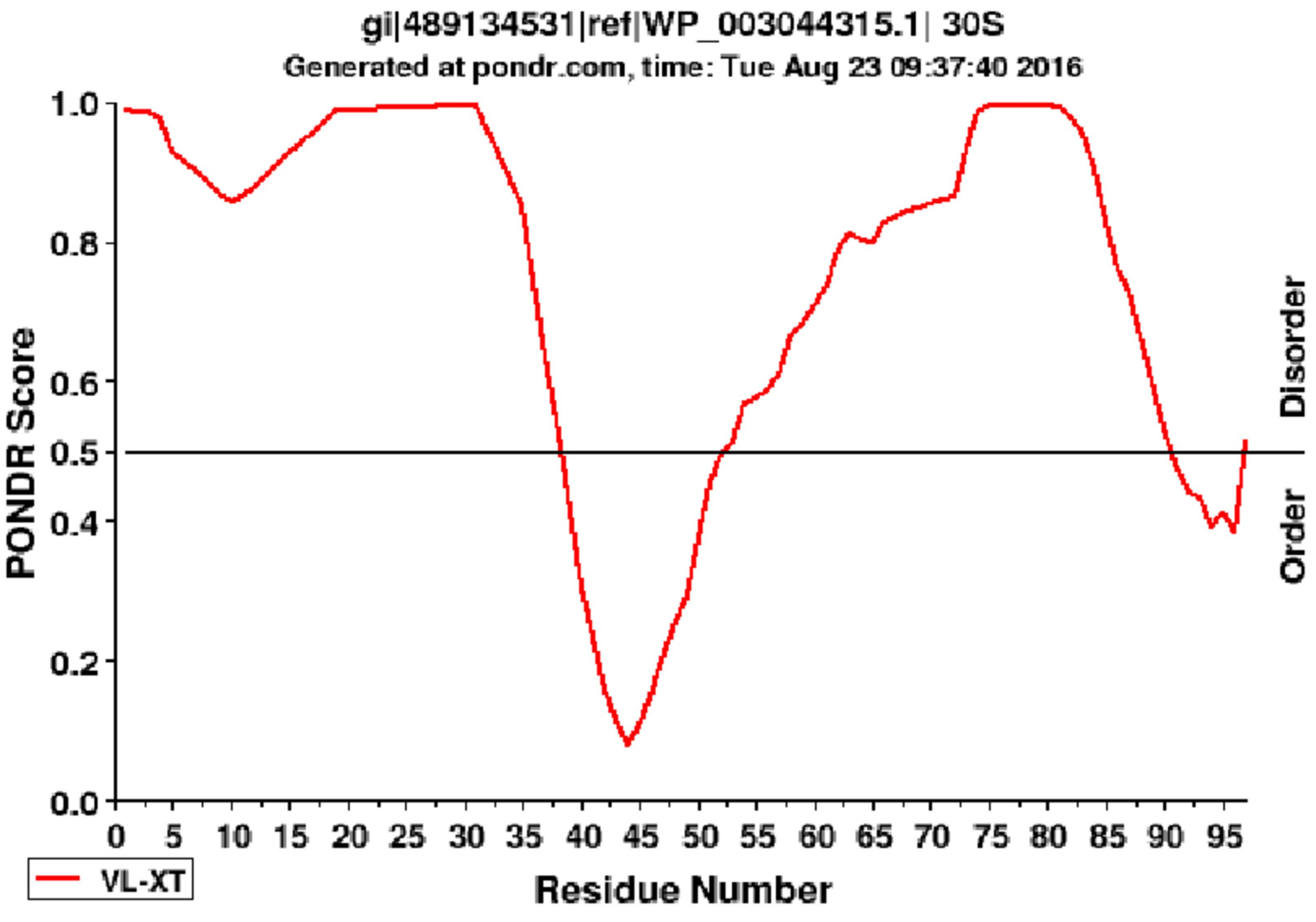
a)

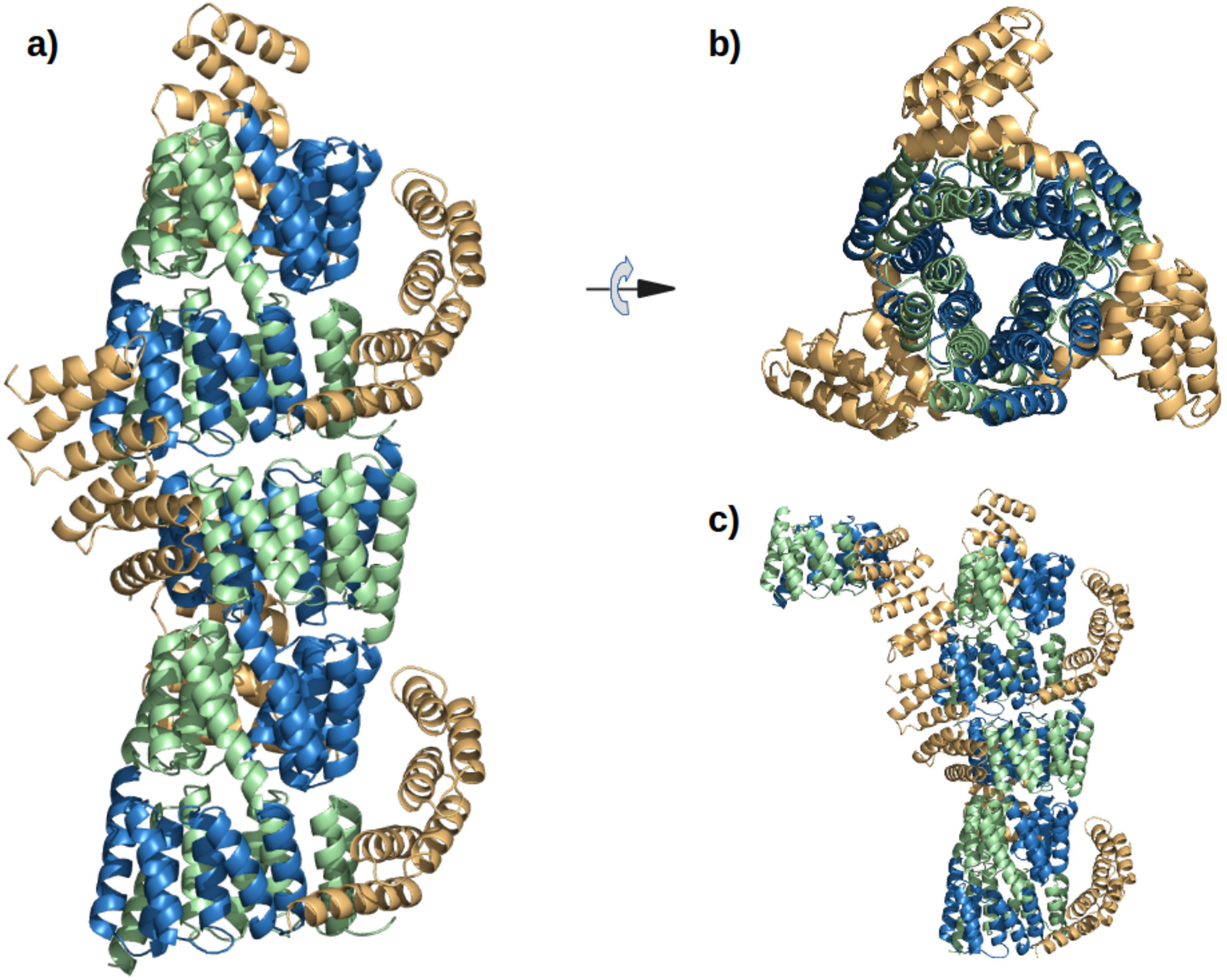

>gi|489134531|ref|WP_003044315.1| 30S ribosomal protein S20 [Thermus aquaticus]
Disorder tendency
1.0
0.8
0.6
0.4
0.2
0.0
1
21
41
61
81
Residue position

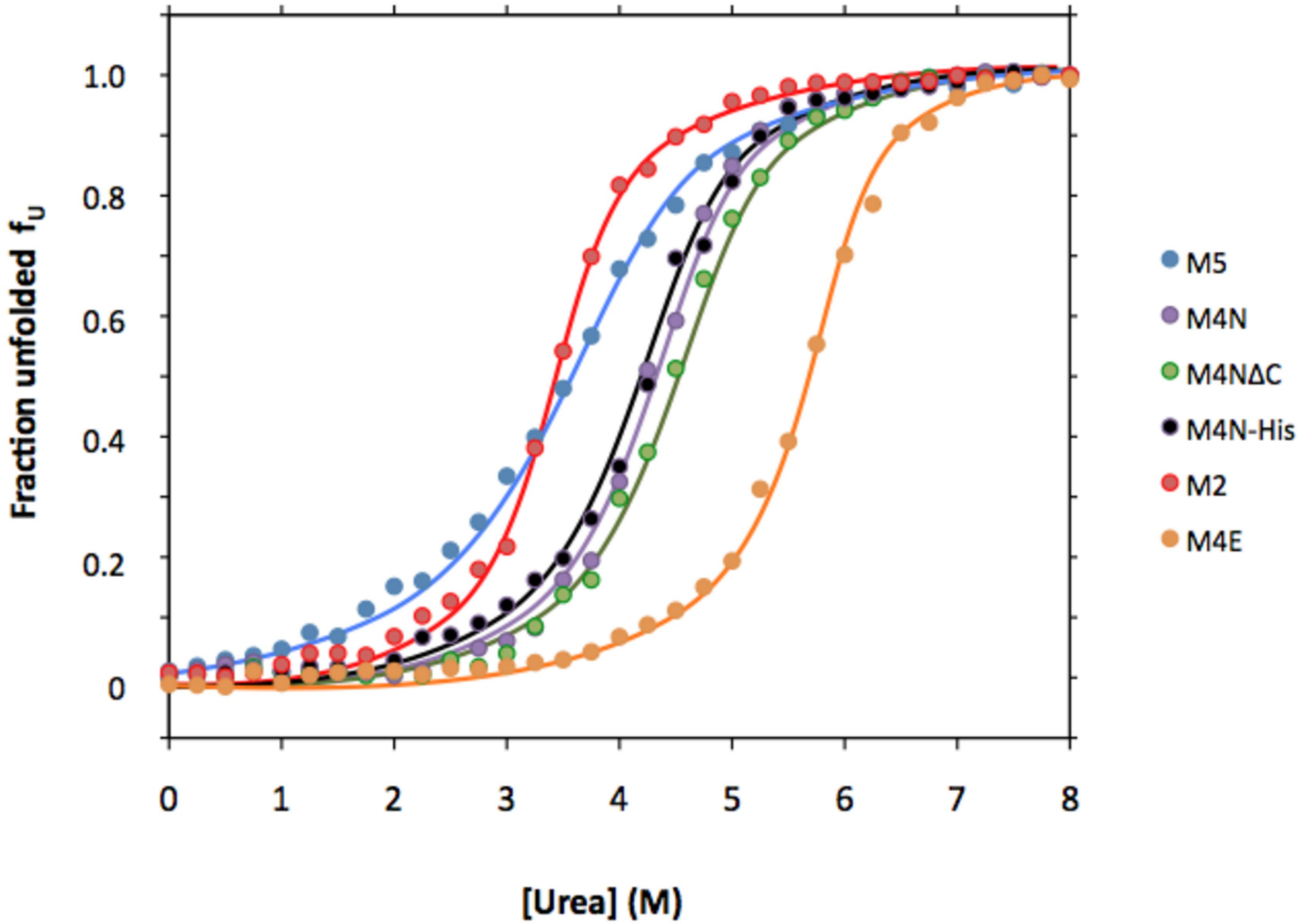
b)

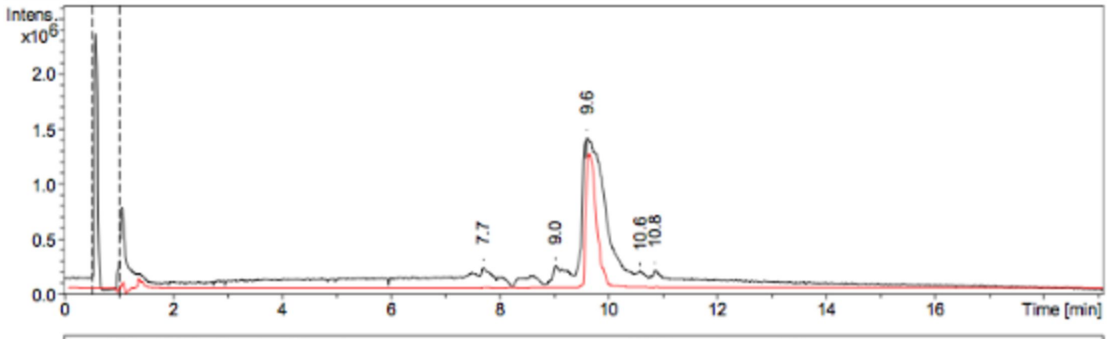

Predictors:
Loops or coil
Remark-465
Hot-loops
Disorder Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70
80
90
Residue

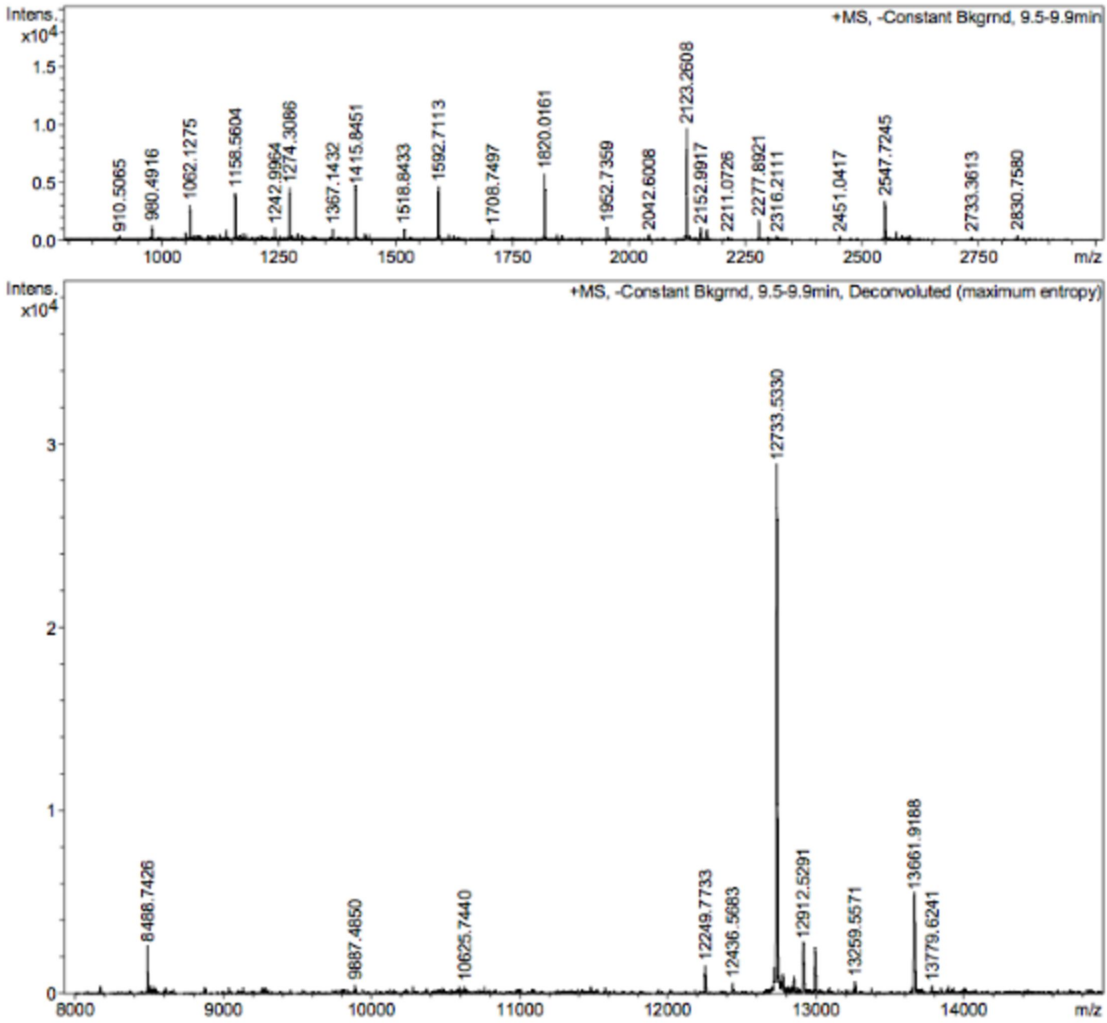
c)


gi|489134531|ref|WP_003044315.1| 30S
Generated at pondr.com, time: Tue Aug 23 09:37:40 2016
PONDR Score
1.0
0.8
0.6
0.5
0.4
0.2
0.0
Disorder
Order
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
Residue Number
VL-XT


a)
b)
c)


Fraction unfolded f_U
1.0
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
6
7
8
[Urea] (M)
M5
M4N
M4NΔC
M4N-His
M2
M4E





>M4N

MGNS

IKTLSNLANLLAQEGKAEEAIKYMRKAVSLDPNN

IKTLSNLANLLAQEGKAEEAIKYMRKAVSLDPNN

IKTLSNLAVLLAQEGKAEEAIKYMRKAVSLIDKA

AKGSTLHKNAAARRKSRLMRKVQKL