BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access

Searching for universal model of amyloid signaling motifs using probabilistic context-free grammars

Witold Dyrka[1*], Marlena Gąsior-Głogowska[1†], Monika Szefczyk[2†] and Natalia Szulc[1]

*Correspondence: witold.dyrka@pwr.edu.pl
†Marlena Gąsior-Głogowska and Monika Szefczyk have contributed equally to this work
[1] Wydział Podstawowych Problemów Techniki, Katedra Inżynierii Biomedycznej, Politechnika Wrocławska, Wrocław, Poland
Full list of author information is available at the end of the article

Abstract

Background: Amyloid signaling motifs are a class of protein motifs which share basic structural and functional features despite the lack of clear sequence homology. They are hard to detect in large sequence databases either with the alignment-based profile methods (due to short length and diversity) or with generic amyloid- and prion-finding tools (due to insufficient discriminative power). We propose to address the challenge with a machine learning grammatical model capable of generalizing over diverse collections of unaligned yet related motifs.

Results: First, we introduce and test improvements to our probabilistic context-free grammar framework for protein sequences that allow for inferring more sophisticated models achieving high sensitivity at low false positive rates. Then, we infer universal grammars for a collection of recently identified bacterial amyloid signaling motifs and demonstrate that the method is capable of generalizing by successfully searching for related motifs in fungi. The results are compared to available alternative methods. Finally, we conduct spectroscopy and staining analyses of selected peptides to verify their structural and functional relationship.

Conclusions: While the profile HMMs remain the method of choice for modeling homologous sets of sequences, PCFGs seem more suitable for building meta-family descriptors and extrapolating beyond the seed sample.

Keywords: Amyloid signaling motif, Functional amyloid, Prion, Sequence motif, Probabilistic context-free grammar, Statistical inference, Amyloid peptide synthesis, ATR-FTIR spectroscopy

Background

Proteins forming amyloid structures have long been a subject of intensive research because of their association with the neurodegenerative diseases. Recently, there has been ever increasing interest in functional amyloids involved in normal physiological processes, for example, in establishing bio-films and membrane-less organelles, and in transmitting molecular signals. In general, the amyloids are defined in terms of the physical structure of cross-β polymer [1–3]. The essential feature of known

amyloid proteins are short amino-acid motifs facilitating aggregation into a beta-sheet-like structure [4, 5]. The templating mechanism of forming amyloid fibrils can be exploited for acting as a prion. The prions are defined in terms of the function of infectious propagation. Indeed, amyloid proteins can act as prions capable of infectious propagation through imposing their own spatial structure on other proteins [3]. A well-known example is the [Het-s] prion from *Podospora anserina* [6, 7]. It is closely related to an ancient signaling pathway of which the amyloid-forming motif is a key element [3, 8, 9]: related motifs were identified in metazoa [10, 11], fungi [12] and bacteria [13]. At least some of these sequence motifs of roughly 20 amino acids form a beta arch fold [14] and often they contain polar amino acids: asparagine and glutamate [15]. Despite these common features, the already identified amyloid signaling motifs (ASM) in bacteria and fungi exhibit high sequence diversity beyond noticeable homology [12, 13]. Is it therefore possible to define universal rules to be obeyed by sequences of functionally-related yet non-homologous ASM? Such a model would allow to identify new amyloid signaling motifs in ever growing data sets of genomic sequences. Moreover, it could facilitate better understanding of mechanisms of conformation transmission and aggregation.

Evolutionary related sequence families are traditionally modeled with the profile Hidden Markov Models (pHMM) [16, 17]. As they assume sequence homology and rely their training process on the multiple sequence alignment (MSA), pHMMs are less suited for modeling diverse collections and meta-family of motifs. Moreover, their discriminative power is limited, especially for short sequences, as amino acid distribution at each position in the alignment is modeled separately.

There exist also various methods dedicated to recognition of amyloidogenic regions of protein sequences [18–22]. They are mainly based on statistical properties of hundreds of hexa-peptides confirmed experimentally to aggregate into amyloid-like fibrils [23]. Unfortunately, these methods often fail to detect functional prion-related amyloid motifs. Indeed, it seems that the amino acid composition of prions has to differ from that of typical amyloids in order to balance water soluble and aggregated state in physiological conditions. This led to developing dedicated prion predictors [15, 24, 25]. One of such algorithms, pWALTZ [15] relies on the model of prion sequence as an amyloidogenic core (or aggregation seed) within a disordered region. Yet, these methods still miss a considerable fraction of HET-s related ASMs. One apparent feature of the motifs that is missing from these models is the propensity to forming the beta-arch structure. This is specifically addressed by ArchCandy [26], a method for detecting beta-arches in protein sequences.

In this piece of research we propose a method aiming at exploiting benefits of beta-arch detection and amyloidogenic composition in a single elegant model. We present a model based on the Probabilistic Context-Free Grammar (PCFG), which extends the profile Hidden Markov Model with capability to capture some dependencies between distant positions in the sequence [27–32]. PCFG is well suited to model nested dependencies resulting from interactions between strands involved in the beta-turn-beta structures (cf. [33]), as recently demonstrated for the HET-s motif [34]. In the same work, the PCFG model was shown to be capable of generalizing between two apparently heterologous architectures of Calcium binding sites [34]. Importantly, the PCFG model does

not assume evolutionary relationship between the sequences, as it does not rely on the multiple sequence alignment.

In [34], the grammars were trained with the genetic algorithm in a setup that significantly limited the number of rules and thus complexity of the model. Here, we propose to use statistical learning, the Inside–Outside (IO) algorithm [35] that allows for training much larger grammars. While the IO algorithm is considered more prone to converge to local minima [36], the benefit of extending the rule set many fold seems to be an overwhelming advantage when describing more complex languages.

The main contributions of the paper are as follows. First, we show that in grammatical modeling of protein sequences, statistical learning leads to comparable or better results than the previously used method of evolutionary learning, while being incomparably quicker. Second, we report on a benefit of smoothing learned profiles of amino-acid emissions represented with the lexical rules. Third, we show that the PCFG model is capable of representing individual families of amyloid signaling motifs, and is practical in searching them in sequence databases. Fourth, we present our main result: the model that generalizes over various amyloid signaling motif families, obtained by training a common PCFG for a set of ten motif families from bacteria. This universal model is then validated by searching for already known fungal ASMs. In this application, the PCFG-based approach is compared to seven other methods. Fifth, we experimentally verify spectroscopic and microscopic characteristics of selected diverse motifs detected with the PCFG model.

Methods

Computational methods

Probabilistic Context-Free Grammar (PCFG) is a generative probabilistic model of sequential categorical data [27]. Under the model, sequences are derived from the start symbol using rewriting rules, associated with some probabilities, until all remaining symbols are non-derivable (or terminal). Formally, PCFG is a quintuple $\mathcal{G} = \langle \Sigma, V, v_0, R, \theta \rangle$, where Σ (alphabet) is a set of terminal symbols, V is a set of non-terminal symbols (variables) disjoint from Σ, $v_0 \in V$ is a start symbol, R is a set of production rules rewriting variables into strings of variables and/or terminals, and θ is a set of corresponding rule probabilities. An illustrative toy example of PCFG modeling a subfamily of beta-hairpin protein sequences can be found in [34]. A probabilistic grammar is *proper* if rule probabilities sum up to 1 over rules rewriting the same variable. A complete derivation is a chain of rules beginning with v_0 and finishing with a string of terminal symbols. Each derivation can be represented as a parse tree. The probability of derivation is the product of probabilities of rules involved. In turn, probability of a sentence x given $\mathcal{G}$ is the sum over all derivations that generate x. The grammar is called *consistent* if the probability mass distributed by the grammar over all sentences sums up to 1. Language is a set of all sentences that can be derived according to the grammar.

Each context-free grammar (whether probabilistic or not) that does not generate empty sentences can be translated to the Chomsky Normal Form (CNF) [37]. This canonical form implies that production rules are either in the form $A \rightarrow a$ (lexical rules) or $B \rightarrow CD$ (structural rules), where lowercase letters denote terminal symbols, while uppercase—non-terminal symbols. In addition, each CNF grammar that does not

generate single letter sentences can be translated to the form where the sets of variables that can be rewritten with the lexical and the structural rules are disjoint. We call such a grammar form *bipartite* CNF, and denote the two groups of variables as *lexical* and *structural* non-terminals, respectively [38].

Context-free grammars are suitable to represent branching and nesting in syntactic description of sequences, but CNF makes the latter unnecessarily lengthy. It is therefore convenient to extend the CNF with *contact* rules, which rewrite variables into triples made of lexical, structural, and lexical non-terminals. This grammar form called Chomsky Form with Contacts (CFC) is especially suitable to represent pairs of amino acids in contact [34]. Since spatial proximity of residues often generates mutual constraints, using a contact rules that generates both residues at once is an effective way to model their dependency.

A parse tree generated with CFG for a protein sequence can be compared to the contact map if corresponding spatial structure is available [38]. However, the PCFG sums the probabilities over all parse trees derivable for the given sequence. Indeed, the most likely parse tree often does not approximate well the most likely shape of all such parse trees [39]. Fortunately, in the case of grammars in the CFC form, it is possible to calculate the probability map of parsing a pair of residues through the contact rules (referred later as probability map of pairing). It can be reasonably expected that residues in contact are often generated with the contact rules. Instead of searching for the most likely shape of parse trees, as used in the RNA structure prediction [40], we propose the probability map of pairing for the best matching sequence fragment as a coarse and partial prediction of spatial distance map for the fragment.

Probabilities of PCFG rules can be inferred from a positive training set of sequences using the Inside–Outside (IO) algorithm [35], which implements the Expectation-Maximization scheme [41]. The algorithm can quickly handle thousands of rules but is prone to converge to local minima [42]. When applied to large generic set of rules constituting a *covering grammar*, the process of optimizing rule probabilities of which most eventually become zero, is akin to learning grammar. The most popular alternatives to IO are based on Genetic Algorithms (GA) using either a fixed set of rules [31, 43], as in the case of IO, or learnable set of rules [36, 44–46].

Efficiency of learning PCFG can be improved when syntactic trees [28, 29, 47–49] or partial syntactic constraints are available [50, 51]. Recently, we proposed using pairwise contacts between amino acids to constrain the GA-based learning of PCFG in CFC form for protein motifs. We showed that even a few relevant contacts led to learning better performing grammars [34].

The outline of the processing pipeline, as proposed and tested in this study, is provided in Additional file 2: Figure S1. Newly added features are described in the "Results" section.

Materials

Computational experiments were carried out using several sets of protein sequences (Additional file 1: Datasets). The collections included existing samples, which were used to benchmark the improved method against the previous approach, and novel samples

of diverse amyloid signaling motifs, which were used to test capability of the current method to generalize.

CaMn A benchmark set of 24 sequences of a Calcium and Manganese binding site from the legume lectins [52] was collected according to PROSITE pattern PS00307 [53] true positive and false negative hits, extended to 27 residues to cover the entire binding site, as in [34, 38]. The motif folds into a stem-like (beta-loop-beta) structure with over 40 internal contacts, many of them forming nested dependencies keeping together beta-strands at the ends of the motif [54]. The sequences were made non-redundant at identity of 70% (nr70) using CD-HIT [55].

HET-s A benchmark set of 160 sequences (nr70) of the HET-s-related motifs r1 and r2 involved in the prion-like signal transduction in fungi was derived from [56]. The largest subset of motifs with length of 21 amino acids was used, as in [34, 38]. The beta-hairpin-like fold of the motif partially relies on interactions between hydrophobic amino acids. HET-s motifs r1 and r2 are known to adopt the beta-hairpin-like fold when templated by the related motif r0 located in the N-terminus of a cooperating NLR protein [57]. While the r0 motifs share a considerable sequence similarity with the interacting r1 and r2 motifs (average identity of around 30%), they contain significantly less aspartic acid, glutamic acid and lysine, and more histidine and serine [56]. A set of 98 HET-s r0 motifs was manually extracted from genes of NLR proteins adjacent to genes encoding proteins containing the r1 and r2 motifs [56]. To test sensitivity of the models trained for the r1 and r2 motifs, we used a subset of 77 non-redundant 21-residue long r0 motifs, as in [34]. HET-s is the only analyzed motif with experimentally solved structure. A high-resolution NMR structure of HET-s amyloid fibrils made of the r1 and r2 motifs from *Podospora anserina* sequence Q03689 is available in the Protein Data Bank (pdb: 2kj3) [58].

BASS Novel families of bacterial amyloid signaling motifs, termed BASS 1 to 10, were identified in neighboring C-termini of Bell domain homologs and N-termini of NLR proteins in bacteria [13]. Each family was defined according to a set of related profile HMMs. For the current piece of research, for each motif family we extracted fragments of Bell-side sequences matched by the motif profile HMMs. Then, we aligned them using Clustal Omega [59] with the *–auto* parameter and submitted to Gremlin [60] to obtain contact constraints. Residue-residue contacts were found for all but three families with the least effective number of sequences (BASS 7, 8, 10). Considering only the most reliable contact pairs, we hand-crafted contact constraints including from 1 to 5 non-overlapping (hence context-free compatible) pairs of residues in contact. We then mapped the contacts onto unaligned sequences. For training, we used the Bell-side motifs samples (nr70) including from 329 (BASS2) to only 7 (BASS10) sequences with length varying from 20 to 40 amino acids. For testing, we used the 143 N-termini of NLR proteins (nr70) with known instances of motifs BASS1-10, according to Supplementary Table 2 in [13].

Other BASS In our previous research [13], a number of pairs of similar amyloid-like patterns were identified in C- and N-termini of neighboring Bell and NLR proteins while being missed with profile HMMs (see Supplementary Table 2 in [13]). We extracted 100 amino-acid long Bell C-termini and 150 amino-acid long NLR N-termini containing these *other* BASS motifs, and made them non-redundant (nr70). This

yielded a set of 18 Bell-side C-termini and 26 NLR-side N-termini, which was used for further testing. In addition to bacterial motifs, the *BASSother* set included 3 related sequences from the Archaea species.

Fungal test motifs Test sets were made of fungal amyloid signaling motifs [12] extracted from a recent set of NLR proteins [61]. The sets included sequence fragments matching Pfam profiles of motifs sigma (Pfam NACHT_sigma, 20 sequences), HET-S (Pfam HET-S, 12 sequences) and PP (Pfam Ses_B, 22 sequences), which were made non-redundant at identity of 70%.

PDBfull and PDBfrag The first negative sample was designed to rather roughly approximate the entire space of protein sequences. It was based on the negative set from [31], which consisted of 829 single chain sequences of 300–500 residues retrieved from the Protein Data Bank [62] at the identity threshold of 30% (accessed on 12th December 2006). In addition, we used the negative sample obtained by cutting the basic negative set into overlapping subsequences of the maximum length of positive sequences and made non-redundant at identity of 70%, as in [34].

NLReff The second negative set was based on a sample of 7901 NLR proteins with N-terminal known to contain non-prion-forming effector domains [61] except for the PNP_UDP_1 domain. The actual negative set consisted of 2411 fragments matching the Pfam profiles of the domains and non-redundant at identity of 70%. Length of the fragments ranged from 41 to 366 amino acids (median: 175). The set was designed to approximate the background encountered when searching positive test motifs in their typical setting in the N-termini of NLRs. The restriction to include only boundaries of domain profiles was based on the fact that putative functional amyloid motifs are sometimes present between the effector and nucleotide-binding domain. PNP_UDP_1 domains were excluded from the set because a fragment of their sequences was predicted to be amyloidogenic according to PASTA2 and AmyloGram, and involved in the beta arch according to ArchCandy. Indeed, available structures show that the fragment consists of two beta strands connected with a loop (e.g. positions 60–100 in pdb:1zos) [63].

DisProt The third negative set was adopted from evaluation of the ArchCandy tool [26] and consisted of 48 sequences (nr70) of soluble disordered protein regions without link to amyloidoses. The set originated from the DisProt database [64]. Lengths of the fragments ranged from 37 to 149 amino acids (median: 101.5). The set was used to check specificity of the tested models against non-amyloidogenic disordered proteins.

Peptides selected for experimental verification Out of sequence fragments identified as ASMs with grammars and other computational methods, we selected four peptides for experimental verification if they form structures consistent with expected features (presence of the beta arch and amyloid-like aggregation). First, to tweak and test the experimental setup, we used two *bona fide* effector-side ASM peptides: BASS3 RHIM-like motif from *Frankia* sp. ORT49035.1 (positions 103 to 123) [13, 65] and *Nectria haematococca* sigma motif from AAS80314.1 (349 to 385) [12, 66]. Then, we analyzed *Methanothrix soehngenii* AEB69175.1 (5 to 29) [67]. This archeal NLR-side motif resembling BASS3 was originally identified through local pairwise alignment of proteins coded by neighboring genes while being missed with the profile HMM-based

search [13]. Finally, we experimentally verified a sequence fragment resembling the sigma motif in *Coleophoma crateriformis* NLR protein RDW70414.1 (382 to 421) [68].

Experimental methods

Peptide synthesis All commercially available reagents and solvents were purchased from Lipo-pharm.pl, Sigma-Aldrich and Merck and used without further purification. Peptides were obtained with an automated solid-phase peptide synthesizer (Liberty Blue, CEM) using rink amide AM resin (loading: 0.59 mmol/g). Fmoc deprotection was achieved using 20% piperidine in DMF for 1 min at 90 °C. A double-coupling procedure was performed with 0.5 M solution of DIC and 0.25 M solution of OXYMA (1:1) in DMF for 4 min at 90 °C. Cleavage of the peptides from the resin was accomplished with the mixture of TFA/TIS/H_2O (95:2.5:2.5) after 3 h of shaking. The crude peptide was precipitated with ice-cold Et_2O and centrifuged (8000 rpm, 15 min, 2 °C). Peptides were purified using preparative HPLC (Knauer Prep) with a C18 column (Thermo Scientific, Hypersil Gold 12 μ, 250 mm × 20 mm) with water/acetonitrile (0.05% TFA) eluent system.

Analytical high-performance liquid chromatography (HPLC) was performed using Kinetex 5μ EVO C18 100A 150 × 4.6 mm column. Program (eluent A: 0.05% TFA in H_2O, eluent B: 0.05% TFA in acetonitrile, flow 0.5 mL/min): A: t = 0 min, 90% A; t = 25 min, 10% A. The peptide purity used for experimental research was ≥95%. Peptides were studied with WATERS LCT Premier XE System consisting of high resolution mass spectrometer (MS) with a time of flight (TOF). Analytical data are provided in Additional file 3: Table S1.

Amyloid-like aggregation To determine aggregation properties of studied peptides Attenuated Total Reflection-Fourier Transform Infrared (ATR-FTIR) experiments were carried out. Vibrational spectroscopy is widely used for protein and polypeptides secondary structure analysis [69, 70] and for monitoring the aggregation processes in amyloids studies [71–73]. The Amide I band (1700–1600 cm^{-1}) corresponding to C=O stretching vibrations and the Amide II band (1600–1500 cm^{-1}) arising mainly from in-plane N–H bending of the peptide bonds are the most useful for secondary structure estimation. For α-helical proteins the maxima of Amide I and Amide II bands are observed at around 1655 cm^{-1} and 1545 cm^{-1}, respectively. Random structures possess the Amide I located at 1645 cm^{-1}. Native β-sheet rich proteins show amide bands maxima near 1635 and 1530 cm^{-1}. When the aggregation occurs, the Amide I band is narrowing and shifting to lower wavenumbers. Rigid and highly ordered amyloid fibrils exhibit the Amide I band below 1625 cm^{-1} [71]. The high water absorption in the Amide I region is main drawback of IR spectroscopy. Subtracting the water absorption spectra may cause significant distortion to the spectral line shape [74]. That is why deuterium oxide is used as an alternative solvent. Due to the frequency of the OH bending mode of D_2O molecules is lowered (from 1635 cm^{-1} for H_2O) to 1210 cm^{-1}. Substitution of water by heavy water causes a relatively small down shift of Amide I band [69].

Spectroscopy For spectroscopic measurements peptides were dissolved in D_2O (deuterium oxide, 99,8% D, Carl Roth, GmbH, Germany) to a final concentration of 2 mg/mL. Peptide solutions were incubated at 37 °C (98.6 °F) for 24 h. ATR-FTIR spectra were collected using a Nicolet 6700 FT-IR Spectrometer (Thermo Scientific, USA) with

Golden Gate Mk II ATR Accessory with Heated Diamond Top-plate (PIKE Technologies). The spectrometer was continuously purged with dry air. All spectra were obtained in the range of 4000–400 cm^{-1}. Directly before sampling, the background spectrum of diamond/air was recorded as a reference (512 scans, 4 cm^{-1}). Spectroscopic measurements were performed at air-dried peptide films. Initially, 10 µl of peptide solution was dropped directly on the diamond surface and was allowed to dry out. For each spectrum, 512 interferograms were coadded, with resolution of 4 cm^{-1}. All spectra were registered at temperature of 37 °C.

All spectra were analyzed using the OriginPro (version 2019, OriginLab Corporation, USA). The analysis included the spectra baseline correction, smoothing using the Savitzky-Golay polynomial filter [75] (polynomial order 2, a window size of 31 points), normalization to 1 for the Amide I' band, and deconvolution into subcomponents using the Lorentz function based on the second derivative spectra. ATR-FTIR spectra were initially preprocessed using OMNIC^{TM} software (version 8, Thermo Fisher Scientific, USA) using the atmospheric and ATR corrections.

Congo red staining The CR method has been widely used to study aggregates in the histopathological samples. While the specificity of this dye is limited due to its ability to bind to proteins with different secondary structures, it is commonly used to study aggregates in vitro [76, 77].

Peptide solutions ($C_{pep} = 50\ \mu M$) were incubated for two months at 37 °C and then used for the CR experiments. A drop of peptide aliquot (10 μL) was allowed to dry on a glass microscope slide. The staining was performed according to the published procedure [78]. Birefringence was determined with an ECLIPSE 50i microscope (Nikon, Japan).

Results

In a previous paper [34], we showed that fungal prion-forming HET-s motifs r1 and r2 can be accurately represented with automatically inferred probabilistic context-free grammars comprising just of three lexical and four structural symbols (*l3s4*). In the cross-validation scheme, the model achieved the average precision (AP) of 0.60 for the negative to positive sample cardinality ratio over 200:1 [34]. Moreover, we found that consensus predictions from grammars comprising of seven structural symbols were practically useful for identifying related HET-s r0 motifs in NLR proteins with AP of 0.82 [34]. However, the PCFGs were outperformed in this task by less expressive profiles of Hidden Markov Models (pHMM). The presumed disadvantage of our PCFG approach was in likely over-simplicity of grammars due to the tiny number of non-terminal symbols. This limit was necessary to make the number of rules manageable in our GA-based scheme for inferring rule probabilities. Indeed, the number of possible rules increases exponentially with the number of available symbols and our implementation of evolutionary approach could not effectively handle search spaces of more than 500–1000 rules [34].

Improved modeling of individual families of protein motifs

Even though the evolutionary scheme could be adjusted [79, 80], in the current project we resorted to the classical statistical learning method, the Inside–Outside (IO)

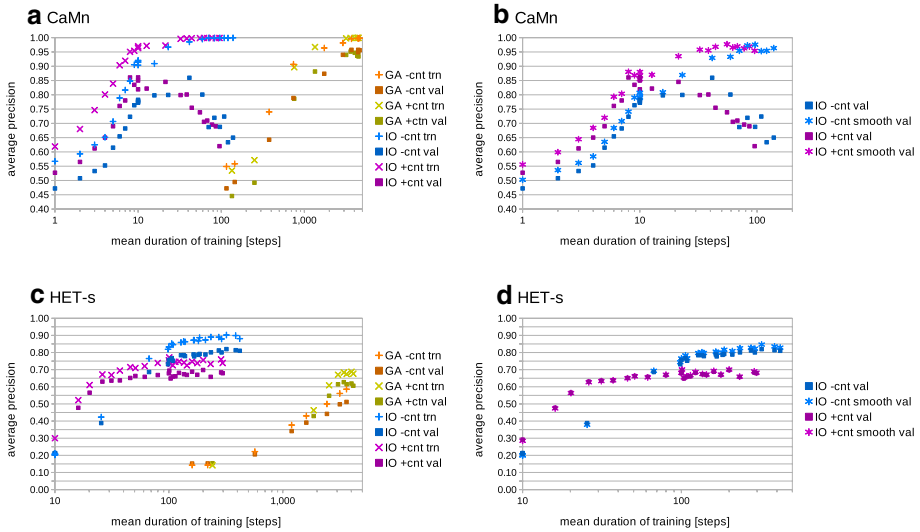


Fig. 1 Performance comparison of grammar learning schemes. Mean average precision with regard to mean training duration achieved for the CaMn (**a**, **b**) and HET-s (**c**, **d**) motifs in the cross-validation experiment. Performance on training and validation sets without smoothing (**a**, **c**) and on validation set with smoothing (**b**, **d**) is shown. Data for evolutionary learning without (with) contact constraints are plotted in shades of orange (green). Data for statistical learning without (with) contact constraints are plotted in shades of blue (magenta). Note logarithmic scale for duration of training

algorithm [35]. The main advantage is relatively quick convergence of the procedure, even for hundred thousands of rules, thus allowing for considerably more non-terminals symbols (for example 40) in the *covering grammar* made of all possible rules.

Smoothing The large size of such a grammar increases the risk of over-generalization, i.e. over-fitting rule probabilities to training data. A viable solution consists on smoothing the probabilities in the course of post-processing, so the grammar can parse sequences with amino acids unseen in the given context during training. While generally not trivial, the smoothing is relatively straightforward when applied to lexical rules modeling amino-acid emissions from lexical variables. Indeed, one can apply one of the classical mutation models such as PAM [81] or BLOSUM [82]. While the latter is typically considered more accurate, the former has the advantage of intrinsic simplicity and elegance of the underlying Markov model, which is why it was chosen for this project. In our implementation, distributions of amino acids modeled by lexical rules can be smoothed according to the requested number of point accepted mutations.

Cross-validation The updated IO-trained PCFG method was tested on two benchmark sets from our previous research, the HET-s motifs r1/r2, and a Calcium and Manganese binding site motif (CaMn). Replicating the procedure from [34], we used a variant of the 8–fold Cross–Validation scheme in which 6 parts were used for training, 1 part was used for validation and parameter selection, and 1 part was used for final testing (the scheme resulted in 56 runs for each sample).

Learning efficiency of IO and GA We first compared efficiency of evolutionary and statistical learning of grammars for the simpler CaMn set. Similarly to [34], we stuck to the *l3s4* grammars in the CFC form and compared training with and without

contact constraints. Setups of evolutionary (GA) and statistical (IO) learning schemes were essentially identical except that the convergence criterion was calculated over 100 epochs for the former and just 10 iterations for the latter, simply because the Inside–Outside procedure converged much quicker: while on average 3000–4000 steps were needed using GA, only 10 or 40 iterations was enough for IO (Fig. 1a). Contact constraints seemed to shorten the training for IO but not for GA. When tested on the validation set against *PDBfrag*, best grammars achieved the maximum AP around 0.95 for evolutionary and 0.85 for statistical learning. This likely resulted from more pronounced over-fitting with IO, where the gap in performance between the training and validation sets was around 0.10–0.15, in comparison to 0.05 for GA (in terms of AP). Closer inspection of the induced best grammars revealed that the Inside–Outside was more prone to suppressing rule probabilities: around half of them was set below $1e{-}5$, and a couple of percent were set to zero. In contrast, the evolutionary scheme suppressed less than 1% of rule probabilities below $1e{-}5$ and did not turn off any rule completely. (Of note is that in practice the structural rules with probability below $1e{-}5$ have at best negligible impact on the sequence probability log scores and can be pruned off in order to improve the speed of parsing.) This finding led us to trying the PAM-based smoothing on the lexical rules probabilities. This turned out to be highly efficient: the average precision of apparently over-fit IO-trained grammars was pushed to over 0.97, almost closing the performance gap of around 0.30 (for the longest training). The best results of smoothing were obtained with PAM values in the range of [5, 20] (Fig. 1b shows results for PAM10). For grammars trained with GA long enough to experience over-fitting, the smoothing pushed the average precision up to above 0.98.

Analogous tests were performed on the larger and more diverse HET-s set. While GA again needed on average 3000-4000 steps to converge, IO needed 200–300 iterations with the convergence criterion calculated over 100 iterations (Fig. 1c). The statistical learning led to the maximum average performance of 0.79, in comparison to 0.63 achieved with the evolutionary scheme. The contact constraints still sped up the convergence of IO, but at the same time they limited the top performance on the training *and* validation sets by around 0.15. This was in contrast to GA-trained grammars which benefited from the contact constraints by around 0.10. The over-fitting gap was moderate: around 0.10 for IO and 0.06 for GA. Consequently, the effect of smoothing was very limited (0.01–0.02, Fig. 1d shows results for the PAM10 smoothing).

Apparently, in our framework, small-number PAM matrices (with no BLOSUM counterparts) stroke the balance between avoiding the harsh penalization of substitutions typical of closely related sequences, and preserving specificity of emissions from lexical variables. In further experiments, the PAM10 matrix was used for smoothing.

Grammar size Saturation of performance on the HET-s training set, together with the lack of significant over-fitting, suggested that grammar size was inadequately small with regard to the diversity of the sample. This prompted us to investigate using larger numbers of non-terminals with the IO algorithm. We opted for relatively long evolution counting on the smoothing to fight back the negative effects of over-fitting. Specifically, we trained rule probabilities of grammars in the CFC form [34, 38] made from 5, 7 and 10 lexical and 10 to 33 structural non-terminals. Consequently, the smallest

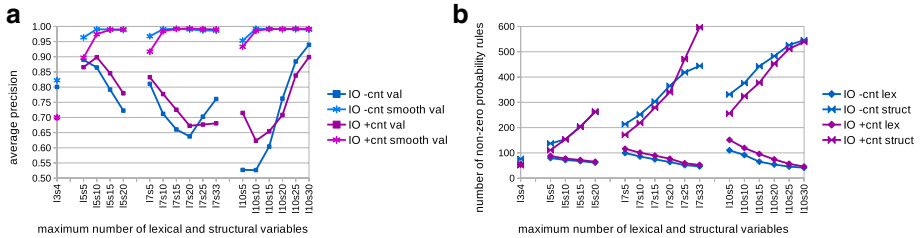


Fig. 2 Number of variables effects on performance and grammar size. Mean number of lexical and structural rules with non-zero probability (**a**) and mean average precision (**b**) in the cross-validation experiment are shown with regard to the maximum number of variables (*l*: lexical, *s*: structural). Data for learning without (with) contact constraints are plotted in shades of blue (magenta)

covering grammar *l*5*s*5 counted 1225 rules, while the largest *l*10*s*30 had 138 200 rules. The stop condition was set to 1.0005 over 100 iterations. At the end of training, grammars retained on average only 198–650 rules (117–482 with probabilities over $1e{-}5$) including 41–151 lexical rules (27–60 with probabilities over $1e{-}5$), see Fig. 2a. Of note, the trained grammars with more structural variables had less lexical variables, which appeared to be decreasing asymptotically to 20 (or one lexical rule per one amino acid). Only grammars with least variables did not achieve virtually perfect fit to the training sample (mean AP over 0.999). Without smoothing, the performance over the validation set dropped for grammars with a medium number of non-terminals, presumably due to the over-fitting (Fig. 2b). Interestingly, the validation performance improved again for grammars with large number of non-terminals. The most plausible explanation is that due to more laborious convergence, the training stopped before significant over-fitting took place. The smoothing of lexical rules probabilities using the PAM 10 model led to the mean AP over the validation set around 0.99 for all grammars with 10 (15) or more structural variables trained without (with) contacts (Fig. 2b).

Practical performance It can be reasonably expected that by using grammars learned for the HET-s r1 and r2 motifs, the r0 motifs can be automatically extracted both from random full length sequences (approximated by *PDBfull*) and from NLR proteins (approximated by *NLReff*). Moreover, grammars have to distinguish amyloid signaling motifs from non-amyloidogenic disordered proteins *DisProt*. For the assessment, we used the measures of recall (sensitivity) of the positive sequences at false positive rates (FPR) of 0.01 and 0.001 (the latter for *NLReff* only), and the average precision (AP) against the *DisProt*. To avoid overestimating performance, the minimal observable FPR due to the negative set cardinality was assumed in averaging over the folds (*PDBfull*: $1.2e{-}3$, *NLReff*: $3.7e{-}4$). The tests were conducted for *l*7*s*15 grammars and their neighboring configurations (*l*5*s*15, *l*7*s*10, *l*7*s*15, *l*7*s*20 and *l*10*s*15).

In the scenario with the r0 motifs searched among non-prionic *NLReff*, the mean recall was in range 0.66–0.73 at FPR of $1e{-}3$, and in range 0.79–0.84 at FPR of $1e{-}2$. Best performance was achieved with *l*5*s*15 grammars trained without the contact constraints. In the search against *PDBfull*, the mean recall at FPR of $1e{-}2$ was in range 0.70–0.76. The mean AP against *DisProt* was between 0.96 and 0.98. Considerable improvement was achieved when scores from grammars obtained in several runs under the same conditions were averaged before classifying hits. The effect was most pronounced when

moving from a single grammar to a pair, but increase was still notable at least up to 6 grammars combined (the largest number tested was 8). In this case the mean recall was in range 0.82-0.84 at FPR of $1e{-}3$ and 0.85–0.91 at FPR of $1e{-}2$ for *NLReff*. Similarly, the mean recall at FPR of $1e{-}2$ increased to 0.81–0.85 against the *PDBfull*. The already high performance against non-amyloid disordered proteins was kept.

Application to bacterial amyloid signaling motifs

Performance of the method was further tested on newly identified bacterial amyloid signaling motifs (*BASS*) [13]. A major difference with regard to the previous tests is the variable length of sequences in the BASS families. The variation could result either from indels, intra-family diversity, or motif truncation in the course of extraction. While sequences in each family were in general alignable, which was exploited in the contact constraints prediction, it is worth noting that the PCFG input was unaligned. In the experiment, we used the set-up established for HET-s and the $l7s15$ covering grammar.

Cross-validation We first applied the 6-fold standard cross-validation scheme to compare performance of grammars trained with and without contact constraints. In the validation phase, positive sets (consisting of sequence fragments of variable length) and the negative set *PDBfrag* (made of 40-amino acid chunks) were scanned with the grammars using the 20-to-40-amino-acid window. Since the average precision is sensitive to the cardinality ratio of positive and negative sets, we used the Youden's Index [83], as a complementary measure, comparable between various BASS families. The results are shown in columns AP and YI in Table 1.

The cross-validation experiment showed that grammars learned the motif pattern in all cases. The mean Youden's index ranged from 0.87 to 0.99, which corresponded to the mean average precision from 0.42 (BASS9) to 0.96 (BASS1). While using the contact constraints was favourable in terms of average precision whenever available, the effect was substantial only for BASS4 (increase from 0.77 to 0.93) and BASS9

Table 1 Average performance of grammars for individual ASMs

motif	**trn**	**val:** *PDBfrag*		**tst:** *PDBfull*	**tst:** *NLReff*		**tst:** *DisProt*
Family	**Size**	**AP**	**YI**	**Recall@ FPR0.01**	**Recall@ FPR0.01**	**Recall@ FPR0.001**	**AP**
BASS1	210	0.93/0.95	0.98/0.98	0.99	0.99	0.97	1.00
BASS2	329	0.92/0.94	0.96/0.97	0.98	0.99	0.97	1.00
BASS3	145	0.92/0.93	0.98/0.97	0.97	0.98	0.92	0.99
BASS4	127	0.77/0.93	0.96/0.97	1.00	1.00	0.96	1.00
BASS5	50	0.83/0.86	0.99/0.99	0.85	0.90	0.81	0.97
BASS6	111	0.89/0.89	0.97/0.98	0.91	0.97	0.89	0.97
BASS7	17	0.85/–	0.95/–	0.97	0.92	0.97	0.98
BASS8	14	0.79/–	0.93/–	1.00	1.00	0.96	1.00
BASS9	38	0.42/0.63	0.87/0.91	0.79	0.90	0.79	0.96
BASS10	7	0.57/–	0.99/–	1.00	1.00	1.00	1.00

AP and YI are given for both training without contacts/with contacts. The test set performance is shown only for the mode with the best validation AP. Notations: trn, val, tst are training, validation and testing positive sets, AP is the average precision and YI is Youden's Index

(increase from 0.42 to 0.63). The best performing constraints comprised of 1 to 3 pairs of residues in contacts.

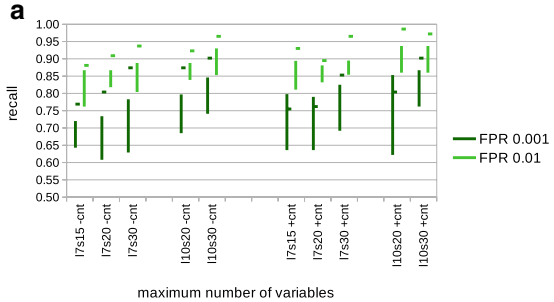
Practical performance Next, we evaluated performance in the practical settings already defined for the HET-s set. The results are shown the last four columns in Table 1. At FPR of 0.01, the grammars found at least 90% of NLR-side motifs among non-prionic NLR effector domains, and at least 79% among sample PDB sequences. At FPR of 0.001 against *NLReff*, the grammars found at least almost 89% of NLR-side motifs for all families except BASS5 and BASS9 (around 80%). Also, grammars allowed for distinguishing ASMs from non-amyloidogenic disordered regions (*DisProt*) with the average precision ranging from 0.96 to 1.00. Overall, single grammars for individual BASS families typically performed better then combined grammars (using the score averaging scheme) for the HET-s motif.

Generalization of bacterial amyloid signaling motifs

Having checked that the PCFG framework can effectively model each motif family, we aimed at assessing whether the method can be used to obtain a general model of the amyloid signaling motifs.

Bacterial motifs The combined set of all ten BASS families was used to train universal BASS grammars in the 6-fold cross-validation scheme (the folds were made by merging the corresponding folds for each motif family). The number of symbols used in the grammars was 7 or 10 for lexical variables and 15, 20 or 30 for structural variables, as it can be reasonably expected that a grammar covering several families requires more complex structures.

The performance of resulting grammars in cross-validation ranged from AP of 0.79 for *l*7*s*15 trained without contact constraints to 0.86 for *l*7*s*30 trained with contact constraints. The best performance was achieved at the log score threshold of 2.4 to 2.9 yielding the Youden's index of 0.87 to 0.92. In the cross-validation experiment, adding more lexical and structural symbols and using the contact constraints improved AP of grammars. The same held for practical performance evaluation on BASS-containing versus



b

recall @ FPR 0.001	l7s15 -cnt	l7s20 -cnt	l7s30 -cnt	l10s20 -cnt	l10s30 -cnt	l7s15 +cnt	l7s20 +cnt	l7s30 +cnt	l10s20 +cnt	l10s30 +cnt
BASS 1	.97	.97	.97	.97	1.0	1.0	.97	.97	1.0	1.0
BASS 2	.98	.98	.98	.98	1.0	.98	.96	1.0	.98	1.0
BASS 3	.79	.84	.90	.90	.95	.79	.74	.90	.95	.95
BASS 4	.89	.89	1.0	1.0	1.0	.89	1.0	1.0	1.0	1.0
BASS 5	.00	.13	.38	.25	.25	.00	.00	.13	.00	.00
BASS 6	.82	1.0	1.0	1.0	1.0	.73	.91	.91	.91	.91
BASS 7	.20	.20	.60	.60	.60	.00	.00	.20	.00	1.0
BASS 8	.00	.00	.00	.00	.25	.00	.00	.00	.00	.25
BASS 9	.38	.38	.75	1.0	.75	.38	.50	.88	.50	.88
BASS 10	.40	.60	.80	.60	1.0	.20	.20	.80	.20	1.0
BASS avg	**.54**	**.60**	**.74**	**.73**	**.78**	**.50**	**.53**	**.68**	**.55**	**.80**
HET-s	.83	.92	.92	.92	.92	.58	.92	.92	.50	.50
sigma	.95	.95	.90	.95	.90	.95	.90	1.0	.90	.90
PP	.64	.64	.86	.68	.68	.41	.27	.46	.18	.23
fungal avg	**.81**	**.83**	**.89**	**.85**	**.83**	**.65**	**.70**	**.79**	**.53**	**.54**

Fig. 3 Performance of the generalized BASS grammars on bacterial and fungal test motifs from NLR proteins. **a** Recall at the false positive rates of 1*e*−2 (pale green) and 1*e*−3 (dark green) for BASS1-10 motifs with regard to the maximum number of variables in grammars (*l*: lexical, *s*: structural) and presence/absence of the contact constraints (±*cnt*). Vertical bars indicate recall ranges achieved using single grammars in six runs. Short horizontal bars indicate recall obtained using the score averaging scheme. **b** Recall at the false positive rates of 1*e*−3 obtained using the score averaging scheme for each BASS class separately and for the three fungal amyloid signaling motifs

non-prionic N-termini of NLRs, except that the benefit of training with the constraints was weaker (Fig. 3a). The best grammars accepted up to around 85% (94%) of the positive test sample at the false positive rate of $1e{-}3$ ($1e{-}2$) against *NLReff*. The mean average precision against the non-amyloidogenic disordered proteins was in the range of 0.988–0.995. We did not notice significant trends with regard to the stop condition.

Fungal motifs Encouraged with the results for the test BASS set, we were curious if the grammars were general enough for searching for novel motifs. Thus, we tested the all-BASS grammars on fungal NLR N-termini with HET-s, sigma and PP/ses_B motif instances against the non-prionic NLR N-termini. The experiment showed considerable performance. The sigma motif was most easily distinguished (average recall of 67–87% at FPR of $1e{-}3$), followed by the HET-s motif (recall 25–63%), and the PP motif (recall 11–46%). Very good performance with the sigma motifs was likely due to the relatively high length and amino acid composition similar to some bacterial motifs (e.g. BASS2, 3 and 7). On the other hand, quite fair sensitivity to the quite distinctive HET-s, as well as relatively poor sensitivity to the RHIM-like PP (seemingly related to BASS3), cannot be easily explained. Unlike the previous BASS tests, training with constraints tended to lower performance for the fungal ASMs, as did adding more symbols and longer training. (With the exception of the sigma motif, for which performance was universally high.) This is not unexpected, since extrapolating outside the training domain has to carefully avoid over-fitting to be successful.

Score averaging We also noticed high variation of performance of grammars trained with the same parameters on different folds. While recall (at FPR of $1e{-}3$) on the NLR-side BASSes in most cases varied only by 10–15% (Fig. 3a), it differed from zero to 75% on the HET-s motif and from zero to 46% on the PP motif test set. While this could be partially due to smaller test sets, it also clearly suggested sub-optimal character of individual grammatical models. Thus, we resorted to the strategy of averaging scores from several grammatical models when scanning sequences. Following previous experiments with grammars for individual motifs, we used average scores of six grammars (one from each training fold). The procedure yielded very good results, with the recall at FPR of $1e{-}3$ increasing up to 86–92% for all fungal amyloid signaling motifs for the $l7s30$ setup trained without contact constraints (Fig. 3b). Performance of grammars trained with the contact constraints was still rather poor for PP and somehow mixed for HET-s, in contrast to universally good performance of grammars trained without the constraints.

The averaging approach also improved the performance on NLR-side BASSes, up to recall of around 90% (99%) at FPR of $1e{-}3$ ($1e{-}2$), as indicated with short horizontal bars on Fig. 3a. In fact, predictors made with the averaging scheme most often outperformed the best single grammars. The breakdown of the results by BASS class indicates that the generalized BASS grammars recognized test samples from the most numerous BASS1-4 and BASS6 classes fairly well (recall of at least 0.90 at FPR of $1e{-}3$ for predictors averaging over $l7s30$ grammars or larger). Performance over classes BASS7, 9 and 10 was mixed, as grammars with 30 structural variables were doing much better than smaller ones. BASS5 and BASS8 classes were apparently not modeled properly (Fig. 3b).

On the set of other BASS motifs, the averaging approach resulted in recall from 50 to 75% (73–89%) at FPR of $1e{-}3$ ($1e{-}2$) against *NLReff*. As with the fungal motifs, grammars trained without the contact constraints performed better.

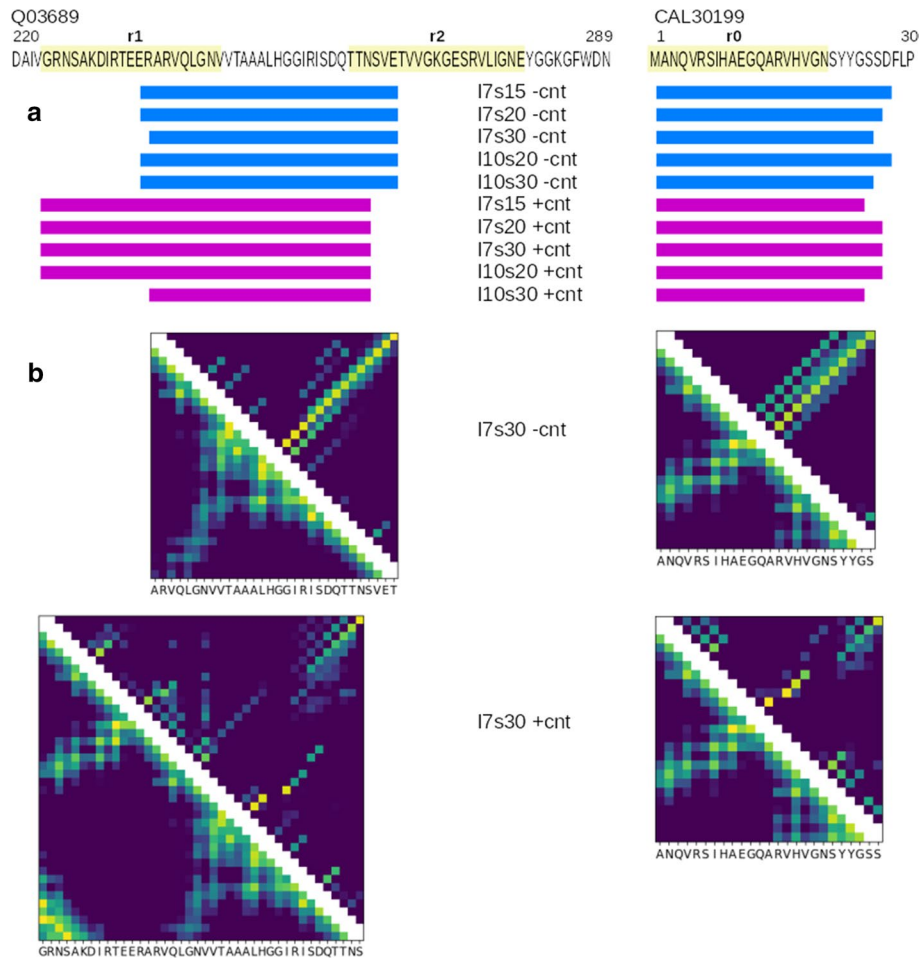


Fig. 4 Qualitative analysis of generalized BASS grammars match to experimentally solved HET-s structure. **a** Best matches in effector-side C-termini and NLR-side N-termini sequences. Actual position of r1, r2 and r0 motifs are marked in pale yellow. **b** Comparison of the structure-derived (pdb:2kj3 [58]) distance map (lower left triangle, color scale from 4 Å or less (yellow) to 16 Å or more (blue) according to the Cβ distance) to grammar-derived pairing probability map (upper right triangle, logarithmic color scale from 0.01 or less (blue) to 1 (yellow))

Pairing potential In order to assess the structure of grammatical descriptors of fungal motifs, we scanned the C-terminal 80 amino-acid fragment of HET effector sequence (accession: Q03689) and the N-terminal 50 amino-acid fragment of its genomic neighbor NLR sequence (accession: CAL30199) using the score averaging approach and the 20-to-40 amino-acids window (Fig. 4a). In CAL30199, the best matches very well covered the r0 motif. In the case of Q03689, the best matches were centered on the loop between the r1 and r2 motifs. Grammars trained without the contact constraints partially overlapped the motifs, while grammar trained with the contact constraints covered the r1 motif using the maximum window size of 40 (except *l*10*s*30). A plausible explanation of discrepancy with the actual motif positions is high content of charged residues in mutually complement r1 and r2 motifs, which is atypical for the BASS motifs. Then, we compared the probability maps of parsing of the best matching sequence fragments with their corresponding spatial distance maps. For the NLR side CAL30199, we used the r2

motif fold, as often assumed in literature [56] (Fig. 4b). The pairing maps generated with grammars trained without the contact constraints were dominated by an apparently artifactual antidiagonal signal resembling the pattern observed previously in mostly likely parse trees [34]. The signal was partially present also on pairing maps generated with grammars trained with the contact constraints. However, in this case, there was also a clear signal corresponding to actual structures of the HET-s fold: the "bulge" from A228 to I231, the tip of the main loop from T233 to V239, and the second loop from Q240 to V244 (positions according to the r1 motif, see Figure 1 in [56] for reference).

Lexical rules Grammars are considered as human readable descriptors. We analyzed grouping of amino acids according to high probability (above 0.1) of being rewritten from given lexical non-terminal symbols. We focused on groupings preserved in at least half of grammars, for each grammar size and the contact constraints option. Almost universally preserved was single non-terminal dedicated overwhelmingly to glycine. Very often grammars consisted of non-terminal symbols dedicated to alanine (sometimes together with serine) and variables likely rewritten to valine and isoleucine (sometimes together with leucine). Next common association was a non-terminal with emissions dominated by glutamine, relatively frequent in the prions. On the other hand, no clear pattern was observed with aspargine and aspartic acid, even though they seem to be relevant for the amyloid signaling motifs. Many grammars included also a non-terminal symbol likely rewritten to a mix of arginine, glutamic acid, lysine, proline, serine, threonine and in some cases histidine, though partition varied. This subset may correspond to a group present in some classical 5-letter alphabets [84, 85] but without amino acids characteristic to the prions. The groupings defined by PCFGs were also partially similar to the best-performing reduced alphabet for amyloid hexa-peptides search from [21], which included groups for glycine alone, isoleucine-leucine-valine hydrophobics, and the lysine-proline-argine mix.

Structural rules For each sequence, the usage of every rule can be recorded, which is the amount of probability mass carried through the rule (as calculated for the Inside–Outside procedure) relative to the overall probability of the sequence given the grammar. If this quantity is summed up for each left-hand-side non-terminal, one obtains usage of each non-terminal symbol. Note that the usage of the start non-terminal is always at least one, and possibly higher if the start symbol is used again in some derivations. We calculated the usage of structural non-terminals in the best matching sequence fragments that achieved positive log probability ratio for the positive test samples (NLR-side BASSes and fungal motifs) and *DisProt*), averaged over the cross-validation folds. For each test set, we identified non-terminal symbols with the usage of at least 1. Then, we compared the test sets in terms of the highly used symbols using the Jaccard distance metric. Since the structural non-terminals represent higher level structures in the grammar (as *noun phrase* in English), similarities in their usage may reflect similarities between motifs. Varying maximum number of non-terminals and the contact constraints option resulted in different sharing of the highly used symbols. As could be expected, smaller grammars resulted in higher shares (50-80%) than large grammars (30–60%). Nevertheless, some clear patterns emerged. BASS1 mostly shared highly used structural non-terminals with BASS9-10 and fungal HET-s, and least with *DisProt*.

BASS2 mostly shared its highly used structural non-terminals with BASS6 and fungal sigma and PP, while least, again, with *DisProt*. Unsurprisingly, BASS3 matched most the PP motif, followed by other fungal motifs. For BASS4 the closest match was HET-s, while for BASS5 it was *DisProt*. Interestingly, the BASS4 motif is relatively often found repeated in the Bell side, as is the case of HET-s. BASS7 motif mostly shared highly used symbols with PP and HET-s, and BASS8—with PP, *DisProt*, sigma, BASS5 and BASS7.

Comparison to alternative approaches

To place our method within the state of the art, we tried identifying sequences with BASS-like motifs using several existing methods. First, we evaluated a diverse yet non-exhaustive selection of ready-made tools devoted to predicting prions, amyloids and beta structures. Importantly, all included tools proved to be relatively easy to employ for scanning large sequence sets:

- **AmyloGram** [21] is a predictor of amyloidogenic hot spots, which is based on the n-gram analysis and the random forest classifier. The tool calculates probability of forming the hot spot over the 6-amino-acid sliding window and returns the maximum value for the input sequence. With the default probability threshold of 0.5, AmyloGram identified amyloidogenic hot spots in 78% of sequences in the BASS test set and in all sequences with fungal ASMs, but also in all negative non-prionic NLR N-termini. Increasing the probability threshold did not improve the outcome.
- **PASTA2** [20] is a popular tool for detecting amyloid structural aggregation, which combines analysis of statistical residue pairing energies with prediction of secondary structures and disordered regions. With the peptide mode settings, PASTA2 reported amyloid-like aggregation regions in 38% of sequences in the BASS test set and in 67% sequences with the fungal ASMs. However, it also found such regions in 91% of the negative non-prionic NLR N-termini set.

  Neither of the tested amyloid prediction methods is therefore suitable for searching the NLR-related amyloid signaling motifs, which is not surprising given it is known that short potentially amyloidogenic regions can be found in proteins never observed to form amyloids [86].
- **ArchCandy** [26] is a method for detecting beta-arches, which is based on quantitative assessment of several sequence features. We used the standalone Java archive executable in version 2.0, kindly provided by the authors. With the recommended score threshold of 0.56, it marked as positive 50% of sequences in the BASS test set, as well as 83% sequences with the fungal ASM motifs. However, it also detected beta-arches in 61% of the negative non-prionic NLR N-termini set. This outcome could also be expected as beta-arches may be present in sequences that neither form amyloid nor act as prions.

  It seemed conceivable that combining ArchCandy and PASTA2 might improve the accuracy of BASS search. For each sequence, we checked if any top 20 amyloidogenic region predicted with PASTA overlapped with any beta-arch predicted with ArchCandy, using the default detection thresholds of both tools. The recall was 24% for

the BASS test set and 41% for *NLReff*, clearly showing that the approach is not viable for searching BASS-like motifs.

- **PrionW** [87] is a web server based on the pWaltz method for detecting prions, which assumes they consist of an amyloidogenic core inside a disordered region, rich in aspargine and glutamine [15]. With default parameters, the tool did not found any prion motif in the BASS and fungal ASM positive test sets. Lowering the pWaltz cut-off to 0.50 resulted in finding 3 out of 54 fungal motifs but also 5 hits in the non-prionic NLR N-termini negative set *NLReff*.
- **PAPA** [24] is a composition-based prion-propensity predictor developed for the asparagine- and glutamine-rich sequences and implemented as a Python script. The window size was set to 20 to cover all motifs and filtering for disordered regions was switched off as this improved accuracy. However, the overall result was poor as only one sigma motif instance could be identified at FPR of $1e{-}3$ against *NLReff*, and 4 BASS instances (including 3 in *BASSother*) at FPR of $1e{-}2$.
  Evidently, both of these prion finders are not sensitive to BASS-like motifs.
- **PLAAC** [25] is a web server for finding sequence with Prion-Like Amino-Acid Composition based on a two-state (prion-like and background) Hidden Markov model [88]. With default parameters except for the minimal contiguous prion-like domain length set to 20, and based on its COREscore, PLAAC identified prions in 3 out 143 BASS, 1 out of 22 fungal PP and 8 out of 20 fungal sigma sequences without any false positive in neither the non-prionic NLR N-termini nor non-amyloidogenic disordered proteins negative sets. In addition, PLAAC returns several quantitative measures for analyzed sequences and by using the raw log-likelihood ratio between the two-state model and the background model calculated over most likely parses (HMMvit) the tool identified around 20% (34%) BASS positive test samples at FPR of $1e{-}3$ ($1e{-}2$) against *NLReff*. For other test sets, optimal results were achieved using the raw log-likelihood ratio between the two-state model and the background model calculated over all parses (HMMall). The sensitivity was 6% (18%) for the BASS test set, 18% (36%) for *BASSother*, 65% (85%) for sigma, 18% (59%) for PP, and 8% (25%) for HET-s, all at FPR of $1e{-}3$ ($1e{-}2$) against *NLReff*. At the same threshold of the log likelihood ratio, the FPR against *DisProt* was 5% (15%).
  Among the tested prion finders, PLAAC proved to be the only method capable of identifying a significant (yet still limited) number of NLR-related amyloid signaling motifs. Interestingly, this required switching off all the heuristics and relying on its core HMM that models the prion-specific amino-acid distribution. Not surprisingly, the highest sensitivity was achieved for the asparagine- and glutamine-rich sigma motif.

Profile hidden Markov models Having tested the ready-made tools, we tried to generalize the amyloid signaling motifs with the profile Hidden Markov Models. The approach required building the multiple sequence alignment for the combined training set comprising of all ten BASS families. It has to be noted that making a single multiple sequence alignment for a diverse collection of rather loosely related motifs requires caution. Here, we evaluated **Clustal Omega** [59] in the *–auto* mode, **Muscle** [89] in the default mode, and **Mafft** [90, 91] in accurate modes (*linsi*, *einsi* and *ginsi*). The modes were chosen arbitrarily based on information displayed by the command-line interface to a regular

user. We first aligned all non-redundant Bell-side motifs from the ten BASS classes using selected tools. The average percent identity between sequences from different BASS classes varied from only 9% for Clustal Omega, through 13% for Muscle in auto/default modes to 18–19% for Mafft in accurate modes. The latter values are comparable with the percent identity of the HET-s set (21%).

To facilitate direct comparison with the PCFG-based method, eventually, MSAs were generated with the aforementioned tools for each of the six training folds separately. Then, the corresponding profile HMMs were trained using **HMMER3** [92] with default parameters. Parsing of the test sequences was performed in the *–max* mode switching all filters off for the sake of accuracy. We found that the sensitivity of the profile HMMs to NLR-side test BASSes was roughly on par with the PCFGs with the recall up to 79% (92%) at FPR of $1e{-}3$ ($1e{-}2$) against *NLReff*. However, the recall of fungal ASMs (averaged over the three classes) was lower, ranging from 17 to 44% (46–67%), in comparison to 38–58% (69–85%) for the PCFGs—all values at FPR of $1e{-}3$ ($1e{-}2$). Expectedly, the accurate MSAs resulted in better performing profile HMMs than the rough MSAs.

In addition, the averaging approach was applied, in which the maximal scores from the profile HMMs trained on the six folds were averaged for each sequence (without checking if the best domain matches overlap). As with the single profiles, the averaged pHMMs performed similarly to averaged PCFGs on the BASS positive test set, up to recall of 92% (97%) versus 90% (99%), and worse on the fungal positive test sets, up to recall of 57% (79%) versus 89% (97%)—all values at FPR of $1e{-}3$ ($1e{-}2$) against the negative set of non-prionic NLR N-termini *NLReff* (Fig. 5). Moreover, up to 41–57% (57–73%) motifs from the *BASSother* positive set were detectable at FPR of $1e{-}3$ ($1e{-}2$). Accuracy against *DisProt* was perfect.

Profile HMMs based on the rough alignment by Muscle were consistently least accurate, while pHMMs based on the accurate Mafft alignments typically performed best. Interestingly, the profiles based on the Clustal Omega auto mode alignment were superior in case of *BASSother* and the fungal sigma motifs.

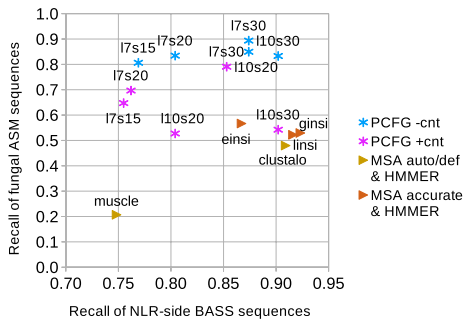


Fig. 5 Performance comparison of generalized BASS grammars and profile hidden Markov models on bacterial and fungal test motifs from NLR proteins. Performance in terms of recall at the false positive rates of $1e{-}3$ obtained using the score averaging scheme. For fungal ASMs, the recall is averaged over the three classes of motifs (sigma, PP, HET-s). Data for PCFGs learned without (with) contact constraints are plotted in blue (magenta). Number of variables (*l*: lexical, *s*: structural) in grammars is indicated. Data for pHMMs learned with rough (accurate) MSAs are plotted in golden (chocolate). Method for generating MSA is indicated ([*egl*]*insi* refer to Mafft modes). See main text for details

Summary Despite that pWaltz was developed using the HET-s experimental structure [15], it was unable to detect more than a few NLR-related ASMs, even with some parameter tweaking. Also PAPA suffered from the lack of sensitivity to BASS-like motifs. ArchCandy, PASTA2 and AmyloGram may perform well in identifying regions of interest or discriminating beta-arches and amyloidogenic regions, respectively, but, by design, they are not specific enough for motif searches in large data sets. PLAAC, or more precisely its core HMM model of the prion composition, identified close to 1/3 of all test motifs at FPR $1e{-}2$. This level of accuracy may be enough, for example, to detect presence of amyloid signaling motifs in taxonomic branches [93], but certainly not for exhaustive searches. Yet, the result is remarkable because of simplicity of the model. Finally, trained profile HMMs performed on par with PCFGs within the modeled meta-class of the ten BASS families. In fact, their accuracy exceeded our expectations given high diversity of the collection making alignment difficult. Extrapolating beyond the ten BASS families, sensitivity of the profile HMMs deteriorated to the level of the least accurate PCFGs for *BASSother* and below for the fungal prions (especially HET-s).

Experimental verification of selected peptides

Eventually, we tested if selected peptides identified as ASMs with grammars and other computational methods (Table 2) form spatial structures consistent with the known HET-s structure [58] and with the experimentally demonstrated amyloid-like aggregation [8, 13, 94–96]. For PCFG and pHMM methods, we report results obtained using the averaging approach.

First, to tweak and test the experimental setup, we used two *bona fide* effector-side ASM peptides: BASS3 RHIM-like motif from *Frankia* sp. ORT49035.1 and *Nectria haematococca* sigma motif from AAS80314.1. Both could be computationally identified using the PCFG and profile HMM methods (at FPR below $1e{-}3$), as well as using AmyloGram [21]. The latter peptide was found prionic using PLAAC and only marginally missed the ArchCandy threshold. Then, we analyzed *Methanothrix soehngenii* AEB69175.1, a NLR-side motif resembling BASS3, which was originally identified through the local pairwise alignment of proteins encoded by neighboring genes, while being missed in searches using profile HMMs of individual BASS1-10 families [13]. In the current study, the motif could be identified with pHMMs at FPR of $1e{-}3$ and with some PCFGs at FPR of around $1e{-}2$; it also scored quite low energy in PASTA2 (yet still above the 95% specificity threshold recommended for peptides). Finally, we experimentally verified a sequence fragment resembling the sigma motif in *Coleophoma crateriformis* NLR protein RDW70414.1 (382 to 421) [68]. The motif is marked as positive according to PCFGs, PLAAC (FPR below $1e{-}3$) and ArchCandy, and as negative according to PASTA2, AmyloGram and some pHMMs ((Table 2).

ATR-FTIR spectra of the four selected peptides (see Materials) in dried form and D_2O solutions with corresponding second derivative spectra in amide bands range (1750–1500 cm^{-1}) are presented in Fig. 6a. Spectral characteristic for peptides ORT49035_103_123, AAS80314_349_385 and AEB69175_5_29 are typical for the aggregates. The Amide I' band maxima are located below 1635 cm^{-1}, what is characteristic for the β-cross structures. While, in ATR-FTIR spectrum of peptide RDW70414_382_421

Table 2 Computational evaluation of peptides selected for experimental verification

Peptide	**Sequence**
ORT49035.1_103_123	VDLRDAKGVQVGDGNVQINRF
AAS80314.1_349_385	SFNNLGSGDQFNTPGGTQNINKGGNEVSGGNFYGSVQF
AEB69175.1_5_29	KSRFDQRGQKVIGQQINVAGDATLP
RDW70414.1_382_421	GAPANNTSNSIQHNNSGSGHQNSGSGQQNIGTNTGSGQQ

Peptide	**PCFG** **FPR↓**	**pHMM** **FPR↓**	**PLAAC** **FPR↓**	**PASTA2** **[PEU]↓**	**AmyloG** **[0-1]↑**	**ArchC** **[0-1]↑**
ORT49035.1_103_123	≤ **0.0004**	< **0.0004**	0.022	− 2.5	**0.63**	0.379
AAS80314.1_349_385	< **0.0004**	< **0.0004**	< **0.0004**	− 1.6	**0.78**	0.555
AEB69175.1_5_29	**0.007**−0.07	< **0.0004**	0.022	− 3.7	0.43	0.488
RDW70414.1_382_421	< **0.0004**	0.01−0.35	<**0.0004**	− 0.9	0.26	**0.599**

Top: List of peptides selected for experimental verification. **Bottom**: results of computational methods. For PCFG, pHMM (both with score averaging) and PLAAC we provide FPR against *NLReff* at which at least a part of the peptide is detected; for PASTA2—Pasta Energy Units; for AmyloGram (AmyloG) and ArchCandy (ArchC)—score ranged from 0 to 1. Arrow indicates if higher or lower value is more positive. Values exceeding the default or suggested thresholds are shown in bold (the thresholds are $1e-2$ for PCFG, pHMM and PLAAC, -5 PEU for PASTA2, 0.5 for AmyloGram, and 0.56 for ArchCandy). For PCFG, we provide the range of values for grammars of different maximum number of symbols; For pHMM—the range of values for profiles made from different accurate MSAs

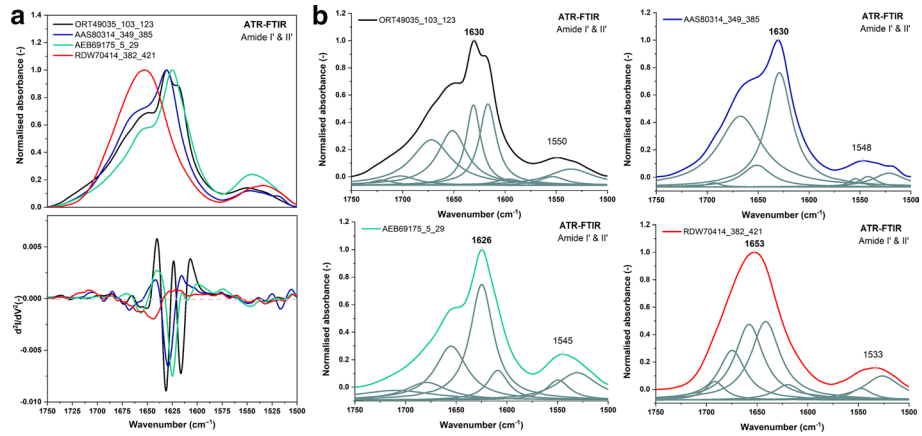


Fig. 6 ATR-FTIR spectroscopy analysis of selected peptides. **a** Normalized ATR-FTIR spectra of air-dried peptides films with second derivative spectra in the amide bands region (1750–1500 cm^{-1}). **b** Particular ATR-FTIR spectra with sub-bands obtained from the curve fitting procedure (Amide I' and II' regions). Color code: ORT49035_103_123—black line, AAS80314_349_385—blue line, AEB69175_5_29—green line, and peptide RDW70414_382_421—red line

no spectral signatures of the aggregation process were found. More accurate information about studied peptides can be obtained from the second derivative and decomposition of the Amide I' band into sub-bands (Fig. 6b). These processes clearly revealed the complex structure of peptide ORT49035_103_123. The Amide I' band can be separated into five components: 1702, 1672, 1652, 1631 and 1617 cm^{-1}, which can be assigned in order to β-sheet or turn, β-turn, α-helix or extended random or loops, β-sheet and aggregates [69, 97]. The origin of component 1652 cm^{-1} is not clear, because the loop and the helical absorption bands overlap in this region [98]. For the other peptides, less components in the Amide I' range were observed, but their assignment is similar.

In the Congo Red staining experiment, the amyloid aggregates were detected for ORT49035_103_123, AAS80314_349_385 and AEB69175_5_29 peptides (Additional file 4: Figure S2). In the case of peptide RDW70414_382_421, the fixation of CR was not observed.

Taken together, the experimental results are compatible with presence of the beta-arch structure and amyloid-like aggregation of peptides ORT49035_103_123, AAS80314_349_385 and AEB69175_5_29. This supports the hypothesis that bacterial, archaeal and fungal NLR-related ASMs share similar structural features. On the other hand, no sign of the amyloid like aggregation was observed for RDW70414_382_421. However, since decomposition of the Amide I' band for monomeric RDW70414_382_421 reveals similar components to other three peptides, it is likely that it assumes the beta-arch structure as well. Thus, this peptide apparently represents a false positive hit of some computational methods, presumably due to over-generalization.

Experimental conditions for ATR-FTIR spectroscopy and CR staining are reported in Additional file 5: Table S2 according to the MIRRAGGE standard [77].

Discussion

In this piece of research, we addressed some of the previously identified challenges in inferring probabilistic context-free grammars for protein motifs [34]. Overall, presented results show a clear advantage of the Inside–Outside training procedure followed with the lexical probabilities smoothing over the previously used evolutionary scheme [31, 34] in learning of the probabilistic context-free grammars for protein motifs. Current procedure allows for generating much larger grammars with sufficient numbers of non-terminal symbols (e.g. 40). Importantly, the inference time with the IO training is relatively short, ranging from a couple of minutes to a couple of hours, depending on the covering grammar size (on 12 cores of the Intel Xeon E5 (Haswell) machines). In accordance with the literature [36], the obtained grammatical models are typically not globally optimal. In practice, we addressed this by combining (averaging) scores of several individual grammars. The averaging scheme turned out to be very effective, usually outperforming best individual grammars in homology searches (Fig. 3a). Yet, it remains a goal for the future research to optimize learning, e.g. by enhancing the statistical learning with the contrastive estimation [99] whenever possible, and by combining with heuristic approaches [46, 100].

Recently, we introduced the use of the contact constraints based on known or predicted spatial proximity of residues to learning PCFGs [34]. Due to the context-freeness of grammars and simplicity of the contact rules in the Chomsky Form with Contacts, only a subset of residue-residue contacts can be used in the training. The contacts are properly chosen if there exist correlations between involved amino acid species that are relevant to modeled structures and functions. In such a case the contact constraints facilitate learning through confining the search space towards the most capable solutions. Obviously, the constraints reduce the amount of information that grammars can learn: correlations incompatible with the constraints cannot be captured in the grammar. So, if the constraints are not properly chosen, they may effectively lead to less capable models. This is not so much a problem when modeling samples of highly homologous sequences where the common signal shared by all sequences is very strong. However, in the case of generalizing over multiple motif families, a suboptimal choice of the constraints is likely to hamper the quality of the model more harshly. This could contribute to weaker performance of the generalizing grammars trained with the contact constraints when tested on *BASSother* and some fungal motif samples (Fig. 3b). Nevertheless, using the contact constraints for training apparently improved compatibility of the residue pairings, generated with grammar using the contact rules, with the actual protein distance map for experimentally solved structure of a signaling amyloid, the HET-s motif (Fig. 4b).

One obvious limitation of the PCFG approach is context-freeness: the property that allows a grammar for considering in a single derivation only the non-overlapping nested and branching dependencies. Yet, the probabilistic parsing of a sequence consists on scoring over all possible derivations, therefore it is capable of covering several overlapping sets of nested (anti-parallel) and branching dependencies. The PCFG model is, however, not suitable for capturing crossing (parallel) dependencies. This is a serious limitation in the context of modeling protein sequences. Unfortunately, more expressive grammar formalisms are also more computationally expensive. For example, computational time complexity for parsing of the mildly context-sensitive linear indexed

grammars [101, 102], which can capture some crossing structures, is $\mathcal{O}(n^6)$. An alternative consists on methods building models from the multiple sequence alignment, such as undirected graphical models or Potts models, which can capture information conveyed in the crossing inter-position correlations [103, 104]. There is ongoing research on using such models for aligning sequences in homology searches, while avoiding the combinatory explosion [105–109]. Interestingly, one Potts-based tool under development has been hitherto outperformed by a PCFG-based tool in RNA homology searches [109]. Even if further development leads to unleashing the full power of Potts-based models, the requirement of MSA for inferring their parameters makes them less suitable for modeling meta-families of motifs whose members do not share relevant homology, yet still share structural or functional principles.

Conclusions

The results obtained in this piece of research show that the proposed method can infer a model capable of generalizing over a diverse set of families of amyloid signaling motifs. While the profile HMMs remain the method of choice for modeling homologous sets of sequences, PCFGs seem more suitable for building meta-family descriptors with the goal of extrapolating beyond the seed sample. (Even if the generalization comes at some price as exposed by the experimentally verified false positive hit.) Indeed, with the score averaging scheme, PCFGs trained without contact constraints outperformed profile HMMs when the BASS-trained models were applied to fungal motifs (sensitivity of 89% vs 57% at FPR of $1e{-}3$). In practice, one can expect even higher specificity of both machine learning methods, since in reporting the results we assumed the conservative upper estimate of the false positive rate when all positive samples scored above every negative sample.

Abbreviations

AP: Average precision; ASM: Amyloid signaling motif; ATR-FTIR: Attenuated total reflection-Fourier transform infrared; BASS: Bacterial amyloid signaling sequence; BLOSUM: Blocks substitution matrix; CFC: Chomsky form with contacts; CNF: Chomsky normal form; CR: Congo red; DIC: Dissolved inorganic carbon; DMF: Dimethylformamide; Fmoc: 9-Fluorenylmethoxycarbonyl; FPR: False positive rate; GA: Genetic algorithm; HMM: Hidden Markov model; HPLC: High-performance liquid chromatography; IO: Inside–outside algorithm; MS: Mass spectrometer; MSA: Multiple sequence alignment; NBS-LRR: Nucleotide-binding site-leucine-rich repeats; NLR: An umbrella term for NOD-like receptor and NBS-LRR; NMR: Nuclear magnetic resonance; NOD: Nucleotide-oligomerization domain; PAM: Point accepted mutation; PCFG: Probabilistic context-free grammar; PDB: Protein Data Bank; PEU: PASTA energy unit; pHMM: Profile HMM; RNA: Ribonucleic acid; TFA: Trifluoroacetic acid; TIS: Triisopropylsilane; TOF: Time of flight; YI: Youden's index.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04139-y.

Additional file 1. Datasets

Additional file 2. Figure S1: The outline of the processing pipeline.

Additional file 3. Table S1: Analytical data for peptide synthesis and purification.

Additional file 4. Figure S2: Representative light microscope images of the peptides stained with CR.

Additional file 5. Table S2: Experimental conditions for the ATR-FTIR spectroscopy and CR staining.

Authors' contributions

WD conceived the study, designed, performed and analyzed the computational experiments, prepared figures and tables, authored the manuscript, developed the software. MG-G designed, performed and analyzed the spectroscopy experiments, prepared figures, authored the manuscript. MS designed and performed the peptide synthesis, prepared tables, authored the manuscript. NS performed and analyzed the staining experiments, prepared figures and tables, authored the manuscript. All authors read and approved the final manuscript.

Funding
This research has been funded by National Science Centre, Poland (ncn.gov.pl), grants no. 2015/17/D/ST6/04054 (WD), 2017/26/D/ST5/00341 (MS) and 2019/35/B/NZ2/03997 (WD, MG-G), by National Centre for Research and Development, Poland (ncbr.gov.pl), project no. POWR.03.02.00-00-I003/16 (NS), and by Politechnika Wrocławska statutory funds, and supported by Wroclaw Centre for Networking and Supercomputing (wcss.pl) grant 98 and the E-SCIENCE.PL infrastructure (e-science.pl). The funders had no role in the design of the study, collection, analysis, interpretation of the data, or in writing the manuscript.

Data availability
The source code is available at git.e-science.pl/wdyrka/pcfg-cm under the GNU General Public License v3.0. The datasets used in the current study (Datasets), the peptide analytical data (Additional file 3: Table S1) and the experimental conditions report for the ATR-FTIR spectroscopy and CR staining (Additional file 5: Table S2) are included as supplementary information files.

Declaration

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Wydział Podstawowych Problemów Techniki, Katedra Inżynierii Biomedycznej, Politechnika Wrocławska, Wrocław, Poland. [2]Wydział Chemiczny, Katedra Chemii Bioorganicznej, Politechnika Wrocławska, Wrocław, Poland.

Received: 9 October 2020 Accepted: 19 April 2021
Published online: 29 April 2021

References

1. Eichner T, Radford SE. A diversity of assembly mechanisms of a generic amyloid fold. Mol Cell. 2011;43(1):8–18.
2. Riek R, Eisenberg D. The activities of amyloids from a structural perspective. Nature. 2016;539:227–35.
3. Saupe SJ. Amyloid signaling in filamentous fungi and bacteria. Annu Rev Microbiol. 2020;74(1):673–91.
4. López de la Paz M, Serrano L. Sequence determinants of amyloid fibril formation. Proc Natl Acad Sci. 2004;101(1):87–92.
5. Chen D, Drombosky KW, Hou Z, Sari L, Kashmer OM, Ryder BD, Perez VA, Woodard DR, Lin MM, Diamond MI, Joachimiak LA. Tau local structure shields an amyloid-forming motif and controls aggregation propensity. Nat Commun. 2019;10(1):2493.
6. Coustou V, Deleu C, Saupe S, Begueret J. The protein product of the het-s heterokaryon incompatibility gene of the fungus podospora anserina behaves as a prion analog. Proc Natl Acad Sci. 1997;94(18):9773–8.
7. Maddelein M-L, Dos Reis S, Duvezin-Caubet S, Coulary-Salin B, Saupe SJ. Amyloid aggregates of the het-s prion protein are infectious. Proc Natl Acad Sci. 2002;99(11):7402–7.
8. Balguerie A, Dos Reis S, Ritter C, Chaignepain S, Coulary-Salin B, Forge V, Bathany K, Lascu I, Schmitter JM, Riek R, Saupe SJ. Domain organization and structure-function relationship of the het-s prion protein of podospora anserina. EMBO J. 2003;22(9):2071–81.
9. Daskalov A, Habenstein B, Martinez D, Debets AJ, Sabate R, Loquet A, Saupe SJ. Signal transduction by a fungal NOD-like receptor based on propagation of a prion amyloid fold. PLoS Biol. 2015;13(2):1002059.
10. Sun X, Yin J, Starovasnik MA, Fairbrother WJ, Dixit VM. Identification of a novel homotypic interaction motif required for the phosphorylation of receptor-interacting protein (rip) by rip3. J Biol Chem. 2002;277(11):9505–11.
11. Kleino A, Ramia NF, Bozkurt G, Shen Y, Nailwal H, Huang J, Napetschnig J, Gangloff M, Chan FK-M, Wu H, Li J, Silverman N. Peptidoglycan-sensing receptors trigger the formation of functional amyloids of the adaptor protein imd to initiate drosophila nf-kb signaling. Immunity. 2017;47(4):635–6476.
12. Daskalov A, Paoletti M, Ness F, Saupe SJ. Genomic clustering and homology between het-s and the nwd2 stand protein in various fungal genomes. PLoS ONE. 2012;7(4):34854.
13. Dyrka W, Coustou V, Daskalov A, Lends A, Bardin T, Berbon M, Kauffmann B, Blancard C, Salin B, Loquet A, Saupe SJ. Identification of nlr-associated amyloid signaling motifs in bacterial genomes. J Mol Biol. 2020;432:6005–27.
14. Kajava AV, Klopffleisch K, Chen S, Hofmann K. Evolutionary link between metazoan RHIM motif and prion-forming domain of fungal heterokaryon incompatibility factor HET-s/HET-s. Sci Rep. 2014;4(1):1–6.
15. Sabate R, Rousseau F, Schymkowitz J, Ventura S. What makes a protein sequence a prion? PLoS Comput Biol. 2015;11(1):1–9.
16. Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis. Probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press; 1998.
17. Eddy SR. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol. 2008;4(5):1000069.
18. Bryan AW Jr, Menke M, Cowen LJ, Lindquist SL, Berger B. Betascan: probable beta-amyloids identified by pairwise probabilistic analysis. PLoS Comput Biol. 2009;5(3):1–11.
19. Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. Bioinformatics. 2009;26(3):326–32.
20. Walsh I, Seno F, Tosatto SCE, Trovato A. PASTA 2.0: an improved server for protein aggregation prediction. Nucleic Acids Res. 2014;42(W1):301–7.
21. Burdukiewicz M, Sobczyk P, Rödiger S, Duda-Madej A, Mackiewicz P, Kotulska M. Amyloidogenic motifs revealed by n-gram analysis. Sci Rep. 2017;7(1):1–10.

22. Wojciechowski JW, Kotulska M. Path-prediction of amyloidogenicity by threading and machine learning. Sci Rep. 2020;10(1):7721.
23. Wozniak PP, Kotulska M. Amyload: website dedicated to amyloidogenic protein fragments. Bioinformatics. 2015;31(20):3395.
24. Toombs JA, Petri M, Paul KR, Kan GY, Ben-Hur A, Ross ED. De novo design of synthetic prion domains. Proc Natl Acad Sci. 2012;109(17):6519–24.
25. Lancaster AK, Nutter-Upham A, Lindquist S, King OD. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. Bioinformatics. 2014;30(17):2501–2.
26. Ahmed AB, Znassi N, Château M-T, Kajava AV. A structure-based approach to predict predisposition to amyloidosis. Alzheimer's & Dementia. 2015;11(6):681–90.
27. Booth TL. Probabilistic representation of formal languages. In: 10th annual symposium on switching and automata theory (swat 1969); 1969. p. 74–81.
28. Sakakibara Y, Brown M, Underwood RC, Mian IS. Stochastic context-free grammars for modeling RNA. In: 27th Hawaii international conference on system sciences; 1993. p. 349–58.
29. Eddy SR, Durbin R. RNA sequence analysis using covariance models. Nucleic Acids Res. 1994;22(11):2079–88.
30. Knudsen B, Hein J. Rna secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics. 1999;15:446–54.
31. Dyrka W, Nebel J-C. A stochastic context free grammar based framework for analysis of protein sequences. BMC Bioinform. 2009;10:323.
32. Sciacca E, Spinella S, Ienco D, Giannini P. Annotated stochastic context free grammars for analysis and synthesis of proteins. In: Pizzuti C, Ritchie M, Giacobini M, editors. Evolutionary computation, machine learning and data mining in bioinformatics. Lecture notes in computer science, vol. 6623. Berlin: Springer; 2011. p. 77–88.
33. Waldispuehl J, Berger B, Clote P, Steyaert J-M. Predicting transmembrane beta-barrels and interstrand residue interactions from sequence. Proteins Struct Funct Genet. 2006;65(1):61–74.
34. Dyrka W, Pyzik M, Coste F, Talibart H. Estimating probabilistic context-free grammars for proteins using contact map constraints. PeerJ. 2019;7:6559.
35. Lari K, Young SJ. The estimation of stochastic context-free grammars using the inside–outside algorithm. Comput Speech Lang. 1990;4(1):35.
36. Keller B, Lutz R. Evolutionary induction of stochastic context free grammars. Pattern Recognit. 2005;38(9):1393–406.
37. Chomsky N. On certain formal properties of grammars. Inf Control. 1959;2(2):137–67.
38. Pyzik M, Coste F, Dyrka W. How to measure the topological quality of protein parse trees? In: Unold O, Dyrka W, Wieczorek W, editors. Proceedings of the fourteenth international conference on grammatical inference. Proceedings of machine learning research, vol. 3; 2019. p. 118–38.
39. Dowell RD, Eddy SR. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinform. 2004;5(1):71.
40. Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res. 2003;31(13):3423–8.
41. Baker JK. Trainable grammars for speech recognition. In: Klatt D, Wolf J, editors. Speech communication papers for the 97th meeting of the Acoustical Society of America; 1979. p. 547–50.
42. Carroll G, Charniak E. Two experiments on learning probabilistic dependency grammars from corpora. In: The workshop on statistically-based natural language programming techniques. The Menlo Park, CA: AAAI Press; 1992. p. 1–13.
43. Tariman K. Genetic algorithms for stochastic context-free grammar parameter estimation. Master's thesis, The University of Georgia, United States; 2004.
44. Kammeyer TE, Belew RK. Stochastic context-free grammar induction with a genetic algorithm using local search. In: Foundations of genetic algorithms IV. San Francisco, CA: Morgan Kaufmann; 1996. p. 3–5.
45. Keller B, Lutz R. Learning scfgs from corpora by a genetic algorithm. In: Artificial neural nets and genetic algorithms. Vienna: Springer; 1998. p. 210–4.
46. Unold O, Gabor M, Wieczorek W. Unsupervised statistical learning of context-free grammar. In: Proceedings of the 12th international conference on agents and artificial intelligence—volume 1: NLPinAI. Setúbal: SciTePress; 2020. p. 431–8.
47. Charniak E. Tree-bank grammars. Technical report CS-96-02, Brown University, Department of Computer Science; 1996.
48. Carrasco RC, Oncina J, Calera-Rubio J. Stochastic inference of regular tree languages. Mach Learn. 2001;44(1):185–97.
49. Cohen SB, Stratos K, Collins M, Foster DP, Ungar L. Spectral learning of latent-variable PCFGs: algorithms and sample complexity. J Mach Learn Res. 2014;15:2399–449.
50. Pereira F, Schabes Y. Inside–outside reestimation from partially bracketed corpora. In: Proceedings of the 30th annual meeting on Association for Computational Linguistics. ACL '92. Stroudsburg, PA: Association for Computational Linguistics; 1992. p. 128–135.
51. Knudsen M. Stochastic context-free grammars and RNA secondary structure prediction. Master's thesis, Aarhus University, Denmark; 2005.
52. Sharon N, Lis H. Legume lectins-a large family of homologous proteins. FASEB J. 1990;4(14):3198–208.
53. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. New and continuing developments at prosite. Nucleic Acids Res. 2013;41(D1):344–7.
54. de Oliveira TM, Delatorre P, da Rocha BAM, de Souza EP, Nascimento KS, Bezerra GA, Moura TR, Benevides RG, Bezerra EHS, Moreno FBMB, Freire VN, de Azevedo WF, Cavada BS. Crystal structure of dioclea rostrata lectin: insights into understanding the ph-dependent dimer-tetramer equilibrium and the structural basis for carbohydrate recognition in diocleinae lectins. J Struct Biol. 2008;164(2):177–82.

55. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2.
56. Daskalov A, Dyrka W, Saupe SJ. Theme and variations: evolutionary diversification of the HET-s functional amyloid motif. Sci Rep. 2015;5:12494.
57. Seuring C, Greenwald J, Wasmer C, Wepf R, Saupe SJ, Meier BH, Riek R. The mechanism of toxicity in HET-S/HET-s prion incompatibility. PLoS Biol. 2012;10(12):1001451.
58. van Melckebeke H, Wasmer C, Lange A, AB E, Loquet A, Böckmann A, Meier BH. Atomic-resolution three-dimensional structure of het-s(218–289) amyloid fibrils by solid-state nmr spectroscopy. J Am Chem Soc. 2010;132(39):13765–75.
59. Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol Syst Biol. 2011;7:539.
60. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife. 2014;3:02030.
61. Daskalov A, Dyrka W, Saupe SJ. NLR function in fungi as revealed by the study of self/non-self recognition systems. In: Benz JP, editor. Genetics and biotechnology. 3rd ed. Cham: The Mycota. Springer; 2020.
62. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TT, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acid Res. 2000;28:235–42.
63. Singh V, Shi W, Almo SC, Evans GB, Furneaux RH, Tyler PC, Painter GF, Lenz DH, Mee S, Zheng R, Schramm VL. Structure and inhibition of a quorum sensing target from streptococcus pneumoniae. Biochemistry. 2006;45(43):12929–41.
64. Hatos A, Hajdu-Soltész B, Monzon AM, Palopoli N, Álvarez L, Aykac-Fas B, Bassot C, Benítez GI, Bevilacqua M, Chasapi A, Chemes L, Davey NE, Davidović R, Dunker AK, Elofsson A, Gobeill J, Foutel NÁSG, Sudha G, Guharoy M, Horvath T, Iglesias V, Kajava AV, Kovacs OP, Lamb J, Lambrughi M, Lazar T, Leclercq JY, Leonardi E, Macedo-Ribeiro S, Macossay-Castillo M, Maiani E, Manso JA, Marino-Buslje C, Martínez-Pérez E, Mészáros B, Mičetić I, Minervini G, Murvai N, Necci M, Ouzounis CA, Pajkos M, Paladin L, Pancsa R, Papaleo E, Parisi G, Pasche E, Barbosa Pereira PJ, Promponas VJ, Pujols J, Quaglia F, Ruch P, Salvatore M, Schad E, Szabo B, Szaniszló T, Tamana S, Tantos A, Veljkovic N, Ventura S, Vranken W, Dosztányi Z, Tompa P, Tosatto SCE, Piovesan D. DisProt: intrinsic protein disorder annotation in 2020. Nucleic Acids Res. 2019;48(D1):269–76.
65. Pesce C, Swanson E, Simpson S, Morris K, Thomas WK, Tisa LS, Sellstedt A. Draft genome sequence of the symbiotic frankia sp. strain kb5 isolated from root nodules of casuarina equisetifolia. J Genom. 2017;5:64–7.
66. Graziani S, Silar P, Daboussi M. Bistability and hysteresis of the "secteur" differentiation are controlled by a two-gene locus in nectria haematococca. BMC Biol. 2004;2:18.
67. Barber RD, Zhang L, Harnack M, Olson MV, Kaul R, Ingram-Smith C, Smith KS. Complete genome sequence of methanosaeta concilii, a specialist in aceticlastic methanogenesis. J Bacteriol. 2011;193(14):3668–9.
68. Wingfield BD, Bills GF, Dong Y, Huang W, Nel WJ, Swalarsk-Parry BS, Vaghefi N, Wilken PM, An Z, de Beer ZW, De Vos L, Chen L, Duong TA, Gao Y, Hammerbacher A, Kikkert JR, Li Y, Li H, Li QK, Liu X, Ma X, Naidoo K, Pethybridge SJ, Sun J, Steenkamp ET, van der Nest MA, van Wyk S, Wingfield MJ, Xiong C, Yue Q, Zhang X. Ima genome-f 9: Draft genome sequence of annulohypoxylon stygium, aspergillus mulundensis, berkeleyomyces basicola (syn. thielaviopsis basicola), ceratocystis smalleyi, two cercospora beticola strains, coleophoma cylindrospora, fusarium fracticaudum, phialophora cf. hyalina, and morchella septimelata. IMA Fungus. 2018;9(1):199–223.
69. Yang H, Yang S, Kong J, Dong A, Yu S. Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy. Nat Protoc. 2015;10(3):382–96.
70. Corujo MP, Sklepari M, Ang DL, Millichip M, Reason A, Goodchild SC, Rodger A. Infrared absorbance spectroscopy of aqueous proteins: comparison of transmission and ATR data collection and analysis for secondary structure fitting. Chirality. 2018;30(8):957–65.
71. Sarroukh R, Goormaghtigh E, Ruysschaert J-M, Raussens V. Atr-ftir: a "rejuvenated" tool to investigate amyloid proteins. Biochim Biophys Acta. 1828;10:2328–38.
72. Ruggeri FS, Longo G, Faggiano S, Lipiec E, Pastore A, Dietler G. Infrared nanospectroscopy characterization of oligomeric and fibrillar aggregates during amyloid formation. Nat Commun. 2015;6:7831.
73. Ruysschaert JM, Raussens V. ATR-FTIR analysis of amyloid proteins. Methods Mol Biol. 2018;1777:69–81.
74. Goldberg ME, Chaffotte AF. Undistorted structural analysis of soluble proteins by attenuated total reflectance infrared spectroscopy. Protein Sci. 2005;14:2781–92.
75. Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. Anal Chem. 1964;36:1627–39.
76. Yakupova EI, Vikhlyantsev IM, Bobyle AG. Congo red and amyloids: history and relationship. Biosci Rep. 2019;39(1):20181415.
77. Martins PM, Navarro S, Silva A, Pinto MF, Sárkány Z, Figueiredo F, Pereira PJB, Pinheiro F, Bednarikova Z, Burdukiewicz M, Galzitskaya OV, Gazova Z, Gomes CM, Pastore A, Serpell LC, Skrabana R, Smirnovas V, Ziaunys M, Otzen DE, Ventura S, Macedo-Ribeiro S. Mirragge—minimum information required for reproducible aggregation experiments. Front Mol Neurosci. 2020;13:222.
78. Azriel R, Gazit E. Analysis of the minimal amyloid-forming fragment of the islet amyloid polypeptide. J Biol Chem. 2001;276:34156–61.
79. Kowalski R. Maszynowe Uczenie Gramatycznych Deskryptorów Sekwencji Białkowych. Engineer's thesis
80. Kowalski R, Pyzik M, Dyrk, W. Towards improved evolutionary learning of probabilistic context-free grammars for protein sequences. In: Mora AM, Esparcia-Alcázar AI, editors. Late-breaking abstracts of EVO* 2019, vol. 1907.12698, arXiv, Ithaca, New York; 2019. p. 10–1.
81. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. Atlas Protein Seq Struct. 1978;5:345–52.
82. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci. 1992;89(22):10915–9.

83. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3:32–5.
84. Solis AD, Rackovsky S. Optimized representations and maximal information in proteins. Proteins. 2000;38(2):149–64.
85. Kosiol C, Goldman N, H. Buttimore N. A new criterion and method for amino acid classification. J Theor Biol. 2004;228(1):97–106.
86. Kotulska M, Unold O. On the amyloid datasets used for training PAFIG—how (not) to extend the experimental dataset of hexapeptides. BMC Bioinform. 2013;14:351.
87. Zambrano R, Conchillo-Sole O, Iglesias V, Illa R, Rousseau F, Schymkowitz J, Sabate R, Daura X, Ventura S. PrionW: a server to identify proteins containing glutamine/asparagine rich prion-like domains and their amyloid cores. Nucleic Acids Res. 2015;43(W1):331–7.
88. Alberti S, Halfmann R, King O, Kapila A, Lindquist S. A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. Cell. 2009;137(1):146–58.
89. Edgar RC. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
90. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30(14):3059–66.
91. Katoh K, Standley DM. Mafft multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
92. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7(10):1002195.
93. Zajkowski T, Lee MD, Mondal SS, Carbajal A, Dec R, Brennock PD, Piast RW, Snyder JE, Bense NB, Dzwolak W, Jarosz DF, Rothschild LJ. The hunt for ancient prions: archaeal prion-like domains form amyloid-based epigenetic elements. Mol Biol Evol. 2021 (**in press**).
94. Sabaté R, Baxa U, Benkemoun L, Sánchez de Groot N, Coulary-Salin B, Maddelein ML, Malato L, Ventura S, Steven AC, Saupe SJ. Prion and non-prion amyloids of the HET-s prion forming domain (2007).
95. Li J, McQuade T, Siemer AB, Napetschnig J, Moriwaki K, Hsiao YS, Damko E, Moquin D, Walz T, McDermott A, Chan FK, Wu H. The RIP1/RIP3] necrosome forms a functional amyloid signaling complex required for programmed necrosis. Cell. 2012;150(2):339–50.
96. Daskalov A, Habenstein B, Sabaté R, Berbon M, Martinez D, Chaignepain S, Coulary-Salin B, Hofmann K, Loquet A, Saupe SJ. Identification of a novel cell death-inducing domain reveals that fungal amyloid-controlled programmed cell death is related to necroptosis. Proc Natl Acad Sci USA. 2016;113(10):2720–5.
97. Khurana R, Fink AL. Do parallel β-helix proteins have a unique Fourier transform infrared spectrum? Biophys J . 2000;78(2):994–1000.
98. Ye M, Zhang Q-L, Li H, Weng Y-X, Wang W-C, Qiu X-G. Infrared spectroscopic discrimination between the loop and α-helices and determination of the loop diffusion kinetics by temperature-jump time-resolved infrared spectroscopy for cytochrome c. Biophys J . 2007;93(8):2756–66.
99. Smith NA, Eisner J. Guiding unsupervised grammar induction using contrastive estimation. In: IJCAI workshop on grammatical inference applications; 2005. p. 73–8.
100. Unold O, Gabor M, Dyrka W. Unsupervised grammar induction for revealing the internal structure of protein sequence motifs. In: Michalowski M, Moskovitch R, editors. Artificial intelligence in medicine–18th international conference on artificial intelligence in medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings, vol. 12299. Lecture notes in computer science. Cham: Springer; 2020. p. 299–309.
101. Gazdar G. Applicability of indexed grammars to natural languages. In: Reyle U, Rohrer C, editors. Nat Lang Parsing and Linguist Theor. Dordrecht: Reidel; 1988. p. 69–94.
102. Weir DJ. A geometric hierarchy beyond context-free languages. Theor Comput Sci. 1992;104(2):235–61.
103. Weigt M, White R, Szurmant H, Hoch J, Hwa T. Identification of direct residue contacts in protein–protein interaction by message passing. Proc Natl Acad Sci. 2009;106:67–72.
104. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CP, Springer M, Sander C, Marks DS. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017;35:128.
105. Lathrop RH. The protein threading problem with sequence amino acid interaction preferences is np-complete. Protein Eng Des Sel. 1994;7(9):1059–68.
106. Talibart H, Coste F. Using residues coevolution to search for protein homologs through alignment of Potts models. In: CECAM 2019—workshop on co-evolutionary methods for the prediction and design of protein structure and interactions; 2019.
107. Muntoni AP, Pagnani A, Weigt M, Zamponi F. Using direct coupling analysis for the protein sequences alignment problem. In: CECAM 2019—workshop on co-evolutionary methods for the prediction and design of protein structure and interactions; 2019.
108. Muntoni AP, Pagnani A, Weigt M, Zamponi F. Aligning biological sequences by exploiting residue conservation and coevolution; 2020. arXiv:2005.08500
109. Wilburn GW, Eddy SR. Remote homology search with hidden Potts models. PLoS Comput Biol. 2020;16(11):1–22.

Publisher's note