Exploring the sequence determinants of amyloid structure using position-specific scoring matrices

Sebastian Maurer-Stroh[1,2], Maja Debulpaep[1], Nico Kuemmerer[3], Manuela Lopez de la Paz[3], Ivo Cristiano Martins[1], Joke Reumers[1], Kyle L Morris[4], Alastair Copland[4], Louise Serpell[4], Luis Serrano[3,5], Joost W H Schymkowitz[1] & Frederic Rousseau[1]

Protein aggregation results in β-sheet–like assemblies that adopt either a variety of amorphous morphologies or ordered amyloid-like structures. These differences in structure also reflect biological differences; amyloid and amorphous β-sheet aggregates have different chaperone affinities, accumulate in different cellular locations and are degraded by different mechanisms. Further, amyloid function depends entirely on a high intrinsic degree of order. Here we experimentally explored the sequence space of amyloid hexapeptides and used the derived data to build Waltz, a web-based tool that uses a position-specific scoring matrix to determine amyloid-forming sequences. Waltz allows users to identify and better distinguish between amyloid sequences and amorphous β-sheet aggregates and allowed us to identify amyloid-forming regions in functional amyloids.

The most common mechanism by which proteins aggregate consists of the incorporation of relatively short sequence segments into β-sheet–like assemblies[1,2]. Although enrichment in β-sheet structure is almost always observed[3,4], aggregates display very disparate macroscopic structures[5]. Most proteins form amorphous aggregates that are characterized by a lack of regular three-dimensional structure. Under physiologically relevant conditions, however, several proteins aggregate into fibrillar amyloids[6]. As with all self-assembly processes, these differences in macromolecular structure reflect differences in packing and molecular structure at the atomic level[7]. These structural differences also reflect biological differences. Whereas misfolded and aggregated proteins are found in perinuclear locations and are generally degraded by the proteasomal system[8,9], amyloids preferentially accumulate in perivacuolar inclusions, where they are degraded by the autophagosome[10]. This segregation directly results from a differential recognition by the protein quality control system[10]. A reduced affinity of amyloids for the protein quality control system is probably also at the root of their higher toxicity[11] as artificial overexpression of chaperones usually leads to decreased toxicity and removal of amyloids from the cell[12]. Contrary to amorphous aggregates, amyloid structures can fulfill biological functions, and functional amyloids are found in organisms from prokaryotes to humans[13].

The amyloid conformation reflects an intrinsic conformational propensity of polypeptides as many proteins can be forced into amyloid fibrils by manipulating external conditions. However, amyloid formation is also a sequence-specific process[14,15]. Under physiological conditions, most peptide sequences derived from proteins will remain soluble even at high concentrations, whereas most hydrophobic sequences will invariably aggregate as amorphous aggregates. The first steps toward understanding the sequence-specific determinants of amyloid structure had been made by analyzing the first microcrystal structures of peptides from amyloid peptides[16,17] and with the observation that protein aggregation correlates with simple biophysical parameters[18]. This triggered the development of several algorithms for predicting aggregation[19–27]. However, owing to the relatively small amount and diversity of experimentally validated amyloidogenic sequences, sequence-based prediction algorithms do not possess sufficient information to distinguish amyloid from amorphous aggregates. They generally also have rather poor predictive capabilities toward amyloid sequences from yeast prions and functional amyloids. In contrast, structure-based homology modeling methods[25,40] can in principle provide very specific predictions. However, as side-chain modeling is very sensitive to even modest changes in backbone conformation, structure-based approaches are also limited by a shortage of experimentally determined structures[28].

To better understand the sequence determinants of amyloid structure, we explored the sequence diversity of amyloid hexapeptides by inspecting more than 200 peptides using various structural and biophysical methods. Using these data we trained a web-based tool called Waltz, which allows users to explore the sequence space of amyloids. Our analysis highlighted the strong position-specific tendencies of the different amino acids for forming amyloid structures observed both in disease-related as well as in functional amyloids.

[1]VIB SWITCH Laboratory, Flanders Institute for Biotechnology and Vrije Universiteit Brussel, Brussels, Belgium. [2]Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), Singapore. [3]Structures Programme, European Molecular Biology Laboratory, Heidelberg, Germany. [4]Department of Chemistry, School of Life Sciences, University of Sussex, Falmer, Brighton, UK. [5]Systems Biology Programme, Centre for Genomic Regulation, Barcelona, Spain. Correspondence should be addressed to J.W.H.S. (jschymko@vub.ac.be), F.R. (froussea@vub.ac.be) or L.S. (luis.serrano@crg.es).

RECEIVED 18 SEPTEMBER 2009; ACCEPTED 16 DECEMBER 2009; PUBLISHED ONLINE 14 FEBRUARY 2010; CORRECTED AFTER PRINT 29 SEPTEMBER 2010; ADDENDUM PUBLISHED ONLINE 29 SEPTEMBER 2010; DOI:10.1038/NMETH.1432

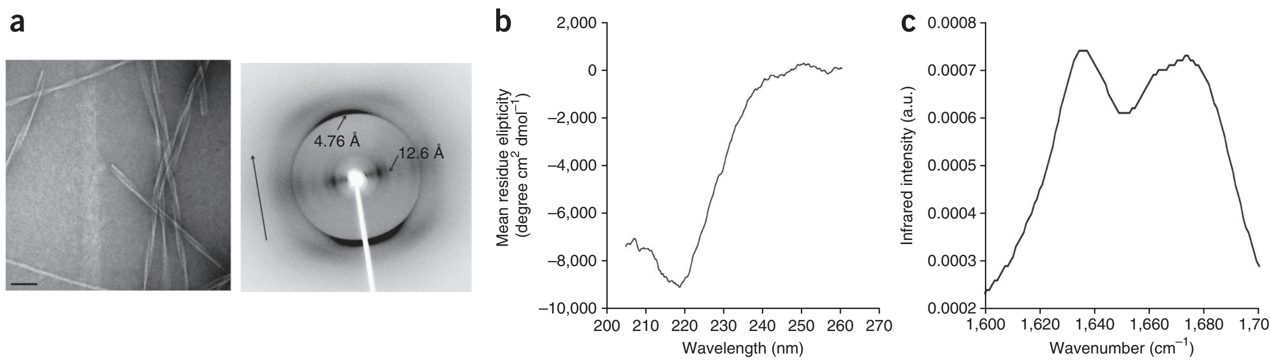

Figure 1 | Confirmation of cross-β structure. (**a**) Transmission electron micrograph (left) of amyloid fibrils formed by hexapeptide HYFNIF, negatively stained with uranyl acetate and X-ray diffraction pattern of unstained fibrils (right). The long arrow indicates the fiber direction and the short arrows show the position of the characteristic meridional (4.76 Å) and equatorial (12.6 Å) reflections in the cross-β-sheet pattern. Scale bar, 100 nm. (**b**) A typical circular dichroism spectrum of the aging peptide solution for the hexapeptide FIVNIV. (**c**) FTIR spectrum of the hexapeptide RVFNIM displays the typical peaks for β-sheet structure around 1,630 and 1,680 cm^{-1}.

RESULTS

Exploration of the sequence space for amyloid propensity

Most of the experimentally characterized amyloid sequences available to this date are hexapeptides. Community-generated, experimentally verified amyloidogenic hexapeptides are available in the AmylHex database[25] consisting of 67 'positive' and 91 'negative' examples that have been used to benchmark new methods. Although there is no doubt that amyloid nucleating sequences in full-length proteins are often longer than six residues[29] and that the full-length protein context can considerably modulate amyloid propensity[30], this length of six residues has been found to represent a minimalistic amyloid nucleating core, which as an insertion is sufficient to induce amyloid conversion of an entire protein domain[2]. The use of short peptides therefore minimizes the risk of alignment errors, whereas for longer amyloid sequences it is often not clear whether all residues participate in the β-sheet or are just tolerated dangling ends.

The AmylHex database, however, has a high sequence redundancy, as it has a strong overrepresentation (51%) of point mutations of the amyloidogenic hexapeptide STVIIE[15]. To unbias this dataset to more generally reflect amyloidogenic properties in naturally occurring peptides, we used the AmylHex set as a starting point to construct a first position scoring matrix, which we used to select increasingly divergent hexapeptides (Online Methods). In this manner, we identified and experimentally validated 49 new amyloid hexapeptide sequences as well as 71 negative (non-amyloid-forming) hexapeptide sequences (**Supplementary Table 1**). To unequivocally distinguish amyloid fibrils from nonfibrillar β-aggregates, we relied on a combination of electron microscopy (**Fig. 1a** and **Supplementary Fig. 1**), circular dichroism and Fourier-transform infrared (FTIR) spectroscopy. For selected cases, we also collected X-ray diffraction data (**Fig. 1a**), which revealed oriented diffraction patterns, all showing the characteristic diffraction signals for cross–β-amyloid fibrils[16]. We collected circular dichroism spectra of an aging hexapeptide solution (**Fig. 1b**), which showed β-sheet–enriched spectrum, and FTIR spectra (**Fig. 1c**), which displayed the typical peaks for β-sheet–enriched structure around 1,630 and 1,680 cm^{-1}.

When we added the 49 new sequences we identified to the existing 67 AmylHex peptides, our training set of amyloid forming hexapeptides was comprised of 116 positive and 103 negative sequence examples (the full training set is available at http://waltz.switchlab.org/). This increases the amount of learning examples compared to the AmylHex database by 70%. More importantly, we increased the number of nonredundant positive sequences by 370%, allowing for a better mapping of the position-specific sequence propensities of amyloidogenic hexapeptides.

Position-specific properties of amyloidogenic hexapeptides

We aligned our learning hexapeptide set and used it to generate a position-specific scoring matrix for amyloid propensity (Online Methods and **Fig. 2**). Almost all amino-acids had highly position-dependent propensities ranging from unfavorable to highly favorable for amyloid formation. Notably, residues with relatively low hydrophobicity such as glutamine and asparagine showed the greatest position dependency. This helps explain why prediction methods relying on average physical properties and

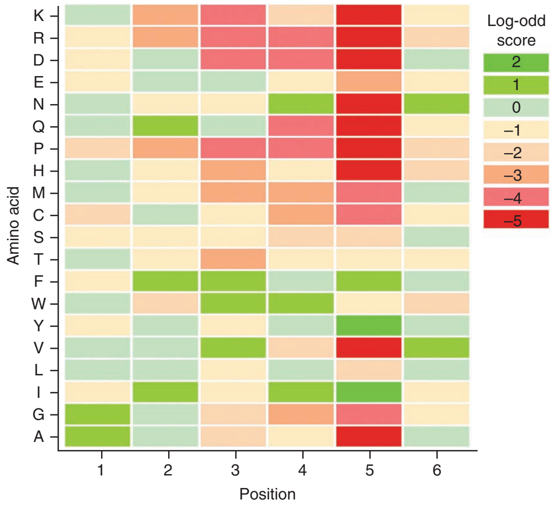

Figure 2 | Amino acid preference per position of the amyloid hexapeptide motif. The position-specific scoring matrix is displayed in color-code (green color indicates the amino acid is favorable for amyloid formation on that position of the hexapeptide mask, red indicates highly unfavorable amino acids and there is a linear color scale in-between).

that are biased toward hydrophobicity have difficulties detecting glutamine- and asparagine-rich amyloids such as yeast prions. As previously observed[15] the edge positions 1 and 6 are very tolerant as they can accommodate virtually any amino acid, whereas the central positions are more restrictive. Position 5 is highly specific: only isoleucine, phenylalanine and tyrosine are strongly favored for amyloid formation, whereas 13 residues are strongly excluded. The restrictions observed at position 5 are reminiscent of the prevalence of the STVIIE hexapeptide family from the original AmylHex set[25]. This means that either our expanded hexapeptide set is still not covering amyloid sequence space widely enough or, alternatively, that the position-specific packing rules for hexapeptides are rather general for a large fraction of the amyloid sequence space. Whatever the case, the considerable increase in sequence diversity in our expanded hexapeptide set relaxed position 5 from 17 excluded residues to 13, allowing four more residues at position 5.

Our analysis also shows that aromatic and hydrophobic residues remain most favored in the core of hexapeptide fibers but again, in a very position-specific manner. Isoleucine and valine, for instance, are largely mutually exclusive. Whereas isoleucine is highly favorable at positions 2, 4 and 5, valine, which is structurally very similar except for lacking the δ-methyl group, is unfavorable in those positions but, in contrast, is favored at positions 3 and 6. Another interesting feature of amyloid hexapeptides is their relative tolerance toward charged and polar residues. Whereas the cores of amorphous β-aggregates are largely devoid of charged and polar residues, amyloid structures accommodate these residues to a considerable extent. This is certainly the case for edge positions 1, 2 and 6, but even at central positions polar sequences can be accommodated or, in the case of glutamine and asparagine, even favorable. The position-specificity emerging from our expanded amyloid hexapeptide set illustrates how amyloid prediction methods that do not consider position-specificities are bound to overpredict strongly hydrophobic sequences and underpredict polar amyloid sequences.

Waltz, a position-specific prediction algorithm

We developed the Waltz algorithm by combining the position-specific sequence information described above with physicochemical as well as structural information described in Online Methods and **Supplementary Notes 1** and **2**. A sequence score $S_{profile}$ was calculated from the log-odd based position-specific scoring matrix (PSSM) (Online Methods). Nineteen selected physical properties, which best describe amyloid propensity enter the scoring function as a physical property term $S_{physprop}$ consisting of the sum of the products of the amino acid frequency with the normalized property value of the respective amino acid for each position (**Supplementary Fig. 2**). The final component of the scoring function is the position-specific pseudoenergy matrix from structural modeling using amyloid backbone structures (S_{struct}). Relative weightings *a* of individual terms was introduced for a balanced scoring function

$$S_{total} = \alpha_{profile}\, S_{profile} + \alpha_{physprop}\, S_{physprop} + \alpha_{struct}\, S_{struct} \quad (1)$$

To test the sensitivity of Waltz, we searched the UniProt database for amyloid-forming hexapeptide sequences and experimentally validated predictions by transmission electron microscopy (TEM) and spectroscopic techniques. We obtained all proteins with known involvement in human disease (9,384 sequences) from the UniProt database and removed those with 90% sequence redundancy, yielding 9,091 proteins. Using Waltz, we identified 21,264 amyloid-forming hexapeptide sequences. We randomly selected 30 sequences from this set and synthesized the peptides. After incubating the peptides at 500 mM in phosphate buffer (pH 7.0) for 3 weeks, we found that 80% of the predicted sequences formed amyloids (**Supplementary Table 2** and **Supplementary Fig. 1**).

The distinction between amyloid fibrils and amorphous β-aggregates appears important as both have very different biological effects. Unfortunately, it is not straightforward to derive bona fide datasets to benchmark such distinction. One possibility is to use a dataset of proteins found in inclusion bodies, but that has little information on the precise nature of aggregation[21]. This dataset is assumed to represent proteins that are generally enriched in aggregating regions. When we compared the ratios of percentage of predicted motifs within this dataset between Waltz and our previously developed β-aggregation predictor Tango[19], we found that Waltz was better at distinguishing amyloids from amorphous aggregating sequences (**Fig. 3a**). However, Tango was better-suited for distinguishing aggregating compared to soluble motifs (**Fig. 3b**).

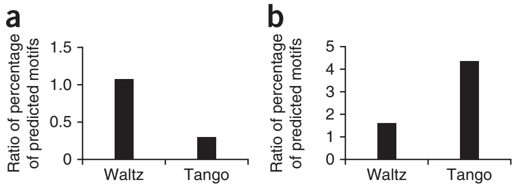

Figure 3 | Specificity analysis of Waltz and Tango for amorphous aggregation and amyloid formation. (**a**,**b**) Comparisons of Waltz and Tango for discerning amyloid fibrils from non-fibrillar aggregates (classification of aggregate morphology; **a**) while for distinguishing aggregating from soluble sequences (**b**). Values are ratios of predicted peptides from soluble, amyloid and generally aggregating protein sets that had been identified previously[21].

Table 1 | Sequences of sup35-derived peptides found to form amyloid

Amino acids	Sequence	Waltz	Packing[23]	Ambiguity[26]	STVIIE[15]	Tango[19]	Ref. 40
7–17	GNNQQNYQQY	+			+		+
16–26	YSQNGNQQQG						
28–38	RYQGYQAYNA	+					
43–53	GGYYQNYQGY	+			+		
46–56	YQNYQGYSGY	+					
52–62	YSGYQQGGYQ						
55–65	YQQGGYQQYN						
94–104	PQGGRGNYKN						
103–113	NFNYNNNLQG	+					
106–116	YNNNLQGYQA	+					
109–119	NLQGYQAGFQ	+					
127–137	NDFQKQQKQA						
Total:	12	7	0	0	2	0	1
Sensitivity:		58%	0%	0%	17%	0%	8%

Benchmarking on independent data

We next addressed whether Waltz covers sufficient sequence space to allow recognition of amyloidogenic sequences that are sufficiently different from the original AmylHex set by comparing Waltz's performance to five existing algorithms that use very different prediction strategies. These algorithms can be accessed on the recently developed AmylPred webserver[26]. These include: (i) the average packing density algorithm[23], (ii) the dual sheet/helix propensity method[26], (iii) the hexapeptide sequence pattern[15], allowing us to directly benchmark the sequence enrichment of our learning set, (iv) Tango[19], which exploits average physical properties such as hydrophobicity, secondary structure and charge to predict β-aggregation, and (v) the template modeling method[40], which uses a scoring function directly derived from structural modeling on microcrystal structures.

We compared the ability of Waltz and these five other algorithms to predict the amyloid-forming propensity of the yeast prion sup35, which is strongly enriched in the polar amino acids glutamine, asparagine and tyrosine. Out of the 73 peptides covering the sup35 N-terminal domain that we experimentally tested, 12 peptides were capable of forming amyloid at pH 7.5 and 150 mM NaCl. Waltz predicted seven of these correctly, yielding a sensitivity of 58%. This is a considerable improvement in comparison with the earlier hexapeptide-based method[15], which predicted only 3 out of 12 peptides (25% sensitivity). The other methods achieved between 0 and 8% sensitivity (**Table 1**) for the sup35 dataset.

Whereas benchmarking by cross-validation (84% sensitivity and 92% specificity; **Supplementary Fig. 3** and **Supplementary Notes 3** and **4**) likely represents the higher-limit performance expectation of Waltz, the very restrictive benchmarking performed on sup35 (58% sensitivity, 90% specificity) probably represents a realistic lower limit performance expectation for Waltz, because we performed the evaluation on independent sequences that have a very different composition from the original AmylHex set. Nevertheless, Waltz outperformed earlier methods in its ability to predict amyloid morphology. Although the sensitivity could still be improved by extension of the training set with ever more divergent sequences, our results showed that Waltz extends amyloid prediction to parts of sequence space that were previously unmapped.

Prediction of functional amyloids

Amyloids are not only associated with disease-related proteins. Nature also exploits the structural and mechanical properties of amyloids to regulate biological functions[13]. For example, in bacteria amyloids such as curli promote biofilm formation and host invasion, and chaplins in bacteria and hydrophobins in yeast act as regulators of surface tension of water. Insects and fish use chorion amyloid as a component of eggshell, and spiders use spidroins in spider silk. Humans possess at least one functional amyloid protein: Pmel17 is used as a structural scaffold in melanin synthesis. The strength and flexibility of amyloids have attracted attention for their use as biomaterials. However, it remains to be understood how functional amyloids avoid the cytotoxic effects observed in disease amyloids. In part, functional amyloids are regulated by chaperones and proteases. But it is also becoming evident that sequence composition modulates critical biophysical parameters determining kinetics of assembly or mechanical strength.

We used Waltz to identify amyloid-forming amino-acid sequences in 22 functional amyloids[6,13]. To maximize our chances of identifying amyloid sequences in these proteins, we used a cutoff ensuring high sensitivity but at the cost of a somewhat lower specificity. In cases in which overlapping hexapeptides were found, we selected a peptide that contained the entire region

Table 2 | Amyloid-forming peptides from known functional amyloids

	UniProt ID	Protein name	Function	Genus and species	Sequence[a]
Mammalia	P40967	PMEL	Control of melanin assembly	*Homo sapiens*	GQVIWVN AEVSIVV LASLIYR
Insects	P07184	Chorion S18	Structural and protective function in the egg shell	*Drosophila melanogaster*	DAQAIAL
	P17110	Chorion S36		*Ceratitis capitata*	RQGNINIVA VQQNYQA
	P07182	Chorion S36		*Drosophila melanogaster*	QQEVINK
Fungi	Q8J0V5	Conidial hydrophobin	Fungal coat formation, modulation of adhesion and surface tension	*Aspergillus fumigatus*	GLINIAT
	060048	Hydrophobin		*Pleurotus ostreatus*[b]	FNGLIVV
	Q8WZI5	Hydrophobin 1		*Pleurotus ostreatus*[b]	GILNIVV
	Q8WZI6	Hydrophobin 1		*Pleurotus ostreatus*[b]	SLLNIVV
	Q8WZJ2	Hydrophobin 2		*Dictyonema glabratum*	TAFTILA
	Q8WZJ3	Hydrophobin 1		*Dictyonema glabratum*	GVVAIGC
	Q6YF32	Hydrophobin 1		*Gibberella moniliformis*	NLSLFDQ
	Q6YF31	Hydrophobin 2		*Gibberella moniliformis*	TVIAFLA
	Q9P8T0	Hydrophobin 1		*Lentinula edodes*[c]	YGNLISL
	Q6YF29	Hydrophobin 4		*Gibberella moniliformis*	EFQEICA
Bacteria	P28307	Major curlin subunit (csgA)	Biofilm formation, host invasion	*Escherichia coli*	AIAAIVF SELNIYQY
	P0A1E5	Major curlin subunit (csgA)		*Salmonella typhimurium*	TLSIYQY
	Q9AD92	Chaplin	Modulation of water surface tension	*Streptomyces coelicolor*	GNVCIN
	B1W2X5	Chaplin		*Streptomyces griseus*	GNTCVN

[a]In several cases more than one non-overlapping sequence was identified in the same protein. [b]Oyster mushroom. [c]Shiitake mushroom.

encompassed by the peptides, resulting in peptides ranging from six to nine amino acids in length and experimentally tested the amyloid-forming propensity of these peptides. As expected from the highly sensitive threshold used, we found that about 50% of the predicted sequences yielded clear amyloid fibrils under the conditions assayed, resulting in the identification of 22 amyloid-forming regions in 17 out of the 22 functional amyloids tested (**Table 2**).

DISCUSSION

Amyloid is a very specific structure in terms of order and orientation. X-ray fiber diffraction from amyloid fibrils reveal a characteristic cross-β-sheet diffraction pattern displaying a sharp 4.7 Å reflection along the fiber axis and a more diffuse reflection at 10–12 Å perpendicular to the fiber axis (**Fig. 4a** and **Supplementary Table 1**). X-ray crystallography of microcrystals from amyloid forming peptides have also allowed investigation of side-chain packing in cross-β amyloid structures[17,28]. As expected, we observed extensive side chain–side chain contacts both along and perpendicular to the fiber axis. Viewed along the fiber axis, side-chain interactions were dominated by in-register interactions of identical residues stacked with a periodicity of 4.7 Å, which is the distance between two adjacent H-bonded β-strands. Side-chain interactions perpendicular to the fiber axis are constituted of interdigitating side chains protruding from backbones that face each other with a distance of 10–12 Å, thereby forming a tightly packed steric zipper[28] that constitutes the core of protofilaments (**Fig. 4b**,**c**).

Amorphous β-sheet aggregation, however, is less position-dependent and can, in principle, be achieved by any sequence that can adopt an extended conformation, is sufficiently hydrophobic and has no unsatisfied hydrogens or electostatic groups[18,19]. Thus β-sheet aggregation can be relatively easily predicted by methods that evaluate aggregation by evaluating biophysical parameters over a sequence segment, without the need for considering position-dependent values of these parameters[19–21]. In contrast, side-chain interactions perpendicular to the fiber axis show that the quaternary packing of two amyloid β-sheets imposes more extensive and complementary steric constraints[28]. β-sheet aggregating sequences that are compatible with the formation of a steric zipper can therefore be stabilized in the amyloid structure by the formation of a protofilament, but sterically incompatible sequences will remain in a 'molten' and thus amorphous β-sheet structure. Thus, predicting amyloid morphology requires knowledge of these position-specific sequence requirements[15]. Given the low abundance and diversity of experimentally validated amyloid-forming sequence sets, we did not have sufficient information to develop a reliable method to differentiate the specific amyloid morphology from amorphous β-sheet aggregates.

To address this issue, we expanded the existing AmylHex hexapeptide set to include increasingly diverging amyloid sequences, increasing the sequence diversity of the existing dataset by almost 400%. Waltz is unique in that it is, to our knowledge, the first position-specific amyloid sequence prediction tool that can be used to distinguish between amyloid and amorphous aggregate sequences. We extensively validated the performance of Waltz by experimentally characterizing additional amyloid-forming peptides both in functional amyloid proteins as well as in proteins hitherto not shown to possess amyloid potential. In total, the training and validation of Waltz resulted in the identification of 111 new amyloid-forming hexapeptides.

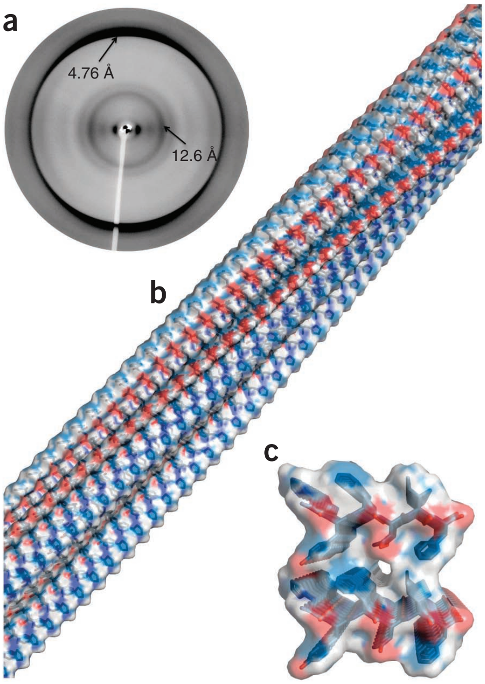

Figure 4 | Cross-β-sheet diffraction pattern and structure. (**a**) X-ray fiber diffraction pattern showing the classical cross-β-sheet diffraction pattern of amyloid fibrils of the hexapeptide sequence HYFNIF. (**b**,**c**) A molecular model showing the β-strands running perpendicular to the fiber axis (**b**) and the fibril cross-section, highlighting the interdigitation of the side chains (**c**).

Waltz also has a better ability to detect amyloid-forming potential involving polar amino acids such as found in prions and functional amyloids. Given the diversity of the current sequence dataset, a realistic performance estimate is in the 60–80% range for both sensitivity and specificity. Although additional sequence information might still improve the accuracy of PSSM-based amyloid prediction, the simplicity of the assumptions (hexapeptide model, no differentiation between parallel and antiparallel orientation) will also have to be revisited and refined as more sequence and structural information becomes available.

Waltz is freely available at http://waltz.switchlab.org/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

S.M.-S. was supported by a Marie Curie Intraeuropean fellowship. I.C.M. was supported by a doctorate scholarship from the Boerhinger Ingelheim Fonds, Foundation for Basic Research in Biomedicine and then the Fundação para a Ciência e Tecnologia. L. Serpell acknowledges the grant support from the Alzheimer's Research Trust and the Biotechnology and Biological Sciences Research Council. L. Serrano was partly supported by the EC grant Apopis. J.W.H.S. and F.R. acknowledge grant support from the Federal Office for Scientific Affairs, Belgium (Interuniversity Attraction Pole P6/43), the Fund for Scientific Research, Flanders, the Institute for Innovation by Science & Technology Flanders, the Stichting Alzheimer Onderzoek and the Alzheimer's Research Trust.

AUTHOR CONTRIBUTIONS

S.M.-S., F.R., L. Serrano and J.S. devised the methods. S.M.-S. implemented software. J.R., S.M.-S., J.W.H.S. and F.R. performed analysis. M.D., N.K., M.L.d.I.P., I.C.M., K.L.M., A.C. and L. Serpell performed experiments. S.M.-S., J.S. and F.R. wrote the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Chiti, F. *et al.* Kinetic partitioning of protein folding and aggregation. *Nat. Struct. Biol.* **9**, 137–143 (2002).
2. Ventura, S. *et al.* Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. *Proc. Natl. Acad. Sci. USA* **101**, 7258–7263 (2004).
3. Carrio, M., Gonzalez-Montalban, N., Vera, A., Villaverde, A. & Ventura, S. Amyloid-like properties of bacterial inclusion bodies. *J. Mol. Biol.* **347**, 1025–1037 (2005).
4. Marshall, K.E. & Serpell, L.C. Structural integrity of beta-sheet assembly. *Biochem. Soc. Trans.* **37**, 671–676 (2009).
5. Rousseau, F., Schymkowitz, J. & Serrano, L. Protein aggregation and amyloidosis: confusion of the kinds? *Curr. Opin. Struct. Biol.* **16**, 118–126 (2006).
6. Chiti, F. & Dobson, C.M. Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.* **75**, 333–366 (2006).
7. Matsumoto, G., Kim, S. & Morimoto, R.I. Huntingtin and mutant SOD1 form aggregate structures with distinct molecular properties in human cells. *J. Biol. Chem.* **281**, 4477–4485 (2006).
8. Kopito, R.R. Aggresomes, inclusion bodies and protein aggregation. *Trends Cell Biol.* **10**, 524–530 (2000).
9. Huyer, G. *et al.* A striking quality control subcompartment in *Saccharomyces cerevisiae*: the endoplasmic reticulum-associated compartment. *Mol. Biol. Cell* **15**, 908–921 (2004).
10. Kaganovich, D., Kopito, R. & Frydman, J. Misfolded proteins partition between two distinct quality control compartments. *Nature* **454**, 1088–1095 (2008).
11. Arrasate, M., Mitra, S., Schweitzer, E.S., Segal, M.R. & Finkbeiner, S. Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. *Nature* **431**, 805–810 (2004).
12. McClellan, A.J., Tam, S., Kaganovich, D. & Frydman, J. Protein quality control: chaperones culling corrupt conformations. *Nat. Cell Biol.* **7**, 736–741 (2005).
13. Fowler, D.M., Koulov, A.V., Balch, W.E. & Kelly, J.W. Functional amyloid–from bacteria to humans. *Trends Biochem. Sci.* **32**, 217–224 (2007).
14. Wang, X. & Chapman, M.R. Sequence determinants of bacterial amyloid formation. *J. Mol. Biol.* **380**, 570–580 (2008).
15. Lopez de la Paz, M. & Serrano, L. Sequence determinants of amyloid fibril formation. *Proc. Natl. Acad. Sci. USA* **101**, 87–92 (2004).
16. Makin, O.S., Atkins, E., Sikorski, P., Johansson, J. & Serpell, L.C. Molecular basis for amyloid fibril formation and stability. *Proc. Natl. Acad. Sci. USA* **102**, 315–320 (2005).
17. Nelson, R. *et al.* Structure of the cross-beta spine of amyloid-like fibrils. *Nature* **435**, 773–778 (2005).
18. Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C.M. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **424**, 805–808 (2003).
19. Fernandez-Escamilla, A.M., Rousseau, F., Schymkowitz, J. & Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **22**, 1302–1306 (2004).
20. Pawar, A.P. *et al.* Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. *J. Mol. Biol.* **350**, 379–392 (2005).
21. Sanchez de Groot, N., Pallares, I., Aviles, F.X., Vendrell, J. & Ventura, S. Prediction of "hot spots" of aggregation in disease-linked polypeptides. *BMC Struct. Biol.* **5**, 18 (2005).
22. Tartaglia, G.G., Cavalli, A., Pellarin, R. & Caflisch, A. Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci.* **14**, 2723–2734 (2005).
23. Galzitskaya, O.V., Garbuzynskiy, S.O. & Lobanov, M.Y. Prediction of amyloidogenic and disordered regions in protein chains. *PLOS Comput. Biol.* **2**, e177 (2006).
24. Saiki, M., Konakahara, T. & Morii, H. Interaction-based evaluation of the propensity for amyloid formation with cross-beta structure. *Biochem. Biophys. Res. Commun.* **343**, 1262–1271 (2006).
25. Thompson, M.J. *et al.* The 3D profile method for identifying fibril-forming segments of proteins. *Proc. Natl. Acad. Sci. USA* **103**, 4074–4078 (2006).
26. Hamodrakas, S.J., Liappa, C. & Iconomidou, V.A. Consensus prediction of amyloidogenic determinants in amyloid fibril-forming proteins. *Int. J. Biol. Macromol.* **41**, 295–300 (2007).
27. Zibaee, S., Makin, O.S., Goedert, M. & Serpell, L.C. A simple algorithm locates beta-strands in the amyloid fibril core of alpha-synuclein, Abeta, and tau using the amino acid sequence alone. *Protein Sci.* **16**, 906–918 (2007).
28. Sawaya, M.R. *et al.* Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature* **447**, 453–457 (2007).
29. Osherovich, L.Z., Cox, B.S., Tuite, M.F. & Weissman, J.S. Dissection and design of yeast prions. *PLoS Biol.* **2**, E86 (2004).
30. Tartaglia, G.G. *et al.* Prediction of aggregation-prone regions in structured proteins. *J. Mol. Biol.* **380**, 425–436 (2008).

ONLINE METHODS

Generation of the hexapeptide learning set. In a first step, the sequence information of the original AmylHex set was used to generate a Prosite[31]-style sequence mask. This was then used to screen and experimentally validate candidate regions in full-length amyloidogenic proteins for which the precise amyloidogenic regions were still unidentified. In parallel, double and triple mutations within the hydrophobic core of the STVIIE sequence have been investigated experimentally for amyloid fibril formation to better map interpositional dependencies. This exercise generated 30 new amyloid hexapeptides which we added to the AmylHex set (**Table 1**). In a second step, an initial crude log-odd based sequence profile has been generated from this extended AmylHex set and used to scan a general set of human proteins previously not related to amyloidosis. The resulting hexapeptide hits were then ranked by their dissimilarity to already experimentally tested ones (as judged with a Hamming distance of BLOSUM62 scores) and the top 30 that shared the least sequence similarity to known amyloidogenic examples were tested experimentally, to enrich the sequence diversity in the learning set. This exercise generated another 19 new amyloid hexapeptides (**Table 1**).

Amyloid peptide analysis. All peptides were obtained from JPT Peptide Technologies GmbH. Peptide stock solutions were prepared by dissolving a weighed amount of peptide (1 mg ml^{-1}) into the buffer indicated in the text. Samples were immediately sonicated using a Branson sonifier for 10 min to dissemble preformed nuclei and centrifuged (5 min at 16,100*g*) to deposit insoluble material. The concentration of the stock solutions was determined by measuring the absorbance at 280 or 220 nm for peptides without aromatic residues. Peptide solutions were incubated at room temperature (25 °C) and checked by circular dichroism or FTIR and electron microscopy at different incubation times (t = 0 and 1 month). In case a peptide sequence did not reveal fibrils by electron microscopy after 3 months, the incubation was repeated with a three times higher concentration.

Electron microscopy. Aliquots (5 µl) of a peptide preparation were adsorbed to carbon-coated FormVar film on 400-mesh copper grids (Plano GmbH) for 1 min. The grids were blotted, washed twice in 50 µl droplets of Milli-Q water and stained with 1% (wt/vol) uranyl acetate. Samples were studied with a FEI Morgagni 268(D) microscope at 120 kV and a JEOL JEM-2100 microscope at 200 kV.

X-ray fiber diffraction analysis. Peptides were dissolved in milliQ 0.2 µm filtered water at a concentration of 10mg ml^{-1} and incubated for a week at room temperature. We suspended 10–20 µl of peptide solution between two wax filled capillaries[32] and allowed them to dry at room temperature to form partially aligned fiber samples. X-ray diffraction data were collected using a RAxis IV++ detector with rotating anode CuKalpha X-ray source with exposure times of 10–20 min and data were examined using Clearer[33].

Spectroscopy. Circular dichroism measurements were recorded on a Jasco Spectropolarimeter J715 using quartz cuvettes (Hellma) with path lengths of 0.2–0.5 mm. A scan rate of 1 nm s^{-1} was used, and 15 spectra were averaged for each measurement. Samples were equilibrated at 25 °C using a water bath. Fourier-transform infrared spectroscopy was performed on a Bruker Tensor 37 FT-IR spectrometer equipped with an AquaSpec flowcell or a bioATR sample cell. The sample compartment was equilibrated to 25 °C; 100 spectra were averaged for a good signal-to-noise ratio.

Alignment of learning set sequences. Besides being reliably verified, the number and variety of bona fide examples is critical for sequence profiles to approximate a general model and sample the sequence space sufficiently. Finally, our strongly increased nonredundant learning set strengthens feasibility of a sequence-based prediction attempt. For a position-specific sequence profile an alignment of the learning set sequences is required that follows the β-sheet arrangement and, therefore, needs to be gapless. Ideally, independent sequence profiles should be derived for parallel and antiparallel β-sheet–forming sequences assuming different interaction types. However, the vast majority of experimental data do not include information on the preferred relative orientation of the β-strands and it is, hence, not possible to separate the data supported by experimental evidence.

One would expect that a cluster analysis of known hexapeptide examples could separate the data into two groups reflecting parallel and antiparallel sheets. Therefore, we modeled a set of hexapeptides from amyloid-related disease proteins into available parallel and antiparallel fibril structures and evaluated their energy using FoldX[34]. Notably, the energy distribution of the tested peptides for both orientations was hardly distinguishable for most cases. Only the classical example for a peptide with opposite charges on each end had a sharp stability increase in the antiparallel sheet, which can clearly be explained by electrostatic effects. Apparently, a certain similarity of residues at corresponding opposite ends can result in sequences that are similarly compatible with both orientations.

Indeed, when comparing the amino acid frequency distributions at the different positions of a nonredundant set of known fibril-forming hexapeptides, we found strong correlations (up to $R > 0.7$) among reciprocal positions both in the center but especially at the ends of the hexapeptide (for example, $R = 0.72$ for positions 1 and 6). Consequently, a sequence profile based on the currently known hexapeptide examples would not suffer from merging a presumed mixture of parallel and antiparallel examples into one learning set owing to the observed symmetry of motif positions flanking the core residues. Eventually, when sufficient reliable experimental data are available to separate parallel and antiparallel fibrils, this issue should be revisited.

Another possible variation in the alignment of β-strands are register shifts (for example, aligning position 1 in one strand with position 2 of the other). However, such shifts are assumed to be of minor importance in the case of isolated hexapeptides because any parallel translation of the strands relative to each other results in increasing solvent exposure of the respective ends which, clearly, is associated with a high energetic cost.

Selection of optimal subset of learning set sequences. Assuming a heterogeneous mixture of interaction types in our learning set consisting of a majority with similar properties and only a few outliers, we opted to test this rationally derived hypothesis by a data-driven approach. Therefore, we progressively knocked out groups of similar sequences (with high identity) from the learning

doi:10.1038/nmeth.1432

set and tested how well the resulting subset predicts the currently available full dataset. This procedure aims to find the most general sequence profile and unselects peptides that are less compatible with the dominant interaction model. For example, in our case, such an outlier-stripped subset profile selected for high specificity can result in a sensitivity increase of up to 40% compared to a profile blindly using the full learning set. Our final learning subset for generating the sequence profile consists of 91 of the 125 positive and 122 of the 138 negative examples contained in the total learning set.

There is the danger that removal of sequences from the profile can create new pseudocounts and give unwanted strength to the implemented pseudocount scheme. This is, however, not the case here, as the amino acids at affected positions are well-sampled by the remaining peptides, and the pseudocount penalty was furthermore chosen to have only little influence on the total score of a peptide.

The majority of unselected positive examples are STVIIE-derived single mutations, including STVIIE itself (accounting for 29 of 34 removed sequences). These peptides featured mutations only at the first and last positions, indicating that their measured effect on fibril formation is strongly dependent on the sequence of the peptide core (TVII), rather than being a general model. Similarly, the unselected negative examples are mainly from the STVIIE-like set. Among them was a group of sequences with opposite charges at the peptide termini that are listed as nonfibril formers in the AmylHex database. However, these were in fact found to form fibrils under special conditions (neutralization of the charged N and C termini but not the side-chain charges, by acetylation and amidation, respectively) and were therefore rightfully unselected from the negative set. Notably, two known fibril-forming peptides from the yeast prion sup35 were also unselected as outliers. They were indeed different from the majority of the learning set through their high glutamine and asparagine content. Nevertheless, the profile of the remaining sequence subset still recognized them as good fibril-formers with high prediction scores, distracting our concerns that the glutamine- and asparagine-rich peptides could have strongly differing interaction modes.

Sequence profile. The sequence profile was calculated as standard log-odd score in a position-specific matrix (the value for each amino acid at each position is the logarithm of the ratio of its frequency in the learning set and the background database). As we have a positive and a negative set that both sample well the amino acid space over the motif positions, we created one profile for each set and subtract the score against the negative profile (compliance with the negative set) from the score against the positive profile. We tested various pseudocount schemes (to avoid undefined logarithm calculation caused by zero frequencies) ranging from a simple constant pseudofrequency, over a background database dependent frequency, to a complex expected frequency derived from BLOSUM substitution probabilities considering all observed amino acid frequencies. Finally, when keeping the influence of the pseudocount model low regarding the contribution to the total score, all tested schemes resulted in a similar prediction performance and, hence, we chose the simple model of a constant pseudofrequency of 0.1% for nonobserved amino acids.

The amino acid variability is not equally distributed among the motif positions and the core positions seem to have higher importance for the amyloidogenic property of a peptide. We accounted for this by adding weights to the individual positions in the sequence profile indirectly proportional to their observed amino acid variability[35]. Again, we tested several different models to estimate this variability and found the Shannon entropy of amino acid group conservation a suitable measure. Thereby, the weight becomes the bit score derived from conservation of amino acids belonging to 6 groups of similar physicochemical properties, as defined previously[36].

Physical property descriptor selection. Besides the sequence profile, we also refined the motif characteristics in terms of physical property descriptors that will be added to the prediction scoring function. From a database of roughly 700 normalized parameter sets of physical properties for each of the 20 natural amino acids[37,38], we selected a subset of candidate properties for individual, pairs and regions of motif positions using various statistical methods, as established previously[39] but following a semiautomated procedure. First, we identify physical property scales that strongly correlated with the amino-acid frequencies of a nonredundant subset of our learning set, separately for the positive and negative examples.

Second, we selected the physical properties with the strongest deviation in positional and regional averages between our positive learning set and UniRef50 representing the expected random property value in large database searches. Third, we identified the strongest normalized property deviations between the positive and negative nonredundant learning sets. For the average deviation analyses, the balance between single positions and regions was reached by dividing the property deviation by the square root of the number of considered positions. We only considered deviations larger than 0.5 of the s.d. of the property in the positive set to reduce the number of candidates based on the significance of the deviation.

Fourth, we considered interpositional dependencies estimated by the Fisher ratio of variances, separately for the positive and negative nonredundant learning sets. The task of identifying interpositional correlations from a multiple alignment of sequences was not straightforward, and we introduced an additional internal cross-validation step. Apart from having a significant Fisher ratio with $P > 0.99$, we required that removing any of the learning set sequences does not alter the average of the property calculated over the correlated positions to fall out of the range determined by the s.d. of the full set. Thereby, we aimed to identify interpositional dependencies that were highly robust and general in regard to the learning set.

The following pruning rules apply to all candidate sets derived from the different analyses described above. Properties that correlate with $R > 0.9$ with an average database composition property are excluded to eliminate unspecific descriptors. To avoid redundant properties for the same positions or regions, only the best properties were selected among those that correlate with each other with $R > 0.4$. To avoid redundancy among positions and regions, we only kept the best position or region for each property.

Altogether, our semiautomated procedure selected 94 property-position or region mappings, which we continue to refer to as physical property regions (PPRs). From this pool of candidate PPRs, we derived an effective subset using a heuristic genetic

doi:10.1038/nmeth.1432

programming-inspired procedure. Therefore, a formula was iteratively 'evolved' by linear combination of PPRs. The 'fitness' of the formula was determined by the Matthew correlation coefficient (MCC) of a receiver-operator characteristic (ROC) curve analysis over the nonredundant positive and negative learning sets, where a set of random peptides from UniRef50 is added to the latter to allow testing for general database overprediction. Analogous to evolutionary selection, combinations were only accepted if the MCC was improved compared to the individual performances. As opposed to stochastic genetic algorithms, all possible combinations of PPRs were tested exhaustively, but we therefore limited the number of generations or iterations. Another difference to classic genetic algorithms is that we did not automatically unselect a fraction of worst performing formulas. Instead, after each iteration, a pruning step removed formulas whose performance was matched by another formula consisting of less individual PPRs, thereby favoring compact formulas. All accepted PPR combinations, as well as all initial single PPRs, were allowed to recombine in the next iteration. However, we only kept the best 7 formulas for each individual starting PPR, to avoid dragging along endless variations of formulas dominated by a single PPR with strong predictive power. After 10 iterations, the above-described procedure presents a formula with only 19 of the 94 PPRs remaining. As we allowed individual PPRs to enter the formula multiple times, a relative weighting of the PPRs among each other is obtained by their relative occurrences. The selected PPRs, listed in **Supplementary Table 3**, were assumed to be characteristic descriptors of the investigated motif as it is presented in the learning sets. More importantly, these semiautomatically selected physical property descriptors matched previous findings and expectations.

Structural modeling. The fibril crystal structure of the GNNQQNY peptide from Sup35 (Protein Data Bank (PDB) code: 1YJP) was first reduced to polyalanine. Then, all possible pair combinations of all 20 natural amino acids at all positions were generated and energy-optimized using FoldX. Energy estimates were calculated with FoldX as the ΔG difference ($\Delta\Delta G$) to the reference polyalanine. To retrieve a position-specific pseudoenergy matrix for the prediction scoring function, we then averaged for each amino acid the energies for all its occurrences at a certain position in combination with all amino acids at other positions.

Composite scoring function. Building up on earlier work[39], Waltz combines sequence, physicochemical as we all as structural information into a composite scoring function (equation 2). A sequence score S_{profile} was calculated from the log-odd based sequence PSSM derived as described in the preceding sequence profile subsection. The 19 selected physical properties enter the scoring function as a physical property term S_{physprop} consisting of the sum of the products of the amino acid frequency with the normalized property value of the respective amino acid for each position. The final component of the scoring function is the position-specific pseudoenergy matrix from the structural modeling (S_{struct}). Relative weightings of the individual terms had to be introduced for a balanced scoring function:

$$S_{\text{total}} = S_{\text{profile}} + S_{\text{physprop}} - 0.2 \times S_{\text{struct}} \qquad (2)$$

Additionally, most methods take advantage of selected physical property scales for their predictions. Typically these properties are characteristic for the sets of known examples they are derived from and, hence, could be a source of overfitting. In our case, the benchmark was executed over the AmylHex dataset, which only had an approximate overlap of 50% with the nonredundant subset of the extended learning set that was used to select physical properties for Waltz. This difference between learning and benchmark set represented a strong cross-validation by itself. In contrast, the structure-derived contributions were independent of the learning set and did not need to be cross-validated.

Indeed, when omitting the physical property and structural descriptors in the prediction function, the sequence profile alone already showed a strong performance increase compared to the competing algorithms, albeit less than with the complete composite scoring function (data not shown). Hence, the sequence component of our scoring function yielded most of the predictive power, and we are confident that we addressed the problem of overfitting appropriately with the rigorous cross-validation.

31. Hulo, N. *et al.* The PROSITE database. *Nucleic Acids Res.* **34**, D227–D230 (2006).
32. Makin, O.S. & Serpell, L. X-ray diffraction studies of amyloid structure. In *Amyloid Proteins: Methods and Protocols* (ed. Sigurdsson, E.M.) vol. 299, 67–80 (Humana Press, 2005).
33. Makin, O.S., Sikorski, P. & Serpell, L. CLEARER: a new tool for the analysis of X-ray fibre diffraction patterns and diffraction simulation from atomic structural models. *Appl. Cryst.* **40**, 966–972 (2007).
34. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382–388 (2005).
35. Maurer-Stroh, S. & Eisenhaber, F. Refinement and prediction of protein prenylation motifs. *Genome Biol.* **6**, R55 (2005).
36. Mirny, L. & Shakhnovich, E. Evolutionary conservation of the folding nucleus. *J. Mol. Biol.* **308**, 123–129 (2001).
37. Eisenhaber, B., Bork, P. & Eisenhaber, F. Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase. *Protein Eng.* **11**, 1155–1161 (1998).
38. Tomii, K. & Kanehisa, M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.* **9**, 27–36 (1996).
39. Eisenhaber, B., Eisenhaber, F., Maurer-Stroh, S. & Neuberger, G. Prediction of sequence signals for lipid post-translational modifications: insights from case studies. *Proteomics* **4**, 1614–1625 (2004).
40. Zhang, Z.Q., Chen, H. & Lai, L.H. Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential. *Bioinformatics* **23**, 2218–2225 (2007).

doi:10.1038/nmeth.1432

Corrigendum: Exploring the sequence determinants of amyloid structure using position-specific scoring matrices

Sebastian Maurer-Stroh, Maja Debulpaep, Nico Kuemmerer, Manuela Lopez de la Paz, Ivo Cristiano Martins, Joke Reumers, Kyle L Morris, Alastair Copland, Louise Serpell, Luis Serrano, Joost W H Schymkowitz & Frederic Rousseau

***Nat. Methods* 7, 237–242 (2010); published online 14 Feburary 2010; corrected after print 29 September 2010.**

In the version of this paper originally published, the name of and reference to the algorithm in the rightmost column of Table 1 were incorrect. The correct reference (ref. 40) has been added in the paper. The error has been corrected in the PDF and HTML versions of the article.

Addendum: Exploring the sequence determinants of amyloid structure using position-specific scoring matrices

Sebastian Maurer-Stroh, Maja Debulpaep, Nico Kuemmerer, Manuela Lopez de la Paz, Ivo Cristiano Martins, Joke Reumers, Kyle L Morris, Alastair Copland, Louise Serpell, Luis Serrano, Joost W H Schymkowitz & Frederic Rousseau

***Nat. Methods* 7, 237–242 (2010); published online 14 Feburary 2010; addendum published after print 29 September 2010.**

After the publication of our paper, we identified a mistake in Table 1 regarding the comparison of our program, Waltz, to the program 3D profile[1] (ref. 25 in our paper); we cited the wrong name and reference of the algorithm in the right column. This error has been corrected after print to refer to the algorithm we actually used, the method described in reference 2 (ref. 40 in the corrected paper). However, as the 3D profile[1] method developed in the Eisenberg laboratory has a long-standing good reputation as an amyloid prediction tool, here we compare it to Waltz. An improved version of 3D profile[3] was published about a week and a half before our paper, so for complete transparency we also compare Waltz to the improved 3D profile algorithm.

In **Table 1** we list all predicted peptides and scores or energies, respectively, comparing Waltz (threshold 77, running on our webserver at http://waltz.switchlab.org/) with the 3D profile[1] scores at the ZipperDB website (http://services.mbi.ucla.edu/zipperdb; energy threshold was –23; additional shape complementarity > 0.7 for the 3D profile 2010 version[3]). The sensitivity of 3D profile on our sup35 positive set was 67% (75% if one includes prediction of a hexapeptide that is almost but not fully included in the tested decapeptide).

However, the higher sensitivity of 3D profile comes at a cost of lower specificity (more false positives). To estimate the rate of false positives, we derived a reliable negative set from our experimental data for sup35, which included all decapeptides that did not form fibers under the unified experimental conditions and did not overlap with any positively tested one (31 in total). However, we cannot draw hard conclusions as the availability of bona fide experimental data is typically limiting and these numbers are too low for a good general comparison. An additional complication is that 3D profile is designed to predict hexapeptides; as next best approximation we defined the best score or energy of a fully included hexapeptide as prediction for the respective peptides. Owing to this limitation and the fact that well-predicted hexapeptides may actually form amyloid fibers and the longer decapeptide does not, it may be wiser to exclude such peptides in an alternative comparison with only 26 'negative' peptides, the reduced benchmark set ('–5') (**Table 1**).

Sensitivities of predictors should either be compared at similar levels of specificity (as should be done in consensus methods, such as AmylPred[4]), or one needs to consider both sensitivity and specificity together. Established measures for this are the Matthew correlation coefficient and the probability excess[5]. Probability excess has the additional advantage that it is also independent of set size inequalities[6], which are not considered in other measures such as accuracy and precision.

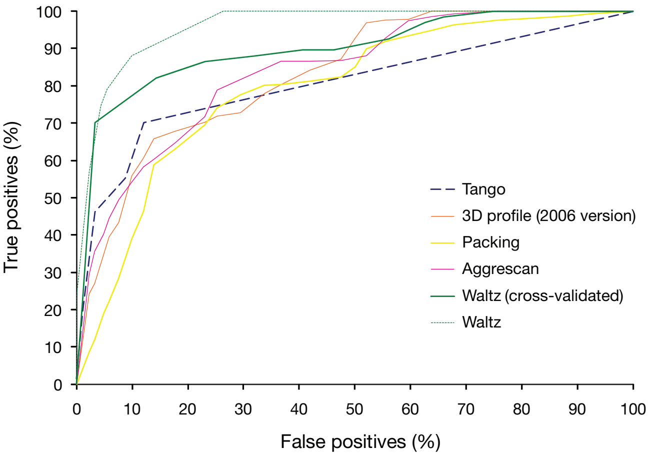

Figure 1 | Comparison of ROC curve performance on the AmylHex dataset.

The resulting performance statistics are reported in **Table 2**. Although 3D profile 2006 version[1] predicted several additional false positives compared to Waltz, the improved 3D profile 2010 version[3] filtered out several of these. Considering the possibility that high-scoring hexapeptides may indeed form fibers outside of the experimentally tested decapeptide context, the performances of Waltz and 3D profile (2010 version)[3] become comparable over the reduced benchmark set ('–5'). In fact, the observed differences may well be within the error of performance estimation given the small benchmark set.

We also performed a receiver operating characteristics (ROC) curve analysis to benchmark the performance of Waltz, 3D profile 2006 version[1], Tango[7], Packing[8] and Aggrescan[9] on the AmylHex dataset[1] (**Fig. 1**). The AmylHex dataset is an experimentally validated set of hexapeptides containing 67 positive (amyloid forming) and 91 negative (non–fiber forming) examples. Although 3D profile and the other methods in this benchmark were not subjected to cross-validation, we additionally scrutinized Waltz using rigorous cross-validation criteria as outlined in Supplementary Notes 3 and 4 of our original paper. We emphasize that the 3D profile method[1] in this ROC curve was the version from 2006; we did not test the performance of the improved 3D profile[3] method.

Our and others' recent work has additionally contributed several new experimentally verified examples, which should form the basis of an enlarged benchmark set to allow standardized ROC comparison of amyloid predictors by all interested groups in the future.

Table 1 | Comparison of true positives and false positives identified by Waltz and 3D profile for sup35-derived peptides

From amino acid	To amino acid	Experimental peptide	Predicted peptide, Waltz	Waltz score	Waltz	Predicted peptide, 3D profile	3D profile energy	3D profile SC (2010 version)	3D profile (2006 version)	3D profile (2010 version)
		Positives								
7	16	GNNQQNYQQY	NQQNYQQY	98.3	+	NNQQNY	−24.8	0.715	+	+
16	25	YSQNGNQQQG	YSQNGNQQQG	65.6	–	GNQQQG	−23.1	0.928	+	+
28	37	RYQGYQAYNA	RYQGYQAYNA	92.8	+	QGYQAY	−23	0.89	+	+
43	52	GGYYQNYQGY	YYQNYQGY	98.0	+	GYYQNY	−26.4	0.827	+	+
46	55	YQNYQGYSGY	NYQGYSGY	80.7	+	QNYQGY	−21.9	0.904	–	–
52	61	YSGYQQGGYQ	QQGGYQ	77.9	+	QQGGYQ	−22.4	0.797	–	–
55	64	YQQGGYQQYN	GYQQYN	92.0	+	GGYQQY	−25.6	0.838	+	+
94	103	PQGGRGNYKN	GRGNYKN	52.2	–	GGRGNY	−19.4	0.901	–[a]	–[a]
103	112	NFNYNNNLQG	NFNYNNNLQG	81.6	+	NYNNNL	−23.4	0.861	+	+
106	115	YNNNLQGYQA	NLQGYQA	82.9	+	NNLQGY	−24.1	0.85	+	+
109	118	NLQGYQAGFQ	NLQGYQAGFQ	81.1	+	GYQAGF	−23.8	0.894	+	+
127	136	NDFQKQQKQA	DFQKQQKQA	57.2	–	QKQQKQ	−22.7	0.665	–	–
		Negatives								
67	76	AGYQQQYNPQ[b]	YQQQYNPQ[b]	92.6	+	GYQQQY[b]	−25.7	0.928	+	+
70	79	QQQYNPQGGY			–				–	–
73	82	YNPQGGYQQY			–				–	–
76	85	QGGYQQYNPQ[b]	GYQQYNPQ[b]	92.0	+	GGYQQY[b]	−25.6	0.838	+	+
79	88	YQQYNPQGGY			–				–	–
82	91	YNPQGGYQQQ[b]			–	GGYQQQ[b]	−24.2	0.84	+	+
139	148	KPKKTLKLVS			–	TLKLVS	−24.6	0.626	+	–
142	151	KTLKLVSSSG			–	LVSSSG	−25	0.526	+	–
145	154	KLVSSSGIKL			–	VSSSGI	−25.5	0.572	+	–
148	157	SSSGIKLANA[b]			–	SSSGIK[b]	−24.7	0.721	+	+
151	160	GIKLANATKK			–	KLANAT	−23.4	0.672	+	–
154	163	LANATKKVGT[b]			–	ANATKK[b]	−24.9	0.735	+	+
157	166	ATKKVGTKPA			–	ATKKVG	−23 (2006) −21.4 (2010)	0.872	+	–
160	169	KVGTKPAESD			–				–	–
163	172	TKPAESDKKE			–				–	–
166	175	AESDKKEEEK			–				–	–
169	178	DKKEEEKSAE			–				–	–
172	181	EEEKSAETKE			–				–	–
175	184	KSAETKEPTK			–				–	–
178	187	ETKEPTKEPT			–				–	–
181	190	EPTKEPTKVE			–				–	–
184	193	KEPTKVEEPV			–				–	–
187	196	TKVEEPVKKE			–				–	–
190	199	EEPVKKEEKP			–				–	–
193	202	VKKEEKPVQT			–				–	–
196	205	EEKPVQTEEK			–				–	–
199	208	PVQTEEKTEE			–				–	–
202	211	TEEKTEEKSE			–				–	–
205	214	KTEEKSELPK			–				–	–
208	217	EKSELPKVED			–				–	–
211	220	ELPKVEDLKI			–				–	–

[a]The 3D profile method predicted the hexapeptide GNYKNF, which is almost fully included in the tested decapeptide. [b]These peptides were optionally removed in set '–5'. SC, shape complementarity.

Table 2 | Performance summary statistics for Waltz and 3D profile

Set	Method (version)	TP	NP	FP	NN	TN	FN	Sensitivity	Specificity	Accuracy	Precision	MCC	PE
All	Waltz	9	12	2	31	29	3	0.750	0.935	0.884	0.818	0.705	0.685
All	3D profile (2006)	8	12	10	31	21	4	0.667	0.677	0.674	0.444	0.313	0.344
All	3D profile (2010)	8	12	5	31	26	4	0.667	0.839	0.791	0.615	0.494	0.505
−5	Waltz	9	12	0	26	26	3	0.750	1.000	0.921	1.000	0.820	0.750
−5	3D profile (2006)	8	12	5	26	21	4	0.667	0.808	0.763	0.615	0.465	0.474
−5	3D profile (2010)	8	12	0	26	26	4	0.667	1.000	0.895	1.000	0.760	0.667

TP, number of true positives; NP, number of positives; FP, number of false positives; NN, number of negatives; TN, number of true negatives; FN, number of false negatives; sensitivity, TP/NP; specificity, TN/NN; accuracy, (TP+TN)/(NP+NN); precision, TP/(TP + FP); MCC, Matthew correlation coefficient = ((TP × TN) − (FP × FN))/√((TP + FP)(TP + FN)(TN + FP)(TN + FN)); PE, probability excess = sensitivity + specificity − 1.

1. Thompson, M.J. *et al*. The 3D profile method for identifying fibril-forming segments of proteins. *Proc. Natl. Acad. Sci. USA* **103**, 4074–4078 (2006).
2. Zhang, Z.Q., Chen, H. & Lai, L.H. Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential. *Bioinformatics* **23**, 2218–2225 (2007).
3. Goldschmidt, L., Teng, P.K., Riek, R. & Eisenberg, D. Identifying the amylome, proteins capable of forming amyloid-like fibrils. *Proc. Natl. Acad. Sci. USA* **107**, 3487–3492 (2010).
4. Hamodrakas, S.J., Liappa, C. & Iconomidou, V.A. Consensus prediction of amyloidogenic determinants in amyloid fibril-forming proteins. *Int. J. Biol. Macromol.* **41**, 295–300 (2007).
5. Sirota, F.L. *et al*. Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. *BMC Genomics* **11**, (Suppl. 1) S15 (2010).
6. Yang, Z.R., Thomson, R., McNeil, P. & Esnouf, R.M. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* **21**, 3369–3376 (2005).
7. Fernandez-Escamilla, A.M., Rousseau, F., Schymkowitz, J. & Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **22**, 1302–1306 (2004).
8. Galzitskaya, O.V., Garbuzynskiy, S.O. & Lobanov, M.Y. Prediction of amyloidogenic and disordered regions in protein chains. *PLOS Comput. Biol.* **2**, e177 (2006).
9. Conchillo-Solé, O. *et al*. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* **8**, 65 (2007).