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 web-server 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.

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

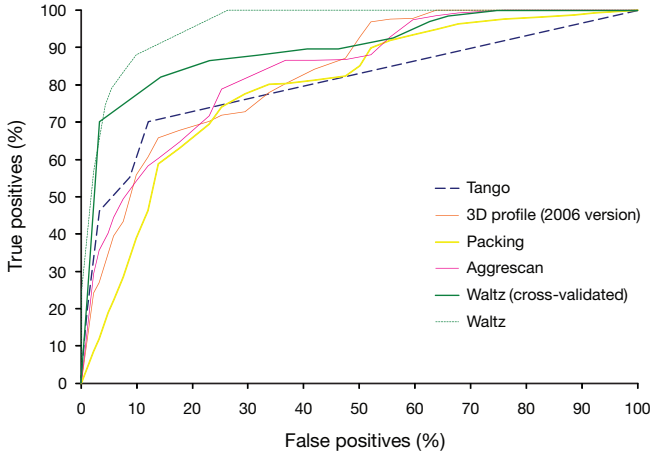

Figure 1 | Comparison of ROC curve performance on the AmylHex dataset.

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)

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 = $((TP \times TN) - (FP \times FN))/\sqrt{((TP + FP)(TP + FN)(TN + FP)(TN + FN))}$; PE, probability excess = sensitivity + specificity − 1.

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.

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).