P. Slama and D. Geman, Identification of family-determining residues in PHD fingers

Supporting material

Table S1 (additional file). List of the proteins used in the PHD finger alignment, resulting from the CD-Hit filtering (see METHODS) of the PFAM dataset.

Figure S2. Relative entropy calculations on alignment subsets and randomized alignment subsets of variable sizes (the size of the subset being the number of protein sequences it contained).

Figure S3. Ratio of the frequency of the significant conditions with respect to the maximum amino-acid frequency observed at their respective positions. This plot shows that our method selects conditions over a range of frequencies.

Table S4. Clustering of the first-percentile conditions when calculating disparity using the L1 (or cityblock) distance.

Figure S5. Conservation scores over our PHD finger sequence alignment when using relative entropy to compare amino-acid distributions with that observed over the Uniprot database. Residue numbering is artificial and as used over this whole study.

Figure S6. List of all conditions having at least one coupling in the first percentile of coupling values when using relative entropy in the analysis of the CD-Hit filtered alignment of 926 PHD fingers.

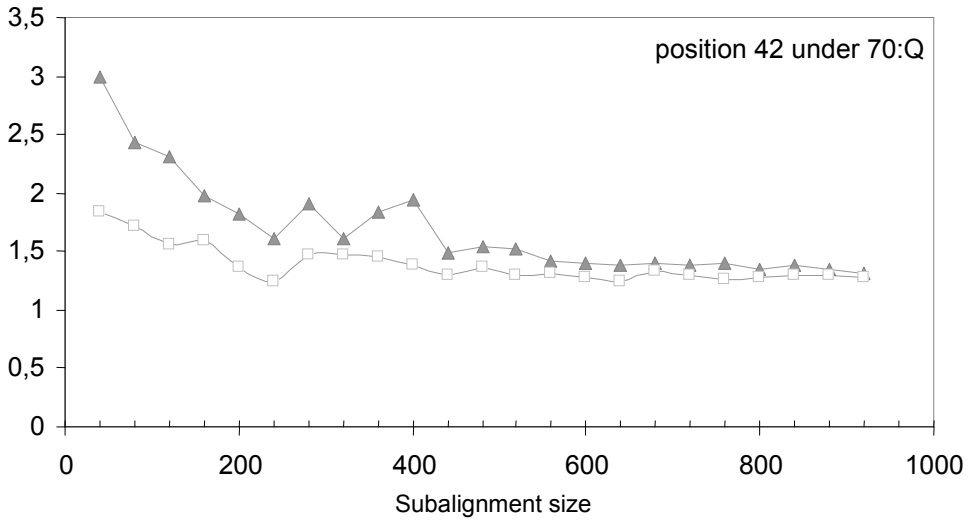
Table S7. Interactions between Histone 3 and calculated PHD structures in docked structures. All hydrogen bonds with length inferior to 3.2 Å are indicated, and are visible as dots in structure images.

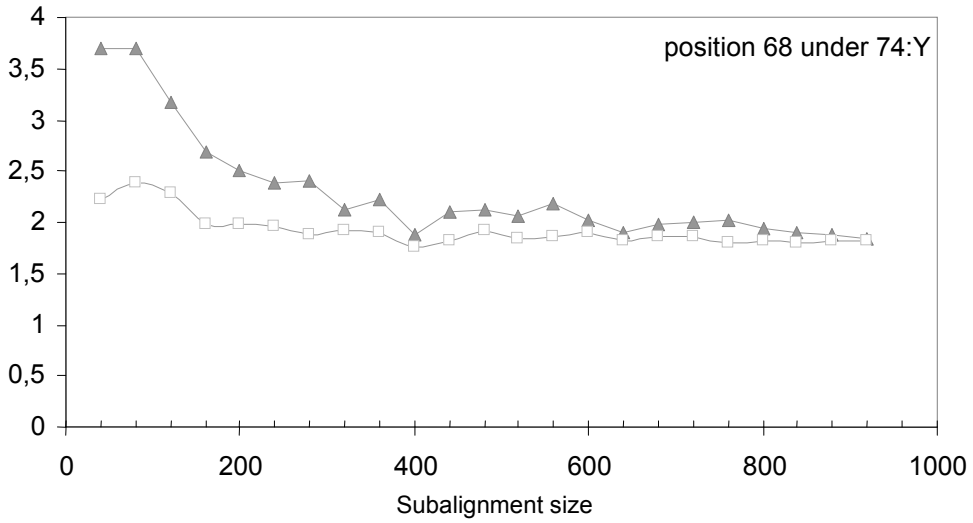
Table S8. Clusters of the first-percentile conditions obtained on the chromodomain alignment.

Table S9. Clustering of the first-percentile conditions obtained on the acyltransferase alignment.

Table S10. Numbering of all positions that defined a first-percentile condition in the acyltransferase dataset with respect to bovine mitochondrial protein GPAT1.

Figure S11. Superposition of the structures of the PHD fingers of AIRE 1 (PDB entry 2KE1) and BRPF1 (predicted, this study).





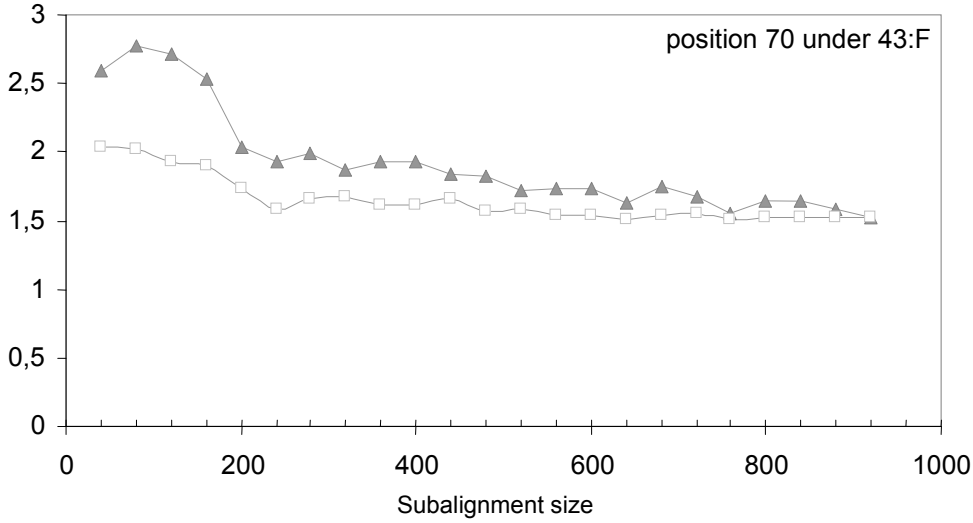
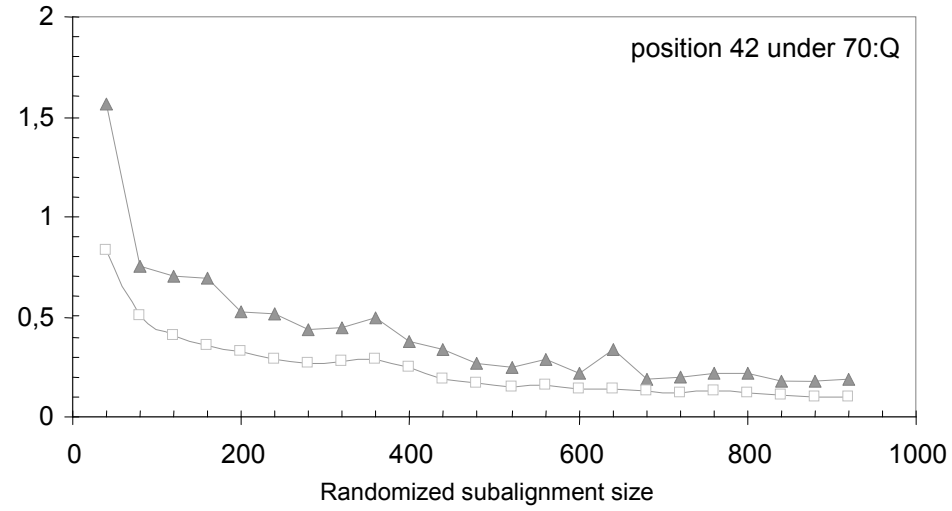
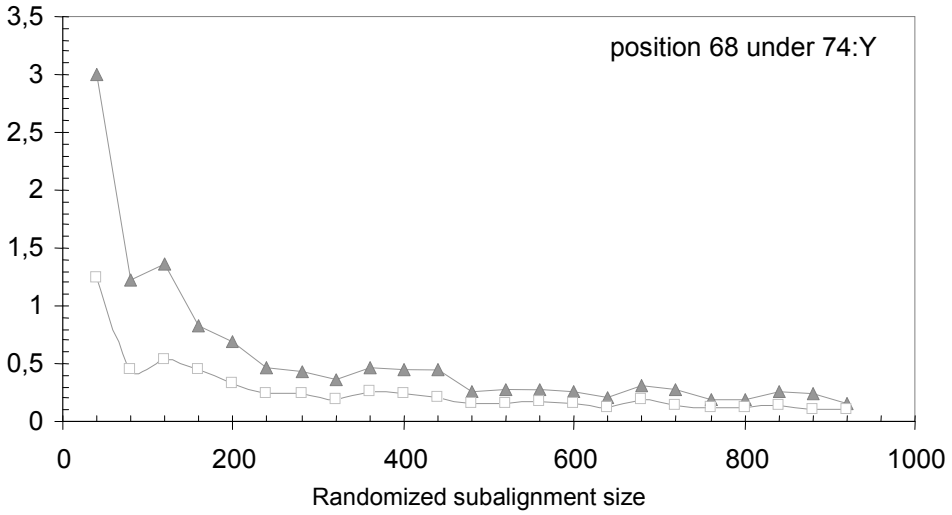


Figure S2.A. Influence of the alignment size on relative entropy values. For three of the five strongest couplings, average (squares) and maximum (triangles) relative entropy values for calculations on 10 random alignment subsets of varying sizes (number of protein sequences) are displayed. The effect of the alignment size vanishes for subalignments containing more than 400 sequences.





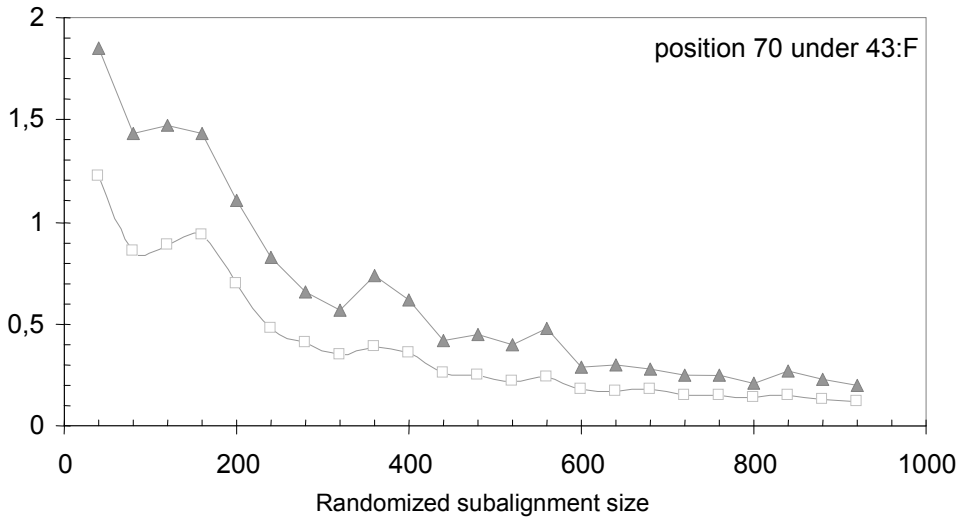


Figure S2.B. Influence of the size of randomized alignments on relative entropy values. Average (squares) and maximum (triangles) relative entropies for calculations on 40 randomized alignment subsets of varying sizes (number of protein sequences) are displayed. Real values for the corresponding couplings are 1.279, 1.814 and 1.523, respectively, as can be read from the final values on non-randomized subsets (Figure S2.A.). These results (A. and B.) show that there is no significant effect due to the size of our alignment on relative entropy values.

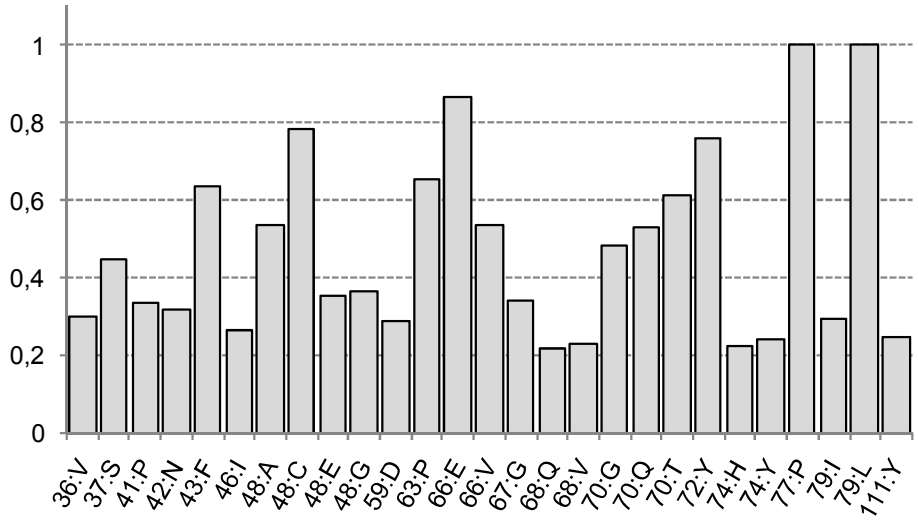


Figure S3. Frequency of the conditions used in the analysis with respect to the maximum amino-acid frequency observed at their respective positions. A value of 1 indicates that the amino-acid used to define the condition is the most frequent at that position in the PHD finger alignment.

Table S4. Families of conditions obtained when using the L1 distance to compare conditions.

I	II	III	IV	V
77:P	48:C	42:N	66:E	111:Y
79:L	63:P	70:Q	37:S	43:F
41:P		46:I	48:A	70:G
67:G		79:I	36:V	68:Q
72:Y		74:H	48:G	66:V
70:T		68:V	59:D	48:E
		74:Y		

Conditions with at least one coupling in the first percentile of all coupling values were compared using a disparity measure based on the L1 (or cityblock, see Equation 4) distance and clustered by affinity clustering. The top line is the centroid of each cluster, and conditions are ranked (top to bottom) by increasing disparity to centroid. The clusters are identical to those obtained using the relative entropy-based disparity measure, except that conditions from family II here were the least central of the cluster with centroid 77: P in the calculation that used symmetrized relative entropy (Table 2).

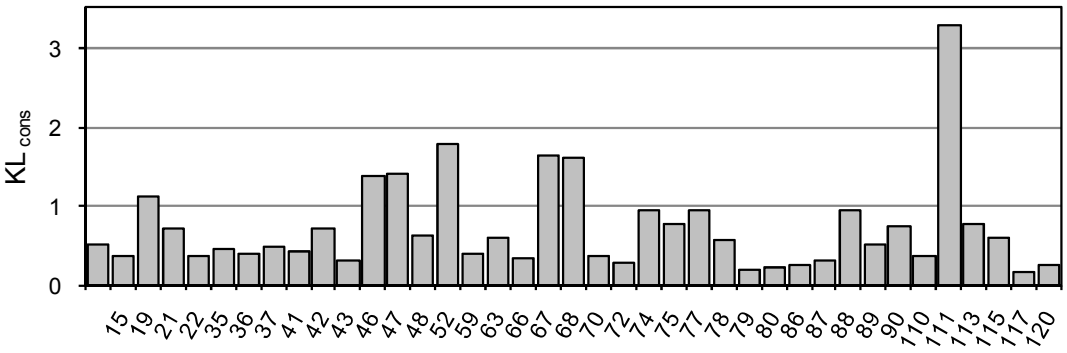


Figure S5. Conservation of all observation positions in the PHD finger sequence alignment, as measured by the relative entropy of their amino-acid distribution to the amino-acid distribution in the Uniprot database (KL_{cons}, see METHODS).

Table S6. Conditions having at least one coupling among the first percentile of couplings values.

Condition	Number of couplings
74:Y	10
68:Q	7
36:V	6
74:H	6
70:Q	5
63:P	5
48:G	4
46:I	4
70:G	3
59:D	3
43:F	3
37:S	3
48:A	2
68:V	2
72:Y	2
42:N	2
48:C	2
79:I	1
77:P	1
111:Y	1
70:T	1
67:G	1
66:V	1
66:E	1
48:E	1
41:P	1
79:L	1

Conditions are ranked by decreasing number of couplings (top to bottom). See Figure 1 or METHODS for a correspondence between the different positions involved in these conditions and their sequence localization.

Table S7. Interactions observed in the docking calculations of the N-terminal of Histone 3 to the predicted PHD finger structures.

BRPF1	Sidechain of protein	Asp214 (Oδ1,Oδ2)	Lys4
		Asp222 (Oδ1,Oδ2)	Arg2
		Glu253 (Oε2)	Arg8
	Main chain of protein	Glu224 (Oε1)	Arg2 (N)
		Cys225 (N)	Thr3 (O)
		Cys225 (O)	Gln5 (N)
		Cys225 (O)	Thr6 (N)
		Cys225 (O)	Thr6 (Oγ1)
		Asn227 (N)	Thr6 (O)
		Gln226 (Nε2)	Ala7 (O)
		Asn227 (Oδ)	Arg8 (N)
TIF1A	Sidechain of protein	Asp823 (Oδ1)	Lys4
		Asp823 (Oδ1,Oδ2)	Arg2
		Asn825 (Oδ1)	Arg8 (Nε)
		Glu842 (Oε2)	Ala7 (N)
		Glu842 (Oε1)	Arg8 (N)
		Phe860 (Ar)	Ala1 (Cβ)
		Phe860 (Ar)	Thr3 (Cγ2)
	Main chain of protein	Gly836 (O)	Arg2 (Nε)
		Leu838 (O)	Thr3 (N)

In parenthesis, the residue atom(s) involved in the interaction in PDB notation.

Table S8. Clusters of the first-percentile conditions obtained on the chromodomain alignment.

I	II	III	IV
43:P	129:H	130:D	53:Q
105:V	128:L	108:R	137:R
108:Q	115:H	106:K	141:L
110:S	137:Y	46:L	153:R
145:R	134:V	137:I	
161 :W	130:C	60:D	
59 :E	113:W		

Positions corresponding to the first-percentile conditions can be read from Figure 6. Conditions are ordered (top to bottom) by increasing distance to the centroid cluster within each cluster. Families are numbered by increasing core size. Underlined positions are those also predicted by method SDR, those in grey boxes are predicted as specificity-determining by proteinkeys.

Table S9. Clusters of the first-percentile conditions obtained on the acyltransferase alignment.

I	II	III	IV	V	VI	VII	VIII
91:T	528:G	71:T	609:G	600:G	141:A	91:G	657:P
90:E	615:A	91:F	145:L	227:Q	153:N	85:G	230:L
671:N	523:T	227:G	208:G	71:L	150:D	391:H	86:H
67:K	227:T	611:Y	591:H	166:T	229:F	523:V	616:L
383:G	665:Q	67:P	227:F	99:P	147:G	71:P	147:A
528:P	169:W	609:F	166:Y	156:Y	361:Y	668:Y	221:G
227:M	86:W	145:G	135:G	169:G	65:P	522:R	383:C
506:W	145:S	150:H	614:A	590:P	356:L	194:P	
	614:Q	614:Y	166:W		134:L	523:L	
	527:D	406:E	615:F		233:R	598:M	
	71:I	394:F	600:P		671:G	616:I	
	71:V	89:W	660:D		522:G	171:G	
	527:N	600:W	214:W		511:F		
	364:P	512:A	99:W		611:G		
	328:R	56:I	352:E		510:W		
		139:W	361:D				
		227:H					

Correspondence between positions and the residues from cow protein GPAT1 is provided in Table S10. Conditions are ordered (top to bottom) by increasing distance to the centroid cluster within each cluster. Families are numbered by increasing core size. Underlined positions are those also predicted by method SDR, those in grey boxes are predicted as specificity-determining by proteinkeys.

Table S10. Sequence correspondence for the positions involved in first-percentile conditions for the acyltransferase dataset to the residues of mitochondrial Glycerol 3-phosphate acyltransferase (GPAT1 or GPAM) from cow.

Position	Residue	Position	Residue	Position	Residue
56	Leu224	169	Ile261	511	Ile312
65	Pro228	171	Pro262	512	Phe313
67	Val229	194	Ile263	522	Thr317
71	Arg231	208	Leu267	523	Arg318
85	Ser232	214	His269	527	Arg320
86	His233	221	Leu271	528	Ser321
89	Ile234	227	Gly273	590	Gly322
90	Asp235	229	Phe274	591	Lys323
91	Tyr236	230	Phe275	598	Ser325
99	Thr240	233	Arg277	600	Ala327
134	Ile248	352	Asp289	609	Leu331
135	Lys249	356	Leu291	611	Leu332
139	Ile253	361	Tyr292	614	Ser333
141	Ala254	364	Arg293	615	Val334
145	Ser255	383	His299	616	Val335
147	Gly256	391	Glu302	657	Thr343
150	Asn257	394	Leu303	660	Asp346
153	Asn258	406	Gln307	665	Leu348
156	-	506	Phe308	668	-
166	Asn260	510	Glu311	671	Ile350

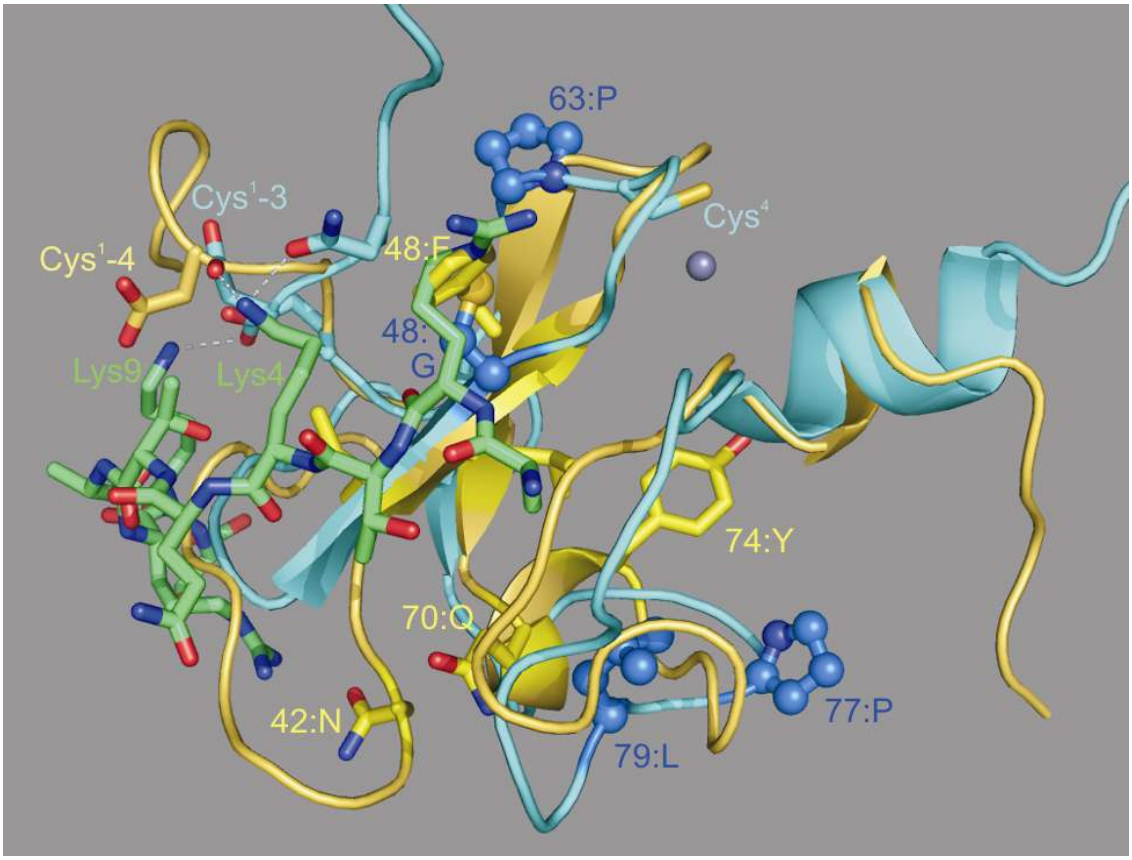


Figure S11. Superposition of the structures of the PHD fingers of AIRE 1 (cyan ribbon; PDB entry 2KE1) and BRPF1 (pale yellow ribbon; predicted, this study). The histone peptide from the AIRE structure is shown as sticks with carbon atoms in green. Conditions for AIRE 1 (family II) are shown as blue ball-and-sticks models, those for BRPF1 (family III) as yellow sticks. The zinc ions from AIRE 1 are shown as grey spheres, and its Cys^4 as sticks. Histone lysine residues are labeled in green, with interactions with AIRE 1 shown as dots. Differences between the two structures are mainly seen at position 48 (Cys^3-1) and by the different locations of the loop that include Cys^4 (top of the image, shift of 2.1 Å between the C_α of the aspartate residues at position Cys^4+1) and that around position 77 (bottom of the image).