Dealing with Small Data: On the Generalization of Context Trees – Supplementary Material –

Ralf Eggeling, Mikko Koivisto, Ivo Grosse

Contents

1 Proof of Theorem 1

Theorem 1. *The expected number of data subsets matched by k-contexts of length ℓ is at most*

$$\binom{n}{r_k - 1}_* + \binom{n}{r_k}\left(1 + \left(\frac{k}{\sigma}\right)^\ell\right)^n,$$

where $r_k = \left\lceil \ln \binom{\sigma}{k} / \ln \frac{\sigma}{k} \right\rceil \leq \left\lceil k\left(1 + \frac{1}{\ln \frac{\sigma}{k}}\right)\right\rceil$.

Proof. For each nonempty data subset S, let X_S be 1 if S is matched by at least one k-context of length ℓ, and 0 otherwise. We show that the expectation $\mathbb{E}[\sum_S X_S]$ has the claimed upper bound.

Using linearity of expectation and the assumption that the data are uniformly distributed, we write $\mathbb{E}[\sum_S X_S]$ as $\sum_{i=1}^n \binom{n}{i} p_{i,\ell}$, where $p_{i,\ell}$ is the probability that $X_S = 1$ given that $|S| = i$.

We next give an upper bound for each $p_{i,\ell}$. To this end, let $x_1, \ldots, x_i \in \Sigma$ be the (random) content of a data subset S of size i in a fixed position j. We say that $(x_1, \ldots, x_i)$ is *covered* by a node $C_j \subseteq \Sigma$ if $\{x_1, \ldots, x_i\} \subseteq C_j$ and $|C_j| \leq k$. Let p_i be the probability that $(x_1, \ldots, x_i)$ is covered by at least one node C_j. Note that the probability is clearly the same for all positions j, and that $p_{i,\ell} = p_i^\ell$. By the union bound we have that $p_i \leq \binom{\sigma}{k}(k/\sigma)^i$. This bound is at most 1 for $i \geq r_k$. For $i < r_k$ we may use the trivial bound $p_i \leq 1$.

We get

$$\mathbb{E}\Big[\sum_S X_S\Big] \;=\; \sum_{i=1}^n \binom{n}{i} p_i^\ell \;\leq\; \binom{n}{r_k - 1}_* + \sum_{i=r_k}^n \binom{n}{i}\left(\binom{\sigma}{k}\left(\frac{k}{\sigma}\right)^i\right)^\ell.$$

It remains to bound the latter term as follows:

$$\begin{aligned}
\sum_{i=r_k}^n \binom{n}{i}\left(\binom{\sigma}{k}\left(\frac{k}{\sigma}\right)^i\right)^\ell &= \sum_{i=0}^{n-r_k} \binom{n}{r_k + i}\left(\binom{\sigma}{k}\left(\frac{k}{\sigma}\right)^{r_k+i}\right)^\ell && (1)\\
&\leq \sum_{i=0}^{n-r_k} \binom{n}{r_k}\binom{n - r_k}{i}\left(\left(\frac{k}{\sigma}\right)^i\right)^\ell && (2)\\
&= \binom{n}{r_k}\left(1 + \left(\frac{k}{\sigma}\right)^\ell\right)^{n-r_k}. && (3)
\end{aligned}$$

The claimed bound now follows because $n - r_k \leq n$.

Finally, the upper bound on r_k follows from the well known bound $\binom{\sigma}{k} \leq \left(\frac{e\sigma}{k}\right)^k$. □

2 Complete running time results

Here, we display all running time results on random data, which are the basis for Section 4.1 of the main manuscript. We study three different tree structures (original PCTs, and 2-GCTs, 2^+-GCTs), and investigate for each of them four different algorithms (basic DP algorithm, enabled fast alphabet partitioning, enabled memoization, and complete enhanced DP algorithm).

For given alphabet size σ and depth d, we sample a sequence of $N + d$ symbols from a uniform distribution, use the $N = 100$ subsequences of length $d + 1$ as context sequences for learning the PCT (or GCT), and measure the running time. We repeat this procedure 10^2 times and take the median of the obtained running times, setting a total time limit of 24 hours. Exceeding it causes the procedure to terminate, resulting in a median of all running times obtained so far.

We visualize the running times of a method using a combination of table and heat map, where the rows of the table correspond to the alphabet size σ, and the columns correspond to the depth d of the PCT. We display the precise median running times in the corresponding cells in seconds, and color cells with a running time of less than a second in yellow. For larger running times the cell color has an increasing content of red according to equivalence classes of problems than run below 1 minute, 1 hour, and 1 day, respectively.

The supplementary figures are related to the main manuscript as follows. Figure 1(a) and Figure 1(d) in this supplement together convey exactly the same results as Figure 4 in the main manuscript. Figure 5(a) and Figure 5(b) in the manuscript display a subset of the results from Figure 2(d) and Figure 3(d) in this supplement.

2.1 PCT

σ	1	2	3	4	5	6	7	8	9	10	11	12
3	0.001	0.002	0.006	0.032	0.13	0.68	4.34	28.5	189	1415	12474	
4	0.001	0.007	0.050	0.57	7.72	108	1792	31653				
5	0.003	0.038	0.84	23.1	714	28316						
6	0.011	0.34	18.6	1294								
7	0.053	4.71	644									
8	0.44	87.4	31248									
9	4.64	2020										
10	39.3	51476										
11	448											
12	8428											

(a) Basic DP algorithm

σ	1	2	3	4	5	6	7	8	9	10	11	12
3	0.002	0.004	0.009	0.029	0.17	0.91	5.83	38.5	257	2188	16325	
4	0.002	0.008	0.062	0.62	8.41	123	1864	34389				
5	0.003	0.031	0.77	19.5	568	13337						
6	0.005	0.12	6.71	478	25345							
7	0.013	0.72	78.1	9752								
8	0.024	3.71	926									
9	0.064	25.2	15802									
10	0.13	123										
11	0.41	993										
12	2.23	6477										
13	6.23	46148										
14	18.4											
15	56.7											
16	177											
17	557											
18	1810											
19	5524											
20	13226											

(b) Fast alphabet partitioning

σ	1	2	3	4	5	6	7	8	9	10	11	12
3	0.001	0.003	0.009	0.025	0.134	0.47	1.74	4.85	13.8	40.7	77.5	345
4	0.001	0.007	0.052	0.49	5.11	34.4	172	1480	5002			
5	0.003	0.044	0.93	21.1	301	3216						
6	0.011	0.36	19.5	723								
7	0.055	5.46	552	44303								
8	0.44	83.61	24654									
9	4.52	1991										
10	41.1	51790										
11	453											
12	7697											

(c) Memoization

σ	1	2	3	4	5	6	7	8	9	10	11	12
3	0.002	0.003	0.008	0.025	0.13	0.56	2.09	6.89	18.5	45.2	99.2	446.5
4	0.002	0.007	0.046	0.53	5.55	38.1	186	1420	2479	11583		
5	0.003	0.039	0.54	14.2	199	2605	17094					
6	0.005	0.12	6.27	279	6420							
7	0.013	0.66	70.2	5183								
8	0.023	3.46	718									
9	0.064	25.5	12055									
10	0.14	126										
11	0.41	704										
12	2.27	5648										
13	5.58	43469										
14	16.2											
15	54.4											
16	161											
17	539											
18	1708											
19	5426											
20	17670											
21	37083											

(d) Enhanced DP algorithm

Figure 1: Running time tables for PCTs on random data.

2.2 2-GCT

σ	1	2	3	4	5	6	7	8	9	10	11	12
3	0.001	0.002	0.005	0.026	0.07	0.39	2.16	12.5	73.5	446	2727	16358
4	0.001	0.006	0.025	0.21	1.92	18.5	187	1905	18660			
5	0.002	0.026	0.25	3.65	50.2	770	11611					
6	0.011	0.14	2.84	62.0	1330	28485						
7	0.054	1.46	39.6	1196	36147							
8	0.44	17.0	648	23748								
9	4.74	215	10275									
10	49.2	2838										
11	563	42828										
12	7994											

(a) Basic DP algorithm

σ	1	2	3	4	5	6	7	8	9	10	11	12
3	0.002	0.004	0.007	0.027	0.092	0.51	2.87	16.9	96.4	552	3559	22240
4	0.002	0.006	0.027	0.23	2.15	21.4	213	2089	45442			
5	0.003	0.024	0.20	2.30	32.2	458	8015					
6	0.005	0.05	0.90	18.7	372	8147						
7	0.012	0.19	4.65	134	3791							
8	0.024	0.70	24.20	878	32061							
9	0.063	3.47	121	5688								
10	0.16	9.78	622	33285								
11	0.50	39.0	2750									
12	2.25	152	11142									
13	6.22	526										
14	18.4	1877										
15	56.7	6352										
16	144											
17	560											
18	1833											
19	5614											
20	12601											

(b) Fast alphabet partitioning

σ	1	2	3	4	5	6	7	8	9	10	11	12
3	0.001	0.002	0.009	0.026	0.077	0.212	0.534	1.20	1.75	3.20	4.07	4.63
4	0.001	0.005	0.12	0.17	0.84	2.35	4.19	6.95	7.23	9.01	10.1	11.6
5	0.003	0.032	0.24	1.81	6.34	11.8	18.4	23.2	28.0	33.0	40.3	43.8
6	0.011	0.15	2.39	19.4	53.4	87.1	116	141.4	163.7	211.8	183.6	214.4
7	0.054	1.55	32.4	205	449	677	875	1071	1159	1318	1453	1565
8	0.45	16.9	428	2236	4216	6099	7872					
9	4.69	208	5653	22557	42885							
10	49.9	2844										
11	563	38364										
12	7660											

(c) Memoization

σ	1	2	3	4	5	6	7	8	9	10	11	12
3	0.002	0.004	0.008	0.029	0.088	0.316	0.71	1.47	2.42	3.71	4.84	5.93
4	0.002	0.006	0.028	0.21	0.96	2.82	4.73	6.47	7.10	8.33	9.94	12.0
5	0.003	0.032	0.17	1.30	5.32	9.98	12.1	14.1	16.3	19.0	20.8	23.2
6	0.005	0.050	0.78	8.21	19.0	29.7	35.3	43.2	49.8	56.3	63.4	69.5
7	0.013	0.18	4.04	25.6	52.2	91.6	112	134	129	147	194	210
8	0.023	0.71	18.6	81.4	190	275	313	376	337	374	538	590
9	0.064	2.77	78.5	318	598	697	959	1172	1370	1544	1748	1899
10	0.18	9.94	296	1031	1687	2351	2969	3670	3504	3958	4373	4827
11	0.51	35.8	1084	3408	5424	7545	10705	10712	12196	14226	16021	16932
12	2.24	143	3904	8905	16552	23627	30663	28676	34249	38593	43197	
13	5.47	443	12588	38676								
14	17.5	1774	46027									
15	54.1	6261										
16	132	18518										
17	529	69080										
18	1661											
19	5444											
20	17874											
21	50107											

(d) Enhanced DP algorithm

Figure 2: Running time tables for 2-GCTs on random data.

2.3 2^+-GCT

σ	1	2	3	4	5	6	7	8	9	10	11	12
3	0.001	0.002	0.006	0.032	0.13	0.86	5.48	35.8	244	1773	12477	
4	0.001	0.006	0.032	0.31	3.19	35.3	301	4527	49466			
5	0.002	0.028	0.31	4.80	77.5	1303	22301					
6	0.011	0.152	3.39	84.1	2023	53511						
7	0.052	1.53	47.1	1702	56750							
8	0.44	17.3	775	32355								
9	4.57	216	12879									
10	49.1	2885										
11	560	43273										
12	8430											

(a) Basic DP algorithm

σ	1	2	3	4	5	6	7	8	9	10	11	12
3	0.002	0.004	0.009	0.03	0.17	1.14	7.54	48.3	320	2172	16357	
4	0.002	0.007	0.035	0.34	3.58	39.4	443	5024				
5	0.003	0.03	0.23	3.18	50.0	782	15403					
6	0.005	0.054	1.12	25.1	566	14157						
7	0.013	0.19	5.71	189	6052							
8	0.023	0.73	30.3	1460	58141							
9	0.065	3.66	159	9809								
10	0.16	9.99	840	56938								
11	0.51	39.3	4109									
12	2.29	154	19502									
13	6.27	529										
14	18.4	1896										
15	56.7	6734										
16	175											
17	571											
18	1824											
19	5595											
20	11927											

(b) Fast alphabet partitioning

σ	1	2	3	4	5	6	7	8	9	10	11	12
3	0.001	0.002	0.008	0.033	0.13	0.57	2.09	6.53	16.3	49.1	100	399
4	0.001	0.006	0.049	0.25	1.53	8.61	30.7	111	286	1051	1503	4483
5	0.002	0.032	0.29	2.70	15.5	65.4	254.6	1353	3672	16975		
6	0.011	0.16	2.91	32.0	197	1203	4124					
7	0.054	1.59	39.3	480	3250	23633						
8	0.44	17.3	542.2	7495	50187							
9	4.68	211	7942									
10	50.4	2904										
11	565	39027										
12	7913											

(c) Memoization

σ	1	2	3	4	5	6	7	8	9	10	11	12
3	0.002	0.004	0.009	0.03	0.16	0.596	2.28	6.96	18.4	55.2	117	526
4	0.002	0.007	0.034	0.30	1.97	8.23	32.6	113	326	1127	1924	5735
5	0.003	0.026	0.21	1.96	13.2	52.8	209	1043	2284	11315		
6	0.005	0.054	0.95	12.5	67.2	454	1762	7058				
7	0.013	0.19	4.84	62.8	423	3244	13125					
8	0.024	0.73	23.7	325	2515	28448						
9	0.064	2.89	103	1507	19308							
10	0.17	10.1	464	9396								
11	0.51	36.3	2206	45451								
12	2.29	143.2	10537									
13	5.63	461	47343									
14	17.9	1807										
15	54.4	6330										
16	161	20074										
17	533	63917										
18	1726											
19	5374											
20	17659											
21	38454											

(d) Enhanced DP algorithm

Figure 3: Running time tables for 2^+-GCTs on random data.

3 Complete memoization results on protein data

We study several well-known proteins of different size and functionality, and extract their sequence from the protein sequence database UniProt [1]: human hormone insulin (Uniprot ID `P01308`, 110 amino acid residues), plant photosynthesis key enzyme RuBisCO (`O03042`, 479 residues), human oxygen-binding proteins myoglobin (`P02144`, 154 residues) and hemoglobin subunit α (abbreviated HG α, `P69905`, 142 residues), human cytoskeleton protein actin (`P68133`, 377 residues), and the green fluorescent protein (abbreviated GFP, `P42212`, 238 residues) from jellyfish. For the representation of these data sets, we use a reduced amino acid alphabet according to the reduction method of Li et al. [2], which offers each desired reduced alphabet size an optimal clustering of amino acids into groups.

We show the effect of memoization w.r.t. the number of visited nodes in the extended tree for computing 2-GCTs, 2^+-GCTs, and original PCTs. In the following (Table 3, Table 5, and Table 1), we show for each of the three structures the results for several non-trivial combinations of alphabet size σ and depth d. For each pair (σ, d), we show the maximal number of nodes in the extended PCT/GCT that have to be visited when memoization is disabled. We then display the fraction of that maximal number that has to be visited when memoization is enabled for the six real world data sets under consideration and random data.

In addition, we also show the raw running times of all combinations of model and data set (Table 4, Table 6, and Table 2) However, since it is here often impossible to solve the problem without memoization at all, the comparison is limited to the running time result on random data from Section 2.

3.1 PCT

Table 1: Number of visited nodes with memoization on original PCTs on protein data.

σ	d	Memoization disabled	Random	RuBisCO	Insulin	Myoglobin	GFP	Actin	HG α
7	3	2.06×10^6	93.25%	100.00%	92.97%	87.06%	94.70%	100.00%	89.14%
8	3	1.66×10^7	90.74%	100.00%	93.75%	84.09%	93.39%	99.22%	89.54%
9	3	1.34×10^8	82.82%	100.00%	92.60%	87.90%	94.71%	99.61%	88.87%
5	4	9.54×10^5	74.42%	96.67%	65.78%	77.95%	81.57%	89.55%	65.15%
6	4	1.60×10^7	62.70%	92.73%	50.41%	49.68%	65.62%	82.18%	55.49%
7	4	2.62×10^8	52.36%	92.15%	43.86%	51.80%	68.02%	85.15%	48.28%
5	5	2.96×10^7	34.45%	80.17%	28.41%	40.69%	53.46%	64.36%	29.90%

Table 2: Running times with memoization on original PCTs on protein data.

σ	d	Random	RuBisCO	Insulin	Myoglobin	GFP	Actin	HG α
7	3	70	75	48	47	56	67	47
8	3	718	1086	794	764	927	1021	822
9	3	12055	9398	7504	7445	8260	9518	7935
5	4	14	26	12	14	18	22	13
6	4	279	492	206	218	317	424	242
7	4	5183	7902	3265	3943	5335	7051	3641
5	5	199	541	157	220	332	420	172

3.2 2-GCT

Table 3: Number of visited nodes with memoization on 2-GCTs on protein data.

σ	d	Memoization disabled	Random	RuBisCO	Insulin	Myoglobin	GFP	Actin	HG α
11	5	1.27×10^9	0.567%	35.020%	2.053%	4.247%	9.520%	22.125%	3.539%
10	6	2.82×10^{10}	0.056%	1.226%	0.067%	0.131%	0.301%	0.765%	0.111%
7	7	1.40×10^{10}	0.018%	0.432%	0.022%	0.044%	0.105%	0.273%	0.036%
5	8	2.75×10^9	0.028%	0.816%	0.035%	0.078%	0.181%	0.487%	0.062%
6	8	3.97×10^{10}	0.004%	0.109%	0.005%	0.010%	0.025%	0.067%	0.009%
5	9	4.12×10^{10}	0.002%	0.076%	0.003%	0.006%	0.016%	0.046%	0.005%

Table 4: Running times in seconds with memoization on 2-GCTs on protein data.

σ	d	Random	RuBisCO	Insulin	Myoglobin	GFP	Actin	HG α
11	5	5425	56057	3578	7425	15767	35823	6208
10	6	2351	32800	1865	3609	8500	20503	2975
7	7	112	2456	85	190	464	1495	159
5	8	14	594	21	55	120	325	38
6	8	43	1484	56	117	286	858	102
5	9	16	957	31	66	148	485	55

3.3 2^+-GCT

Table 5: Number of visited nodes with memoization on 2^+-GCTs on protein data.

σ	d	Memoization disabled	Random	RuBisCO	Insulin	Myoglobin	GFP	Actin	HG α
12	3	4.26×10^7	63.10%	93.48%	59.76%	66.94%	77.73%	86.59%	63.41%
7	4	4.04×10^6	33.55%	76.60%	31.04%	38.10%	51.59%	66.08%	34.43%
8	4	1.78×10^7	26.12%	70.85%	26.48%	31.43%	41.51%	60.45%	31.49%
10	4	2.99×10^9	18.94%	62.86%	18.29%	28.91%	37.74%	53.58%	27.74%
6	5	2.00×10^8	12.79%	43.25%	10.91%	13.11%	22.00%	31.91%	14.00%
8	5	7.51×10^9	4.21%	27.87%	4.88%	7.21%	11.66%	21.61%	6.83%
6	6	4.74×10^9	2.68%	15.53%	2.35%	3.18%	6.39%	10.83%	3.18%
5	7	7.57×10^9	1.40%	12.74%	1.39%	2.56%	4.89%	8.25%	1.82%
4	8	3.97×10^9	1.54%	13.80%	1.73%	2.59%	5.77%	10.38%	2.29%

Table 6: Running times with memoization on 2^+-GCTs on protein data.

σ	d	Random	RuBisCO	Insulin	Myoglobin	GFP	Actin	HG α
12	3	10537	10523	6920	7475	8902	9771	7292
7	4	63	114	41	50	74	102	46
8	4	325	652	227	270	359	577	259
10	4	9396	17481	5711	8755	11081	15017	8418
6	5	67	231	58	68	116	173	70
8	5	25145	9635	1702	2556	4054	7747	2465
6	6	454	1771	244	327	681	1235	322
5	7	209	1744	164	298	616	1164	227
4	8	113	934	85	138	310	669	114

4 Complete prediction results

For the comparison of different types of context trees w.r.t. their predictive performance, we use inhomogeneous PMMs and adopt a hyperparameter-free learning scheme [3], that is, BIC [4] as structure score and fsNML [5] as parameter estimation method. As data sets, we use the CEBP data set of a previously publication [6], for which PCTs have been demonstrated to predict better than CTs, and four additional data sets from the JASPAR database [7], namely DAF-12 from *C. elegans*, BZR1 and PIL5 from *A. thaliana*, and human NR2C2. For all data sets, all structural variants, and different maximal depths $d = 1, \ldots, 7$ we compare the prediction performance using leave-one-out cross validation, and we also include the simple independence model in the comparison. We plot the mean log predictive probability in Figure 4. Error bars indicate double standard error.

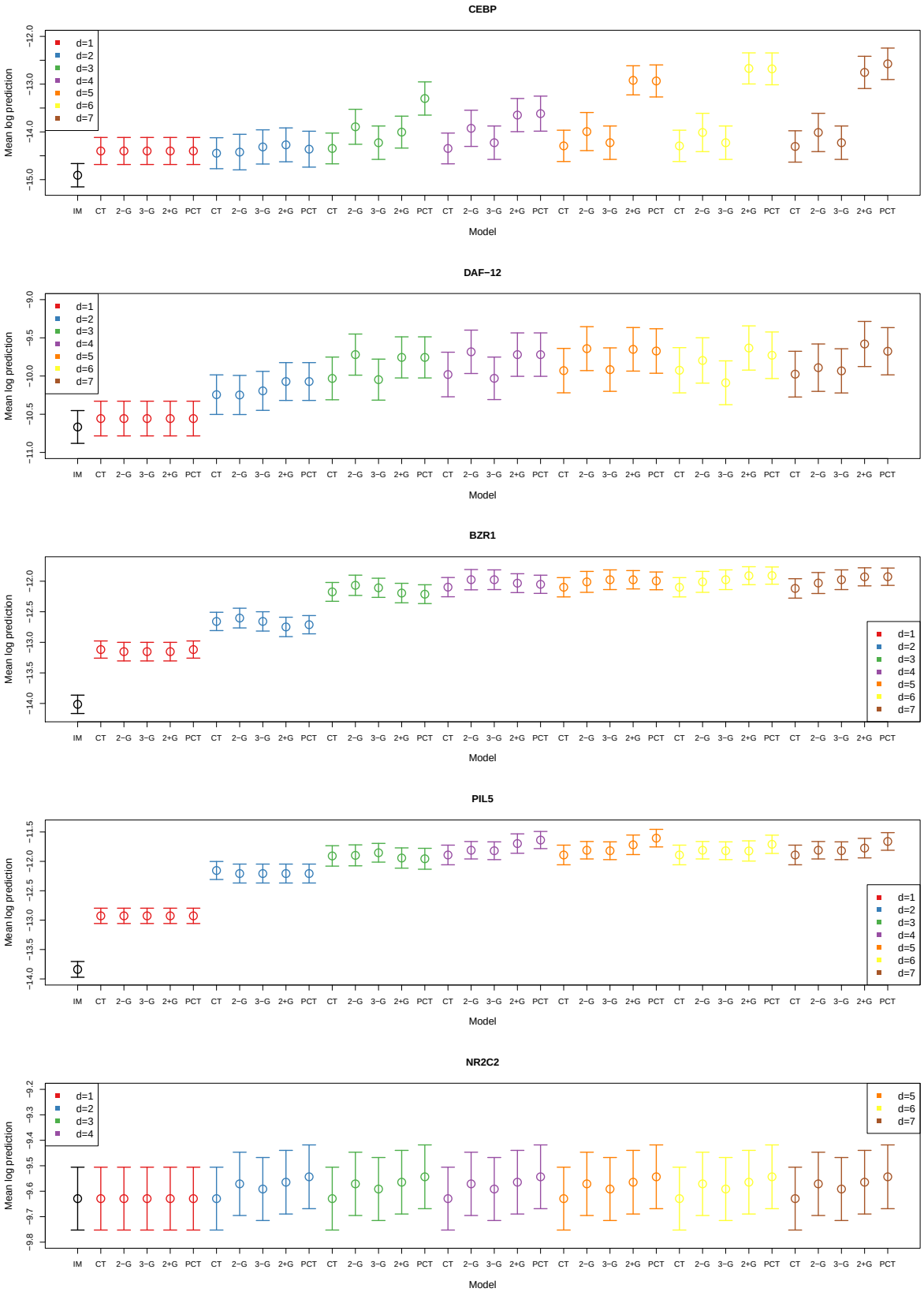

Figure 4: Complete prediction results for five data sets, five different types of context trees (with 2-G, 3-G, and 2+G being abbreviations for 2-GCTs, 3-GCTs, and 2^+-GCTs), and seven different maximal tree depths, supplemented by the result of the plain independence model (IM).

References

[1] The UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research*, 41:D43–D47, 2013.

[2] T. Li, K. Fan, J. Wang, and W. Wang. Reduction of protein sequence complexity by residue grouping. *Protein Engineering*, 16:323–330, 2003.

[3] R. Eggeling, T. Roos, P. Myllymäki, and I. Grosse. Robust learning of inhomogeneous PMMs. In *Proc. AISTATS*, volume 33 of *JMLR: W&CP*, pages 229–237, 2014.

[4] G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 2:461–464, 1978.

[5] T. Silander, T. Roos, and P. Myllymäki. Locally minimax optimal predictive modeling with Bayesian networks. In *Proc. AISTATS*, volume 5 of *JMLR: W&CP*, pages 504–511, 2009.

[6] R. Eggeling, A. Gohr, P.-Y. Bourguignon, E. Wingender, and I. Grosse. Inhomogeneous parsimonious Markov models. In *Proc. ECMLPKDD*, volume 1, pages 321–336. Springer, 2013.

[7] A. Sandelin, W. Alkema, P. Engström, W.W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32:D91–D94, 2004.