Supplementary Materials

Building a Profile Hidden Markov Model for the Kunitz-type protease inhibitor domain

Laia Torres Masdeu[1,*]

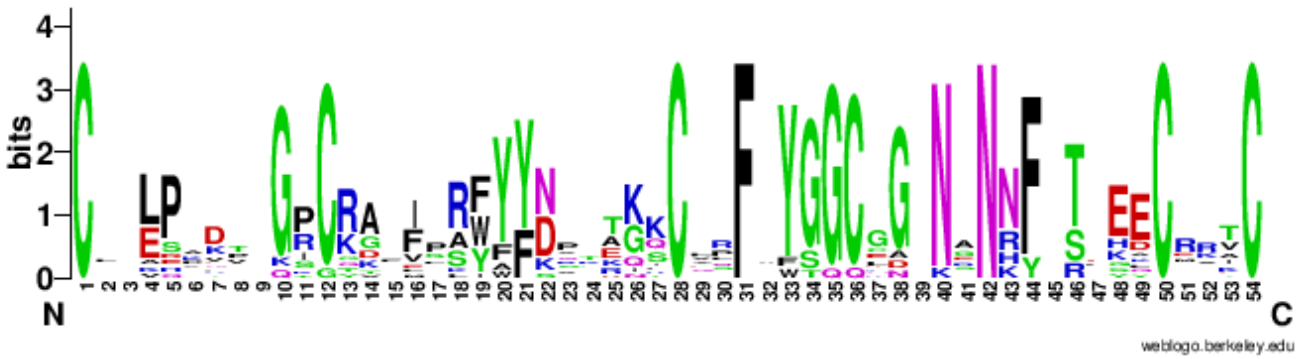
[1]Department of Pharmacology and Biotecnology, Alma Mater Studorum – Università di Bologna, Bologna, Italy

[*]Corresponding author. laia.torresmasdeu@studio.unibo.it

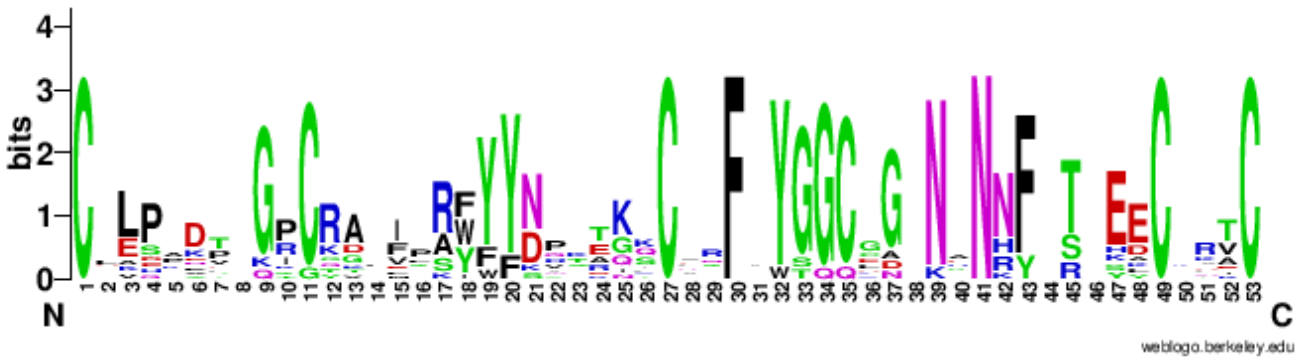
To access the supplementary computational materials (code and file outputs), please go to the following url: `https://github.com/torresmasdeu/kunitz_HMM_project`.

For Supplementary Figures, scroll to the next page.

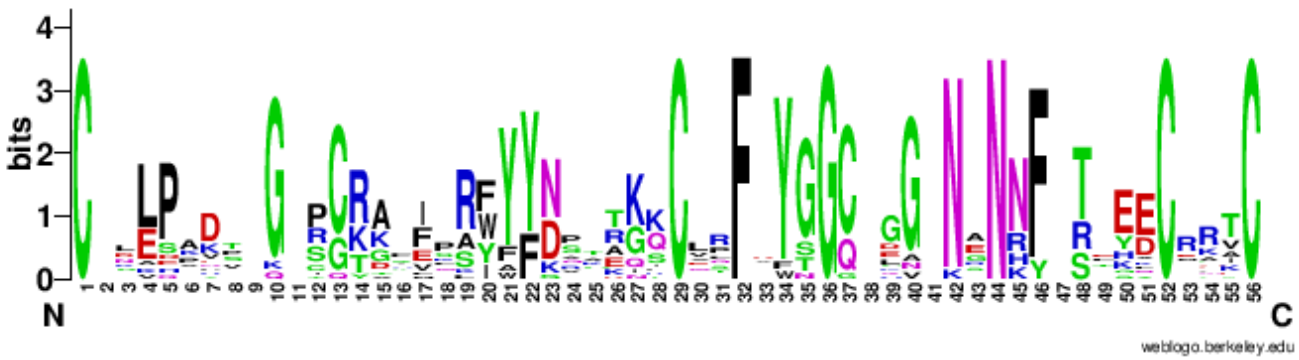
Supplementary Figure 1



(a) Sequence Logo of M2 (CATH). This graphical representation shows the conservation of amino acids at each position in the sequence alignment. Observe how the six cysteines are highly conserved.



(b) Sequence Logo of M3 (SCOP). This graphical representation shows the conservation of amino acids at each position in the sequence alignment. Observe how the six cysteines are highly conserved.



(c) Sequence Logo of M4 (Pfam+CATH+SCOP). This graphical representation shows the conservation of amino acids at each position in the sequence alignment. Observe how the six cysteines are highly conserved.

Supplementary Figure 2

RMSD

structure:		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	PDB 1aap:A		1.147	3.328	0.632	0.891	0.611	0.460	2.399	2.717	0.567	0.883	0.456	0.610	1.464	0.899	0.382	0.828	0.676	0.694	0.593
2	PDB 1bun:B	1.147		3.037	1.208	1.051	1.077	1.194	2.389	2.776	1.106	1.211	1.181	0.974	0.702	1.235	1.066	1.078	1.164	1.076	1.228
3	PDB 1d0d:A	3.328	3.037		3.211	2.966	3.160	3.292	3.572	3.646	3.292	3.171	3.310	3.117	2.879	3.253	3.226	2.991	3.295	3.228	3.424
4	PDB 1dtx:A	0.632	1.208	3.211		0.706	0.573	0.556	2.496	2.756	0.563	0.636	0.533	0.602	1.425	0.750	0.554	0.544	0.875	0.449	0.649
5	PDB 1kth:A	0.891	1.051	2.966	0.706		0.737	0.981	2.491	2.864	0.750	0.931	0.782	0.679	1.204	1.009	0.770	0.666	1.047	0.728	0.920
6	PDB 1yc0:I	0.611	1.077	3.160	0.573	0.737		0.590	2.268	2.694	0.600	0.624	0.578	0.302	1.344	0.797	0.423	0.535	0.606	0.455	0.646
7	PDB 1zr0:B	0.460	1.194	3.292	0.556	0.981	0.590		2.411	2.621	0.637	0.720	0.516	0.677	1.491	0.781	0.441	0.779	0.715	0.616	0.567
8	PDB 2uuy:B	2.399	2.389	3.572	2.496	2.491	2.268	2.411		3.250	2.423	2.495	2.420	2.309	2.477	2.472	2.387	2.445	2.305	2.355	2.455
9	PDB 2w8x:A	2.717	2.776	3.646	2.756	2.864	2.694	2.621	3.250		2.774	2.744	2.763	2.719	2.873	2.813	2.682	2.726	2.750	2.777	2.779
10	PDB 3byb:A	0.567	1.106	3.292	0.563	0.750	0.600	0.637	2.423	2.774		0.892	0.478	0.594	1.400	0.981	0.555	0.701	0.876	0.594	0.654
11	PDB 3m7q:B	0.883	1.211	3.171	0.636	0.931	0.624	0.720	2.495	2.744	0.892		0.794	0.733	1.389	0.523	0.737	0.549	0.869	0.685	0.839
12	PDB 4dtg:K	0.456	1.181	3.310	0.533	0.782	0.578	0.516	2.420	2.763	0.478	0.794		0.626	1.460	0.849	0.485	0.724	0.756	0.600	0.552
13	PDB 4iso:B	0.610	0.974	3.117	0.602	0.679	0.302	0.677	2.309	2.719	0.594	0.733	0.626		1.252	0.884	0.430	0.536	0.641	0.531	0.747
14	PDB 4ntw:B	1.464	0.702	2.879	1.425	1.204	1.344	1.491	2.477	2.873	1.400	1.389	1.460	1.252		1.427	1.387	1.251	1.515	1.345	1.528
15	PDB 4u30:X	0.899	1.235	3.253	0.750	1.009	0.797	0.781	2.472	2.813	0.981	0.523	0.849	0.884	1.427		0.781	0.792	0.888	0.806	0.884
16	PDB 4u32:X	0.382	1.066	3.226	0.554	0.770	0.423	0.441	2.387	2.682	0.555	0.737	0.485	0.430	1.387	0.781		0.679	0.622	0.546	0.570
17	PDB 5m4v:A	0.828	1.078	2.991	0.544	0.666	0.535	0.779	2.445	2.726	0.701	0.549	0.724	0.536	1.251	0.792	0.679		0.938	0.627	0.894
18	PDB 5pti:A	0.676	1.164	3.295	0.875	1.047	0.606	0.715	2.305	2.750	0.876	0.869	0.756	0.641	1.515	0.888	0.622	0.938		0.809	0.757
19	PDB 5yv7:A	0.694	1.076	3.228	0.449	0.728	0.455	0.616	2.355	2.777	0.594	0.685	0.600	0.531	1.345	0.806	0.546	0.627	0.809		0.621
20	PDB 6q61:A	0.593	1.228	3.424	0.649	0.920	0.646	0.567	2.455	2.779	0.654	0.839	0.552	0.747	1.528	0.884	0.570	0.894	0.757	0.621	

(a) RMSD values of the first MSA performed for M2 (CATH). This table shows, in yellow, those RMSD values higher than 2Å, which correspond to entries 2UUY and 2W8X.

RMSD

structure:		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	PDB 1aap:A		1.147	3.328	0.632	0.891	0.611	0.460	2.399	2.717	0.567	0.883	0.456	0.610	1.464	0.899	0.382	0.828	0.676	0.694	0.593
2	PDB 1bun:B	1.147		3.037	1.208	1.051	1.077	1.194	2.389	2.776	1.106	1.211	1.181	0.974	0.702	1.235	1.066	1.078	1.164	1.076	1.228
3	PDB 1d0d:A	3.328	3.037		3.211	2.966	3.160	3.292	3.572	3.646	3.292	3.171	3.310	3.117	2.879	3.253	3.226	2.991	3.295	3.228	3.424
4	PDB 1dtx:A	0.632	1.208	3.211		0.706	0.573	0.556	2.496	2.756	0.563	0.636	0.533	0.602	1.425	0.750	0.554	0.544	0.875	0.449	0.649
5	PDB 1kth:A	0.891	1.051	2.966	0.706		0.737	0.981	2.491	2.864	0.750	0.931	0.782	0.679	1.204	1.009	0.770	0.666	1.047	0.728	0.920
6	PDB 1yc0:I	0.611	1.077	3.160	0.573	0.737		0.590	2.268	2.694	0.600	0.624	0.578	0.302	1.344	0.797	0.423	0.535	0.606	0.455	0.646
7	PDB 1zr0:B	0.460	1.194	3.292	0.556	0.981	0.590		2.411	2.621	0.637	0.720	0.516	0.677	1.491	0.781	0.441	0.779	0.715	0.616	0.567
8	PDB 2uuy:B	2.399	2.389	3.572	2.496	2.491	2.268	2.411		3.250	2.423	2.495	2.420	2.309	2.477	2.472	2.387	2.445	2.305	2.355	2.455
9	PDB 2w8x:A	2.717	2.776	3.646	2.756	2.864	2.694	2.621	3.250		2.774	2.744	2.763	2.719	2.873	2.813	2.682	2.726	2.750	2.777	2.779
10	PDB 3byb:A	0.567	1.106	3.292	0.563	0.750	0.600	0.637	2.423	2.774		0.892	0.478	0.594	1.400	0.981	0.555	0.701	0.876	0.594	0.654
11	PDB 3m7q:B	0.883	1.211	3.171	0.636	0.931	0.624	0.720	2.495	2.744	0.892		0.794	0.733	1.389	0.523	0.737	0.549	0.869	0.685	0.839
12	PDB 4dtg:K	0.456	1.181	3.310	0.533	0.782	0.578	0.516	2.420	2.763	0.478	0.794		0.626	1.460	0.849	0.485	0.724	0.756	0.600	0.552
13	PDB 4iso:B	0.610	0.974	3.117	0.602	0.679	0.302	0.677	2.309	2.719	0.594	0.733	0.626		1.252	0.884	0.430	0.536	0.641	0.531	0.747
14	PDB 4ntw:B	1.464	0.702	2.879	1.425	1.204	1.344	1.491	2.477	2.873	1.400	1.389	1.460	1.252		1.427	1.387	1.251	1.515	1.345	1.528
15	PDB 4u30:X	0.899	1.235	3.253	0.750	1.009	0.797	0.781	2.472	2.813	0.981	0.523	0.849	0.884	1.427		0.781	0.792	0.888	0.806	0.884
16	PDB 4u32:X	0.382	1.066	3.226	0.554	0.770	0.423	0.441	2.387	2.682	0.555	0.737	0.485	0.430	1.387	0.781		0.679	0.622	0.546	0.570
17	PDB 5m4v:A	0.828	1.078	2.991	0.544	0.666	0.535	0.779	2.445	2.726	0.701	0.549	0.724	0.536	1.251	0.792	0.679		0.938	0.627	0.894
18	PDB 5pti:A	0.676	1.164	3.295	0.875	1.047	0.606	0.715	2.305	2.750	0.876	0.869	0.756	0.641	1.515	0.888	0.622	0.938		0.809	0.757
19	PDB 5yv7:A	0.694	1.076	3.228	0.449	0.728	0.455	0.616	2.355	2.777	0.594	0.685	0.600	0.531	1.345	0.806	0.546	0.627	0.809		0.621
20	PDB 6q61:A	0.593	1.228	3.424	0.649	0.920	0.646	0.567	2.455	2.779	0.654	0.839	0.552	0.747	1.528	0.884	0.570	0.894	0.757	0.621	

(b) RMSD values of the first MSA performed for M2 (CATH). This table shows, in yellow, those RMSD values higher than 3Å, which correspond to entry 1D0D.

Supplementary Figure 3

```
>> sp|D3GGZ8|BLI5_HAECO  Kunitz-type protein bli-5 OS=Haemonchus contortus OX=6289 GN=bli-5 PE=3 SV=1
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 ?   15.2   0.4   0.00043      0.63      16      51 .]     156     190 ..     148     190 .. 0.83

  Alignments for each domain:
  == domain 1  score: 15.2 bits;  conditional E-value: 0.00043
            kunitz_mod  16 rfYYnakekkCerFvYgGcgg.NeNnFkteeeCrrtC 51
                           r+ ++    +C++F +   +   +NnFkt+ +C+++C
  sp|D3GGZ8|BLI5_HAECO 156 RWGFDG--SECVKFKWDPEKPsSANNFKTKLQCESYC 190
                           555554..47******9877679**********9 PP

>> sp|O62247|BLI5_CAEEL  Kunitz-type protein bli-5 OS=Caenorhabditis elegans OX=6239 GN=bli-5 PE=1 SV=1
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 ?   14.8   0.8   0.00055       0.8      13      51 .]     147     184 ..     135     184 .. 0.83

  Alignments for each domain:
  == domain 1  score: 14.8 bits;  conditional E-value: 0.00055
            kunitz_mod  13 fikrfYYnakekkCerFvYgGcgg.NeNnFkteeeCrrtC 51
                           +++r+ ++   ++C +F +   +   +NnFkt  +C+++C
  sp|O62247|BLI5_CAEEL 147 YLSRWGFDG--EQCIEFKWNPERPsSANNFKTRAHCEDYC 184
                           567777765..68******9977779**********9 PP
```

(a) hmmsearch| output of O62247 and D3GGZ8 UniProtIDs. 36 and 39 residues of the proteins (respectively) have aligned (to some extent) to the model, out of 51 positions that it has. Moreover, among these partially aligned residues, 2 of the 6 C that are crucial for the Kunitz domain folding, are not present in the protein sequences.