Isofunctional Protein Subfamily Detection using Data Integration and Spectral Clustering

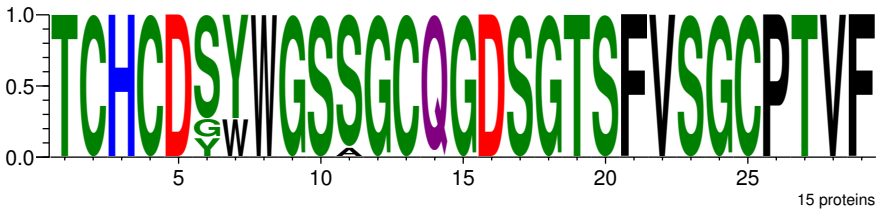
Elisa Boari de Lima[1,2,*], Wagner Meira Júnior[2], Raquel Cardoso de Melo-Minardi[2]

1 Department of Biochemistry and Immunology, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil
2 Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

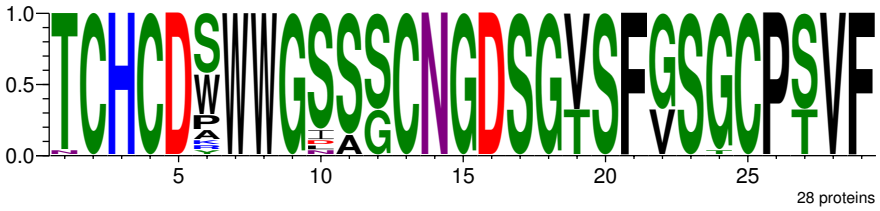
*** eblima@dcc.ufmg.br**

S9 Text: Dividing the serine proteases into eleven clusters

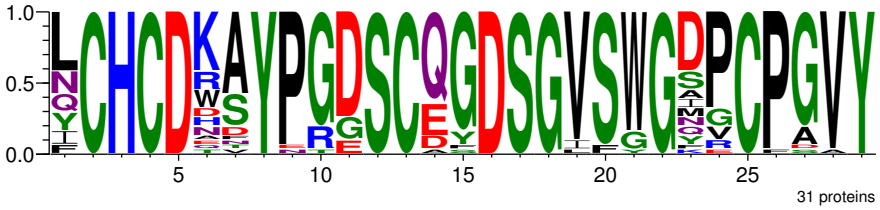
The second level of the hierarchical clustering produced by ASMC divided the serine proteases into eleven clusters, whose logos and compositions according to subfamily labels are presented in Fig. S9.1. This clustering has MI = 10.59.



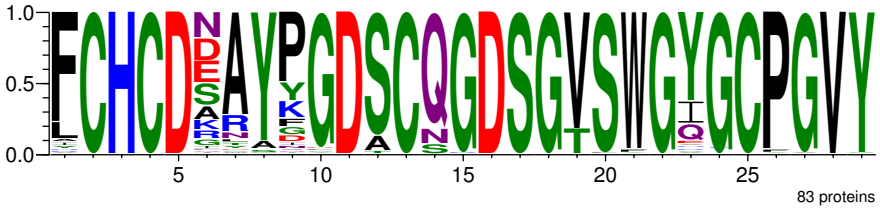
(a) Cluster I.A: 15 elastases



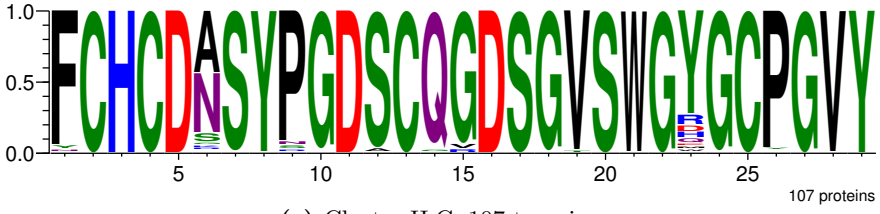
(b) Cluster I.B: 28 elastases



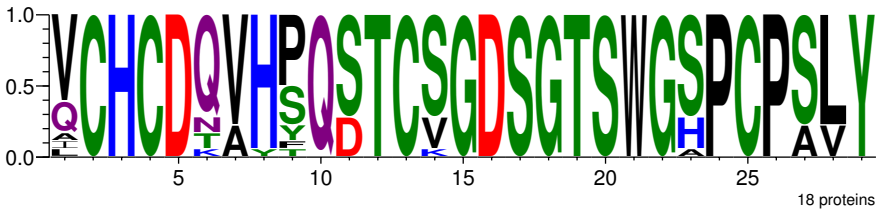
(c) Cluster II.A: 31 trypsins



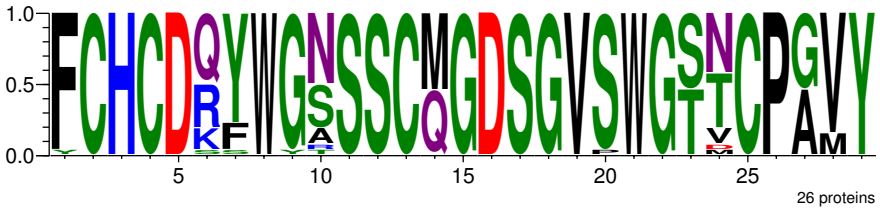
(d) Cluster II.B: 83 trypsins



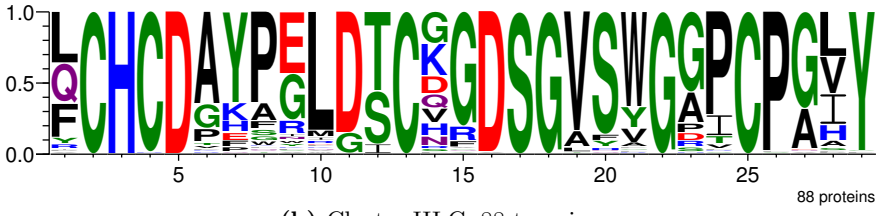
(e) Cluster II.C: 107 trypsins



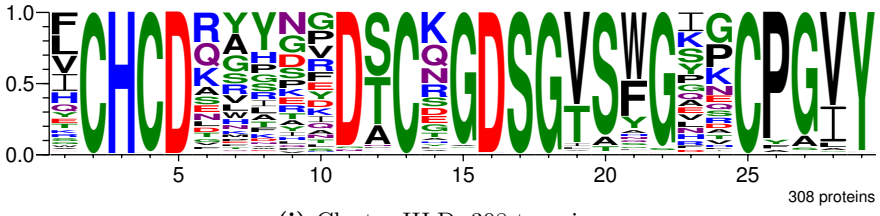
(f) Cluster III.A: 13 kallikreins and 5 trypsins



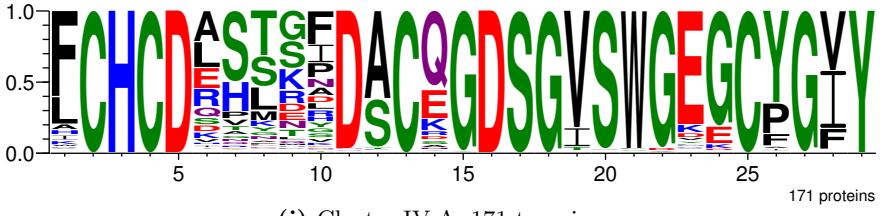
(g) Cluster III.B: 26 chymotrypsins



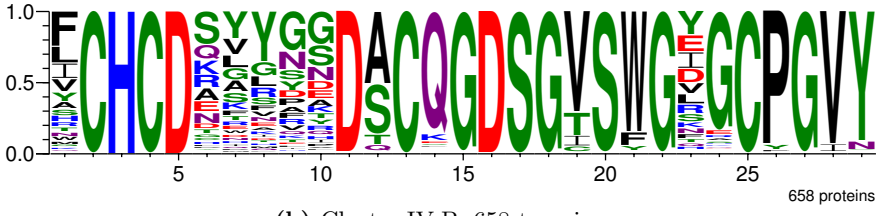
(h) Cluster III.C: 88 trypsins



(i) Cluster III.D: 308 trypsins



(j) Cluster IV.A: 171 trypsins



(k) Cluster IV.B: 658 trypsins

Figure S9.1. Serine protease division into eleven clusters in the second level of ASMC's hierarchical clustering.

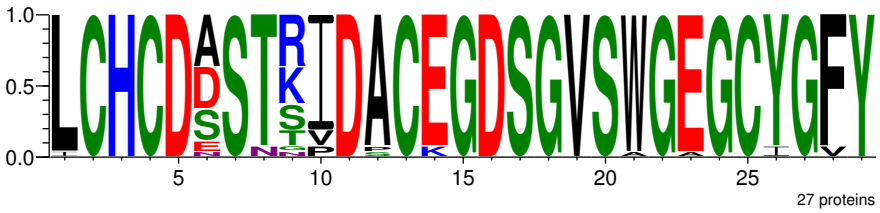
Given ASMC's criteria for considering as specificity determining positions (SDPs) those positions with p-values smaller than 0.0001 [1], the SDPs per cluster for this clustering are presented in Table S9.1, in which one may notice the presence of the known SDPs for this family, which were listed in the S8 Text.

Table S9.1. Cluster SDPs for the eleven serine protease clusters produced by ASMC.

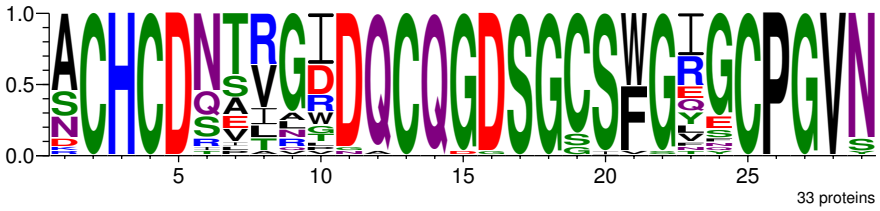
Cluster	Positions
I.A	$\mathbf{11_{189}}$, $\mathbf{22_{216}}$, $\mathbf{27_{226}}$, 29_{173}
I.B	$\mathbf{8_{172}}$, $\mathbf{11_{189}}$, 14_{192}, $\mathbf{27_{226}}$, 29_{173}
II.A	-
II.B	-
II.C	7_{171}, 9_{173}
III.A	$\mathbf{11_{189}}$, $\mathbf{27_{226}}$, 28_{227}
III.B	$\mathbf{11_{189}}$, 14_{192}
III.C	$\mathbf{8_{172}}$, 10_{174}, 24_{219}, 28_{227}
III.D	21_{215}, 24_{219}
IV.A	23_{217}, 26_{225}, 28_{227}
IV.B	12_{190}, 14_{192}, 28_{227}

Listed in order of active site position. Positions in bold correspond to known SDPs. Subscripted positions correspond to those in PDB structure 5PTP:A.

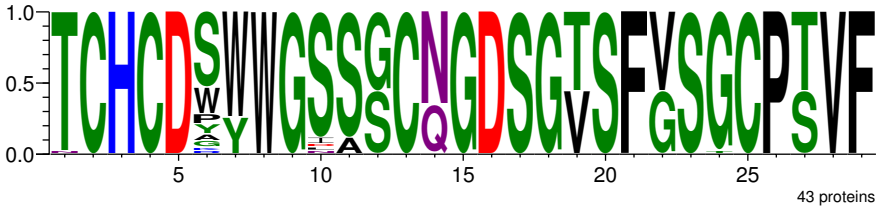
The best result obtained by the GP system for eleven clusters has MI = 12.09 and uses a combination of three data types, as shown in the main text. Cluster logos and compositions are presented in Fig. S9.2, while the residues which most distinguish each cluster are listed in S9.2, in which one can note the presence of the known SDPs for the cases in which the residues occurring in such positions distinguish the corresponding cluster from the others.



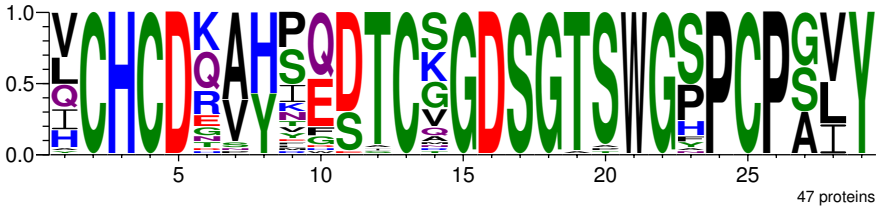
(a) Cluster I: 27 trypsins



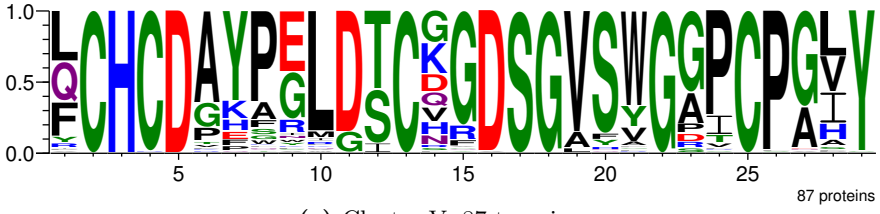
(b) Cluster II: 33 trypsins



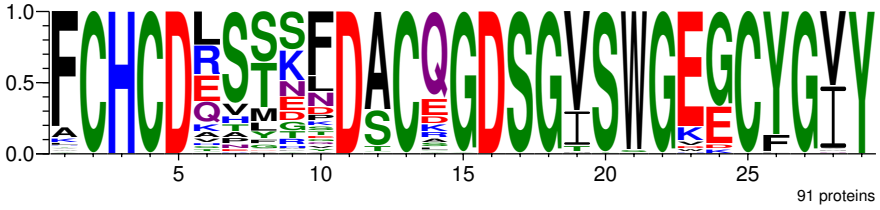
(c) Cluster III: 43 elastases



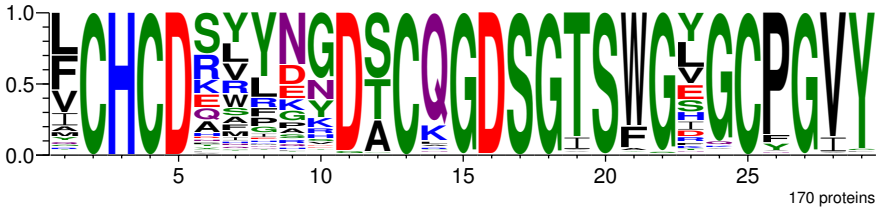
(d) Cluster IV: 34 trypsins and 13 kallikreins



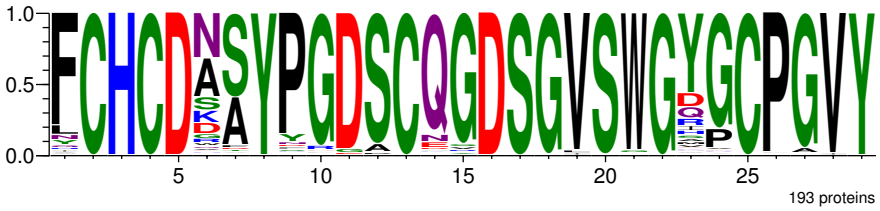
(e) Cluster V: 87 trypsins



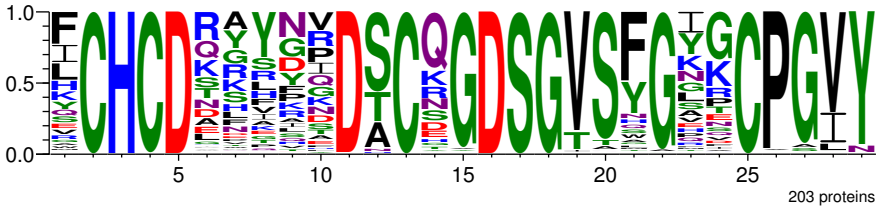
(f) Cluster VI: 91 trypsins



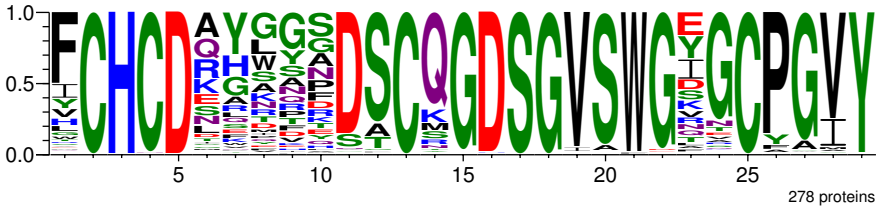
(g) Cluster VII: 170 trypsins



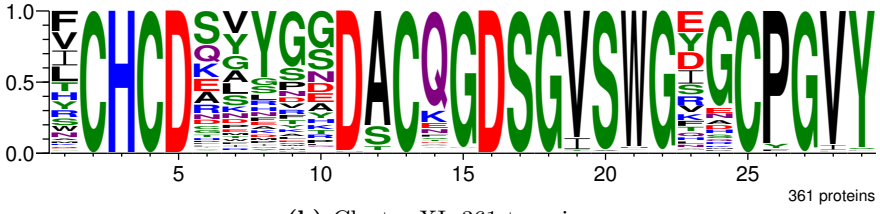
(h) Cluster VIII: 193 trypsins



(i) Cluster IX: 203 trypsins



(j) Cluster X: 252 trypsins and 26 chymotrypsins



(k) Cluster XI: 361 trypsins

Figure S9.2. Serine protease division into eleven clusters by the GP system.

Table S9.2. Most important residues for the eleven serine protease clusters produced by the GP system.

Cluster	Residues
I	F28$_{227}$, E14$_{192}$, **T8$_{172}$**, I10$_{174}$, Y26$_{225}$, L1$_{41}$, E23$_{217}$, **S8$_{172}$**
II	Q12$_{190}$, C19$_{213}$, N29$_{229}$, A1$_{41}$
III	F29$_{229}$, **W8$_{172}$**, T1$_{41}$, **S11$_{189}$**, F21$_{215}$, W7$_{171}$, **V22$_{216}$**, S10$_{174}$, **T27$_{226}$**
IV	P24$_{219}$, T19$_{213}$, T12$_{190}$, **H8$_{172}$**
V	L10$_{174}$, **P8$_{172}$**, P24$_{219}$, G23$_{217}$, A6$_{170}$, E9$_{173}$, Y7$_{171}$
VI	Y26$_{225}$, E23$_{217}$, F10$_{174}$, S8$_{171}$
VII	T19$_{213}$, N9$_{173}$
VIII	P9$_{173}$, G10$_{174}$, **Y8$_{172}$**, S8$_{171}$, F1$_{41}$, S12$_{190}$, Y23$_{217}$
IX	F21$_{215}$, Y21$_{215}$
X	S12$_{190}$, V19$_{213}$, W21$_{215}$
XI	A12$_{190}$, W21$_{215}$, G9$_{173}$

Listed in decreasing order of partial MI value. Residues in bold correspond to known SDPs. Subscripted positions correspond to those in PDB structure 5PTP:A.

References

1. Melo-Minardi RC, Bastard K, Artiguenave F. Identification of subfamily-specific sites based on active sites modeling and clustering. Bioinformatics. 2010 Dec;26(24):3075–3082.