# 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


## S10 Text: Dividing the serine proteases into twelve clusters

The best result found by the genetic programming (GP) system for dividing the serine proteases into twelve clusters is obtained using equation 3 ASid +2 ASscr + go. Cluster logos and compositions according to subfamily labels are presented in Fig. S10.1. One may observe that a single elastase cluster is maintained, along with the interesting kallikrein and prothrombin clusters mentioned in the paper. Additionally, the GP system was able to isolate the chymotrypsins in their own cluster. The residues most important to distinguish each cluster are listed in Table S10.1, in which we note the presence of the known SDPs for the family, listed in the S 8 Text , for the cases in which the residues occurring in such positions distinguish the corresponding cluster from the others.


(g) Cluster VII: 88 trypsins

(h) Cluster VIII: 161 trypsins

(i) Cluster IX: 191 trypsins

(j) Cluster X: 192 trypsins

(k) Cluster XI: 303 trypsins

(1) Cluster XII: 341 trypsins

Figure S10.1. Serine protease division into twelve clusters by the GP system.

Table S10.1. Most important residues for the twelve serine protease clusters produced by the GP system.

| Cluster | Residues |
| :---: | :---: |
| I | S11 $\mathbf{1 8 9}^{\text {, }} \mathbf{W} \mathbf{8 1 7 2}_{\mathbf{1 7 2}}$, M14 ${ }_{192}$, G9 $9_{173}$ |
| II | $\mathrm{Q} 12_{190}, \mathrm{C} 19_{213}, \mathrm{~N} 29_{229}, \mathrm{A1} 1_{41}$ |
| III |  |
| IV |  |
| V | $\mathrm{P} 24_{219}, \mathrm{~T} 12_{190}, \mathrm{~T} 19_{213}, \mathbf{H 8} \mathbf{1 7 2}$ |
| VI | Y26 $2_{25}, \mathrm{E} 23_{217}, \mathrm{~F} 10_{174}, \mathrm{~S} 7_{171}, \mathrm{~F} 1_{41}, \mathbf{S} 9_{172}$ |
| VII | $\mathrm{L} 10_{174}, \mathbf{P} 8_{172}, \mathrm{P} 24_{219}, \mathrm{G} 23_{217}, \mathrm{~A} 6_{170}, \mathrm{E} 9_{173}, \mathrm{Y} 7_{171}$ |
| VIII | $\mathrm{I} 28_{227}, \mathbf{L 8} \mathbf{8}_{172}$ |
| IX | Y8 $\mathbf{1 7 2}^{2}, \mathrm{~V} 7_{171}, \mathrm{S6} \mathbf{1 7 0}$ |
| X | $\mathrm{P} 9_{173}, \mathrm{G10} 1_{174}, \mathbf{Y} 8_{172}, \mathrm{F1}_{41}, \mathrm{Y} 23_{217}, \mathrm{~S} 7_{171}, \mathrm{~S} 12_{190}$ |
| XI | S12 $1_{190}$, F21 ${ }_{215}$ |
| XII | A12 190 |

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.

