

Protein structural robustness to mutations: an *in silico* investigation

Mounia Achoch^a, Rodrigo Dorantes-Gilardi^b, Chris Wymant^c, Giovanni Feverati^d, Kave

Salamatian^a, Laurent Vuillon^b and Claire Lesieur^{e†}

^a *Laboratoire d'informatique systèmes, traitement de l'information et de la connaissance (LISTIC), Université de Savoie, Annecy le Vieux, France*

^b *Laboratoire de mathématiques (LAMA UMR 5127), Université Savoie Mont Blanc, CNRS, Le Bourget du Lac, France*

^c *Medical Research Council Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom*

^d *Federation de recherche Fr3405, Modelisation, Simulations, Interactions Fondamentales, Annecy-le-Vieux, France*

^e *CNRS-UCBL, IXXI-ENS-Lyon, Laboratoire AMPERE, Lyon, France*

† *Corresponding author: claire.lesieur@ens-lyon.fr*

The electronic supplementary Information contains: supplementary method, supplementary Figure S1 and the Amino Acid rank pseudo-code.

Supplementary methods.

Fold X. The run parameters are as followed. The PDB of CtxB5 is 1EEI and only two chains (Chains D and E) are considered for generating the mutation. Backbone and side chain atoms were allowed to move as well as neighbour atoms upon mutation (parameter `moveNeighbours` set to true). The “`pdbHhydrogens`” option was set to false so that hydrogen atoms were not included in the PDB output, because the positions of hydrogen atoms are not always calculated in a X-ray structure. Temperature was set to 298 K, the pH to 7.0, the ionic strength to 0.05M. The “`crystalwaters`” parameter was set to true, so crystallographic water bridges are considered in the PDB output if available in the crystal. `OutPDB` was set to true in order to generate a mutated PDB. `Complex with DNA` was set to false.



Supplementary Figure S1. Spheres of influence as seen on the X-ray structures of the 58 mutants. Two adjacent chains of the CtxB₅ are shown in strands, the interface is shown in dark and light grey ribbons to distinguish both chains (PDB WT 1EEI). The hotspots modified by the mutation are shown in spacefill, the mutated residue in red. The mutations are ordered per type of amino acids and decreasing arankr values within each type.

Amino Acid Rank Pseudocode.

The following procedure is used to define the hotspot network $G = (V, E)$ where V and E are the vertice set (also called hotspot node set) and the edge set (also called link set) respectively and the weighthed adjacency matrix $W = (w_{i,j})$ and the adjacency matrix $A = (a_{i,j})$.

```

Function SpectralPro(Protein PDB).
for Chain in Protein do
  for Residue i in Chain do
    for Residue j in  $\{Protein \setminus Chain\}$  do
       $w = 0$ 
       $w =$  Cardinality of the set  $\{(atom, atom') \mid atom \in Residue\ i, atom' \in$ 
Residue j and  $dis(atom, atom') \leq 5\text{\AA}\}$ 
      if  $w > 0$  then
         $Residue\ i \in V, Residue\ j \in V$  and  $edge(Residue\ i, Residue\ j) \in$ 
 $E, w_{i,j} = w$  and  $a_{i,j} = 1$ 
      else  $w_{i,j} = 0$  and  $a_{i,j} = 0$ 
      end if
    end for
  end for
end for
return  $G, W, A$ 

```

In the algorithm *Residue j* in $\{Protein \setminus Chain\}$ means *Residue j* belongs to a different chain than *Residue i*.

Next, we present the procedure to define the difference matrix D_r in order to compute the structural changes induced by a single hotspot mutation at position r . The terminology *wt* and *mut* is used to refer to the wild-type-protein weighted adjacency matrix W^{wt} and the mutated-protein adjacency matrix W_r^{mut} associated with a mutation of residue in position r .

```

Function Arank( $W^{wt}, W_r^{mut}$ ).
for ( $i, j$ ) in  $N \times N$  do
     $d_{i,j} = w_{i,j}^{mut} - w_{i,j}^{wt}$ 
end for
 $arank_r = \sum_{i,j} |d_{i,j}|$ 
 $local_r = \sum_j |d_{r,j}|$ 
return  $D_r, arank_r, local_r$ 

```

Finally, we state the procedure to define the function backup, $B = (b_{i,j})$ where $(i, j) \in E(G^{wt})$ (The set of edges of the wild type protein graph). Be careful the set of nodes C_1 and C_2 are amino acids along 2 given chains of the whole protein. The PDB of the protein gives the numbering from the first to the last amino acid of the protein. And we define σ the renumbering of amino acid in the hotspot network from 1 to N . Namely a hot spot amino acid in position n in the whole sequence has position $\sigma(n)$ in the hot spot networks. If we have a non hot spot residue m in C_1 then the value $a_{\sigma(m),\sigma(n)}^{wt}$ is 0 in the sum below or if we have a non hot spot residue n in C_2 then the value $a_{\sigma(m),\sigma(n)}^{wt}$ is 0 in the sum below .

Function Backup(*ProteinPDB*, G^{wt}).

for (i, j) in $E(G^{wt})$ **do**

$C_1 = \{\text{Residue } i - 4, \dots, \text{Residue } i, \dots, \text{Residue } i + 4 \mid \sigma(\text{Residue } i) \in V(G^{wt})\}$

$C_2 = \{\text{Residue } j - 4, \dots, \text{Residue } j, \dots, \text{Residue } j + 4 \mid \sigma(\text{Residue } j) \in V(G^{wt})\}$

$b_{i,j} = -1 + \sum_{m \in C_1, n \in C_2} a_{\sigma(m),\sigma(n)}^{wt}$

end for

return B

Remark that residues i and j belong to different chains and therefore $C_1 \cap C_2 = \emptyset$ for any two vertices i, j .