PocHunter: A New Pocket Clustering Scheme

We propose a new pocket searching scheme, named as PocHunter, which can easily find the ATP-binding pocket(s) based on the set (**S**) of the predicted ATP-binding residues. The PocHunter scheme uses two sub-algorithms to cluster the binding residues: (a) PocHunter can firstly employ a SRH (special residues hunter) to search the special residue set, denoted as **R**, where the distance between any two special residues is consistently greater than the distance threshold, $T_{Cluster}$. (b) The clustering scheme then divides the predicted ATP-binding residues into |**R**| clusters based on a simply assignment algorithm, named as ASSIGN.

The process of the proposed SRH is described as follows:

Step I: Searching the maximum distance (d_{max}) between any two residues in **S**, and the corresponding two residues, which are denoted as r_i and r_j. If d_{max} is greater than $T_{Cluster}$, the two special residue r_i and r_j will be put into the special residue set **R**, and then go to Step II. Otherwise, only r_i is put in **R** and the SRH program will terminate.

Step II: Firstly, searching all the other special residues $\{r_k\}$ in **S**, where the distance between each r_k and each residue in **R** is always larger than $T_{Cluster}$. We then select the best residue r_k from $\{r_k\}$ which can achieve the maximum average distance against all residues in **R**. Lastly, we update the special residue set **R**, i.e., putting the selected r_k to **R**. Notice that if there is no other new special residues can be put in **R**, SRH program will terminate.

Step III: Repeating Step II until no other special residue could be selected to update **R**.

Algorithm 1 summarizes the special residues hunter (SRH). In this study, the clustering threshold $T_{Cluster}$ is set to 39.4Å, which is calculated based on the average ATP-binding pocket diameter.

The process of the ASSIGN is described as follows:

In order to conveniently describe the ASSIGN algorithm, we denote the special residue set **R** as follows:

$$\mathbf{R} = \{r_1^s, r_2^s, \cdots, r_{|\mathbf{R}|}^s\} \quad (1)$$

where r_i^s represents the *i*-th special residue. Then, ASSIGN will divide the residues in **S** into |**R**| clusters as follows:

For each residue *r* in **S**, ASSIGN will search the best special residue from **R**, which closest to the residue *r*. Supposing r_i^s is the best special residue, the residue *r*

will be then assigned to the *i*-th cluster (or pocket). ASSIGN will terminate until all residues in **S** are assigned.

Algorithm 1	Special Residues Hunter (SRH)
Input	**S**: the set of predicted ATP-binding residues; $T_{Cluster}$: the threshold parameter for hunting.
Output	**R**: a set of the special residues, different residue belongs to different pockets.
1	Searching the maximum distance (d_{max}) between any two residues in S, and the two corresponding residues (r_i and r_j).
2	**R** $\leftarrow \{r_i\}$
3	**IF** $d_{max} > T_{Cluster}$
4	**R** $\leftarrow$ **R** $\cup \{ r_j\}$
5	**WHILE TRUE**
6	$d_{ave} \leftarrow 0$, $r_k \leftarrow r_i$
7	**FOR** each residue r_x in **S**
8	**D** $\leftarrow \varnothing$
9	**FOR** each residue r_y in **R**
10	Calculate the distance (d) between r_x and r_y.
11	**IF** $d < T_{Cluster}$
12	**Break**
13	**END IF**
14	**D** $\leftarrow$ **D** $\cup \{d\}$
15	**END FOR**
16	**IF** $\mid$ **D** $\mid = \mid$ **R** $\mid$ and $average(\mathbf{D}) > d_{ave}$
17	$d_{ave} \leftarrow average(\mathbf{D})$ and $r_k \leftarrow r_x$
18	**ELSE**
19	**break**
20	**END IF**
21	**END FOR**
22	**IF** $d_{ave} \neq 0$
23	**R** $\leftarrow$ **R** $\cup \{r_k\}$
24	**ELSE**
25	**break**
26	**END IF**
27	**END WHILE**
28	**END IF**
29	**RETURN R**