Search and alignment

Fabio Zanini

June 4, 2012

The given sequence is taken from the PDB entry 1YJE (http://www.rcsb.org/pdb/explore/explore.do?structureId=1YJE). The paper that describes it is Ref. [1]. I checked this by diff the text files, exactly the same label even.

CSBlasting it against the non-redundant PDB using Lupas's toolkit reports quite a few sequences with a reasonably low e-value (number of entries that are reported by chance, according to Ref. [2]). CSBlast also outputs an MSA of the sequences, so I could check directly how closely related they were to our "unknown" sequence:

1. the first sequence is ours, found back from the PDB;
2. the next ones, up to $e \sim 0.01$, have reasonable chunks in common with our sequence;
3. the remaining ones have only a domain in common, either at the N-terminus or in the middle.

I therefore selected only the first ones ($e < 0.01$) and aligned them using MUSCLE with 6 iterations. It took 3 seconds, meaning that they were already aligned `:-)`. Now we should go for modeller.

References

[1] Ralf Flaig, Holger Greschik, Carole Peluso-Iltis, and Dino Moras. Structural basis for the Cell-Specific activities of the NGFI-B and the nurr1 Ligand-Binding domain. *Journal of Biological Chemistry*, 280(19):19250–19258, May 2005.

[2] S Karlin and S F Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences*, 87(6):2264 –2268, March 1990.