生物資訊學-第四次作業

選擇的蛋白質序列:

這段序列來自人類 FUT1 基因所編碼的蛋白，**a-1,2 fucosyltransferase**，屬於醣基轉移酶類，具有很強的生物功能。

```
>sp|XPC16722.1| a-1, 2 fucosyltransferase [Homo sapiens]
MWLRSHRQLCLAFLLVCVLSVIFFLHIHQDSFPHGGLGLSILCPDRRLVTPPVAIFCLPGTAMGPNASSSCPQHPASLSG
TWTVYPNGRFGNQMGQYATLLALAQLNGRRAFILPAMHAALAPVFRITPLVLAPEVDSRTPWRELQLHDWMSEEYADLRD
PFLKLSGFPCSWTFLHHLREQIRREFTLHDHLREEAQSVLQGLRLGRTGDRPRTFVGVHVRRGDYLQVMPQRWKGVGDSA
YLRQAMDWFRARHEAPVFVTNSNGMEWCKENIDTSQGDVTFAGDGQEATPWKDFALLTQCNHIMTIMTGTFGFWAAYLAG
GDTVYLANFTLPDSEFLKIFKPEAAFLPEWVGINAIDLSPLWTLAKP
```

以下參數為我所使用之參數:

比對組別	Word Size	Matrix	備註
嚴苛組	6	BLOSUM80	找保守序列
寬鬆組	2	BLOSUM45	找遠親序列

比對完成後，將兩組結果進行比較，我將會使用簡單的python進行資料處理，以觀察以下指標:

- **命中數（Hit Count）**：比對到的相似序列數量。
- **Bit score**：比對的得分，數值越高表示比對越好。
- **E-value**：期望值，數值越小表示比對越有意義。
- **% Identity**：序列相似度的百分比。

我已經將兩組 BLASTp 並對結果做了簡單整理，如下:

設定	Hits 數量	平均 %Identity	平均 Bitscore	最小 E-value	最大 E-value
嚴苛 (Word=6, BLOSUM80)	36	73.95%	502.94	0.0	2.7×10^{-2}
寬鬆 (Word=2, BLOSUM45)	34	76.61%	553.24	0.0	7.63×10^{-8}

重點比較:

1. **Hits 數量**
   - 嚴苛設定比對到 36 筆序列，寬鬆設定反而比對到 34 筆。
   - 意外地，兩者命中數相差不大，嚴苛設定稍微多出 2 筆。
2. **平均相似度（%Identity）**
   - 嚴苛組的平均相似度約 73.95%，
   - 寬鬆組平均相似度約 76.61%。
   - 寬鬆組反而有略高的平均相似度，表示雖然放寬參數，但找到的序列相對也相當相似。
3. **平均 Bitscore**

- 嚴苛組平均 Bitscore ≈ 502.9，
- 寬鬆組平均 Bitscore ≈ 553.2。
- 寬鬆組的 bitscore 較高，表明在整體得分上，寬鬆設定也能搜尋到高質量比對。

4. **E-value 範圍**
   - 兩組最小 E-value 都是 0（代表非常顯著）；
   - 最小和最大值差異：
     - 嚴苛組最大 E-value 約 2.7×10^{-2}，
     - 寬鬆組最大 E-value 更低（7.6×10^{-8}），
   - 表明寬鬆組雖然放寬檢索條件，卻同時篩選出了更顯著的序列。

結論:

- 雖然預期「寬鬆參數」會比「嚴苛參數」找到更多低相似度的遠緣序列，但實際上兩組的 Hits 數量非常接近，且寬鬆組反而在平均相似度與 Bitscore、E-value 上都略勝一籌。
- 這可能是因為該蛋白質本身在 Swiss-Prot 中已有相當多高相似度的保守家族成員，無論用哪組參數都能抓到這些熱門的對應序列。