Transzkripciós faktorok kötőhelyeinek előrejelzése gépi tanulási módszerekkel

Szalai-Gindl János Márk, Információs Rendszerek Tanszék,
szalaigindl@inf.elte.hu

Leírás

Egy DNS szekvencia - informatikai nézőpontból - az {'A','C','G','T'} karakterhalmaz elemeinek sorozata. Egy egyed teljes DNS szekvenciájának bizonyos szakaszai, - a gének, - kitüntetett jelentőséggel bírnak: magukba foglalják a fehérjék előállításához szükséges információkat (a „tervrajzukat"). Egy élőlénynek majdnem az összes sejtje ugyanazt a DNS-t tartalmazza a sejtmagban. A sejtek azért különböznek mégis egymástól, mert eltérő típusú sejtekben különféle gének lehetnek „bekapcsolt" vagy „kikapcsolt" állapotokban. Ezeket az úgynevezett transzkripciós faktorok befolyásolják, amelyek jellegzetes szekvencia motívumokhoz kötődnek.

Különböző molekuláris biológiai technikák léteznek arra, hogy a transzkripciós faktorok kötődési helyeit meghatározzák. Azonban hiába nőnek folyamatosan a kötődés adathalmazok, az továbbra sem lehetséges, hogy az összes sejttípusra és transzkripciós faktorra el lehessen készíteni a kísérleteket. Emiatt számítási megközelítések szükségesek, hogy a kísérleti eredményeket ki tudjuk egészíteni, és meg tudjuk jósolni a kötődési helyeket.

A probléma hátterét az *ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge*[1] adja, ahol azt a célt tűzték ki a szervezők, hogy bizonyos tanítóhalmazok (és további segédhalmazok) alapján előre lehessen jelezni egy adott sejttípus teljes DNS szekvenciájának mindenegyes pozíciójára, hogy milyen valószínűséggel tud bekötődni egy adott transzkripciós faktor.

A hallgató feladata, hogy megismerkedjen a Python, illetve R programnyelv számítógépes molekuláris biológiai eszköztárával, áttekintse a szakirodalomban fellelhető predikciós módszereket, illetve implementációkat készítsen. A kutatás lehetséges további célja: az elérhető módszerek teljesítményének javítása. Angol nyelvismeret szükséges.

Hivatkozások

- Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015) Predicting the Sequence Specificities of DNA-and Rna-Binding Proteins by Deep Learning. *Nature Biotechnology*. 33, 831–838.
- Agius, P., Arvey, A., Chang, W., Noble, W. S., & Leslie, C. (2010) High Resolution Models of Transcription Factor-DNA Affinities Improve In Vitro and In Vivo Binding Predictions. *PLoS Computational Biology*. 6.
- Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., & Pritchard, J. K. (2011). Accurate Inference of Transcription Factor Binding from DNA Sequence and Chromatin Accessibility Data. *Genome Research*, 21, 447–455.

[1] https://www.synapse.org/#!Synapse:syn6131484/wiki/