Lineaarne mudel geeni ekspressiooni andmete analüüsi jaoks

Bakalaureusetöö

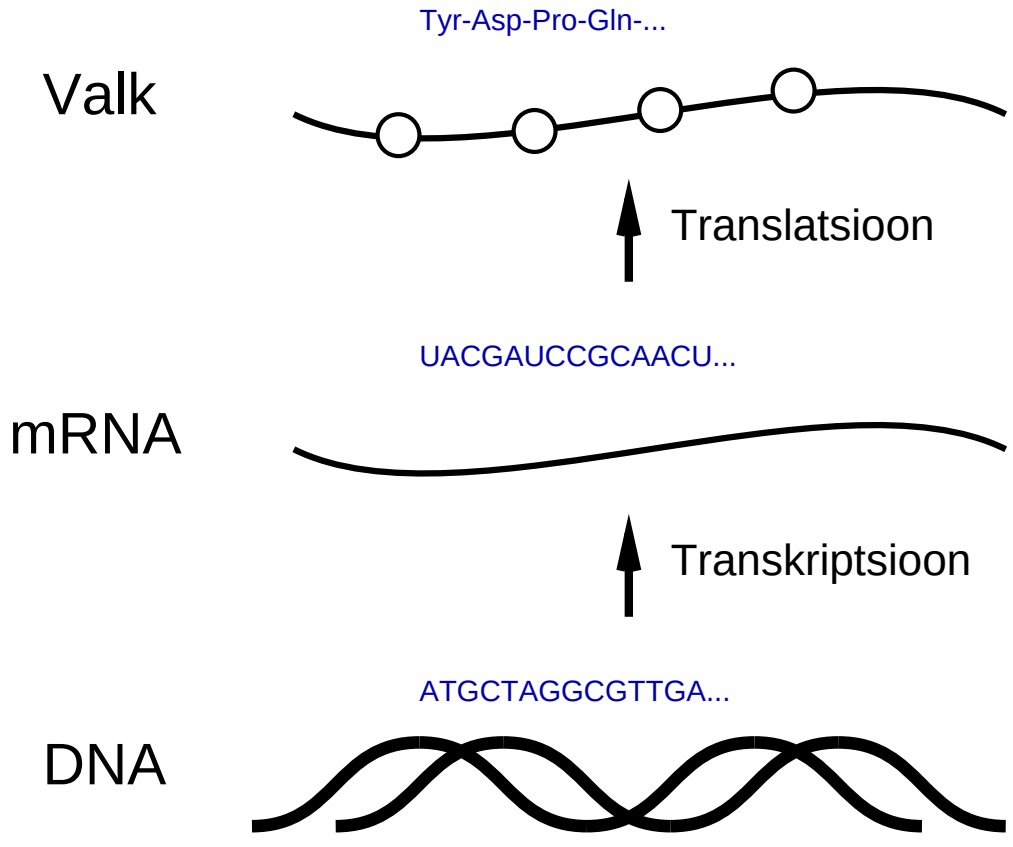
Konstantin Tretjakov Juhendaja: Jaak Vilo

7. juuni 2005

Sisukord

- Konteksti kiirtutvustus
  - Bioloogiline sissejuhatus (DNA, RNA, Mikrokiip-eksperimendid)
  - Statistiline sissejuhatus (Lineaarsed mudelid, SVM)
- Lineaarne mudel geeni ekspressiooni andmete analüüsi jaoks
- Praktiline eksperiment

DNA, RNA, Transkriptsioon, Translatsioon

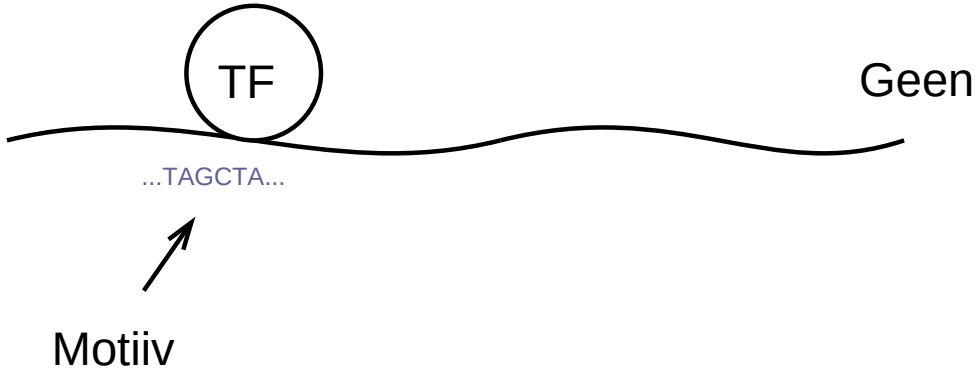


Geenide ekspressioon

- Geen — DNA jupp, kodeerib ühe või mitut valku.
- Geeni ekspressioon — vastava geeni valkude tekitamine raku poolt.
- Sõltuvalt tingimustest võivad ühed geenid olla ekspresseeritud rohkem kui teised.
- Geenide ekspressiooni regulatsioon on keeruline protsess.

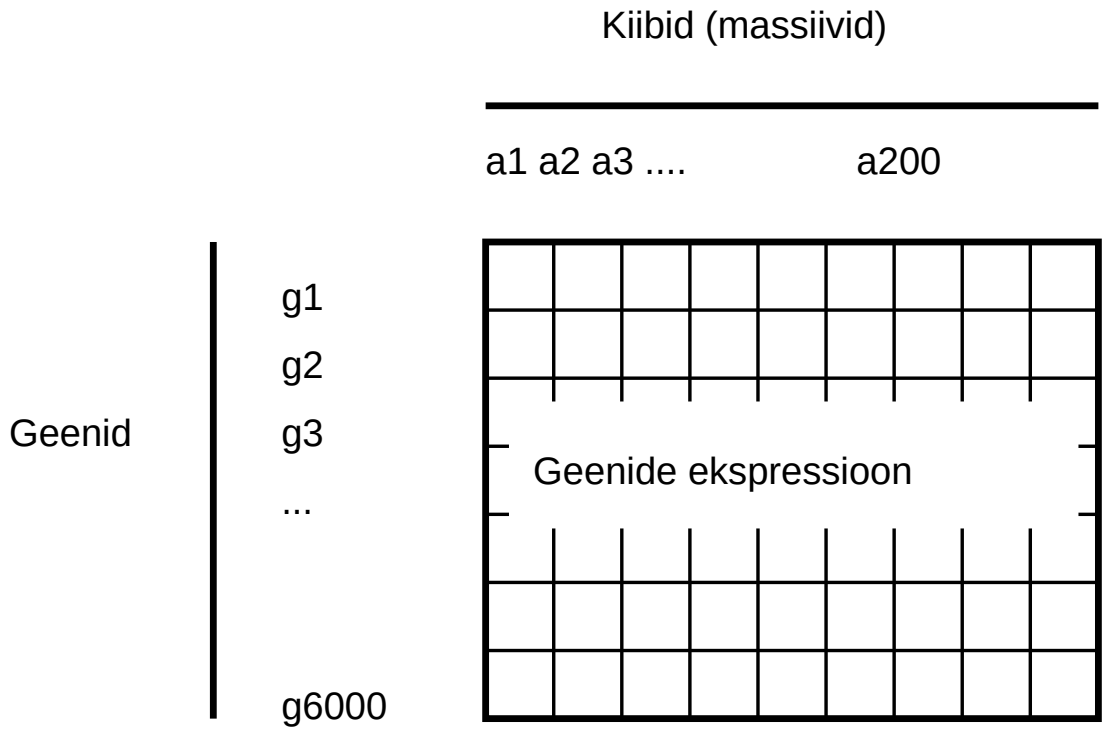
Geenide ekspressioon

Teatud valgud (transkriptsiooni faktorid, TF) saavad mõjutada teiste geenide transkriptsiooni ühendades ennast teatud *motiivide* külge geeni peal.



Geenide ekspressiooni andmed

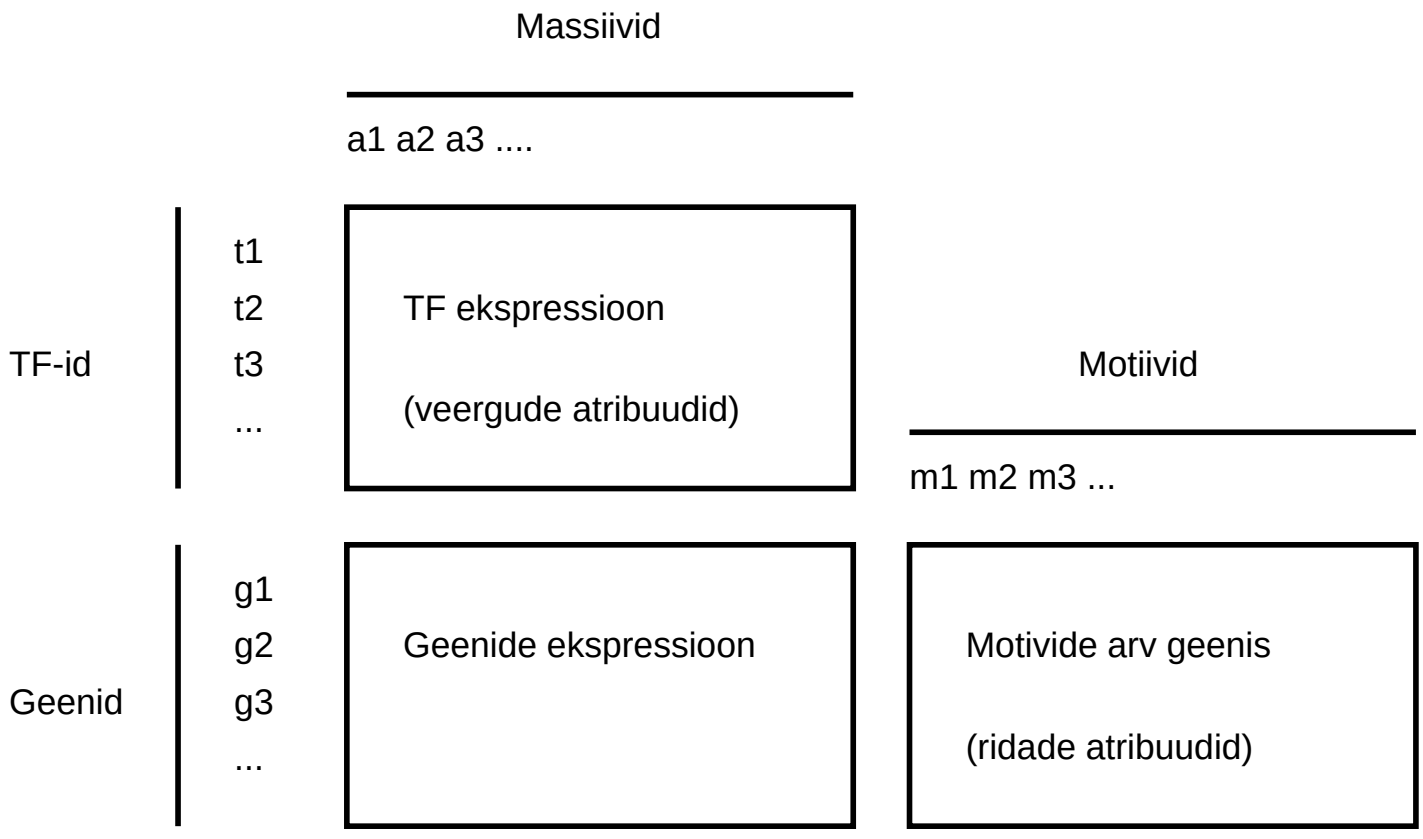
Geenide ekspressiooni saab mõõta mikrokiibi eksperimentide abil. Mitme eksperimendi tulemusena saadakse nn *ekspressioonimaatriksi*:



Ekspressioonimaatriksi analüüs

- Kuidas tuletada ekspressioonimaatriksist kasulikku infot?
  - Klasterdamine
  - Ennustavad mudelid
    - Otsustuspuud
    - Lineaarsed mudelid
- Kuidas kombineerida ekspressiooniandmeid muu bioloogilise infoga, näiteks DNA sekventsidega?

Ekspressiooniandmed + DNA motiivid



Tähistused

- G_{ij} — geeni g_i ekspressioon kiibil a_j.
- T_{kj} — TF-i t_k ekspressioon kiibil a_j.
- M_{il} — motiivi m_l esinemiste arv geenis g_i.
- $\mathbf{G}$, $\mathbf{M}$, $\mathbf{T}$ — maatriksid (G_{ij}), (M_{il}) ja (T_{kj}).

Lisaks fikseerime mõned i ja j, ning tähistame

$$G := G_{ij} \qquad T_k := T_{kj} \qquad M_l := M_{il}$$

Geenide ekspressiooni ennustamine

$$G = f(T_1, T_2, \ldots, T_{n_t}, M_1, M_2, \ldots, M_{n_m})$$

- Oli juba tehtud otsustuspuude abil

```
if upregulated(t1) and
   not upregulated(t2)
   and hasmotif(g, m1)
      then upregulated(g)
```

- Pakume lineaarse mudeli.

Lineaarne mudel

$$G = \alpha_{11} M_1 T_1 + \alpha_{12} M_1 T_2 + \cdots + \alpha_{lk} M_{n_m} T_{n_t} =$$

$$= \sum_{l,k} \alpha_{lk} M_l T_k$$

- Bioloogiliselt motiveeritud
- α_{lk} näitab kui oluline on seos motiivi m_l ning TF t_k vahel.

Mudeli vähimruutude hinnang

Tähistagu $\mathbf{A}$ parameetrite maatriksit (α_{lk}). Osutub, et parameetrite vähimruutude hinnangu leidmiseks tuleb leida selline $\mathbf{A}$, mille puhul

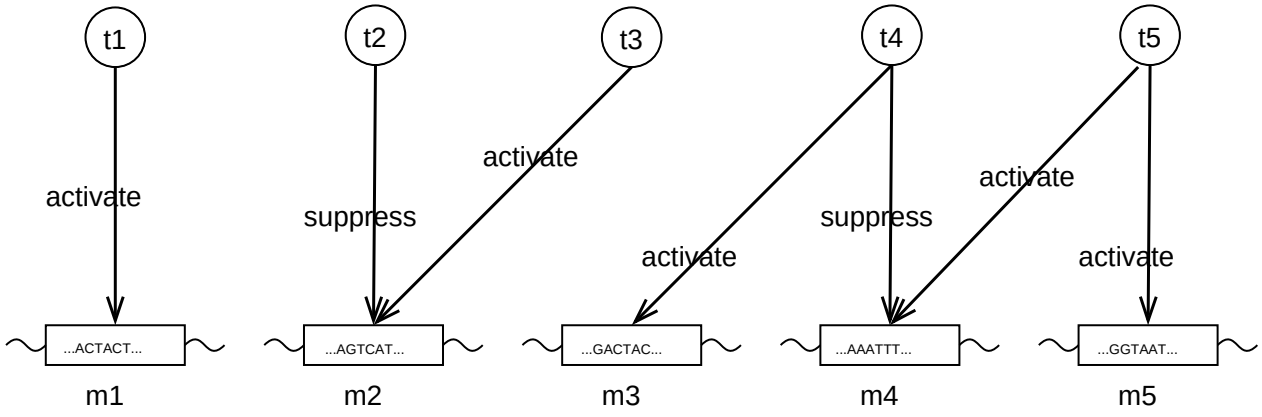
$$\mathbf{G} \approx \mathbf{MAT}$$

Lahendus avaldub kujul

$$\mathbf{A} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{G}\mathbf{T}^T(\mathbf{T}\mathbf{T}^T)^{-1}$$

See on suhteliselt efektiivselt arvutatav.

Näide



$$\mathbf{A} = \begin{array}{c|ccccc} & t_1 & t_2 & t_3 & t_4 & t_5 \\ \hline m_1 & 1 & 0 & 0 & 0 & 0 \\ m_2 & 0 & -2 & 1 & 0 & 0 \\ m_3 & 0 & 0 & 0 & 1 & 0 \\ m_4 & 0 & 0 & 0 & -1 & 1 \\ m_5 & 0 & 0 & 0 & 0 & 2 \end{array} .$$

Klassifitseerija versioon

Võib üritada sobitada mudelit kujul

$$G = \text{sign}\left(\sum_{l,k} \alpha_{lk} M_l T_k\right)$$

Selle mudeli sobitamiseks efektiivselt algoritmit ei leitud. Prooviti kasutada tugivektormasinaid (SVM), kuid midagi huvitavat välja ei tulnud.

Katsed sünteetilisel andmestikul

- Kahjuks tegelikke andmete analüüsini ei jõutud. Selle asemel genereeriti sünteetilist andmestiku.
- Genereeritud andmestik kasutas veidike teist mudelit, oli mürane, ning puuduvate andmetega.
- Sellest suutis meetod taastada 25% tegelikke seoseid, 15% spetsiifilisusega.
- See on 10 korda parem juhuslikust valikust.

Tegemata

- Katsetada meetodit reaalsete andmete peal
- Tuletada efektiivse algoritmi klassifitseerija versiooni jaoks
- Proovida *kernelized* versiooni (näit SVM).
- Üritada leida meetodi abil olulisi TF-eid.

Aitäh tähelepanu eest!

Küsimused?