ШИНЖЛЭХ УХААН ТЕХНОЛОГИЙН ИХ СУРГУУЛЬ
Мэдээлэл, Холбооны Технологийн Сургууль

*F.CS213 Биоалгоритм*

Hidden Markov Models

# Марковын далд загварууд

*Лекц 11*

*Багш* ***Ганбаатарын ГАНБАТ.*** *Өрөө* ***#304,*** *ganbatg@must.edu.mn,* ***99999467***

# Лекцийн агуулга

- Удиртгал
- Марковын далд загвар?
  - Дарааллын магадлал
  - Элементүүд
  - Даалгавар/Task-ууд
- HMMs for Database Search

# Удиртгал

- *Тэмдэгт мөр харгалзуулах асуудал (String matching problem)*–ын абстракт загварын Төгсгөлөг төлөвт автомат (Deterministic Finite Automaton - DFA) – 5р бүлэг
  - Паттернаас автоматыг бүтээх санаан дээр суурилдаг.
  - Паттерний тохиолдлыг илрүүлэх зорилготой (олон дахин ашиглах боломжтой)
  - Төгсгөлөг тооны төлвүүдийн графаар дүрслэгдэнэ (төлвийн диаграмм)
  - Паттерны тэмдэгт бүр нь төлөв бөгөөд тааралдсан ижил тэмдэгт бүр нь шинэ төлөвт шилжүүлнэ.
  - Эхлэх, дуусах төлөвтэй ба төгсөх төлөвт орсон бол тэмдэгт мөр танигдсан байна
  - Тооцооллын үр ашиг өндөр - Оролтын тэмдэгт мөрийг нэг л удаа нэвтрүүлнэ
  - Регуляр илрэхийлэл ашиглахтай адил хүчин чадалтай
  - Сканнердсан дараалалд паттерн илэрч байгаа эсэхийг шалгах ***детерминик*** хайлт
- Гэхдээ биологийн дараалалд хамаарах үйл явц нь ерөнхийдөө ***магадлал***-д суурилдаг.
  - Жнь өгөгдсөн уургийн дэд эсэн зохион байгуулалт (цитозол, цөм, бүрхүүл...)-ын ангилал;
    - Дараалал дээрх уургийн домэйны оролцоо
    - Генийн промотор бүсийн дараалал дээр Транскрипцын фактор залгагдах магадлал.
- *Марковын далд загвар (Hidden Markov Models – HMMs)*
  - Шугаман дараалал дээрх статистик зүй тогтлыг таних
  - Ген илрүүлэх, профайл хайх, олон дараалал дээрх зэрэгцүүлэлт, зохицуулалтын хэсгийг тодорхойлох, уургийн боломжит хэлбэрүүдийг таамаглах.

# Марковын далд загвар

- HMM нь дарааллыг сканнердаж, тэмдэгт бүрийг харгалзах бүсээр нь шошгожуулна.
  - Шошгонд харгалзсан төлвүүд + дарааллын эхлэх ба төгсөх төлвүүд.
  - Эхлэх төлөв нь шошгонд харглзах төлвүүдийн нэгнийх нь дараалал эхлүүлдэг байх магадлал байна.

- Боловсорсон (Mature) mRNA-г тодорхойлох дүрмийн бүтэц
  - *Давтагдсан гурвал бүтэц:* (5'UTR), (CDS), (3'UTR) – бүсүүд буюу ***үгүүд (word)***
  - ***Өгүүлбэр****:* 5’UTR нь CDS-ийн өмнө, 3'UTR нь CDS-ийн ард байх ёстой.
  - ***Зорилго***: Дарааллыг сканнердаж, түүний тэмдэгт бүрийг харгалзах бүсээр шошгожуулна.

- *Дарааллын магадлал*-ыг тодорхойлохын тулд тухайн төлвийн тэмдэгтийн болон шилжилтийн магадлалыг үржүүлнэ.
  - Шошгожуулалтын цэг бүр нь магадлалын хамааралтай.

  *Шилжилтийн магадлал (Transition probabilities):*
  - Дараагийн тэмдэгт одоогийн төлөвт хамаарах эсэх.
  - 5'UTR төлөвт байхад дараагийн тэмдэгт нь 5’UTR байх магадлал 0.8, CDS руу шилжих магадлал 0.2.

  *Ялгаралтын магадлал (Emission probabilities):*
  - mRNA-ийн бүсүүдийн тогтоц нь биологийн төрөл зүйлээс хамаарч өөр өөр байдаг.
  - G+C агуулга 5'UTR-д 60%, CDS-д 50%, 3'UTR-д 30% байна гэвэл уг нийлмэл давтамжууд нь төлвөөр ялгарах тэмдэгтийн магадлалыг тодорхойлно.

HMMs

# Дарааллын магадлалыг тооцоолох

- CDS-ын, 3'UTR-ын тус бүр 3 нуклеотидтэй TAGTTA тэмдэгт мөр байг.
  - $S$ төлөв дээрх $N$ тэмдэгтийн $P_S(N)$ ялгарлын магадлал
  - $S_A$-ээс $S_B$ гэсэн $P_{S_A,S_B}$ шилжилтийн магадлал

$$P_{CDS}(T) \times P_{CDS,CDS} \times P_{CDS}(A) \times P_{CDS,CDS} \times P_{CDS}(G) \times P_{CDS,3U} \times P_{3U}(T) \times P_{3U,3U} \times P_{3U}(T) \times P_{3U,3U} \times P_{3U}(A)$$
$$= 0.25 \times 0.9 \times 0.25 \times 0.9 \times 0.25 \times 0.1 \times 0.35 \times 0.8 \times 0.35 \times 0.8 \times 0.35 = 3.47287 \times 10 - 5$$

*Логарифмчилбол (–1-ээр үржүүлэх) оноог **нийлбэрээр** тооцоолж болно (Үр дүн нь* $10.26$*).*

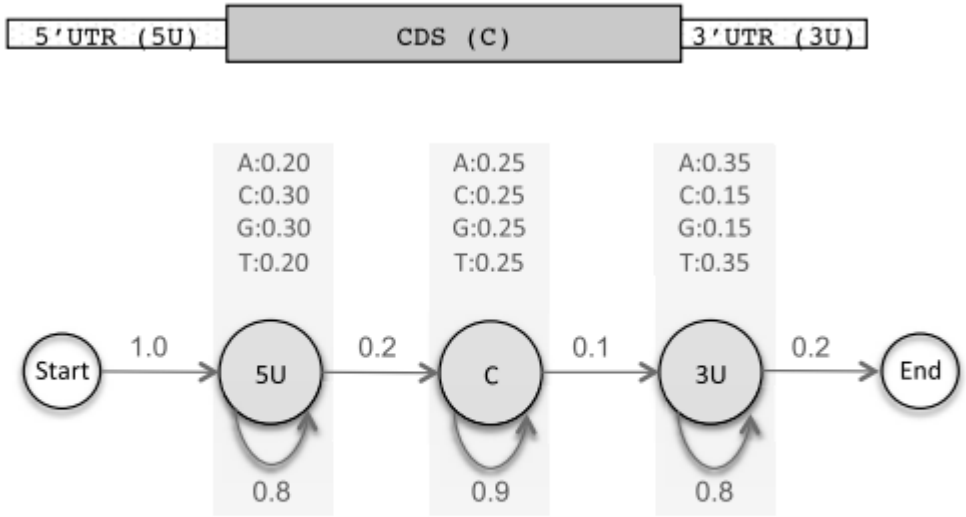
- HMM нь мэдээллийн 2 тэмдэгт мөрийг үүсгэнэ
  - Төлөв бүрт ялгарсан тэмдэгтүүдийн ***ажиглалтын(observed) дараалал*** буюу ***төлвийн зам (state path)***
    - Тухайн төлөв дээрх магадлал нь ажиглалтын дарааллын *одоогийн тэмдэгт* дээр үндэслэнэ.
  - Тэмдэгт бүрийг нь шошгоруу харгалзуулсан төлвүүдийн ***далд дараалал (hidden sequence)***.
    - Дараагийн төлөв рүү хийх шилжилт нь одоогийн төлөв дээрх *шилжилтийн магадлалуудын хуваарилалтаас* хамаарна.
- Магадлалын дээрх тооцоолол нь ***Марковын цуваа (Markov chain)*** гэдэг процессын дагуу явагддаг
  - Цуварсан хамааралтай төлвүүдийн төгсгөлөг олонлогт суурилсан үйл явцын дараалал
  - Дараагийн төлөв нь зөвхөн одоогийн төлвөөс хамаардаг.

HMMs

# Элементүүд ба магадлалын нөхцөл

- HMM нь 5 элементтэйгээр, DFA шиг тодорхойлогдоно.
  - *Цагаан толгой*, $A = \{a_1, a_2, \ldots, a_k\}$.,
  - *Төлвийн олонлог*, $S = \{S_1, S_2, \ldots, S_n\}$.
  - *Анхдагч төлвийн магадлал* $I(S_i)$: $t = 0$ хугацааны $S_i$ төлөвт байх магадлал.
  - *Ялгаралтын магадлал*: $i$ төлөвт $a$ тэмдэг ялгарах магадлал $e_{i,a}$-ийн хувьд $\Sigma_{a \in A} e_i(a) = 1$ байх.
  - *Шилжилтийн магадлал*, $T_{i,j}$ нь $i$ -ээс $j$ төлөв рүү шилжилтийн магадлал, өөрийг нь оруулаад $\Sigma_{i \in S, j \in S} T_i(j) = 1$ байх.
- Эхний 3 нь 1D вектор/жагсаалт, ялгаралт ба шилжилтийн магадлал нь 2D матрицууд.
- HMM-ийн параметрүүдийг ерөнхийдөө (frequently) $\theta$-өөр тэмдэглэдэг.
  - Зарим тохиолдолд параметруудыг зөвхөн дэд олонлогоор нь тохируулж болно,
  - Зөвхөн ялгаралт ба шилжилтийн магадлал нь тухайн хэсэгт хамааралтай: $\theta = (e_{i,a}; T_{i,j})$ .
- Тусгайлсан магадлалыг тохруулахад дөрвөлжин хаалт ([ ]) бүхий том $P$ тэмдэглэгээг ашиглана.
  - $P[O, \pi | M]$ нь $M$ загварт өгөгдсөн $O$ дараалал болон $\pi$ төлвийн замыг ажиглах магадлал.
    - $T$ урттай ажиглалтын дараалал $O = o_1, o_2, .., o_T$: $O_t$ бүр нь $t$ байрлалын тэмдэгт.
    - Төлвийн дараалал/замыг $\pi_t$ нь $t$ байрлалд харгалзах төлөв байх $\pi$-ээр тэмдэглэнэ.

# Үндсэн даалгаврууд




```
Sequence: ...GTGCCTAGTCTAGTTATCAAATA...
  States: ...CCCCCCCCCCCCC3333333333...
```

Боловсорсон mRNA-ийн бүсийн шошгололтын хялбар жишээний HMM бүтэц.

- 55CCCC33 төлөв шилжилтийн боломжит дараалал болон ATCGCGAA ялгарсан дарааллын
- эхлэх төлвийн магадлал: $I(5) = 1, I(C) = 0, I(3) = 0$.
- хувьд бид харгалзах магадлалыг дараах байдлаар тооцоолж болно.

$$P_T[55CCCC33] =$$
$$I(5) \times T_{5,5} \times T_{5,C} \times T_{C,C} \times T_{C,C} \times T_{C,C} \times T_{C,C} \times T_{C,3} \times T_{3,3}$$
$$= 1 \times 0.8 \times 0.2 \times 0.9 \times 0.9 \times 0.9 \times 0.1 \times 0.8$$
$$\approx 0.0093312$$

$$P_e[ATCGCGAA] =$$
$$0.2 \times 0.2 \times 0.25 \times 0.25 \times 0.25 \times 0.35 \times 0.35$$
$$\approx 1.9141e{-}05$$

$$P[e(ATCGCGC) \wedge T(55CCCC33)] =$$
$$0.0093312 \times 1.9141e - 05 \approx 1.78e - 07$$

- HMM-ын дараах даалгаврууд ыг *тодорхой нэг төлөвийн замд* эсвэл *төлөвийн бүх боломжит замуудын* хэмжээнд хэрэглэж болно.
  - *scoring* буюу магадлалын тооцоолол, *код тайлах (decoding)*, *суралцах (learning)*

# Онооны функц

- Онооны функцууд нь ажиглалтын дарааллын магадлалуудын тооцооллыг явуулдаг.
  - $O$ ажиглаглалтын дараалал болон $\pi$ төлвийн замын хувьд
  - HMM $M$ доторх $S$ ба $\pi$-ийн хамтарсан магадлал нь $P[O, \pi | M]$ байна.
  - Үүнийг бид ялгаралтын болон шилжилтийн магадлалын хамтарсан үржвэрийг дээрх жишээн дээрх шиг тооцоолж болно.
  - Хэрэв төлвийн зам өгөгдөөгүй бол дарааллын нийт магадлал $P(O|M)$-ийг тооцоолно.
  - Энэ магадлалыг бүх боломжит замуудын магадлалыг нийлбэрээр гаргаж болох бөгөөд энэ нь экспоненциал тооны зам боловч *Forward алгоритм*-аар үр дүнтэй тооцоолох боломжтой.
- *Давших (Forward) магадлал* $\alpha_t(S_i)$ нь $P[O = o_1 \dots o_t | \pi_t = S_i | M]$-тэй эквивалент
- *Гэдрэг (Backward)* магадлал $\beta_t(S_i)$ нь $P[O = o_t \dots o_T | \pi_t = S_i | M]$-тэй эквивалент

# Код тайлах

- Ажиглалтын дарааллаар өгөгдсөн төлийн хамгийн их магадлалтай дарааллыг тодорхойлох.
  - Ажиглалтын дарааллын тэмдэгт бүрийн хувьд 5U, CDS, 3U төлвүүдийн дарааллыг тодорхойлох.
  - HMM $M$ ба $O$ ажиглагдсан дараалал өгөгдсөн бол хамгийн их магадлалтай $\pi$ төлвийн замыг олох код тайлагчийн үүрэг болно.
  - Энд бүх замын хувьд хамгийн их оноотой замыг олохын тулд динамик програмчлалын Viterbi алгоритмыг ашиглана.
  - Trace-back функцийг ашиглан зочилсон бүх замыг хянаж, хамгийн өндөр магадлалтай замыг илрүүлэх боломжтой.
- Өөр нэг чухал асуулт бол бүх төлвийн замыг авч үзэх үед $k$ төлөвөөс ялгарах $o_t$ тэмдэгтийн нийт магадлалыг тодорхойлох явдал юм.
  - Үүнийг ***хожуу (posterior)*** магадлал гэж нэрлээд $P[\pi_t = k|O, M]$ гэж тэмдэглэнэ.
  - Энэ $P(O)$ магадлалыг $\alpha_t(k)$ ба $\beta_t(k)$-ийн үржвэрийг $O$-ийн магадлалд хуваах замаар олно.

# Сургалт

- mRNA HMM-ийн жишээн дээр тус загварын параметрүүдийг *priori(эрэмбэ)*-ийг тодорхойлсон.
  - Энэ нь өмнөх мэдлэгийг ашиглах уу эсвэл ажиглагдсан өгөгдлөөр хэмжих уу гэдгийг шийдэх.
  - Хэрэв шинжилгээний явцад илүү их өгөгдөл олж авбал эдгээр параметрүүдийг тохируулж, загвараа оновчтой болгож болно.
- Өгөгдсөн датасетийг шинжилснээр бид эдгээр параметрүүдийг сурган авч болно.
  - $M$ загвар дотор ажиглалтын дарааллын $P\ (O|M)$ магадлалыг нэмэгдүүлэх $\theta = (e_{i,a};\ T_{i,j})$ параметрүүдийг олох.
  - Хэрэв ажиглалтын дарааллыг шошгожуулсан буюу төлвийн зам нь мэдэгдэж байвал *хяналттай (supervised)* сургалт явагдана.
    - Энэ тохиолдолд ялгаралт болон шилжилтийн магадлалыг сургалтын өгөгдлөөс шууд гаргадаг.
  - Хэрэв сургалтын дараалал шошгогүй бол *хяналтгүй (unsupervised)* сургалт явагдана.
- *Баум-Вэлч (Baum-Welch)* гэдэг Expectation-Maximization алгоритмыг суралцахад ашиглана.
  - Алгоритм нь $M$ загварын $\theta$ параметруудад зориулсан хамгийн сайн таамаглалтайгаар эхэлдэг.
  - Дараа нь ялгаралтын болон шилжилтийн магадлалыг өгөгдлөөс тооцоолдж, загварыг шинэчилдэг.
  - $M$ загварын $\theta$ параметрүүд давхцах хүртэл энэ процедур дээр давтагдана.

# HMMs for Database Search

- Хамаарал бүхий дараалал ба бүтэцтэй уурагуудыг нэг бүлэгт оруулж болно.
  - Тэд ихэвчлэн ижил төстэй үүрэг бөгөөл тодорхой хувьслын харилцаатай байдаг.
  - Нэг бүлгийн олон дараалал дээр хадгалагдсан бүсүүдийг зэрэгцүүлэх тухай 11-р бүлэгт үзсэн
    - Байрлалын жингийн матрицыг (PWM) гаргаж авах боломжтой.
    - Зэрэгцүүлэлт дээрх өөр өөр байрлалуудтай давтагдсан паттернийг олж авч болно.
- Биологийн дарааллын шинжилгээнд HMM-ийн хамгийн түгээмэл нэг хэрэглээ: ***HMM-профайл (HMM-profile)***.
- Уургийн бүлгийн магадлалын профайлыг өгдөг
  - PWM-тэй төстэй боловч оруулах, устгах (insertions and deletions) явуулахад илүү уян хатан.
  - HMM-профайлын топологи нь эхлэл ба төгсгөлөөс гадна ***main, insertion, deletion*** 3 өөр бүлэг төлвүүдтэй.
  - ***main:*** баганууудыг загварчлах бөгөөд эдгээр нь зэрэгцүүлэлтийн байрлалтай адил олон байна.
  - ***insertion:*** зэрэгцүүлэлтийн өндөр хувьсалтай бүсүүдийг загварчлах.
  - ***deletion:*** зэрэгцүүлэлтийн нэг буюу хэд хэдэн баганын хооронд шилжих боломжтой болгодог.
- Био-информатикийн ердийн хэрэглээ бол бүлэг уургаас HMM-профайл үүсгэх, бүлгийн бусад мэдэгдэхгүй байгаа гишүүдийн дарааллыг өгөгдлийн сангаас хайх.
  - Бүлэг уургийгийн HMM-профайл $M$-ийн хувьд $D$ өгөгдлийн сангийн бүх $W$ дарааллыг сканнерддаг.
  - Тодорхой босгооос өндөр магадлал/оноотой дараалал нь бүлгийн гишүүд болно.
  - $P[W|M]$ магадлалыг Forward алгоритмаар тооцоолж болно.

# АНХААРАЛ ТАВЬСАНД БАЯРЛАЛАА