Gibbsovo samplování a jeho využití

Regulace genů

- Hlavní pozornost výzkumů DNA je většinou věnována analýze genů
- Geny tvoří pouhá 3% lidské DNA
- Ukazuje se, že zbývající „junk DNA“ má také velký význam
- Obsahuje sekvence, které regulují přepis genů
  - Krátkodobé efekty – regulace syntézy nebo potlačení enzymů pro adaptaci buňky při změně vnějšího prostředí
  - Dlouhodobé efekty – důležité pro samotné vytváření buňky a její chování

Regulace genů 2

- Nesprávná funkce regulačních prostředků může vést k vrozeným defektům
- Může být způsobena také dědičnými chorobami
- Rakovinné buňky mají potlačenou regulaci, která u normálních buněk zajišťuje ukončení dělení
- Nalezení sekvencí DNA, které regulují přepis genů, může osvětlit takové změny

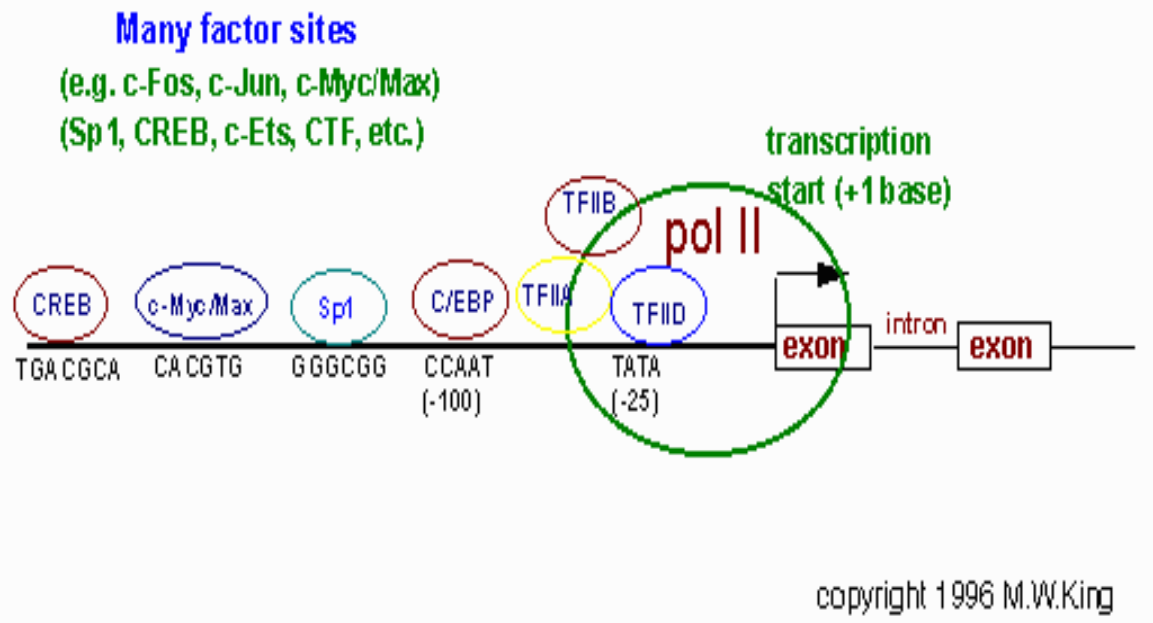
Transcription factor binding sites

- Motiv – součást regionu před genem, tzv. promotoru, který reguluje, zda se bude gen přepisovat
- U jednoduchých organismů (např. kvasinky) – transkripční faktor (protein) navázaný na TFBS (Transcription factor binding site) umožní přepis DNA
- Různé geny mohou sdílet stejný motiv – jsou regulovány stejným transkripčním faktorem

Transcription factor binding sites 2

- U jednodušších organismů jsou motivy relativně neměnné
- U vyšších živočichů jsou motivy často degenerované – obsahují různé báze, ale váže se na ně stejný transkripční faktor
- U vyšších organismů transkripční faktory často spolupracují – dva nebo více transkripčních faktorů je nutné pro přepis
- V případě více transkripčních faktorů má každý vlastní TFBS

Transcription factor binding sites 3



Struktura regionu před typickým mRNA genem eukaryotické buňky

Gibbsovo samplování (opakování)

- Vstup: n sekvencí DNA $\{s_1, \ldots, s_n\}$
- Náhodně vyber z každé sekvence s_i jeden l-mer a_i

- Vyber náhodně jednu ze sekvencí s_h
- Vytvoř profil X velikosti 4 x l z $a_1,\ldots,a_{h-1},a_{h+1},\ldots,a_n$
- Vypočítej četnosti Q vstupních sekvencí $s_1,\ldots,s_{h-1},s_{h+1},\ldots,s_n$ (“pozadí”)
- Pro každý l-mer a z s_h spočítej $w(a)=\frac{P(a|X)}{P(a|Q)}$

- Polož a_h = a pro nějaké a vybrané z s_h s pravděpodobností $\frac{w(a)}{\sum_{a' \in s_h} w(a')}$

- Opakuj, dokud nezkonverguje

Modifikace

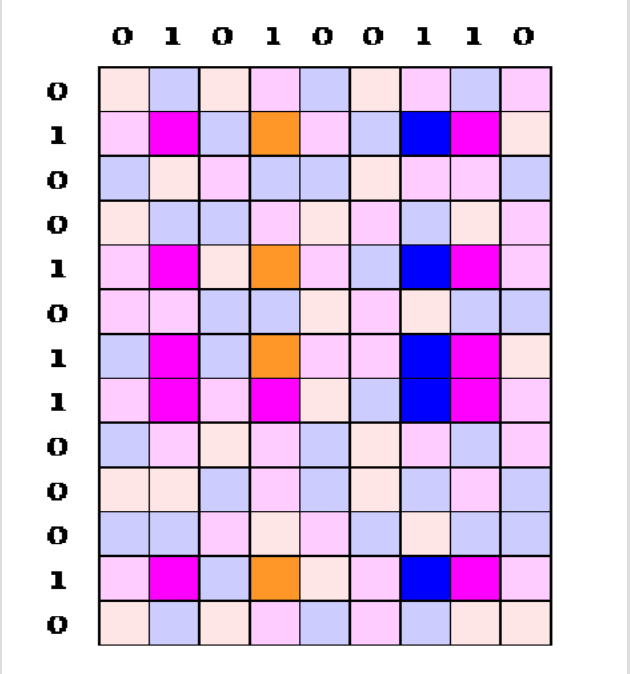
- vyhledávání více motivů – už nalezené motivy jsou nějakým způsobem maskovány
- lepší modelování “pozadí”, např. pomocí skrytých Markovovských modelů

Konvergence algoritmu

- počet iterací nutných ke zkonvergování výsledku se může velmi lišit
- záleží na počáteční volbě l-merů
- l-mery by měly být vybrány náhodně z celého vstupního prostoru
- existují různé míry konvergence tohoto algoritmu

Použití

- hledání TFBS (Transcription Factor Binding Sites)
- klastrování výsledků z DNA čipů – hledání podmnožiny genů, které se v některých situacích chovají obdobně



Parkinsonova choroba

- Příznaky
  - Problémy s motorikou – třes, ztráta rovnováhy, problémy s polykaním ...
  - Problémy se spánkem
  - Zpomalené reakce, demence, halucinace, krátkodobé ztráty paměti

Parkinsonova choroba

- Příčiny
  - Většinou idiopatické (bez známé příčiny)
  - Genetické – mutace v 13 různých genech => 13 druhů Parkinsonovy choroby PARK1-13
  - Toxiny – pesticidy, mangan, železo
  - Úrazy hlavy
  - Léky – antipsychotika (léky na schizofrenii a psychózy)

Slavné osobnosti

- Jan Pavel II.
- Adolf Hitler
- Mao Ce-tung
- Francisco Franco
- Muhammad Ali
- Salvador Dali
- Mervyn Peake

DJ-1

- Zaměříme se na studium genu DJ-1, známého také jako PARK7 (podle typu choroby, který mutace v něm může způsobit)
- Výskyt
  - Člověk – chromozom 1, 7944380-7967926
  - Pes – chromozom 5, 64574993-64590963
  - Kráva – chromozom 16, 41159243-41176000
  - Šimpanz – chromozom 1, 8033422-8064999

Gibbs Sampler

- Kvůli rychlosti jsme použili sampler, který je volně ke stažení na
  `http://bayesweb.wadsworth.org/gibbs/gibbs.html`

Použité sekvence

- Použili jsme sekvence z člověka, psa a krávy, které obsahovaly gen DJ-1 a navíc 10000 bází na začátku a na konci
- Celkem každá sekvence obsahovala přibližně 40000 bází
- Celková délka sekvencí byla 116276 bází

Kódující části

- Sekvence obsahují kódující i nekódující části
- Podle GenBanku první kódující část v každé sekvenci začíná přibližně na pozici 1100
- Když se nám podaří najít motiv někde na pozici kolem 10-11 tisíc bází od začátku, pravděpodobně jsme lokalizovali gen

Nalezené motivy

- Motiv se 40 pevnými pozicemi (60 pozic celkově)
- `CTGGTCATCCTGGCTAAAGGAGCAGAGGAAATG GAGACGGTCATCCCTGTAGATGTCATG`
- Nachází se na pozici 11011, 11072 a 11130 u krávy, člověka respektive psa
- Pěkný motiv, ale podle GenBank těsně za koncem první kódující sekvence

Nalezené motivy

- 16 pevných pozic (22 celkem)
- `GACGGCGCGCGTGCGTGCCGGC`
- Na pozicích 9894, 9955 a 10251 (kráva, člověk, pes)
- Odpovídá tomu, co bychom najít měli

Nalezené motivy

- 18 pevných pozic (celkem 26)
- `ggcgc GCGCCTGCGCAGTGCGGGGCTGAAGG ccaag`

- `ggcgt GAGTCTGCGCAGTGTGGGGCTGAGGG aggcc`
- `ggcgt GCGTCTGCGCAGTGCGGCGCCGAGGG ctcgc`
- `      **  * ** ********* ** *  **`
- Hvězdičky označují místa, která za motiv označil sampler
- Na pozicích 9860, 9923, 10004 (kráva, člověk, pes)
- Nejlepší nalezený motiv

Závěr

- Podařilo se nám relativně přesně určit začátkek genu DJ-1
- Jeden výpočet trval přibližně 5 minut
- Při hledání dlouhých motivů je možné najít podobné části i uvnitř genu

☺ Při jednom z běhů se nám podařilo najít motiv TTTTTTTTTTTTTTTTTTTT