SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br.000

Fraktalna vizualizacija evolucijskim algoritmima

Mirjam Škarica

Zagreb, svibanj 2014.

Umjesto ove stranice umetnite izvornik Vašeg rada.
Da bi ste uklonili ovu stranicu obrišite naredbu \izvornik.

Sadržaj

1. Uvod

2. Uvod u bioinformatiku

Bioinformatika je širok pojam, sljedeća definicija je predložena od strane NCBI-a (*The National Center for Biotechnology Information).*

> *Bioinformatika svodi pojmove biologije na razinu makromolekula te potom primjenjuje informatičke tehnike (izvedene iz disciplina kao što su primijenjena matematika, računarska znanost i statistika) kako bi se razumjele i organizirale informacije povezane s tim molekulama, u velikim razmjerima.* [1]

Svaka stanica je zapravo dinamičan sustav koji se sastoji od molekula, kemijskih reakcija i kopije genetskog materijala, odnosno genoma tog organizma. Makromolekule nukleinskih kiselina, proteina i ugljikohidrata su presudne za funkcioniranje svih poznatih živih organizama.

DNA (eng. *deoxyribonucleic acid*) i RNA (engl. *ribonucleic acid*) su nukleinske kiseline. DNA kontrolira aktivnosti u stanici određivanjem enzima i drugih proteina koji će se sintetizirati.

DNA se sastoji od dvostruke uzvojnice koje se sastoje o manjih gradivih jedinica nukleotida. Svaki nukleotid se sastoji od fosfata, šećera i nukleinskih baza (adenin, citozin, gvanin i timin) koje označavamo slovima A, C, G i T. DNA je organizirana upakiran u strukture koje nazivamo kromosomima.

RNA obično ima samo jedan lanac iako postoje i RNA s dva lanca. Uz šećer i fosfatne, nukleinske baze koje grade lanac su adenin, citozin, gvanin i uracil koje označavamo sa slovima A, C, G i T.

Proteini su makromolekule sastavljene od jednoga ili više lanaca aminokiselina, čine većinu biomase organizma. Obavljaju različite funkcije npr. ubrzavanje metaboličkih reakcija, umnažanje i prepisivanje DNA, i mnoge druge.

---

1 *„Bioinformatics is conceptualizing biology in terms of macromolecules (in the sense of physical-chemistry) and then applying "informatics" techniques (derived from disciplines such as applied maths, computer science, and statistics) to understand and organize the information associated with these molecules, on a large-scale."* [1]

Objašnjene toka genetskih informacija između DNA, RNA i proteina naziva se centralna dogma molekularne biologije.[2] Opći prijenos je umnažanje DNA, kopiranje informacije iz DNA u RNA. Ova sekvenca RNA dalje nosi informaciju u obliku, tzv. kodona. Kodon je slijed 3 uzastopne nukleinske baze koje određuju umetanje određene aminokiseline u polipeptidni lanac tijekom sinteze proteina, ili signaliziraju početak i prestanak sinteze istih. Postoji 64 različitih kodona, npr. AGU, CUC, GUU.

3. Uvod u evolucijske algoritme

Evolucijski algoritmi traže rješenje problema simulirajući proces evolucije. Ideju EA je predstavio I. Rechenberg 1960-tih godina u svom djelu „*Evolution strategie*" („Evolutionsstrategie"). Danas postoje mnoge verzije EA, ali je većina bazirana na istim principima.

Algoritam započinje stvaranjem određenog broja jedinki. Jedinke su izgenerirane nasumično i svaka od njih predstavlja moguće rješenje problema. Skup svih jedinki jedne generacije čini populaciju. Sljedeći korak je evaluacija populacije koja se radi pomoću funkcije dobrote gdje se svakoj jedinki pridružuje faktor dobrote, bolje jedinke imaju veći faktor dobrote. Izdvajajući određeni broj rješenja (jedinki), formira se nova populacija. Početna nova populacija mijenja se operatorima križanja i mutacije. Algoritam ponavlja ovaj proces sve dok nije pronađeno rješenje problema, odnosno dok jedna jedinka nije zadovoljila uvjete evolucije.

4 . IFS fraktali

Fraktal, kao pojam, skovao je Benoit Mandelbrot 1975. Fraktal je prirodna tvorevina, a opisujemo ga kao geometrijski uzorak koji se ponavlja (barem približno) na svim skalama umanjenosti tvoreći nepravilne oblike i površine koje se ne mogu predstaviti klasičnom geometrijom. Odnosno, fraktali su samoslični bilo da ih gledamo iz bliza ili iz daleka. Osobito se koriste u računalnom modeliranju nepravilnih uzoraka i struktura u prirodi.

IFS (engl. *iterated function systems*) su metode konstrukcije fraktala. Rezultirajuće konstrukcije su uvijek samoslične. IFS fraktali mogu biti bilo koje dimenzije, ali se najčešće računaju i crtaju u 2D. U principu, koristi se skupina jednostavnih transformacija kao što su rotacija, skaliranje i translatiranje kako bi se pomicala točka. Orbita koju dobijemo izborom i primjenom tih transformacija mnogo puta na tu točku je upravo IFS fraktal.

Ako želimo omeđenu orbitu, odnosno fraktal konačne površine, ne mogu se koristiti bilo kave transformacije već samo one za koje vrijedi da se za bilo koji par točaka njihova međusobna udaljenost smanjuje s primjenom te transformacije. Takve transformacije nazivamo kontrakcijskim mapama (engl. *contraction maps*). Formalno, $f : \mathbb{R}^2 \to \mathbb{R}^2$ je kontrakcijska mapa ako vrijedi:

$$d(p,q) \geqslant d(f(p), f(q)) \; \forall \, p, q \in \mathbb{R} \qquad (4.1)$$

Jedan od najpoznatijih primjera je IFS za trokut Sierpinskog. Neka su dane sljedeće transformacije (svaki redak predstavlja jednu od transformacija).

a	b	c	d	e	f	p_i
0.5	0.0	0.0	0.5	0.0	0.0	1/3
0.5	0.0	0.0	0.5	1.28	0.0	1/3
0.5	0.0	0.0	0.5	0.64	0.8	1/3

Tablica 4.1: transformacije za trokut Sierpinskog

Ove transformacije točku $T_i = (x_i , y_i)$ preslikavaju u točku $T_{i+1} = (x_{i+1} , y_{i+1})$, gdje

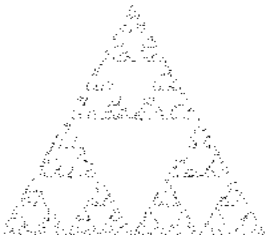
se x_{i+1} , y_{i+1} računaju pomoću funkcija:

$$x_{i+1} = a \cdot x_i + b \cdot y_i + e \tag{4.2}$$

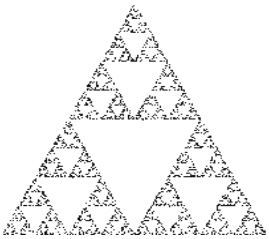
$$y_{i+1} = c \cdot x_i + d \cdot y_i + f \tag{4.3}$$

Fraktal, u ovom slučaju trokut Sierpinskog, dobijemo tako da na početnu točku iterativno primjenjujemo jednu od *3* transformacije iz tablice 4.1, s vjerojatnošću p_i. Općenito mora vrijediti $\sum_{i=1}^{n} p_i = 1$. Počevši od T_o= *(0,0),* i ponavljajući postupak preslikavanja velik broj puta dobiju se rezultati na slikama 4.1- 4.4.

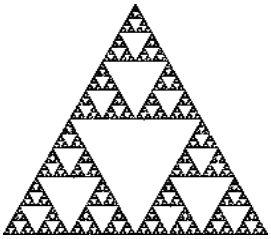
Slika 4.1: 200. iteracija



Slika 4.2: 800. iteracija



Slika 4.3: 3000. iteracija



Slika 4.4: 20000. iteracija

Lako se da zaključiti kako se korištenjem drukčijih vrijednosti koeficijenata, vjerojatnosti ili broja transformacija mogu dobiti vrlo različiti rezultati, pa time i fraktali.

4.1. IFS fraktali generirani ulaznim podacima

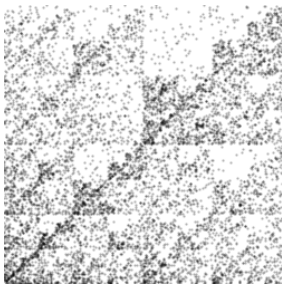
Na primjeru IFS za trokut Sierpinskog je pokazan način generiranja fraktala pomoću transformacija čiji izbor određuje slučajnost, odnosno njihove vjerojatnosti. Ovaj rad se na dalje bavi generiranjem fraktala pomoću transformacija čiji izbor određuju ulazni podaci. Da povučemo paralelu s DNA podacima, najjednostavniji primjer bio bi da imamo definirane neke 4 transformacije, da nam je ulazni niz upravo niz nukleinskih baza (A, C, G, T) te da pojava svake od baza uvijek rezultira primjenom iste transformacije. Algoritam bi tad slijedno prolazio kroz niz te na točku primjenjivao transformaciju koja odgovara zadnjem pročitanom znaku, $znak \in \{A, C, T, G\}$.

Primjer[2] ovih transformacija prikazan je tablicom *4.2*.

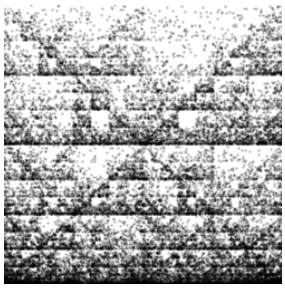
a	*b*	*c*	*d*	*e*	*f*	*znak*
0.5	0.0	0.0	0.5	0.0	0.5	A
0.5	0.0	0.0	0.5	0.5	0.5	T
0.5	0.0	0.0	0.5	0.5	0.0	G
0.5	0.0	0.0	0.5	0.0	0.0	C

Tablica 4.2: Transformacije za IFS fraktal generiran DNA nizom

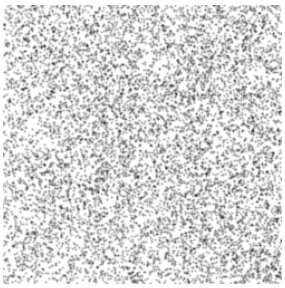
Što ove transformacije rade možemo lakše predočiti slikom 4.8. Svaka nukleinska baza "privlači" točku u svoj kut, tako da trenutnu točku preslika u novu na pola puta između "svog" kuta i trenutne pozicije točke. Na slikama 4.5-4.6 prikazani su IFS fraktali generirani ulaznim podacima HIV genoma, Methanococcus jannaschii genoma te s nasumično generiranim DNA podacima.



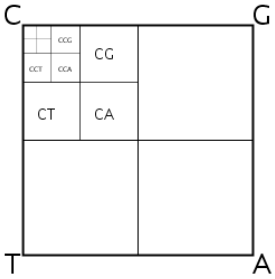
Slika 4.5 HIV



Slika 4.6 Methanococcus jannaschii



Slika 4.7 nasumično generirani podaci



Slika 4.8 legenda

Zanimljivo je primijetiti uzorke dijagonala, izostanak istih, te tzv. *double scoop* koji se može lako primijetiti na slici 4.5 kao relativno prazan prostor u svakom podkvadrantu GC. Pojava *double scoop* uzorka je prvi put zabilježena u ljudskoj beta-globin regiji i ukazuje na relativnu rijetkost uzastopne pojave gvanina i citozina.[4]

Upravo ovo brzo i jednostavno dolaženje do informacija je glavni motiv vizualiziranja DNA podataka, odnosno podataka općenito.

---

2 U literaturi se može pronaći pod nazovim DNA *driven four-cornered chaos game* ili c*haos game representation algorithm.*

4.2. Prikaz jedinke

Idealno bi bilo kada bi ista fraktalna reprezentacija radila ujedno s DNA, proteinima i kodonima. Iako se svi ovi oblici mogu dobiti iz bilo kojeg drugog, svaki od njih se pojavljuje u drugom stadiju biološkog procesa. Sirov DNA ima najviše informacija, ali najmanji stupanj interpretabilnosti, dok dijeljenjem DNA podataka u kodone, ona postaje je interpretabilnija (npr. kodon sadrži informaciju o termalnoj stabilnosti DNA)[3]. Po prijedlogu, [Ashclock, Golden] dalje u radu, izbor transformacije će ovisiti o kodonima koji dijele DNA u 64 moguće trojke.

Spomenimo još jednom kako svaka jedinka u evolucijskim algoritmima predstavlja jedno moguće rješenje problema. Zato je izbor i modeliranje strukture iste jedno od najvažnijih koraka u EA.

4.3. Sintetiziranje DNA podataka Markovljevim modelom

4.4. IFS fraktali generirani evolucijskim algoritmima

4.5. Korištene evolucijske strategije

5. Rezultati

6. Zaključak

Literatura

1. N. M. Luscombe, D. Greenbaum, M. Gerstein (2001) *What is Bioinformatics? A Proposed Definition and Overview of the Field*. http://www.ebi.ac.uk/luscombe/docs/imia_review.pdf
2. F. Crick (1958), *Central Dogma of Molecular Biology* http://cs.brynmawr.edu/Courses/cs380/fall2012/CrickCentralDogma1970.pdf
3. D. Ashlock, J. Golden, *Evolutionary Computation and Fractal Visualization of Sequence Data* http://eldar.mathstat.uoguelph.ca/dashlock/eprints/biochapter.pdf
4. *Achuthsankar S. Nair, Vrinda V. Nair, Arun K S, Bio-sequence Signatures Using Chaos Game Representation* http://deity.gov.in/hindi/sites/upload_files/dithindi/files/Bio-sewuence_AlpanaDey.pdf
5.

Fraktalna vizualizacija evolucijskim algoritmima

Sažetak

Sažetak na hrvatskom jeziku.

Ključne riječi: ključne riječi, odvojene zarezima.

Fractal visualization with evolutionary algorithms

Abstract

Abstract.

Keywords: Keywords.