VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

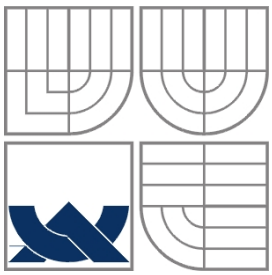
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

HĽADANIE GÉNOV V DNA

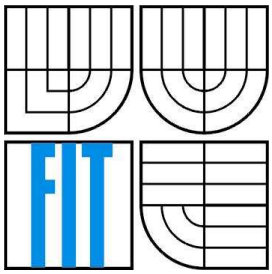
BAKALÁŘSKÁ PRÁCE
BACHELOR´S THESIS

AUTOR PRÁCE Zuzana Piovarčiová
AUTHOR

BRNO 2007



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

HĽADANIE GÉNOV V DNA

DNA GENE FINDING

BAKALÁŘSKÁ PRÁCE
BACHELOR´S THESIS

AUTOR PRÁCE — Zuzana Piovarčiová
AUTHOR

VEDOUCÍ PRÁCE — Ing. Ivana Rudolfová
SUPERVISOR

BRNO 2007

Abstrakt

Bioinformatika je rýchlo sa rozvýjajúcim vedným odborom, ktorý využíva techniky a metódy aplikovanej matematiky, informatiky, štatistiky, umelej inteligencie, chémie a biochémie na vyriešenie problémov týkajúcich sa biológie na molekulárnej úrovni. Táto práca ponúka vlastne štúdiu základných faktov, ktoré sú potrebné pri práci s bioinformatickým materiálom a následného vývoja aplikácií týkajúcich sa bioinformatiky. Je mnoho rôznych metód, ktoré s av bioinformatiek používajú,z čoho najznámejšie pri hľadaní génov v DNA sú parsovanie, HMM modely a rôzne druhy algoritmov, ako napríklad Viterbiho algoritmus.

Klíčová slova

Proteosyntéza, gén, genóm, DNA, metódy, ktoré sa používajú na hľadanie génov v DNA,DNA databázy, Hidden Markov Model, HMM, Viterbiho algoritmus, FASTA formát, GenBank, Microarrays, parsing

Abstract

Bioinformatics is a very quickly developing discipline, which uses technicques and methods of applied mathematics, computer science, artificial inteligence, chemistry and biochemistry to solve biological problems on the molecular level. This work offers accualy study of basic facts and knowledges which are needed in work with bioinformatic material and folowing bioinformatics aplication developement. There exists a lots of methods that are wide used in bioinformatics, from which the most known for gene finding in DNA are parsing, models of HMM and lots of other different algorithms like Viterbi algorithm.

Keywords

DNA,Genetic transcription, gene, genome, methods used for searching for genes in DNA, DNA databases, Hidden Markov Model, HMM, Viterbi algorithm, FASTA format, Gen Bank, Microarrays, parsing

Citace

Zuzana Piovarčiová: Hľadanie génov v DNA. Brno,2008, bakalářská práce, FIT VUT v Brně.

Hľadanie génov v DNA

Prohlášení

Prehlasujem, že som túto tuto bakalársku prácu vypracovala samostatne pod vedením Ing. Ivany Rudolfovej.
Uviedla som všetky literárne pramene a publikácie, z ktorých som čerpala.

……
Zuzana Piovarčiová
25. júla 2008

Obsah

1 Úvod

Bioinformatika je rýchlo sa rozvýjajúcim vedným odborom, ktorý využíva techniky a metódy aplikovanej matematiky, informatiky, štatistiky, umelej inteligencie, chémie a biochémie na vyriešenie problémov týkajúcich sa biológie na molekulárnej úrovni. Táto práca ponúka vlastne štúdiu základných faktov, ktoré sú potrebné pri práci s bioinformatickým materiálom a následného vývoja aplikácií týkajúcich sa bioinformatiky. Je mnoho rôznych metód, ktoré sa v bioinformatike používajú, z čoho najznámejšie pri hľadaní génov v DNA sú parsovanie, HMM modely a rôzne druhy algoritmov.

V tejto práci som mala za úlohu naštudovať a spracovať témy týkajúce sa DNA a proteosyntézy, získať prehľad o metódach, ktoré sa používajú na hľadanie génov v DNA. Zoznámiť sa s DNA databázami. Preštudovať HMM a Viterbiho algoritmus a pokúsiť sa nejaký model HMM navrhnúť a skúsiť ho čiastočne implementovať a odtestovať na dátach v DNA databází. Cieľom malo byť taktiež samozhodnotenie svojej práce a vypracovanie stručného plagátu reprezentujúceho moju prácu.

Radenie kapitôl

V úvodnej kapitole stručne charakterizujem obor bioinformatika, oblasti, ktorými sa zaoberá a jej hlavné ciele.

Nasleduje kapitola venovaná molekulárnej biológii, v ktorej sa hlavne zameriavam na oblasti, ktoré sú nutné na pochopenie pri mojom projekte. Tymito oblasťami je DNA a proces zvaný proteosyntéza.

Taktiež sa daju nájsť v tejto kapitole menšie odstavčeky venované skutočným základom biológie.

Nasleduje kapitola s názvom Laboratórne prostriedky na získavanie dát a ako už samotný názov naznačuje, sú tu opísané fyzické prostriedky na získavanie biologických dát (napr. Sequencery,microarrays …) a ich fungovanie v kocke.

Taktiež sú tu spomenuté jednolitvé metódy, ktorými sa dané dáta získavajú(rentgen etc.).

V ďalšej kapitole, už v poradí 5 je uvedený stručný prehľad algorimov a programov, ktoré sa používajú v rôznych oblastiach bioinformatiky.

Šiesta kapitola je venovaná genetickým databázam. Je tu uvedený stručný prehľad jednotlivých typov databáz,taktiež je tu spomenutá najväčšia existujúca databáza genetického materiálu GenBank.

Nasleduje stručný prehľad formátov, v ktorých sa môžu vyskytovať bioinformaticko dáta a bližšie roobraný je tu FASTA formát, kedže je to formát, s ktorým pracujem vo spojej práci.

Je tu tiež spomenutý menší návod,ako čítať jednotlivé dáta z GenBanku.

V siedmej kapitole som sa zamerala na HMM a Viterbiho algoritmus, ktoré sú tam detailne vysvetlené a nachádzajú sa tam aj moje HMM modely k danej práci.

Osma kapitola je venovaná záveru a zhodnoteniu práce.

2 Bioinformatika

Bioinformatika je rýchlo sa rozvýjajúcim vedným odborom, ktorý využíva techniky aplikovanej matematiky, informatiky, štatistiky, umelej inteligencie, chémie a biochémie na vyriešenie problémov týkajúcich sa biológie na molekulárnej úrovni.
V bioinformatike sa stretávame s obrovským množstvom dát, ktoré sa vyprodukujú pri výzkome v rôznych oblastiach biológie. Preto je nevyhnutné tieto dáta uchovávať a efektívne ich spracovávať.
Z tohto dôvodu medzi hlavné úlohy bioinformatiky patrí ukladanie, obnova, organizácia a analýza rozsiahlych súborov biologických dát.

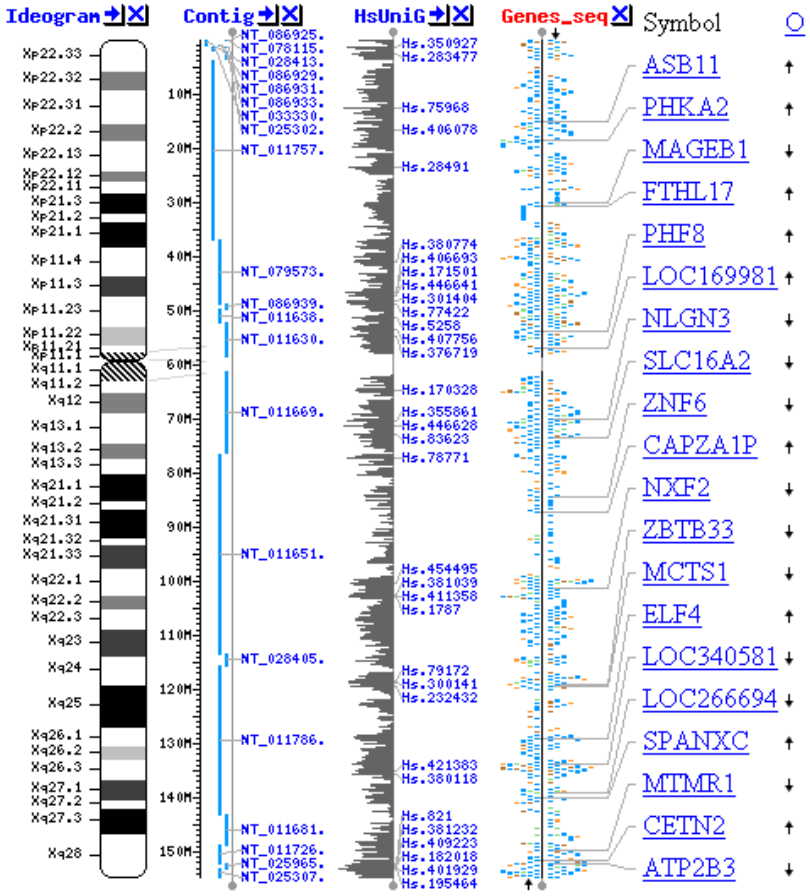
2.1.1 Hlavné oblasti, ktorými sa bioinformatika zaoberá

- **Analýza sekvencií** – Využíva sequencory. Ďalej sa o nich zmieňujem v kapitole XX a aj o základnom princípe ich fungovania. Ďalšou úlohou analýzy sekvecií je automatické hľadanie génov a regulačných sekvencií v genóme
- **Rozpoznávanie génov** [17]– Vyhľadávanie konkrétkych génov v dĺhých sekvenciách DNA/RNA
- **Počítačová evolučná biológia** – Evolučná biológia je veda o pôvode druhov, z čoho informatike jej v tomto smere trošku vypomáha. Jej cieľom je vytvorenie úplneho stromu života na základe bilogických dát.
- **Merania biodiverzity** – Informácie o biodiverzite našej planéty sú zhromažďované do obrovských databáz, z ktorých špecializovaný software nájde, zvizualizuje a zanalyzuje dostupné informácie. Týmto spôsobom sa modeluje napríklad rast populácie.
- **Analýza expresie génov** – Merajú sa pomocou microarrays- ktoré sú spomenuté delej v kapitole XX a vyexprimovanej sekvenčnej značky (EST),alebo postupnej analýzy génov (SAGE),alebo metódou zvanou massively paralel signature sequencing(MPSS)[18].
- **Analýza regulácie** – Regulácia je komplexný systém udalostí počnúc mimobunečnými signálmi,ako sú napriklad hormóny až po zvýšenie alebo zníženie aktivity proteinov. V tejto časti bioinformatiky sa vedci zaoberajú skúmaním jednotlivých príčin regulácie v živých organizmoch.
- **Analýza mutácie rakoviny**
- **Predpovedanie proteínovej štruktúry**
- **Porovnávanie genóv** – porovnávajú sa gény jednotlivých organizmov medzi sebou.
- **Modelovanie biologických systémov**

- **Vysokovýkonnostná analýza obrazu**
- **Spájanie proteínov (tzv. Protein docking)** - určovanie 3D štruktúry proteínu pomocou rentgénovej krystalografie a magnetickej rezonancie
- **Vývoj sofwaru a náastrojov na prácu s bilogickými dátami** – najznámejší program bioinformatike je asi BLAST[19]. Program na zarovnávanie sekvencií.
- **Webové služby v bioinformatike** – Základné služby v bioinformatike sú zoradene podľa EBI [20] do týchto kategórií: SSS[22] (Sequence search services – služby na hľadanie sekvencií), MSA[23] (Multiple Sequence Alignment – Viacnásobné zarovnanie sekvencií) a BSA (Biological Sequence Analysis – Analýza biologických sekvencií)

2.1.2 Momentálne ciele bioinformatiky

Momentálna roľa bioinformatiky je pomoc biológom pri zhromažďovaní a spracovaní genomických dát na štúdium funkcie proteínov. Jedna z daľších dôležitých úloh bioinformatiky je pomoc vedcom z farmaceutických spoločností v detailnej štúdii štruktúry proteínov, ktoré by uľahčilo návrh nových liekov.



Obrázok 1.1.[21]:
Mapa ľudského genómu

3 Základy molekulárnej biológie

V tejto kapitole získáme náhľad do základov molekulárnej biologie, ktoré sú nevyhnutné pri pochopení jednotlivých textov a práci v oblasti bioinformatiky. Budeme sa tu zaoberať základnými pojmami v oblasti molekulárnej biologie a taktiež si vysvětlíme jednolivé principy, ktoré sú veľmi dôležité na spávne pochopenie fungovania živých organizmov.

3.1 História

Ako na začiatku každého vedného odboru,aj na začiatku molekulárnej biológie stálo pár zvlášnych ľudí. Celé to začalo v roku 1665, keď jeden nekonformný individualista prevádzajúci verejné pitvy zvierat pod záštitou Royal Society v Londýne, menom Robert Hooke, objavil, že organizmy sú zložené z jednolivých častí, ktoré sa nazývajú bunky.
Neskôr vzniklá bunečná teória, ktorej tvorcami boli nemeckí botanici Mathias Schleiden a Theodor Swann v roku 1930 sa stala dôležitým mílnikom v dejinách molekulárnej biológie.
Základom tejto bunečnej teórie je, že v prírode existuje obrovská rozmanitosť buniek, ale všetky majú isté spoločné znaky. Všetky bunky majú svoj cyklus a tým je : zrod, získavanie energie, rozmnožovanie a smrť.

3.1.1 Genetický kód

Život na tejto planéte sa skladá z buniek. Tie sa delia na prokaryotické organizmy (organizmy, kroým obyčaje chýba jadro, sú to obyčajne baktérie a archey – organizmy, ktoré prežívajú v extrémnych podmienkach) a eukariotické organizmy, ktoré tvoria všetko zložitejšie živé na tejto planéte (od kvasiniek až po nás ľudí,zvieratá a rastliny).Každý organizmus musí však mať schopnosť uchovávať, využívať a ďalej posielať životne dôležité informácie. Tieto informácie sú uložené v jadre bunky (u eukaryotických buniek) alebo priamo v cytoplazme u prokaryotických buniek. Úplny genetický materiál každej bunky sa označije ako genóm a je uložený v DNA (deoxyribonukleova kyselina). Informácie, ktoré sú uložené vo vnutri DNA umožňujú zorganizovať všetky anorganické molekuly tak, aby dokázali vytvoriť živé bunky a z nich živé organizmy.

3.1.2 Štruktúra DNA[25]

DNA má tvar dvojitej pravotočivej šroubovice a obsahuje 2 vlákna.

Po mnohých pokusoch bolo dokázané, že genetický kód je tripletový, čo znamená, že jedna trojica bází kóduje jednu aminokyselinu. Tieto trojúseky na mRNA sa nazývajú **kodóny.** Dokopy existujú 4 bázy (nukleotidy). Tieto báze tvoria základnústavebnú štruktúru DNA. Jednotlivé bázy sú:

Adenín(A),guanín (G), cytozín (C) a tymín (T). Všetka genetická informácia je zapísaná v tomto štvorpísmenovom kóde. Poradie bází v reťazci sa označuje ako primárna štruktúra DNA.

Na kombináciu máme vcelku 4x4x4=64 možnosti. Dôležitý je triplet **ATG**, pretože tento triplet je inciačný (zároveň však kóduje methionín) a triplety **TAA**, **TAG** a **TGA**, ktoré sú označované ako triplety terminačné,

Z chemického hľadiska sú tieto nuklotidy tvorené fosfátovou skupinou, ribózou (cukor) a dusíkovou bázou, ktorou sa jednotlivé nukleotidy líšia. Nukleotidy sa spojujú za pomocou väzieb fosfátových skupín jednoho nuleotidu a cukru druhého nukeotidu. Tento polynukleotidový reťazec má dva rôzne konce, ktoré sa označujú ako 5’ koniec a 3’koniec. Označenie koncov reťazcov je veľmi dôležité, z toho dôvodu že väčšina procesov na týchto reťazcoch prebieha od 5’ konca k 3’ koncu. V rovnakom poradí sa tiež zapisujú a čítajú sekvencie nukleotidov.

Poradie nukleotidov v oboch vláknach nie je tovnaké,avšak sa riadi princípom komplemetarity. To znamená, že molekula guanínu tvorí pár s molekulou cytozínu a naopak. Tam, kde sa na jednom vlákne bude na chádzať molekula adenínu, bude sa na druhom vlákne nachádzať molekula thymínu.

Každá molekula adenínu a thymínu je schopná vytvoriť dve vodíkové väzby a každá molekula guanínu a cytozínu je schopná vytvoriť 3 vodíkové väzby. Vďaka týmto väzbám držia obe vlákna molekuly DNA dokopy.

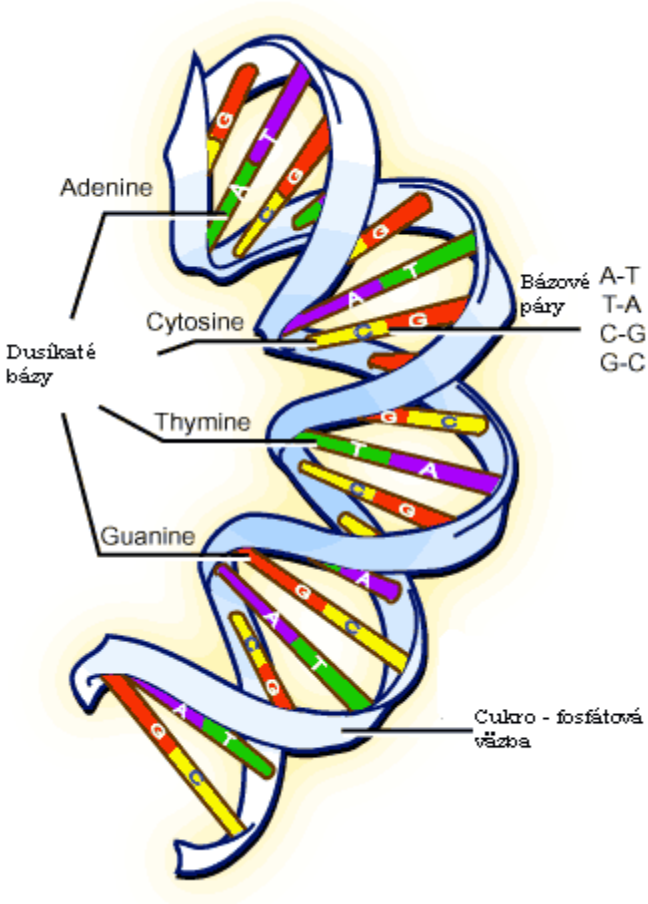
Ďalšou schopnosťou melekuly je jej schopnosť replikácie. Replikácia prebieha tak, že na určitom mieste dôjde k uvoľneniu oboch vlákien pôvodnej molekuly a na ne sa naviažu nové nukleotidy (na základe princípu komplemetarity). Tým vzniknú dva dcérske podreťazce molekuly DNA.

Jednotka zodpovedná za vznik dedičnej vlastnosti se označuje ako gén. Touto jednotkou je súvislý úsek molekuly DNA. Potom genóm je vlastne súbor všetkých génov organizmu. Ľudský genóm obsahuje približne 30000 génov a tvorí $3*10^9$ bází.

V DNA sa okrem kódujúcich oblastí (exony)tiež vyskytujú nekódujúce oblasti (introny). Bohužiaľ sa doteraz nepodarilo presne určiť funkciu intronov v genetickom kóde.

Ako som už vyššie zmienila DNA je uložená v jadre bunky u eukaryotických buniek a tam je rozdelená do niekoľkých chromozómov.

Náš genetický kód je degenerovaný, pretože asi pre 20 rôznych aminokyselín existuje oveľa viac kódujúcich kodónov. Z toho vyplýva, že niektoré aminokyseliny sú kódované viacerými tripletmi.
Na druhou stranu tento genetický kód je platný pre všetky organizmy na Zemi (existujú ale aj výnimky - napr. v genetickom kóde ľudských mitochondrií).



Obrazok 3.1. : Dvojvlákno DNA

3.2 Centrálna dogma molekulárnej biologie

Aby sa prejavili funkcie zakódované v DNA, musí dôjsť k expresii génov. Pri expresii génov dochádza k vytváraniu proteínov, ktoré umožňujú prejavenie informácie uloženej v DNA. Proces označovaný jako centrálna dogma molekulárnej biologie je procesom, pri ktorom dochádza k vyzdvihnutiu informácie uloženej v sekvenciú nukleotidov a vytvoreniu proteínov. Tento proces sa skladá z dvoch krokov: transkripcia a translácia.

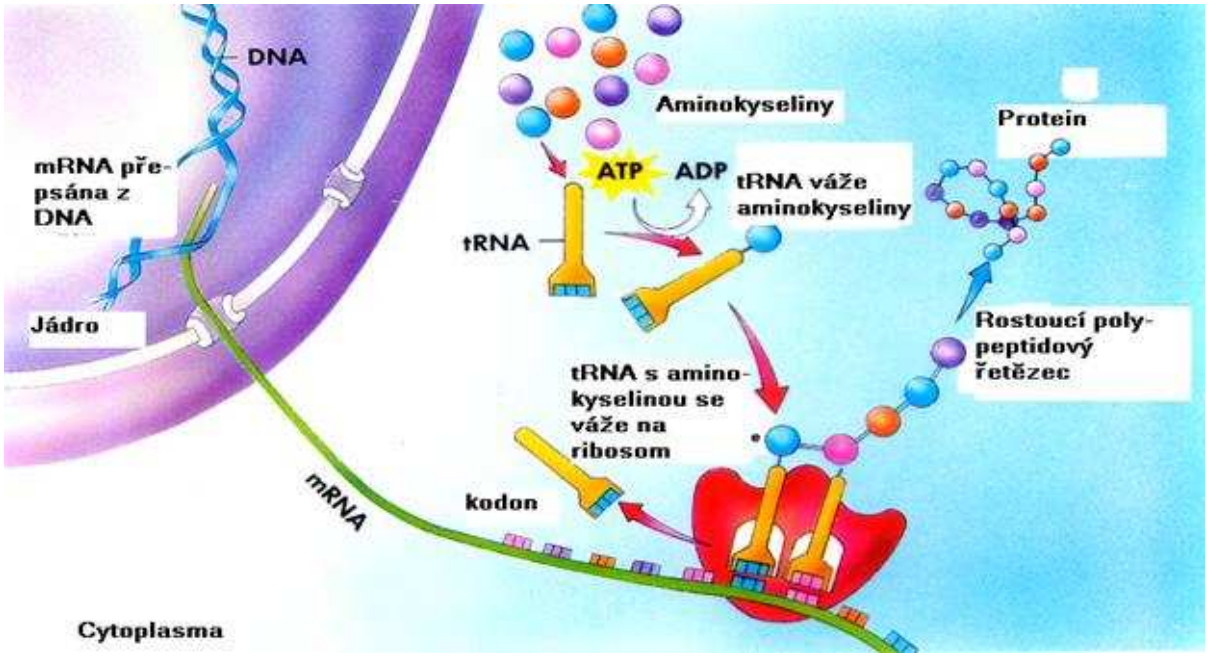
Transkripcia

Pri transkripcií dochádza k vztvorenie kópie génu do melekuly RNA (ribonukleova kyselina) za pôsobenia enzýmu RNA – polymeráza.RNA ma podobnú štruktúru ako DNA,avšak namiesto báze thymínu obsahuje uracil (U). Pri transkripcii sa z jadra uvoľní DNA a na základe princípu komplementarity sa vytvorí molekula RNA. Vzniknútá molekula RNA už neobsahuje introny,ke+d6e sa vystrihli (splicing).Hotové vlákno RNA vycestuje z jadra bunky.

Translácia

Proces pri ktorom sa preloží sekvencia nukleotidov v RNA na sekvenciu aminokyselín, ktorá vytvára proteín.

Takiež sa tento proces nazýva proteosyntéza.Je zahájená iniciačnou tRNA, to znamená, tá ktorá nese methionin. Tá sa naviaže na malú ribozomálnu podjednotku a začne pomaly prechádzať molekulu mRNA od 5' konce. Akonáhle objaví inicializačnú sekvenciu AUG - naviaže sa a translácia začína. Na ďalšie sekvencie (kodony) nasadajú daľšie tRNA podľa komplementarity bází (systém kodón na mRNA - antikodón na tRNA). Medzi prenesenými aminokyselinami vznikajú peptidové väzby. Za túto časť translácie je zodpovedná predovšetkým veľká ribozomálna podjednotka. Akonáhle zostáva už len kodón, ktorý je tzv terminačný je proteosyntéza ukončená a vzniknuté polypeptidové vlákno môže byť ďalej v bunke upravené na požadovanú bielkovinu.



Obrázok 3.2.1 : Proces translácie.

3.2.1.1 Tabuľka genetického kódu

Táto tabuľka nám určuje druh aminokyseliny, ktorý je pri proteosyntéze prenesený pomocou tRNA na ribozóm. Určujíce je pritom poradie nukleotidov v triplete na mRNA. Každému z nich zodpovedá nejaká aminokyselina. Niektoré aminokyseliny sú kódované viac než jednou možnosťou. Špeciálny význam majú pritom triplety **AUG**, ktore zahajujú proteosyntézu a triplety **UAA, UAG a UGA**, ktoré ju

ukončujú.

	U		C		A		G	
U	UUU	fenylalanin	UCU	serin	UAU	tyrosin	UGU	cystein
	UUC	fenylalanin	UCC	serin	UAC	tyrosin	UGC	cystein
	UUA	leucin	UCA	serin	UAA	**stop**	UGA	**stop**
	UUG	leucin	UCG	serin	UAG	**stop**	UGG	tryptofan
C	CUU	leucin	CCU	prolin	CAU	histidin	CGU	arginin
	CUC	leucin	CCC	prolin	CAC	histidin	CGC	arginin
	CUA	leucin	CCA	prolin	CAA	glutamin	CGA	arginin
	CUG	leucin	CCG	prolin	CAG	glutamin	CGG	arginin
A	AUU	izoleucin	ACU	treonin	AAU	asparagin	AGU	serin
	AUC	izoleucin	ACC	treonin	AAC	asparagin	AGC	serin
	AUA	izoleucin	ACA	treonin	AAA	lysin	AGA	arginin
	AUG	**metionin**	ACG	treonin	AAG	lysin	AGG	arginin
G	GUU	valin	GCU	alanin	GAU	kys.	GGU	glycin
	GUC	valin	GCC	alanin	GAC	asparagová	GGC	glycin
	GUA	valin	GCA	alanin	GAA	kys.	GGA	glycin
	GUG	valin	GCG	alanin	GAG	glutamová	GGG	glycin

.

4 Laboratórne prostriedky na získavanie biologických dát

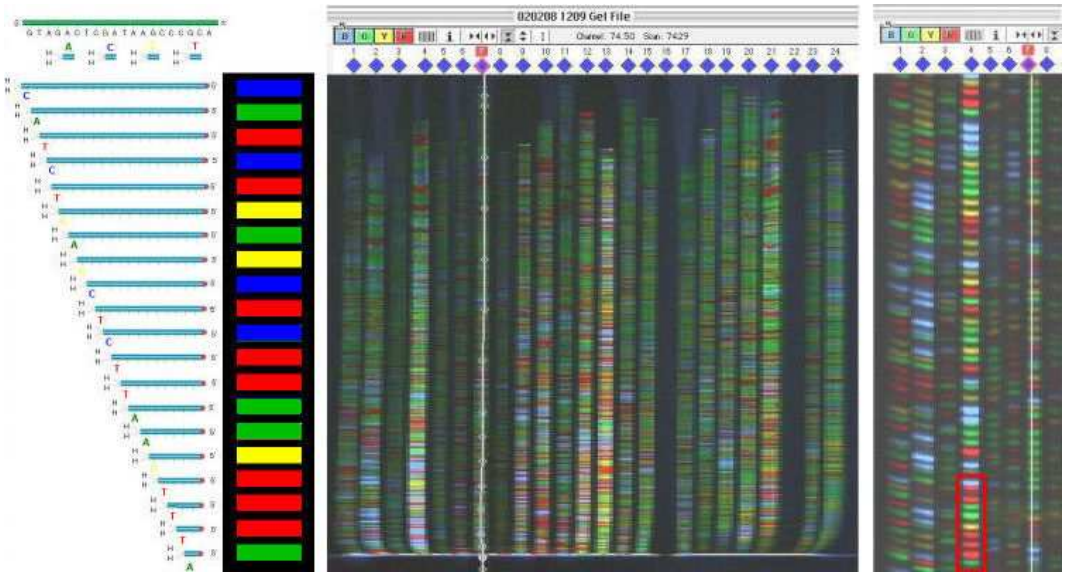
Prostriedky, ktoré v tejto podkapitole nie sú všetky, ktoré existujú, avšak by som rada predstavila aspoň menši prehľad tých základných. Tieto prostriedky sú totiž dôležité na správne pochopenie toho, aké dáta máme vlastne k dispozícii a aké ďalšie nástroje je potrebujeme použiť, na to, aby sme z nich získali čo najviac informácií.

Sequencers :

Sú to prístroje schopné vyčítať sekvenciu nukleotidov z vlákna DNA. Tieto prístroje sú priamo prepojené s počítačom, ktorý zároveň na to prevádza analýzu. Tento počítačový program, ktorý prevádza analýzu taktiež určuje stupeň vierohodnosti identifikovanej informácie z danej DNA. Sequencery dnešnej doby sú schopné prečítať a zanalyzovať viac ako 300 000 báz za deň za veľmi slušnú cenu.

Nevýhodou daných sekvencerov je, že sekvencie získané týmto spôsobom sú veľmi krátke a následne sa musí použiť počítačový program na rekonštrukciu celého genómu. (tzv. ho poskladá) Toto sa nazýva tzv. "shotgun method" (brokovnicová metóda) sekvencovania. Keďže DNA obsahuje mnoho sekvencií, ktoré sa opakujú, skladanie DNA môže byť trošku záludná záležitosť.

Obraok 4.1. [16] : Obrázok z automatizovaného sequencera



Spektrografia hmoty :

Táto technika zahŕňa určovanie génov a následne určenie korešpondujúcich proteínovv čistej forme.(určujú sa tým hlavne peptidy,lebo ich molekulárna váha sa dá ľahko určiť molekulárnym spektrografom. Z toho sa dajú potom ľahko vyvodiť zložky peptidov podľa ich molekulárnej váhy.

Rentgénová krystalógia a magnetická rezonancia (nuclear magnetic resonace - NMR):

Pomocou týchto dvoch metód sa získava 3D štruktúra proteínov. Obidve metódy sú veľmi náročné na čas a finančné prostriedky. V rentgénovej krystalografii sa vedci snažia zistiť 3D pozíciu každého atómu proteínu, z obrazu získaného prechodom rentgénových lúčov cez skryštalizovanú vzorku. Najzložitejšie na tejto metóde je získať dobrú vzorku kryštálu proteínu.

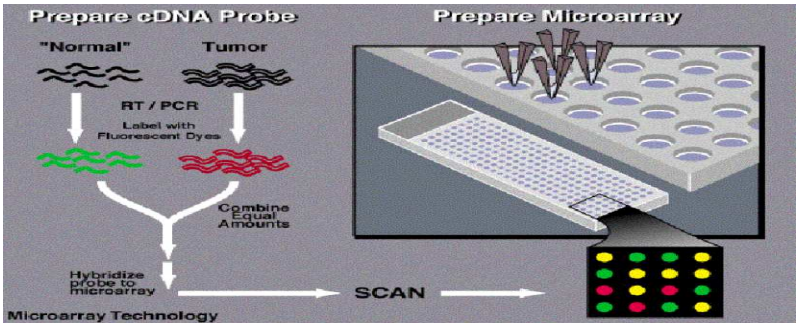
Pomocou NMR sa získa veľké množstvo matíc, ktoré vlastne určujú vzdialenosť dvoch atómov proteínu (tieto dva atómy sa však nesmú nachádzať v jednom a tom istom vlákne proteínu). Z týchto matíc sa dá vydedukovať 3D štruktúra.Veľkou výhodou NMR je, že dovoľujú štúdium pohyblivých častí proteínov, čo je nemožné v prípade predošlej metóde, kde je nutné použiť kryštály proteínov.

Microarrays:

Mikročipy, ktoré umožňujú určenie veľkého množstva mRNA, ktorá sa vytvorí z tisícov génov v tom istom čase (expresia génov). Následne RNA vytvára isté množstvo proteínov.

Tieto experimenty sa skladajú z troch fází. V prevej fázi sú umiestnené tisíce rôznych jednovláknových zhlukov RNA do miniatúrnych priehlbiniek na povrch malého skleneného čipu.

Táto DNA zodpovedá už určitému známemu génu. Druhá fáza sa skladá z rozšírenia jednovlákovnej DNA na povrch skleneného čipu, získaná pri experimente, ktorý chceme previesť. Táto druhá RNA môže byť z infikovanej bunky, bunky vystavenej vyhladoveniu a pod. Tieto dve RNA sa v každej priehlbinke čipu spoja. Stupeň skombinovaného materiálu získaného spájaním nukleotidov je indikátorom, koľko RNA je vytvorené každým génom, ktorý je študovaný. Tretia fáza spočíva v tom, že laserový skener pripojený k počítaču odmeria množstvo skombinovaného materálu v každej priehlbinke čipu a určí stupeň expresie génu. Dáta z týchto mikročipov sú široko dostupné vo veľých množstvách, avšak sú veľmi zašumené,takže niekedy spôsobujú ťažkosti pri interpretácii. Microarrays sú dôležitým prvkom,ktorý nám pomáha pochopiť fungovanie niekorých chorôb.

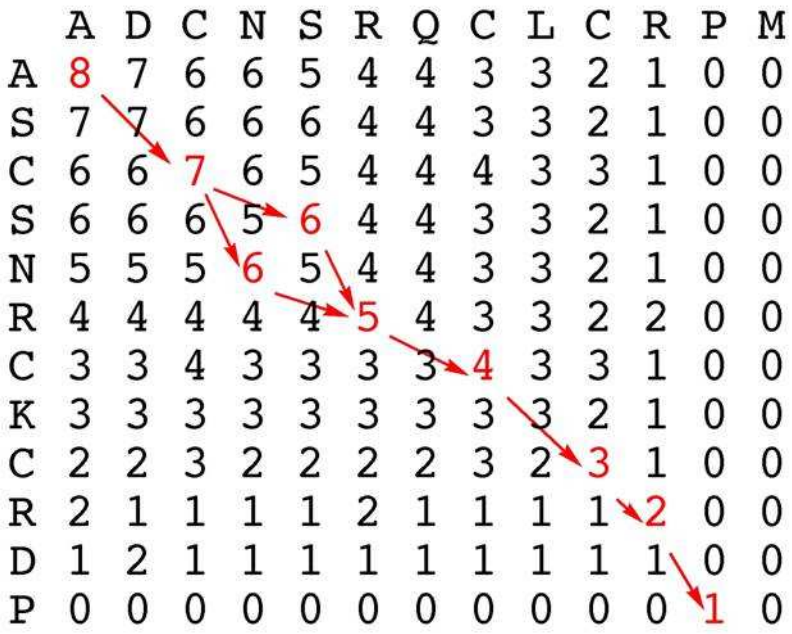


Obrazok 4.2.[15]:
Princíp použitia Microarrays.

5 Algoritmy často používané bioinformatike

5.1.1.1 Porovnávanie sekvencií

Z biologického hradiska je porovnávanie sekvencií motivované faktom, že všetky živé organizmy sú nejakým spôsobom navzájem spojené evolúciou.Z toho vyplýva, že druhy, ktoré sú si druhovo blízke by mali vykazovať podobnosti v DNA z čoho by sa dali vyvodiť aj jednotlivé funkcie génov.
Toto sa vykonáva podľa zarovnania skúmaných sekvecií a miery ich podobnosti.
Vzhľadom na velké množstvo sekvencií, ktoré sú verejne dostupné, vznikla nutnosť vivinúť algoritmy, ktoré by boli schopné porovnávať velké množstvo dlhých sekvencií. Tieto algoritmy by mali umožňovať vymazávanie, vkladače a výmenu symbolov reprezentujúcich nukleotidy alebo aminokyseliny.
Najznámejšiej dva algoritmy v porovnávaní sekvencií nukleotidov sú **Needleman-Wunsch** a **Smith – Waterman.** Needleman-Wunsch sa od Smith-Watermana líši tým, že používa váhy aj na vonkajšie okraje a tým sa snaží dosiahnúť najlepšieho grlobálneho zarovnania. Napriek tomu Smith-Waterman vychádza z výhody blízkosti zarovnaných segmentov.
Na zarovnanie sekvencií aminokyselín sa používajú trojúholníkové matice o veľkosti, v ktorých sú váhy na porovnanie zhodnosti, alebo rozdielnosti aminokyselín zakomponované, takisto ako aj označené medzery. Majčastejšie používané matice sú **PAM** (Percent Accepted Mutation) a **BLOSUM** (Blocks Substitution Matrix).
Najznámejšie programy na zarovnanie sekvencií je **BLAST** a **FASTA**.BLAST (Basic Local Alignment Search Tool) bol vyvinutý spoločnosťou National Center for Biotechnology information.Princíp fungovania Blatsu je, že sa sekvencia, ktorú hľadáme v obrovskej databázi sa rozdelí a menšie podsekvencie. Následne sa vyhľadáva pomocou hashovania a indexovania. BLAST sa neskôr pokúša o zhodu tým,že rozširuje z ľavej a z pravej strany danú sekvenciu (ktorú si pred tým rozkúskoval).Kde nenájde zhodu, tie sekvencie necháva a už sa k nim nevracia. S BLASTom sa dajú porovnávať ako sekvencie nukleotidov tak aj sekvencie aminokyselín. Používa tzv, p-value – je to hodnota, ktorá nám udáva mieru dôkazu, ktorý máme oproti hypotéze nula (Hypotéza nula nám určuje hodnotu, ktorú by sme dostali,keby náš hľadaný výsledok sme dostali iba náhodou). FASTA používa podobnú stratégiu ako BLAST. Ďalším takýmto programom je **Pipmaker**.



Obrázok 5.1 : Znázázorňuje fungovanie algoritmu Needleman – Wunsch (hladanie cesty s najnižšími hodnotami)

V mnohonásobnom zarovnávanie, ktoré sa obyčajne používa při aminokyselinách, o kterých sa předpokládá, že majú podobnú strukturu. Při mnohonásobnom zarovnaní sa veľmi často používajú Hiden Markov Models (HMMs). Taktiež existuje softwarový balík s názvom CLUSTAL a jeho varianty, ktorý sa veľmi často používa na mnohonásobné zarovnanie.

5.1.1.2 Evolučné (filogenické) stormy

Evolučné stromy sa často vytvárajú hneď po skončení porovnávania sekvencií DNA patriacich dvom rozličným organizmom. V týchto stromoch sa zoskupujú sekvencie podľa stupňa ich vzájomnej podobnosti. Toto nám slúži ako vodítko k tomu, ako sa vlastne sekvencie menili počas procesu evolúcie.Výsledkom takéhoto binárneho stromu, korého koreň reprezentuje prvotnú sekvenciu. Predpokladá sa, že táto prvotná sekvencia vygenerovala všetky ostatné. Existuje niekoľko druhov filogenických stromov v bioinformatike. Ja sa ymienim len o **Unrooted trees** (tzv. stormy bez koreňa). Sú to stromy, ktoré nám určujú vzdialenosti(rozdielnosti) medzi druhmi. Dĺžka cesty medzi dvoma listami nám určuje nahromadené rozdiely. Najčastejší algoritmus, ktorý sa používa na vytváranie Unrooted Trees je **UPGMA** (Unweighted Pair Group Method Using Arithmetic averages).
Ďalej sa môžem zmieniť o **Kladogramoch**. Sú to stromy, ktoré obsahujú koreň a ktorých dĺžka vetví nemá žiaden zvláštny význam.
Ďalej máme **Filogramy** – sú to vlastne rozšírené cladrogramy, v ktorých dĺžka vetví určuje veľkosť genetickej zmeny, ktorá sa objavila medzi daným uzlom a jeho priamym predchodcom.Pri zostavovaní filogramov sa veľmi často používa pojem parsimony(skúposť). Parsimony je založená na

hipotéze, že mutácie sa objavujú len veľmi zriedkavo. Z toho vyplýva, že počet mutácií, ktoré sa predpokladá, že sa objavia by mal byť minimálny.

Poslednými, o ktorých trošku napíšem budú **ultrametrické stromy**. Sú to filogramy, v ktorých nahromadené dĺžky od koreňa až po ktorýkoľvek list je vyčíslený tým istým číslom.. Ultrametrické stromy sú preto tie, ktoré nám dávajú najviac informácií o evolučných zmenách.

Najznámejším softwarovým balíčkom, ktorý sa používa pre konštrukciu filogenických stromov je **PHYLIP** (Phylogenic Interface Package)

5.1.1.3 Detekovanie vzorov v sekvenciách

Existujú isté časti DNA a aminokyselín, ktoré je potrebné detekovať.Dva hlavné príklady sú hľadanie génov v DNA(ako napr exony,introny a promoter regiony v DNA a intronov v RNA) a určovanie ich častí zo sekvencie aminokyselín (z ich sekundárnej štruktúry hraníc α- helixov a β- sheetov). Existuje niekoľko spôsobov, ako to docieliť. Niektoré sú založené na strojovom učení a obsahujú pravdepodobnostné gramatiky alebo neutrónové siete.Existujú dva prístupy, ktoré sa tu používajú a to gramatiky a parsovanie a Hidden Markov Models (tzv.HMMs). Aj v tu sa pri HMM používa Baum- Welchov algoritmus.

5.1.1.4 Určovanie 3D štruktúry zo sekvencií

Toto je veľkým problémom v bioinformatike. Určenie tvaru RNA zo sekvencie vyžaduje algoritmy kubickej komplexnosti. A v konečnom dôsledku určovanie tvaru proteínov z aminokyselín nie je doposiaľ úplne vyriešené.

5.1.1.5 Analýza regulácie bunky

Funkcia génu, alebo proteínu sa najlepšie ukáže v jeho roli v metabolických alebo signalizačných ciest. Gény medzi sebou reagujú. Proteíny môžu predchádzať alebo pomáhať pri produkcií iných proteínov. Dostupné modely bunečnej regulácie môžu byť diskrétne, alebo spojité.Najčastejšie sa tu využíva metód simulácie a modelovania. Pri diskrétnych modeloch sa zo štatistickým metód používajú k analýze regulácie bunky Bayesonove siete a support vector machines. Tieto dve metódy sa hlavne používajú na určenie správania génov z microarrays. Taktiež sa používajú clustering lgorithms (zhlukovacie algoritmy) na zhlukovnie dát o viditeľnej podobnosti správania génov.

E-CELL – ambiciózny japonský projekt, ktorý sa zameriava na simuláciu buniek s použitím stochastických systémov neliárnej diferenciálnej rovnosti. E-CELL simulátor je dostupný pre niekoľko platform.

5.1.1.6 Zisťovanie funkcie proteínov a metabolických ciest

Jedna z oblastí, ktoré sú zároveň najväčšou výzvou v bioinformatike a v ktorej neexistuje dostatok dostupných dát. Úlohou v tejto oblasti je vývoj databáz(napr japonský KEGG[24] Kyoto Encyclopedia of genes and genomes) reprezentujúcich grafy, ktoré sú špecifické kvôli existencii uzlov(špecifikujú reakcie) a ciest (špecifikujú postupnosti reakcií).

Taktiež bol vyvinutý meta-systém Metacyc an generovanie metabolických ciest z genomických dát rovnakých organizmov.

5.1.1.7 Skladanie fragmentov DNA

Fragmenty dodané sequencormi sa spájajú za pomoci výpočetnej techniky. Najtažšia časť v tejto úlohe je, že DNA obsahuje mnoho častí, ktoré sa opakujú a ten istý fragment môže patriť do rozdielnych oblasí genómu.

5.1.1.8 Skriptovacie Jazyky

Mnohé z už spomenutých algoritmov sú dostupné v aplikáciach, ktoré sú verejne dostupné na internete. Ich použitie si vyžaduje znalosť niektorého zo skriptovacích jazykov, ktorý priamo dodá dáta aplikácii, obdrží výsled a následne ho zanalyzuje.

Najčastejšie skriptovacie jazyky, ktoré sa používajú v bioinformatike sú Perl(a jeho obdoba pre bioinformatiku BioPerl)a Python.

6 Genetické databázy

Sekvenčné biologické databázy sú v skutečnosti obrovskými zbierkami DNA a proteínov,alebo aj iných sekvencií uložených na počítači. Daná databáza môže obsahovať sekvencie jedného organizmu, jako aj databázu všetkých dostupných proteínov, alebo taktiež môže obsahovať sekvencie vrtkých organizmov,kterých DNA bola kedy rozluštěná.
Sekvenčné biologické databázy sú neuveriteľným nástrojom, ktorý nám ponúka jedinečnú možnosť nahliadnuť do minulosti. Ponúkajú nám odpovede na dnešné otázky týkajúce sa biológie tým, že nám umožňujú analyzovať sekvencie, ktoré môžu byť až 25 rokov staré, kedy bola celá táto technológia vytvorená. Pomocou genetických databáz sa prepojila dnešná molekulárna biológia s molekulárnou biológiou minulosti.

Prvé databázy boli v podstate vytvorené ako nejaký druh sekvenčného múzea, kde by sekvencie mohli byť uchované na celú večnosť v pôvodnej podobe, presne tak, ako boli objavené, interpretované a publikované ich pôvodnými autormi. Toto historické stanovisko (časové púzdro) je stále zachované v databázach ako je GenBank, jeden z najhlavnejších skladov nukeotidových sekvencií udržiavané ako konzorcium medzi U.S. National Center for Biotechnology Information (NCBI), European Molecular Biology Laboratory (EMBL) a DNA Data Bank of Japan (DDBJ).

Typ skladiskovej databázy je vynikajúci, ak sa človek chce dostať k záznamom použitej literatúry k určitej sekvencii, ale neponúkajú jednoduchý prístup k sekvenčným dátam ako samotným. Z tohto dôvodu sa v druhej generácii genetických sekvenčných databáz prijal pohľad viac orientovaný na gény, kde všetky relevatné informácie o sekvenciách daných génov sú prístupné naraz. Jedným z príkladov databázy druhej generácie je tiež NCBI Entrez Gene.

Databáze biologických dát môžeme podľa informácií, ktoré obsahujú rozdeliť do nasledujúcich skupín[11]:

- **Databáze sekvencií nukleotidov** (EMBL-Bank, DDBJ)
- **Databáze sekvencií protenínov**(SWISS-PROT,TrEMBL)
- **Genómové database** (Ensembl)
- **Databáze štrukúr proteínov** (PDB,MSD)
- **Databáze obsahujúce informácie o expresii génov** (ArrayExpress)

Okrem týchto značne obsiahlych databáz existuje veľké množstvo ďalších databáz biologických dát.
Všetky tieto database zahŕňajú okrem už spomínamých sekvenčných dát aj ďalšie informácie týkajúce sa štrukúry génov, transkripčných mechanizmov a jednotlivé funkcie proteínov. Všetky tieto informácie sú integrované z mnohých rôznych zdrojov. Medzi jednotlivými databázami existuje

veľké množstvo odkazov, ktoré umožňujú veľmi jednoduché hľadanie vzťahov medzi molekulami rôznych typov.

*Zoznam existujúcich biologických databáz je uvedený v prílohe č.6

GenBank[13]

GenBank (Genetic Sequence Data Bank) je rýchlorastúce medzinárodné skladisko známych genetických sekvencií rôznych organizmov. Je to v podstate najdôležitejšia génová banka v modernej biológii a informatike. GenBank má otvorený prístup. Patrí pod National Center for Biotechnology Information (NCBI) a taktiež pod International Nucletide Sequence Database Collaboration(INSDC) Táto databáza spolupracuje s ostanými na celom svete a má uložených viac ako 100 000 genómov rôznych organizmov. Je to databáza rastúca exponenciálne,tzv sa zdvojnásobuje jej obsah každých 18 mesiacov.

6.1.1 Stručný prehľad formátov ktoré sa môžu vyskytovať bioinformatické dáta.

Existuje mnoho fomátov,v ktorých môžeme nájsť zapísané dáta z DNA zapísané.(približne 20 rôznych formátov). V nasledujúcich riadkoch ponúkam stručný prehľad najpoužívanejších formátov.

FASTA:

Programy FASTA a BLAST (Basic Local Alignment Search Technique) sú v podstate najznámejšími v bioinformatickom svete. Obydva používajú FASTA formát. Vďaka svojej jednoduchosti je FASTA najrožšírenejším a najpouýívanejším formátom, popri formáte GenBank.

Genetic Sequence Data Bank (GenBank):

GenBank je zbierkou verejne dostupných genetických dát. V tejt genetickej banke môžeme nájsť aj mnoho informácií navyše, nie len čisto informácie o genetickom kóde.

EMBL (European Molecular Biology Laboratory)

Databáza EMBL má veľmi pdobné dáta ako GenBank a DDBJ (Dna Databank of Japan), ale formát v ktorom sú zapísané je trošku iný.

Jednoducké dáta alebo inak Applied Biosystems (ABI) sekvenčný výstup

Tieto sekvenčné dáta DNA nemajú žiadne formátovanie, sú zapísane znakmi, ktoré reprezentuju jednotlivé bázy. Je to vlastne výstup, ktorý dávajú prístroje na čítanie sekvencií z ABI a iných prístrojov.

Protection Identification Resource (PIR)

Je to jeden z najväčších skladov sekvenčných dát proteínov.

Genetics Computer Group (GCG)

Program GCG (GCG Wisconsin Package) od firmy Accelrys sa používa na mnohých výskumých inštitúciách. Všetky dáta musia byť následne vo formáte GCG, aby mohli byť použiteňé ich programami.

Z týchto 6 už spomenutých formátov, GenBank a Fasta sú najznámejšie a najpoužívanejšie.

6.2 Fasta formát:

Je to textový formát slúžiaci na reprezentáciu sekvencií nukleových kyselín alebo sekvencie peptidov, v ktorej sú reprezentované bázové páry alebo aminokyseliny pomocou jednopísmenkového kódu. Tento formát tiež povoľuje používať mená sekvencií a komentáre, ktoré uvádzajú danú sekvenciu.

Jednoduchosť formátu FASTA dovoľuje ľahkú manipulovateľnosť a rozbor sekvencií pomocou nástrojov na spracovanie textu a skriptovacích jazykov ako je Python a Perl.

6.2.1 Opis

Genetická sekvencia vo FASTA formáte začína jednoriadkovým opisom, nasleduje riadkami obsahujúcimi už danú sekvenciu genetických dát. Opisný riadok je oddelený od sekvenčných dát znamienkom väčšítka (">") v prvom riadku. Slovo, ktoré nasleduje za väčšítkom ">" je identifikátor danej sekvencie a zbytok riadku je vlastne len opis (obidve sú voliteľné) . Odporúča sa, aby žiaden riadok nebol dlhší než 80 znakov. Sekvencia končí práve tam, kde znova narážame na znak väčšítka, ktorý určuje začiatok novej sekvencie.

Tu je jednoduchý príklad sekvencie vo FASTA formáte:

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

6.2.2 Zmeny formátu

Súbory FASTA môžu byť po dávkach premenené na ale z formátu MultiFASTA pomocou nástrojou jako sú **FASTA to multi-FASTA converter** a **multi-FASTA to FASTA converter**. Taktiež existujú nástroje na konverziu formátu chromatogramu, ktoré sú voľne dostupné na Internete. (ABI/SCF) na FASTA: **ABI2FASTA converter, Chromatogram explorer**.

6.2.2.1 Hlavička

Hlavička začína ako sme si povedali už pár riadkov dozadu znakom '>' a udáva meno a/alebo jedinečný identifikátor sekvencií a mnohokát aj iné informácie. Mnoho rôznych sekvenčných databází používa štandardizované hlavičky, ktoré pomáhajú automaticky extrahovať infomáciu priamo z hlavičky. Jedna hlavička môže obsahovať okrem samej seba ešte aj inú hlavičku oddelenú znakom ^A (Control-A).

Príklad hlavičky:

```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAKKADRLAAEG
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKDPNKPEHK
IPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQLDSKLTL
MGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKTEDFAAEVAAQL
>SEQUENCE_2
SATVSEINSETDFVAKNDQFIALTKDTTAHIQSNSLQSVEELHSSTINGVKFEEYLKSQI
ATIGENLVVRRFATLKAGANGVVNGYIHTNGRVGVVIAAACDSAEVASKSRDLLRQICMH
```

6.2.2.2 Reprezentácia sekvencie

Sekvencie,ktoré sa nám môžu vyskytnúť vo FASTA formáte môžu byť sekvencie proteínov,alebo nukleových kyselín. Tieto sekvencie môžu obsahovať medzery alebo znaky zarovnania. Sekvencie by mali byť zapísané v štandardnom IUB/IUPAC[12] kóde aminokyselín alebo nukleových kyselín. Existujú tieto výnimky:

- malé písmená sú akceptované a následne premené na veľké
- Čiarka,alebo pomlčka určujú medzeru (tzv. chýbajúce písmeno)
- V aminokyselinách sa akceptuje U a * ako písmeno
- Nie sú povolené číslice, avšak sa používajú v niektorých databázach na určenie pozície v sekvencii

6.2.2.3 Identifikátory sekvencií

Štandard definovaný NCBI unikátneho identifikátoru na sekvenciu (SeqID) sa nachádza v hlavičkovom riadku.

GenBank	gi|*gi-number*|gb|*accession*|*locus*
EMBL Data Library	gi|*gi-number*|emb|*accession*|*locus*
DDBJ, DNA Database of Japan	gi|*gi-number*|dbj|*accession*|*locus*
NBRF PIR	pir||*entry*
Protein Research Foundation	prf||*name*
SWISS-PROT	sp|*accession*|*name*
Brookhaven Protein Data Bank (1)	pdb|*entry*|*chain*
Brookhaven Protein Data Bank (2)	*entry*:*chain*|PDBID|CHAIN|SEQUENCE
Patents	pat|country|number
GenInfo Backbone Id	bbs|number
General database identifier	gnl|database|identifier
NCBI Reference Sequence	ref|accession|locus
Local Sequence identifier	lcl|identifier

6.2.2.4 Koncovky

Súbory formátu FASTA majú obyčajne koncovky ako .fa, .mpfa, .fna,.fsa, .fas alebo .fast

Podporované znaky v nukleových kyseinách sú:

Kód nukleovej kyseliny	Význam
-	Medzera neurčitej dĺžky
A	Adenozín
B	G T C (ale nie A) (**B** ide hneď po A)
C	Cytozín
D	G A T (ale nie C) (**D** ide hneď po C)
G	Guanín
H	A C T(ale nie G) (**H** ide hneď po G)
K	G T (Ketón)
M	A C (Amino skupina)
N	A G C T (a**N**y)
R	G A (Purín)
S	G C (Silná interakcia)
T	Timín
U	Uracil
V	G C A (nie T, ani U) (**V** ide hneď po U)
W	A T (Slabá interakcia)
X	Maskovanie
Y	T C (p**Y**rimidín)

Podporované kódy aminokyselín sú:

Kód aminokyseliny	Význam
A	Alanín
B	Kyselnina ašpargová alebo tzv. Aspargín
C	Cysteín
D	Ašpargová kyselina
E	Kyselina glutamová
F	Fenilanín
G	Glycín
H	Histidín
I	Izoleucín
K	Lysín
L	Leucín
M	Methionín
N	Ašpargín
O	Pyrolysín
P	Prolín
Q	Glutamín
R	Arginín
S	Serin
T	Threonín
U	Selenocystein
V	Valin
W	Tryptopan
Y	Tyrosin
Z	Kyselina glutamová alebo Glutamín
X	Akákoľvek
*	Koniec translácie
-	Medzera neurčitej dĺžky

Ako prečítat hlavičku v GenBank u prokaryotických buniek

Typickými kľúčovými slovami u hlavičky sú :

LOCUS : dáva nám meno miesta, kde sa gén v genóme nachádza, veľkosť sekvencie nukleotido v bázových pároch, charakter molekuly (v našom prípade je to DNA) a jeho topológiu (či je to priama alebo ohrúhla molekula).

DEFINITION : ponúka krátku definíciu vstupnej sekvencie.

ACCESSION : uvádza vstupné číslo – jedinečný identifikátor, ktorý platí v rôznych databázach (ako napríklad proteínová databáza) .

VERSION : uvedie rovnaké, alebo ID čísla, ktoré mala daná v sekvencia v minulosti.

KEYWORDS : predstavuje list výrazov, ktoré obšírnejšie charakterizujú našu vstupnú sekvenciu. Tieto výrazy sa dajú použiť ako kľúčové slová na hľadanie v určitých databázach.

SOURCE : nám prezrádza klasické meno daného organizmu, ktorému patrí sekvencia, ktorú sme si dali vyhľadať.

ORGANISM :udáva komplexnejšiu identifikáciu organizmu, komplexnú aj jej technickou taxonomickou klasifikáciou.

REFERENCE : uvádza časť, kde sú uvedené mená(zásluhy za objavenie) objaviteľov danej sekvencie (rôzne časti sekvencie mohli byť objavené rôznymi autormi). Táto časť obsahuje rôzne celky, za ktoré hovorí sám nazov: AUTHORS,TITLE,JOURNAL, PUBMED.

COMMENT : Obsahuje take veci ako podakovanie,alebo informácie, ktoré priamo nesedia ani do jednej z vyššie spomínaných častí

FEATURES : Presne vysvetľuje oblasti génova ich biologické vlastnosti , ktoré boli identifikované z danej sekvencie nukleotidov. Táto časť obsahuje pár ďalších kľúčových slov, ktoré tu ozrejmím:

- **source:** označuje pôvod špecifických oblastí v danej sekvencii. To je veľmi užitočné, ak chce niekto rozoznať klonovací vektor od hosťovskej sekvencie.
- **promoter :** indikuje presné súradnice promoter oblasti.
- **misc feature :** indikuje domnelé oblasti, kde by mohla transkripcia začínať (mRNA syntéza)
- **RBS** (Ribosome Binding Site) **:** indikuje oblasť časti 3'konca
- **CDS** (CoDing Segment) **:** uvádza časť, v ktorej je opis génov uvedený vo formáte *"open reading frame"* (ORF)
  - Prvý riadok určuje súradnice ORF z počiatočného **A**TG do posledného nukleotidu prvého stop kodónu TA**A**

- Každý z nasledujúcich riadkov (odsadený na rovnakej úrovni) pomenúváva proteínový výsledok a určuje tzv. reading frame, využiteľný genetický kód, a mnoho IDs pre proteínové sekvencie.
- /translation – posledné dôležité slovo CDS sekcie, uvádza aminokyseliny kódovacieho segmentu. Táto sekvencia je v podstate počítačový preklad, ktorý využíva už vyššie spomínané súradnice, reading frame a genetický kód uvedený v predchádzajúcich riadkoch.

❖ Další **misc feature –** nasleduje po CDS sekcii. Obsahuje riadky, ktoré by mohliobsahovať stem-loop struktúru a opakovania. Toto sú potenciálne regulačné zložky dUTPase génu.

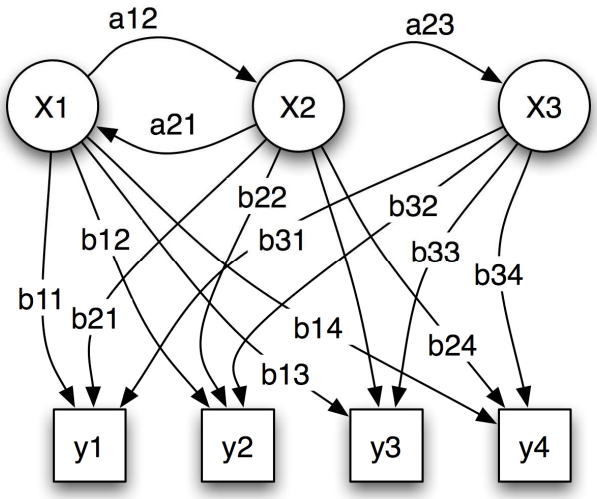
7 HMM a Viterbiho algoritmus

7.1 HMM (Hidden Markov Model)

Hidden Markov Model („schovaný markovovský model) je stochastický model, ktorý zachytáva štatistický charakter dát v pozorovanom reálnom svete. V HMM sa modelovaný systém pokladá za Markovský process s neznámimi parametrami a úlohou je určiť neznáme parametre z premenných, ktoré môžme vypozorovať. Vyextrahované parametre modelu môžu byť neskôr využité na ďalšiu analýzu, napríklad v aplikácii rozoznávajúce obrazce. HMM môže byť tiež považované za najjednoduchšiu Bayesonovu dynamickú sieť.
Spávne vymodelovaný HMM presne modeluje zdoroj z reálneho sveta, ktorý nám dáva pozorované dáta a má schopnosť imitovať(simulovať) zdroj.

V obyčajnom Markovovom modeli, každý stav je priamo viditeľný pozorovateľovi, a preto pravdepodobnosti stavových prechodov sú jeho jedinými parametrami. V skrytých markovových modeloch, stav nie je priamo viditeľný, ale premenne ovplyvnené daným stavom viditeľné sú.
Každý stav má pravdepodobnostné royloženie nad možnými výstupnými symbolmi. Preto vlastne postupnosť daných symbolov generovaná našim HMM nám dáva informácie o postupnosti stavov.

Hidden Markov Models sú obvzvlášť známe pre ich použitie v časovom rozoznávaní vzorov ako napríklad reč, písmo, rozoznávanie giest, prepis hudby, čiastočná priepustnosť elektrických obvodov a bioinformatika.



Pravdepodobnostné parametre HMM

x = stavy
y = možné pozorovania
a = pravdepodobnosti prechodu stavov
b = výstupné pravdepodobnosti

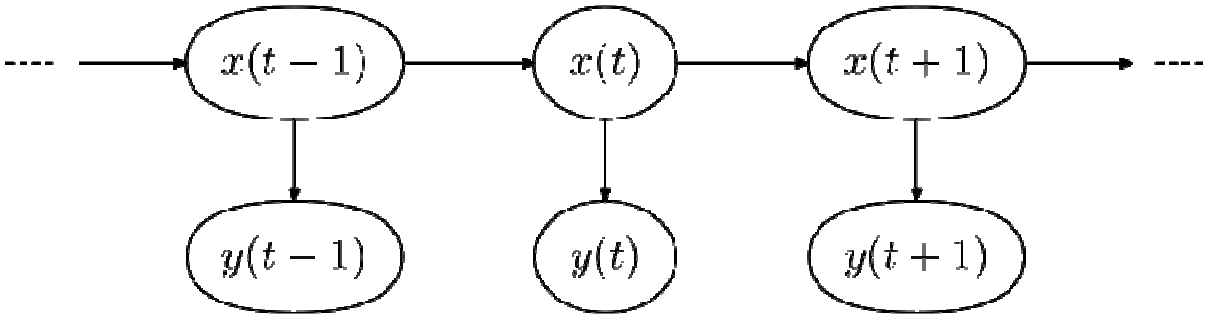
Architektúra skrytého Markovovho modelu

V diagrame, ktorý je uvedený pod týmto textom je ukázaná základná architektúra ukážkového HMM. Každá oválna značka reprezentuje náhodnú premennú, ktorá môže nadobúdať niekoľko rôznych hodnôt. Náhodná premenná *x(t)* je skrytý stav v čase *t* (s modelom, ktorý sa nachádza na predošlej stane, $x(t) \in \{x_1, x_2, x_3\}$

Náhodná premenná *y(t)* je pozorovanie v čase *t* ($y(t) \in \{y_1, y_2, y_3, y_4\}$

Šípky v diagrame (často taktiež nazývaný aj mriežkový diagram) vyznačujú podmienkové závisloti.

Z diagramu taktiež veľmi jasne vyplýva , že hodnota skrytej premennej *x(t)* (v čase *t*) záleží čisto iba na hodnote skrytej premennej *x(t – 1)* : hodnoty v čase *t* – 2 a pred tým nemajú na hodnotu skrytej premennej *x(t)* (v čase *t*) žiaden vplyv. Toto sa taktiež nazýva markovovská vlastnosť. Pdobne hodnota pozorovanej premennej *y(t)* záleží iba na hodnote skrytej premennej *x(t)* (obydve v čase *t*)



Pravdepodobnosť sledovanej postupnosti

Pravdepodobnosť spozorovania sekvencie $Y = y(0), y(1), \ldots, y(L-1)$ ke L je daná

$P(Y) = \sum_X P(Y \mid X) P(X)$, de suma vlastne poníma všetky uzlové sekvencie $X = x(0), x(1), \ldots, x(L-1)$

Hrubý mechanický výpočet *P*(*Y*) je neriešiteľný pre väčšinu skutočných problémov, takisto ako aj počet možných skrytých uzlových sekvencií je veľmi vysoký.

Napriek tomu môže byť výpočet veľmi urýchlený použitím forward algoritmu alebo ekvivaletneho backward algoritmu.

História HMM

Skryté Markovové modely boli po prvý krát charakterizované v seriáli štatistických novín Leonardom E. Baumom a inými autormi v druhej polovici rokov šesťdesiatych minulého storočia. Jedna z prvých aplikácii,využívajúcich HMM, bola práve na rozoznávanie reči počiatkom polovice 70tych rokov minulého storočia.

V druhej polovici rokov osemdesiatych minulého storočia, HMM začal byť využívaný k analýze biologických sekevencií, presnejšie DNA. Od tej doby sa stal všadeprítomný v oblasti bioinformatiky.

Problémy spojené s HMM

Existujú tri základné problémy spojené s použitím HMM:

- Pomocou daných parametrov modelu spraviť výpočet pravdepodobnosti istej výstupnej sekvencie a pravdepodobnosť hodnôt skrytého stavu daných výstupnou sekvenciou. Tento problém rieši **forward – backward algoritmus**.
- Pomocou daných parametrov modelu nájsť najpravdepodobnejšiu sekvenciu skrytých stavov, ktorá by bola schopná vygenerovať daný výstup sekvencie.
  Tento problém rieši Viterbiho algoritmus.
- Pomocou danej výstupnej sekvencie alebo množiny takejto sekvencie, nájsť najpravdepodobnejšiu množinu prechodov stavov a výstupných pravdepodobností.
  Inými slovami povedané, nájsť parametre HMM určeného dátovou množinou sekvencií.
  Tento problém rieši Baum – Welchov algoritmus.

HMM v Bioinformatike

Obor bioinformatiky zahŕňa využitie informatických teórií a prístupov v problematike biológie a medicíny. V bioinforatike sa už v tejto dobe existuje mnoho molekulárnych databáz zhromaždených z rôznych druhov organizmov. Tieto databázy sú verejne prístupné z ktoréhokoľvek miesta tejto planéty, čo je veľkou motiváciou pre bioinformatikov. Hoci prístupnosť týchto dát vyznačuje nový vek v biologickom výzkume, sámo osebe to neposkytuje žiadne biologicky významné informácie. Cieľom bioinformatiky je v podstate vysvetliť dodatočné informácie týkajúce sa kódovania proteínov, funkcie proteínov a fungovania mnohých iných bunečných mechanizmov

pomocou analýzy dát z už spomínaných databáz. Tieto nové informácie sú dôležité pri vývoji nových liekov, lekárskych diagnóz, následného liečenia a nespočetných odvetví výzkumu.

Väčšina prvotných (nijak nespracovaných dát) molekulárnych dát používaných v bioinformatike zodpovedá sekvencii nukleotidov zodpovedajúcich primárnej štruktúre DNA a RNA alebo sekvencii aminokyselín zodpovedajúcich primárnej štruktúre proteínov. Z tohto dôvodu problém dosahu vedomostí z týchto dát patrí do širšej triedy problémov analýzy sekvencií.

Dva z najštudovanejších problémov sekvenčnej analýzy sú rozpoznávanie reči a spracovanie reči.
Biologické sekvencie majú rovnaký lineárny vzhľad ako postupnosti zvukových signálov vyskytujúce sa v reči a sekvencie slov reprezentujúcich jazyk. Následne hlavný bioinformatický problém v analýze sekvencií môže byť vykreslený ako linguistický problém. Bežné linguistické porekaldlo je v bioinformatike, že akákoľvek skupina proteínov sa dá namapovť ako rozoznávanie reči.
Toto rčení navrhuje interpretovať rôzne proteíny patriace do tej istej skupiny ako rôzne znenie toho istého slova. Iná metafora sa odkazuje na hľadanie sekvencií génov v DNA ako rozoberanie jazyka do slov na semantický zrozumiteľné vety. Z toho vyplýva, že so sekvenciami génov môžeme zaobchádzať linguisticky a použiť tie isté techniky, ktoré by sme použili na rozoznávanie reči alebo spracovanie reči.

Použitie HMM v špecifických problémoch

Je isté, že prístup založený na HMM je logický na riešenie problémov v bioinformatike. Na základe tohto prístupu už vzniklo veľmi veľa aplikácií. Baldi a Bunk [2] definujú 3 hlavné druhy problémov.

- HMM sa dá využiť na mnohonásobné zarovnanie DNA sekvencií. Toto je v celku zložitá úloha, pri ktorej je nutné použiť dynamické programovanie.
- Štruktúra nami skúšaného HMM môže odhaliť podobnosti v biologických dátach. Tieto v podobnosti sa používajú na odhalenie periodickosti v istých oblastiach dát a pomáhajú predpovedať oblasti dát, ktoré majú sklon formovať špecifické štruktúry.
- Veľké množstvo klasifikačných problémov.
  Prístupy založené na HMM boli využité v predpovedaní štruktúry, konkrétnejšie sa snažili vyriešiť problém s klasifikovaním každého nukleotidu podľa štruktúry, do ktorej patrí. HMM bolo tiež využité na profilovanie proteínov z dôvodu rozlíšenia medzi rôznymi druhmi proteínov a predpovedanie nového proteínového druhu alebo poddruhu.
- Hľadanie génov v DNA.
  Tento problém je založený na klasifikácii DNA bází podľa ich ich funkcie počas transkripcie.

Problém zvaný hľadanie génov

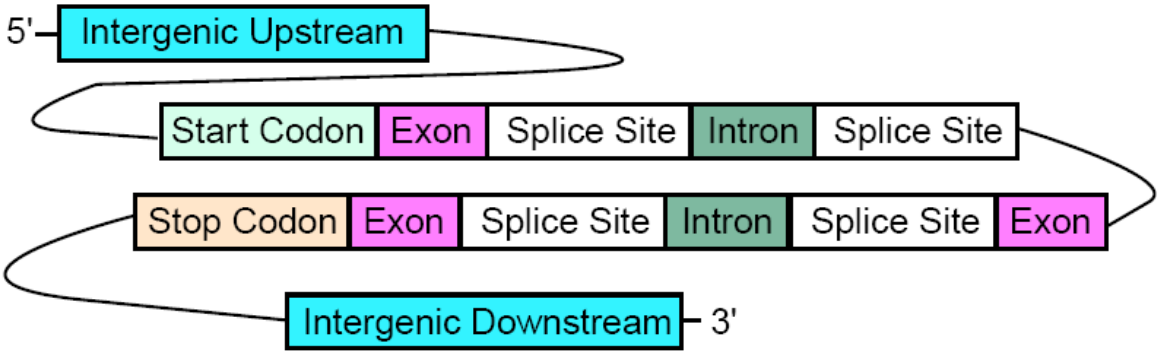
Problém hľadania génov v DNA je problémom už po niekoľko rokov. Bol to jeden z prvých problémov vypichnutých hneď ako vyšiel na svetlo sveta a bolo dostupné dostatočné množtvo genetických dát. V danom probléme máme danú nejakú sekvenciu DNA a máme určiť, na ktorých miestach sa vlastne gény nachádzajú, ktoré sú vlastne oblasti obsahujúce informáciu o kódovaní proteínov. Najvšeobecnejšie sad á povedať, že nukleotidy môžu patriť do úsekov, ktoré kódujú gén, nekódujúce oblasti alebo medzigenetické oblasti. Potom by problem hľadania génov mohol byť formulovaný následovne :

Vstup: Sekvencia DNA $X = (x_1, \dots x_n) \in \sum^*$, kde $\sum$ = A, C, G, T

Výstup : Správne označenie každého prvku z množiny X ako prvku patriaceho do kódujúcej oblasti,neódujúcej oblasti a medzigenetickej oblasti.

Hľadanie génov sa stáva oveľa zložitejšie v prípade, že k nemu budeme pristupovať v širšom biologickom kontexte. Eukaryotické bunky obsahujú kódujúce regióny – exony, ktoré sú prerušované nekódujúcimi regiónmi – intronmi. Exony a introny sú oddelené takzvanými spojovacími časťami. Oblasti mimo génov sa nazývajú medzigénové. Hlavným cieľom hladania génov je označiť množiny genetických dát, špecificky a obvzvlášť oblasti, kde sa nachádzajú exony, introny a tzv. promoter regions.

Bolo už vytvorených mnoho nástrojov na hľadanie génov. Jednými z najznámejších sú Genie[6], GeneID [4], a HMMGene[5]. [3] Napriek tomu, že každý z nástrojov má trošku odlišný model, všetky využívajú techniku kombinácie niekoľkých špecializovaných submodelov do väčšej sústay. Submodely zodpovedajú priamo rôznym oblastiam DNA definovaných podľa ich funkcie počas transkripcie. Väčina nástrojov použíaných na hľadanie génov sú hybridné modely, ktoré obsahujú tiež prvky neutrónových sietí. V týchto nástrojoch, narozdiel od HMM, neutrónová sieť modeluje isté oblasti, ako napríklad spojovacie oblasti.[7] Rôzne programy založené na HMM na hľadanie génov kombinujú submodely do väčších modelov zodpovedajúcich organizácii génu v DNA a jeho funkčnosti. Následný obrázok ukazuje rôzne oblasti DNA rozdelené podľa ich funkcie v transkripcii a translácii.



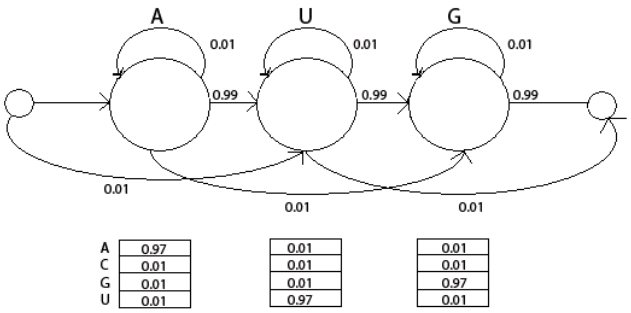
Splice site : genetická spojovacia časť

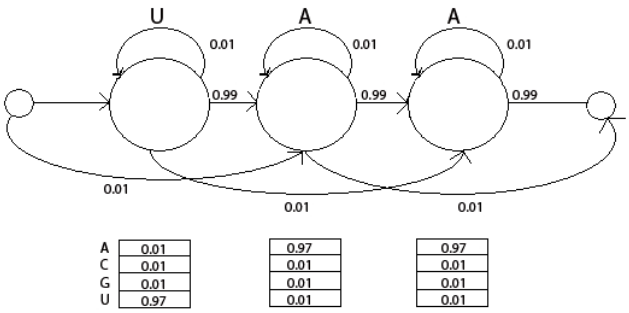
Integenic upstream : tzv. 5’ začiatok

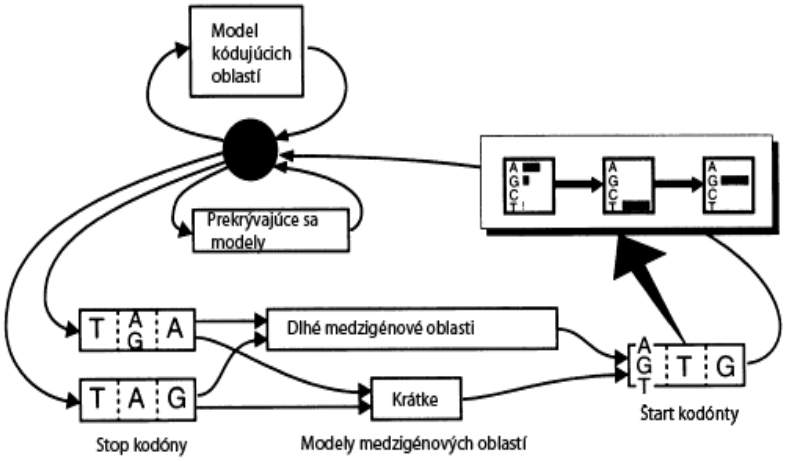
Integenic downstream : tzv. 3’ koniec

Obrazok 7.1.1.[8]: Základná štruktúra génu je taktiež základnou štruktúrou modelu na hľadanie génov.Jednotlivé časti zodpovedajú funkčne rozdielnym oblastiam v DNA

Konkrétne modely na hľadanie génov







Obrazok vľavo hore: HMM pre hľadanie počiatočného kodónu v gene.

Obrázok vpravo hore: HMM

Pre hľadanie jednoho z koncových kodónov (UAA, UAG, UGA). Modely pre ostatné konce sú rovnaké.

Obrázok vľavo dole: HMM architectúra pre kompletný génóm.

7.2 Viterbiho algoritmus

Viterbiho algoritmus je dynamický programovací algoritmus na hľadanie najpravdepodobnejšej sekvencie v skytých stavoch – takzvaná Viterbiho cesta. Ako výsledok viterbiho cesty dostaneme sekvenciu pozorovaných stavov, používa sa hlavne v súvisloslosti so skytými Markovovmi modelmi,presnejšie s Markovovými informačními zdrojmi.
Algoritmus používa techniku zvanú "forward dynamic programming" . Tento algoritmus sa hlavne používa na výpočty pravdepodobnosti sekvencie pozorovaných javov. Tento algoritmus patrí do podmnožiny tvoriacej informačnú teóriu.

Algoritmus vytvorí niekoľko predpokladov. V prvom rade, obidva javy (či už skrytý,alebo pozorovaný) musia za sebou nasledovať. Toto poradie často zospovedá určitému času. V druhom rade, tieto dve sekvecie musia byť zarovnané a zároveň každá pozorovaná udalosť musí korešpondovať presne s jednou skrytou udalosťou. Za tretie, výpočet najpradepodobnejšej skrytej sekvencie do určitého bodu t musí záležať iba na pozorovanej udalostiv bode t a najpravdepodobnejšej sekvencii v bode t-1. Tieto všetky predpoklady by mal zahŕňať už spomínaný Hidden Markov model.

Pojmy ako Viterbiho cesta a Viterbiho algoritmus sa takisto používajú v príbuzných dymanických programovacích algoritmoch, ktoré odhaľujú najradepodobnejšie vysvetlenie pozorovania.
Na príklad: v štatistickom gramatickom rozbore sa môže použiť dynamický programovací algoritmus na objavenie bezkontextovej odchylky v reťazci, čo sa taktiež niekedy nazýva „Viterbiho rozbor".

Viterbiho algoritmus po prvý krát formuloval a navrhol Andrew Viterbi ako schému na opravovanie chýb na zašumených digitálnych komunikačných linkách. Avšak netušil,že tento algoritmus nájde svoje využitie aj v dekódovaní konvolučných kódov, ktoré používajú ako CDMA a GSM, tak aj dial-up modemy, satelity a 802.11 bezdrôtové LAN siete.
V poslednej dobe sa veľmi často využíva v oboroch ako je rozoznávanie reči, počítačová linguistika a bioinformatika.

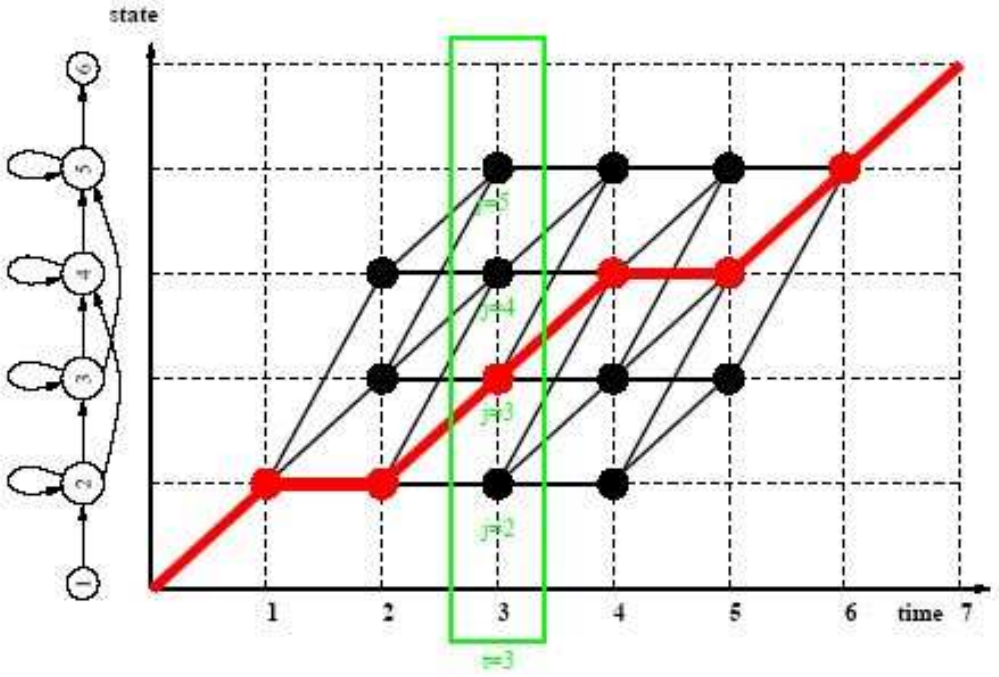
Prehľad

Viterbiho algoritmus pracuje na základe predpokladu stavu stroja. Znamená to, že systém, ktorý modelujeme sa nachádza v každom čase v určitom stave. Existuje konečný počet stavov, aj napriek tomu, že je obyčajne dosť vysoký. Každý stav je reprezentovaný ako uzol. Mnohé sekvencie stavov (ciest) môžu viesť k danému stavu, ale iba jedna je najpravdepodobnejšou cestou k nami zadanému

stavu. Táto cesta sa nazýva tzv. “survivor path” (cesta najsilnešieho). Toto tvorí základný predpoklad algoritmu, pretože náš algoritmus vyskúša všetky možné cesty cesty vedúce k danému stavu a nechá iba tú najpravdepodobnejšiu. Týmto spôsobom si algoritmus nemusí uchovávať všety možné cesty, uchováva iba jednu na daný stav, ktorým prechádza.

Druhoradý predpoklad je, že prechod z predošlého stavu do nového stavu je označený icrementom, obyčajne číslom. Tento prechod sa počíta z udalosti. Tretoradý predpoklad je, že udalosti v určitom zmysle sa kumulujú nad cestou. Takže podstata algoritmu je uchovať si číslo pre každý stav. Keď sa objaví nejaká udalosť, algoritmus to vyšetrí posunom vpred k danej množine stavov kombináciou čísel z možných predchádzajúcich stavov s incrementom prechodov podľa udalosti a vyberie tie najlepšie. Inkrement spájajúci sa s udalosťou závisí na prechodovej možnosti prejsť z predošlého stavu do nového. Napríklad v dátovej komunikácii, je možné preniesť iba polku symbolov z nepárneho počtu stavov a druhú polovicu z párneho počtu stavov. Navyše v mnohých prípadoch stavový prechodový graf nie je úplne prepojený. Najjednoduchším príkladom je auto, ktoré má 3 stavy – dopredu, zastaviť a spiatočka a prechod od vpred k spiatočke nie je povolený. Musí najprv prejsť cez stav zastaviť. Po výpočte kombinácií incrementov a stavov, iba najlepšia prežije, všetky ostatné cesty sa eliminujú.
Existujú modifikácie tohto základného algoritmu, ktoré dovoľujú prechod od hľadaniu dopredu, hľadanie po spiatku.

Prechod hľadanej cesty musí byť niekde uložený. V niektorých prípadoch môžeme dokonca naraziť na úplny záznam cesty, pretože stavový stroj,pretože kóder začína na už známom mieste a zároveň má dosť pamäte, aby si uchoval všetky cesty. V iných prípadoch sa musi nájsť výpočetné riešenie z nedostatku fyzických zdrojov. Jedným z príkladov je konvolučné kódovovanie, kde dekóder je nútený skátiť záznam v hĺbke, ktorá ešte dovoľuje rekonštrukciu do prijateľnej úrovne. Napriek tomu je Viterbiho algoritmus veľmi efektívny a existujú modifikácie, ktoré redukujú výpočetné zaťaženiea skutočnosť, aby nároky na pamäť zostali zachované.



Obrázok 7.2.1.[1] : Ilustrácia Viterbiho algoritmu. Rozhodnutie o najlepšej ceste sa robí v každom čase *t* a v každom stave *j*. Po spracovaní času t, zelené okno sa posunie do nasledujúceho času *t + 1*.

- Inicializácia:

$$\Phi_j(1) = a_{1j} b_j [o(1)] \text{ pre } 2 \leq j \leq N - 1$$

- cyklus pre všetky časy t a všetky stavy j (ilustrácia viz Obr. X.x):

$$\Phi_j (t) = \max \{ \Phi_i(t - 1) a_{ij} \} \; b_j [o(t)] \text{ pre } 2 \leq j \; N \leq 1$$

- Uzavretie algoritmu:

$$\Phi_N(T + 1) = \max \{ \Phi_i (T) a_{iN} \} \quad (11.23)$$

Rozšírenia

Pomocou algoritmu zvaného iteratívne Viterbiho dekódovanie je možné nájsť subsekvenciu daného pozorovania, ktorá zodpovedá najlepšie (v priemere) danému HMM. Iteratívne Viterbiho dekódovanie funguje postupným iteratívnym volaním modifikovaného Viterbiho algoritmu a nasleduje nové oceňovanie výsledkov až kým nekonverguje.

8 Záver

Bioinformatika je stále sa rozvíjajúci veľmi mladý odbor, v ktorom je veľmi veľa možností výskumu a inovácie. Aj napriek tomu, že sa objavili nové technológie ako je BLAST alebo FASTA, stále je čo zlepšovať a inovovať.

Počas práce na tomto projekte som si oprášila pamäť už takmer zabudnutej molekulárnej biologie a genetiky, ktorá bola mojou obľúbenou časťou biologie na strednej škole. Taktiež som si prehĺbila moje znalosti v oblasti technológií, ktoré sa používajú na získavanie génov, ich spracovanie a ukladanie a vyhľadávanie vo veľkých verejne prístupných databázích biologických dát.

Snaha pochopiť fungovanie Hidden Markov Model sa dúfam čiatsočne zdarila a moje modely budú korektné,aj keď sa mi ich korektnosť nepodarilo overiť u žiadneho odborníka na túto tému v oblasti bioinformatiky.

Počas svojho pátrania zdrojov k svojmu projektu som sa dopátrala k neuveriteľnému množstvu informácií, o ktorých som před tým nemala ani tušenia. Bohužiaľ som tiž zistila, že nie v na každej univerzite a niekedy ani v danom štáte příliš výzkum v tejto oblasti nie je. Aj napriek tomu, že predmet bioinformatika býva vo volitelných predmetoch na daných informatických fakultách, tak jeho náplň v podstate spočíva v obeznámení studentov so zakladmi biologie. Ďalšie prehĺbenie znalostí v danej oblasti nie je příliš časté a v hlavnom meste jako je Madrid, člověk zrazu zistí, že nieto človeka, který by rozumel danej oblasti. Toto nasaj svedčí o tom, že je to obor veľmi mladý a skýva veľa neprebádaného pre nás ostatných, ktorých zaujíma aké tajomstvá sa skrývajú v kóde, který kóduje všetok život na tejto planéte.

Práca na tomto projekte bola pre mňa velkým prínosom a rada by som prehĺbila svoje znalosti v tejto oblasti a pokračovala v tomto projetke aj v diplomovej práci.

V závere tiež musím podotknúť, že sa mi podarilo implementovať gene_parser. Čo je vlastne program, ktorý hľadá gény v súboroch DNA z databázy GenBank. Čiže program na hľadanie génov sa v tejto práci podarilo uskutočniť.

Náveznsoť v diplomovej práci by mohla byť sprevádzkovanie už vytvořeného HMM modelu a jeho úplna funkčnosť.Poprípade iné rozšírenia.

Literatura

[1] Honza Černocký – "Zpracováví řečových signálů – studijní opora " . Obrázok a strana 122. Fakulta Informačních technologií, Vysoké učení technické v Brně

[2] Baldi, P. & Brunak S. " Bioinformatics – The Machine Learning Approach ". Massachusetts Institute of Technology,

[3] Haussler, D. "Computational Genefinding" . Trends in Biochemical Sciences, Supplementary Guide to Bioinformatics, Strany 12 – 15

[4] Viac informcií o projekte GeneID na http://genome.imim.es/software/geneid/index.html#top

[5] Viac informácií o projekte HMMGene na http://www.cbs.dtu.dk/services/HMMgene/

[6] Viac info na Genie http://www.fruitfly.org/seq_tools/genie.html

[7] Kulp, D. Haussler, D. Reese, M.G. Eeckman, F.H. " A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA". *Proceedings of the Fourth International Conference on Intelifent Systems for Molecular Biology, AAAI Press, Menlo Park, CA,* strany 134 -142

[8] Kaleigh Smith, Hidden Markov Models in Bioinformatics with Application to Gene Finding in Human DNA, Machine Learning Project, strana 5

[9] Podlad čepaný z http://www.scripps.edu/~cliff/snap/abst.html

[10] http://www.uic.edu/classes/bios/bios100/summer2002/rna-loop.jpg

[11] doc.Ing. Jaroslav Zendulka,Ing. Vladimír Bártík Ph.D., Ing. Roman Lukáš Ph.D.,Ing. Ivana Rudolfová -Získávaní znalostí z databází – Studijní opora, verzia 31.10.2006, strany 146-

[12] http://www.bioinformatica.crs4.org/help/iub-iupac-code

[13] http://en.wikipedia.org/wiki/GenBank

[14] http://www.kokocinski.net/bioinformatics/databases.php
http://en.wikipedia.org/wiki/Biological_database

[15] http://www.genome.gov/Pages/Hyperion/DIR/VIP/Glossary/Illustration/Images /microarray_technology.gif

[16] http://www.mun.ca/biology/scarr/377_gel_files_narrow.jpg

[17] http://en.wikipedia.org/wiki/Gene_finding

[18] http://mpss.licr.org/

[19] http://en.wikipedia.org/wiki/BLAST

[20] http://www.ebi.ac.uk/

[21] http://en.wikipedia.org/wiki/Image:Genome_viewer_screenshot_small.png

[22] http://en.wikipedia.org/wiki/Sequence_alignment_software

[23] http://en.wikipedia.org/wiki/Multiple_sequence_alignment

[24] KEGG – Kyoto Encyclopedia od Genes and Genomes - http://www.genome.jp/kegg/

[25] doc.Ing. Jaroslav Zendulka,Ing. Vladimír Bártík Ph.D., Ing. Roman Lukáš Ph.D.,Ing. Ivana Rudolfová -Získávaní znalostí z databází – Studijní opora, verzia 31.10.2006, strany 141-145

Iné zdroje:

1.) Jones and Pevzner - An introduction to bioinformatics, ISBN: 0262101068, MIT Press, 2004.

2.) Jean Michel Claverie,Phd.,Cedric Notredame, Phd. – Bioinformatics for Dummies, Second edition.

3.) James Tisdall - Beginig Perl for Bioinformatics, vydavateľstvo O'Reilly

4.) T.K. Attwood & Parry Smith – Introduction to Bioinformatics

5.) Arthur Lesk – Introduction to Bioinformatics, Oxford Press

6.) Jacques Cohen – Bioinformatics – An introduction for Computer Scientists

7.) http://www.rcsb.org/pdb/home/home.do

8.) http://www.ncbi.nlm.nih.gov/

9.) http://www.ensembl.org/Homo_sapiens/index.html

10.) http://www.ebi.ac.uk/

11.) http://www.sanger.ac.uk/

12.) http://wiki.bioinformatics.org/Bioinformatics_FAQ#education

13.) http://www.e-cell.org/ecell/

14.) http://vision.ai.uiuc.edu/dugad/hmm_tut.html

Seznam příloh

Príloha č. 2

Manuál ako sa dostať k dátam v GeneBank

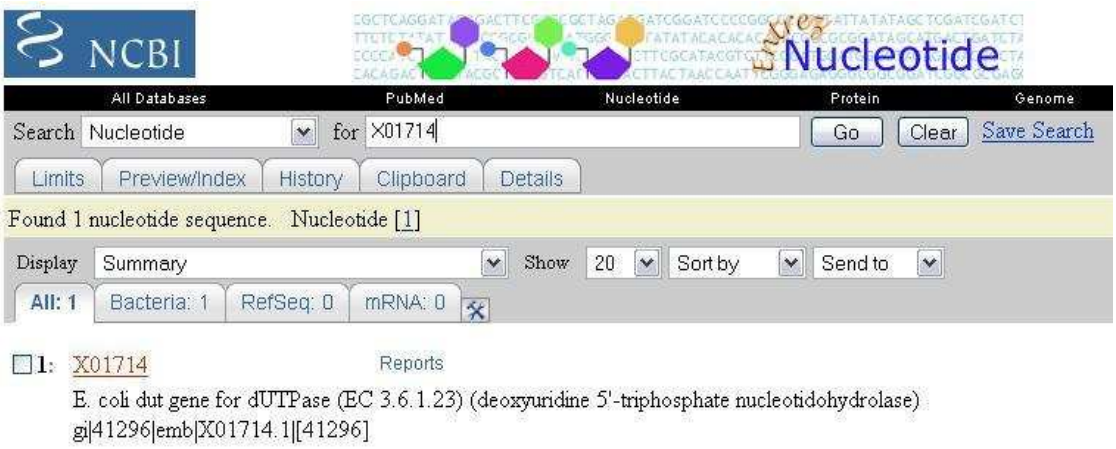
Postup si ukážeme pomocou toho, že budeme hľadať informácie o Escherichia coli dUTPase gene krok za krokom. (Genank X01714)

1.) Vložte do svojho prehliadača adresu NCI www.ncbi.nlm.nih.gov
Objaví sa vám domáca stránka National Center for Biotechnology Information

2.) V hornej lište pri Search (hľadaj) vyberte možnosť Nucleotide

3.) Napíšte do dalšej lišty hneď vedľa, kďe pred tým stojí for GenBank ID X01714
Následne sa Vám ešte nemusí zobraziť požadovaný výsledok, ktorý by mal byť podtrhnutý odkaz X01714.

4.) Objavia sa Vám nejaké výsledky,ktoré však nie sú absolútne korektné.
Treba tu zmeniť pole Search znova na Nucleotide do poľa for dať X01714 a následne do políčka Summary naklikať GeneBank. A stlačiť Go. Vytúžený výsledok vy sa mal dostaviť v podobe odkazu.



5.) Následne po zakliknutí hypertextového odkazu, ktorý môžeme vidieť na obrázku, objaví sa nám klasický GeneBank formát

6.) Obyčajne sa to nazýva GeneBank flat – file format, pretože sa dá všetko čítať v riadkoch a neobsahuje žiadne indexy, pointre, ani žiadne doplnkové informácie (ktoré sú všadeprítomné pre štruktúru sofistikovanejších databáz)

7.) Vyberte z možnosť Text zo skrolovacieho políčka napravo hore Text. Ono nám to pekne vygeneruje čistý textový formát len s dátami X01714.

8.) Následne si tieto informácie dáme pekne uložiť klasickým sposobom na náš pevný disk. Súbor-> Uložiť ...

Príloha č. 3

Záznam z databáze GenBank

```
LOCUS       X01714    1609 bp    DNA     linear   BCT 12-SEP-1993
DEFINITION  E. coli dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine
            5'-triphosphate nucleotidohydrolase).
ACCESSION   X01714
VERSION     X01714.1  GI:41296
KEYWORDS    dUTPase; unidentified reading frame.
SOURCE      Escherichia coli
  ORGANISM  Escherichia coli
            Bacteria; Proteobacteria; Gammaproteobacteria;
            Enterobacteriales; Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 1609)
  AUTHORS   Lundberg,L.G., Thoresson,H.O., Karlstrom,O.H. and Nyman,P.O.
  TITLE     Nucleotide sequence of the structural gene for dUTPase of
            Escherichia coli K-12
  JOURNAL   EMBO J. 2 (6), 967-971 (1983)
   PUBMED   6139280
COMMENT     Data kindly reviewed (25-NOV-1985) by L. Lundberg.
FEATURES             Location/Qualifiers
     source          1..1609
                     /organism="Escherichia coli"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:562"
     promoter        286..291
                     /note="-35 region"
     promoter        310..316
                     /note="-10 region"
     misc_feature    322..324
                     /note="put. transcription start region"
     RBS             330..333
                     /note="put. rRNA binding site"
     CDS             343..798
                     /note="unnamed protein product; dUTP-ase (aa 1-151)"
                     /codon_start=1
                     /transl_table=11
                     /protein_id="CAA25859.1"
                     /db_xref="GI:41297"
                     /db_xref="GOA:P06968"
                     /db_xref="InterPro:IPR008180"
                     /db_xref="InterPro:IPR008181"
                     /db_xref="PDB:1DUD"
                     /db_xref="PDB:1DUP"
                     /db_xref="PDB:1EU5"
                     /db_xref="PDB:1EUW"
                     /db_xref="PDB:1RN8"
                     /db_xref="PDB:1RNJ"
                     /db_xref="PDB:1SEH"
                     /db_xref="PDB:1SYL"
                     /db_xref="UniProtKB/Swiss-Prot:P06968"
```

```
                     /translation="MKKIDVKILDPRVGKEFPLPTYATSGSAGLDLRACLNDAVELAP
                     GDTTLVPTGLAIHIADPSLAAMMLPRSGLGHKHGIVLGNLVGLIDSDYQGQLMISVWN
                     RGQDSFTIQPGERIAQMIFVPVVQAEFNLVEDFDATDRGEGGFGHSGRQ"
     misc_feature    831..851
                     /note="put.stem-loop structure"
     repeat_unit     831..838
                     /note="inverted repeat A"
     repeat_unit     844..851
                     /note="inverted repeat A'"
     misc_feature    866..893
                     /note="put. stem-loop structure"
     repeat_unit     866..872
                     /note="imp. inverted repeat B"
     repeat_unit     888..893
                     /note="imp. inverted repeat B'"
     RBS             889..895
                     /note="pot. rRNA binding site"
     CDS             905..1540
                     /note="unnamed protein product; unidentified reading
                     frame"
                     /codon_start=1
                     /transl_table=11
                     /protein_id="CAA25860.1"
                     /db_xref="GI:41298"
                     /db_xref="GOA:P0C093"
                     /db_xref="UniProtKB/Swiss-Prot:P0C093"
                     /translation="MAEKQTAKRNRREEILQSLALMLESSDGSQRITTAKLAASVGVS
                     EAALYRHFPSKTRMFDSLIEFIEDSLITRINLILKDEKDTTARLRLIVLLLLGFGERN
                     PGLTRILTGHALMFEQDRLQGRINQLFERIEAQLRQVLREKRMREGEGYTTDETLLAS
                     QILAFCEGMLSRFVRSEFKYRPTDDFDARWPLIAASCSNMTPDDFSSGEFL"

ORIGIN
1 cagagaaaat caaaaagcag gccacgcagg gtgatgaatt aacaataaaa atggttaaaa
61 accccgatat cgtcgcaggc gttgccgcac taaaagacca tcgaccctac gtcgttggat
121 ttgccgccga aacaaataat gtggaagaat acgcccggca aaaacgtatc cgtaaaaacc
181 ttgatctgat ctgcgcgaac gatgtttccc agccaactca aggatttaac agcgacaaca
241 acgcattaca ccttttctgg caggacggag ataaagtctt accgcttgag cgcaaagagc
301 tccttggcca attattactc gacgagatcg tgacccgtta tgatgaaaaa aatcgacgtt
361 aagattctgg acccgcgcgt tgggaaggaa tttccgctcc cgacttatgc cacctctggc
421 tctgccggac ttgacctgcg tgcctgtctc aacgacgccg tagaactggc tccgggtgac
481 actacgctgg ttccgaccgg gctggcgatt catattgccg atccttcact ggcggcaatg
541 atgctgccgc gctccggatt gggacataag cacggtatcg tgcttggtaa cctggtagga
601 ttgatcgatt ctgactatca gggccagttg atgatttccg tgtggaaccg tggtcaggac
661 agcttcacca ttcaacctgg cgaacgcatc gcccagatga tttttgttcc ggtagtacag
721 gctgaattta atctggtgga agatttcgac gccaccgacc gcggtgaagg cggctttggt
781 cactctggtc gtcagtaaca catacgcatc cgaataacgt cataacatag ccgcaaacat
841 ttcgtttgcg gtcatagcgt gggtgccgcc tggcaagtgc ttattttcag gggtattttg
901 taacatggca gaaaaacaaa ctgcgaaaag gaaccgtcgc gaggaaatac ttcagtctct
961 ggcgctgatg ctggaatcca gcgatggaag ccaacgtatc acgacggcaa aactggccgc
1021 ctctgtcggc gtttccgaag cggcactgta tcgccacttc cccagtaaga cccgcatgtt
1081 cgatagcctg attgagttta tcgaagatag cctgattact cgcatcaacc tgattctgaa
1141 agatgagaaa gacaccacag cgcgcctgcg tctgattgtg ttgctgcttc tcggttttgg
1201 tgagcgtaat cctggcctga cccgcatcct cactggtcat gcgctaatgt ttgaacagga
1261 tcgcctgcaa gggcgcatca accagctgtt cgagcgtatt gaagcgcagc tgcgccaggt
1321 attgcgtgaa aagagaatgc gtgagggtga aggttacacc accgatgaaa ccctgctggc
1381 aagccagatc ctggccttct gtgaaggtat gctgtcacgt tttgtccgca gcgaatttaa
1441 ataccgcccg acggatgatt ttgacgcccg ctggccgcta attgcggcca gttgcagtaa
1501 tatgacgccg gatgactttt catccggcga gtttctttaa acgccaaact cttcgcgata
1561 ggccttaacc gccgccagat gttccgccat ttccggcttc tcttccagg
//
```

Príloha č. 4

Zoznam existujúcich biologických databáz [13]

8.1.1 Primárne sekvenčné databázy

Medzinárodná sekvenčná nukleotidová databáza (INSD) pozostáva z nasledujúcich databáz.

1. **DDBJ** (DNA Data Bank of Japan) - http://www.ddbj.nig.ac.jp/Welcome-e.html
2. **EMBL Nucleotide DB** (European Molecular Biology Laboratory) http://www.ebi.ac.uk/embl/index.html
3. **GenBank** (National Center for Biotechnology Information) http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide

Tieto databanky reprezentujú doterajšiu znalosť o sekvenciách nukleotidov všetkých organizmov, ktorých genetický kód bol doposiaľ prečítaný. Medzi sebou si navzájom vymieňajú informácie a sú zdrojom pre mnohé iné databázy.

8.1.2 Meta-databázy

Sú to takzvané databázy databáz, ktoré podporujú vyhľadávanie v niekoľkých databázach naraz. Tieto databázy sú navzájom poprepájané a dáta sú neustále aktualizované.Sú tu uložené dáta z rôznych zdrojov a obyčajne sú sprístupnené v novej a pohodlnejšej forme na spracovanie, niektoré sú zamerané na určité konkrétne choroby alebo organizmy.

1. **Entrez** (National Center for Biotechnology Information) http://www.ncbi.nlm.nih.gov/sites/gquery
2. **euGenes** (Indiana University) http://eugenes.org/
3. **GeneCards** (Weizmann Institute) http://www.genecards.org/
4. **SOURCE** (Stanford University) http://smd-www.stanford.edu/cgi-bin/source/sourceSearch
5. **mGen** – obsahuje štyri najväčšie svetové databázy – GenBank , Refseq, EMBL a DDBJ
6. **Bioinformatic Harvester** (Karlsruhe Institute of Technology) – Integruje 26 hlavných proteinových a génových zdrojov.
7. **MetaBase (KOBIC)** – Biologická databáza, ktorá je založená na príspevkoch užívateľov.

8.1.3 Databázy genómov

Tieto databázy obsahujú okomentované a zanalyzované sekvencie genómov rôznych organizmov. Sú verejne dostupné. V týchto databázach môžeme nájsť genómy rôznych organizmov, alebo naopak môžu obsahovať len jeden jediny model genómu organizmu.

1. **Ensembl** – obsahuje genómy rôznych druhov organizmov. Môžme tu napríklad nájsť kompletný ľudský genóm,alebo genóm myši. http://www.ensembl.org/
2. **JGI Genomes of the DOE**-Joint Genome Institute – táto databáza obsahuje mnohé genómy aukariotických a mikróbnych organizmov. http://genome.jgi.doe.gov/
3. **CAMERA** – Databáza genómov mikróbov. http://camera.calit2.net/index.php/
4. **MGI Mouse Genome** (Jackson Lab.) http://www.informatics.jax.org/
5. **Corn** – genetická databáza plodín (napr.kukurice) http://www.maizegdb.org/
6. **Saccharomyces Genome Database** - je tu uložený model genómu kvasiniek. http://www.yeastgenome.org/
7. **Wormbase** – je tu uložený model genómu Caenorhabditis elegans http://www.wormbase.org/
8. **Zebrafish Information Network** – tu je uložený model rybky Danio pruhované. http://zfin.org/cgi-bin/webdriver?MIval=aa-ZDB_home.apg
9. **Viral Bioinformatics Resource Center** – databáza obsahujúca okomentované genetické dáta jedenástich druhov vírusov. http://troy.bioc.uvic.ca/

8.1.4 Genómové prehliadače

Bádateľom dávajú možnosť vizualizácie a prehliadania celých genómov, s komentármi, ktoré sú k nim vytvorené. Majú funkcie, ktoré umožňujú predikciu génov a proteínov, ich štruktúry, expresiu génov atď. Okomentované dáta sú obyčajne z rôznych zdrojov.

1. **Integrated Microbial Genomes** (IMG) system vytvorený DOE-Joint Genome Institute http://img.jgi.doe.gov/cgi-bin/pub/main.cgi
2. **UCSC Genome Bioinformatics Genome Browser and Tools** (UCSC) http://genome.ucsc.edu/
3. **Ensembl The Ensembl Genome Browser** (Sanger Institute and EBI) http://www.ensembl.org/
4. **GBrowse The GMOD GBrowse Project** http://www.gmod.org/?q=node/71
5. **Pathway Tools Genome Browser** http://bioinformatics.ai.sri.com/ptools/
6. **X:Map –** Genómový prehliadač, ktorý ukazuje, kde sa v danom mikročipe uskutečnila expresia génov, dajú sa prehliadať dáta z exonov v jednotlivých génov. http://xmap.picr.man.ac.uk/#1/16950/15/h
7. **Viral Genome Organizer** (VGO) – Genómový prehliadač poskytujúci vizualizačné a analytické prostředky na sprcovanie úplneho genómu jedenástich druhov vírov z databáz VBRC (Viral Bioinformatics Resource Center).

8.1.5 Proteínové databázy

1. **UniProt –** Verejná databáza proteínov (UniProt Consortium: EBI, Expasy, PIR) http://www.uniprot.org/
2. **PIR Protein Information Resource** (Georgetown University Medical Center (GUMC)) http://pir.georgetown.edu/
3. **Swiss-Prot - Protein Knowledgebase** (Swiss Institute of Bioinformatics) http://www.expasy.org/sprot/
4. **PEDANT Protein Extraction, Description and ANalysis Tool** (Forschungszentrum f. Umwelt & Gesundheit) http://pedant.gsf.de/
5. **PROSITE** Databáze rôznych druhov proteínov a ich domén http://www.expasy.org/prosite/
6. **DIP Database of Interacting Proteins** (Univ. of California) http://dip.doe-mbi.ucla.edu/
7. **Pfam** – Je to databáza proteínov podľa rozloženia a HMM k daným génom. (Sanger Institute) http://www.sanger.ac.uk/Software/Pfam/
8. **ProDom** Všeobecná databáza hlavných druhov proteínov(INRA/CNRS) http://prodom.prabi.fr/prodom/current/html/home.php

9. **SignalP 3.0 –** Server slúžiaci na signalizáciu predpokladnaých peptidov.Je založený na umelých neutrónových sietiach a modelech HMM. http://www.cbs.dtu.dk/services/SignalP/
10. **SUPERFAMILY** – knižnica rôznych modelov HMM reprezentujúcich nadčeľade. Taktiež je to databáza, kde môžeme nájsť okomentované celé genómy organizmov. http://supfam.org/SUPERFAMILY/

8.1.6 Databázy proteínových štruktúr

1. **Protein Data Bank**(PDB) (Research Collaboratory for Structural Bioinformatics (RCSB)) http://www.pdb.org/pdb/home/home.do
2. **CATH** – Klasifikácia štruktúr proteínov http://www.cathdb.info/index.html
3. **SCOP** - Klasifikácia štruktúr proteínov
4. **SWISS-MODEL** - Server a skaldisko pre modely proteínových struktur. http://swissmodel.expasy.org//SWISS-MODEL.html
5. **ModBase Database of Comparative Protein Structure Models** (Sali Lab, UCSF) http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi

8.1.7 Databázy medziproteínových interakcií.

1. **BioGRID** – Hlavný sklad informácií o interakciách v datasetoch. (Samuel Lunenfeld Research Institute) http://www.thebiogrid.org/
2. **STRING**: Databáza známych a predpovedaných interakcií proteínov. (EMBL) http://string.embl.de/
3. **Database of Interacting Proteins** http://dip.doe-mbi.ucla.edu/

8.1.8 Databázy metabolických ciest

1. **BioCyc Database Collection** obsahuje **EcoCyc** a **MetaCyc** http://biocyc.org/ http://ecocyc.org/ , http://metacyc.org/
2. **KEGG PATHWAY Database**(Univ. of Kyoto) http://www.genome.ad.jp/kegg/pathway.html
3. **MANET database** (University of Illinois) http://www.manet.uiuc.edu/
4. **Reactome** (Cold Spring Harbor Laboratory, EBI, Gene Ontology Consortium) http://www.reactome.org/

8.1.9 Databázy založené na znalostiach získaných z microarrays

1. **ArrayExpress** (European Bioinformatics Institute) http://www.ebi.ac.uk/microarray-as/ae/
2. **Gene Expression Omnibus** (National Center for Biotechnology Information) http://www.ncbi.nlm.nih.gov/projects/geo/
3. **maxd** (Univ. of Manchester) http://www.bioinf.man.ac.uk/microarray/maxd/index.html
4. **SMD** (Stanford University) http://smd-www.stanford.edu/
5. **GPX**(Scottish Centre for Genomic Technology and Informatics) http://www.pathwaymedicine.ed.ac.uk/GPX

8.1.10 Databázy matematických modelov

1. **CellML** www.cellml.org/
2. **Biomodels Database** http://www.ebi.ac.uk/biomodels/

8.1.11 Databázy PCR / PCR v reálnom čase

1. **PathoOligoDB**: Voľne dostupná QPCR multidatabáza rôznych patogénov. http://www.pathooligodb.com/

*PCR – polymerase chain reaction – reťazová reakcia polymerázy

*QPCR – real-time polymerase chain reaction – reťazová reakcia polymerázy v reálnom čase

8.1.12 Špecializované databázy

1. **BIOMOVIE** (ETHZurich) Videá , ktoré súvisia s novými trenadmi v biológii a biotechnológiách. http://lmc-risc2.ethz.ch/myvideo/Tracer.jsp?command=MainVideo&category=biology&search=science
2. **CGAP Cancer Genes** (National Cancer Institute) http://cgap.nci.nih.gov/Genes/GeneFinder
3. **Clone Registry Clone Collections** (National Center for Biotechnology Information) http://www.ncbi.nlm.nih.gov/projects/genome/clone/
4. **DBGET H.sapiens** (Univ. of Kyoto) http://www.genome.ad.jp/dbget-bin/www_bfind?h.sapiens
5. **GDB Hum. Genome Db** (Human Genome Organisation) http://www.gdb.org/gdb
6. **SHMPD** The Singapore Human Mutation and Polymorphism Database http://shmpd.bii.a-star.edu.sg/
7. **NCBI-UniGen**e (National Center for Biotechnology Information) http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene
8. **OMIM** – Databáza dedičných chorôb (Online Mendelian Inheritance in Man) http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim
9. **Off. Hum. Genome Db** (HUGO Gene Nomenclature Committee) http://www.gene.ucl.ac.uk/
10. **HGMD –** Databáza mutácií spôsobujúcich ochorenia (HGMD Human Gene Mutation Database) http://www.hgmd.cf.ac.uk/ac/index.php
11. **PhenCode** – Databáza spájajúca mutácie v ľudskom genóme s fenotypom. http://globin.bx.psu.edu/phencode/
12. **List with SNP**
13. **p53** http://p53.bii.a-star.edu.sg/
14. **Edinburgh Mouse Atlas** http://genex.hgu.mrc.ac.uk/
15. **HvrBase++** - nachádza sa tu ľudská mitochondriálna DNA a mitochodnriálna DNA primátov. http://www.hvrbase.org/
16. **PolygenicPathways** – Gény a rizikové faktory, ktoré sú zodpovedné za Alzehimerovu chorbu, maniodepresívnu psychózy alebo schyzofréniu. http://www.polygenicpathways.co.uk/
17. **Connectivity map** – Sú tu uložené dáta z génovej transkripcie a nástroje pomocou, ktorých je možné zistiť ich vzájomný vzťah s jednotlivými liekmi. http://www.broad.mit.edu/cmap/
18. **CTD The Comparative Toxicogenomics Database –** Opisuje interakcie chemických látok, génov a chorôb. http://ctd.mdibl.org/

Príloha č. 5

Manuál k program gen_parser

1. Preložíme pomocou make
2. Spustíme pomocou príkazu ./gen_parser vstupny_subor
3. Gen parser nám následne na obrazovku vypíše všetky gény v danom súbore (daný program bol testovaný na súboroch získaných z databázy http://www.ncbi.nlm.nih.gov/ a sú v zásade v textovom formáte)

   Daný súbor, ktorý vkladáme ako vstupný, musí byť v tom istom adresári ako gen_parser.exe

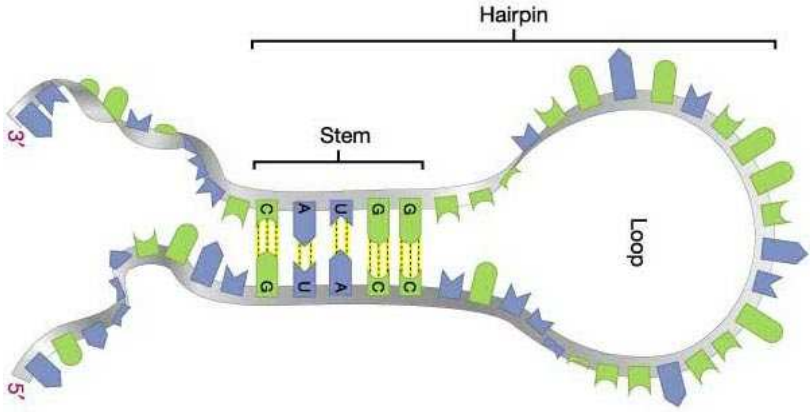
Príloha č. 6

Vysvetlivky

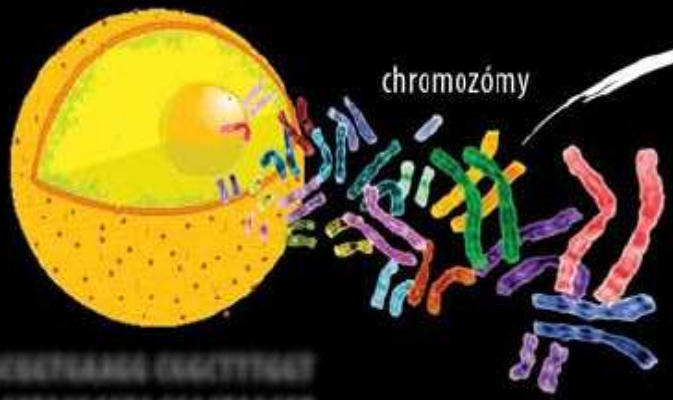
dUTPase[9] - Základný enzým dUPT, zvaný tiež dUPT pyrofosfatáza. Je obzvášť špecifický pre dUPT a je veľmi dôležitý pre vernú replikáciu DNA a jej opravu.

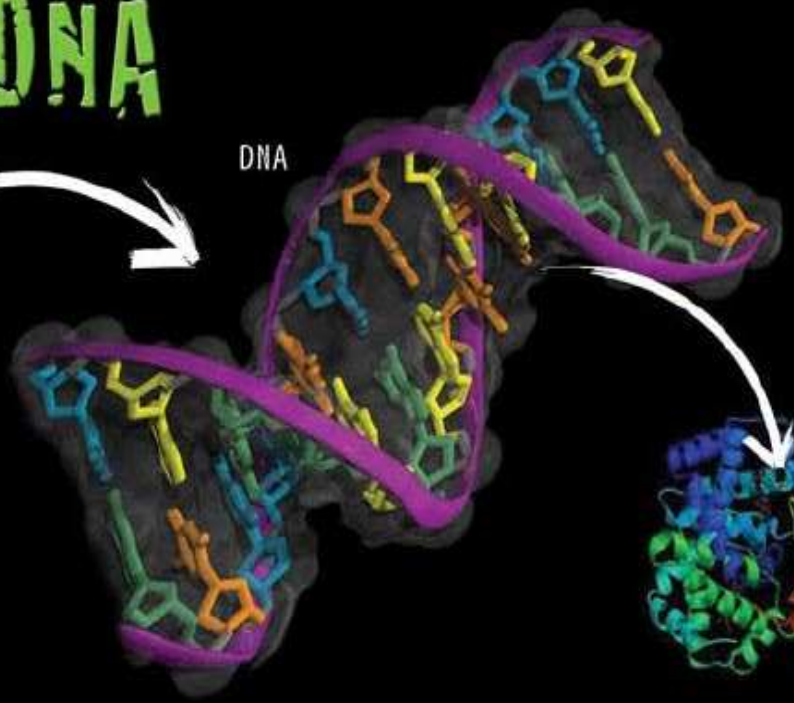
Stem-loop [10] –Je medzimolekulárne párovanie bází, ktoré vytvára štruktúru, ktorá sa môže vyskytnúť aj v jednovláknovej DNA,častejšie sa však vyskytuje v RNA. Táto štruktúra je tiež známa ako "hairpin" alebo " hairpin loop". Vytvorí sa hlavne vtedy, keď sa dvom oblastiam rovnakej molekuly,(obyčajne,ak sa nachádza palidrom v nukleotidovej sekvencii) spárujú bázy, ktoré vytvoria dvojitú šroubovicu, ktorá končí v nespárovanej slučke. Výsledkom je tvar štruktúry, ktorý pripomína lízatko, ktorý je základným stavebným blokom mnohých sekundárnych štruktúr RNA.

Obrazok x.x.[10]: RNA Stem-loop



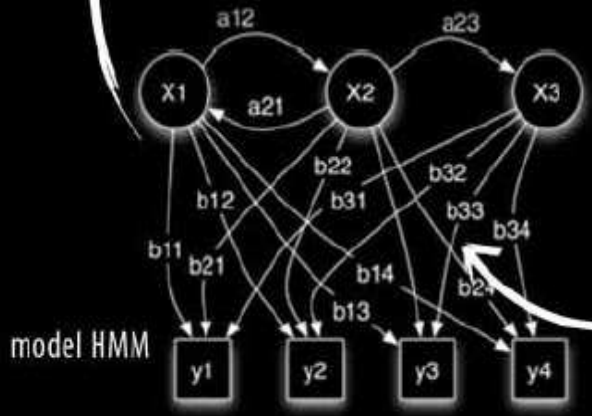
HĽADANIE GÉNOV V DNA






```
TTCGAC GCCACCGACC GCGGTGAAGG CGGCTTTGGT
TCTGGTC GTCAGTAACA CATACGCATC CGAATAACGT
TGGCA GAAAAACAAA CTGCGAAAAG GAACCGTCGC
TCACCA TTCAACCTGG CGAACGCATC GCCCAGATGA
TGTTCC GGTAGTACAG GCTGAATTTA ATCTGGTGGA
TTCGAC GCCACCGACC GCGGTGAAGG CGGCTTTGGT
AAATAC TTCAGTCTCT GGCGCTGATG CTGGAATCCA
TGGAAG CCAACGTATC ACGACGGCAA AACTGGCCGC
TGTCGGC GTTTCCGAAG CGGCACTGTA TCGCCTAGTC
TTGTTCC GGTAG   AG GCTGAATTTA ATCTGGTGGA
TTCGAC GCCA  CGACC GCGGTGAAGG CGGCTTTGGT
TCTGGTC GTCAGTAACA CATACGCATC CGAATAACGT
```

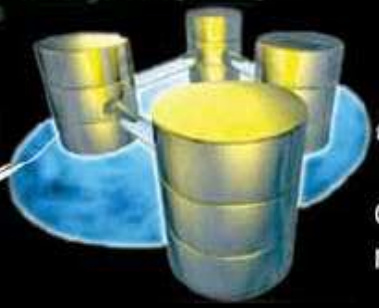
vyhľadaný gén



model HMM

```
if __name__ == '__main__':
    import sys
    import string
    import math
    try:
        import psyco
        psyco.full()
    except:
        print 'Warning: No psyco available'

    # COMMANDLINE: viterbi.py
    # States
    states = xrange(3) # [0,1,2]
    nice_names = ['start','heads','tails']  ## more mnemonic names for sta
                                            ## [0,1,2]


    s0 = 0  # start state
    # transition probs A[from][to]
    # A an array of arrays.
    A = [[0.0, 0.5, 0.5],   # From State 0
         [0.0, 0.5, 0.5],   # From State 1 ...
         [0.0, 0.5, 0.5]]

    # emission probs a dictionary: key is a member of input vocab (a word)
    #                              val is an array of length len(states)
    B = { 'h':[ [ 0.0, 0.5, 0.5 ],  # from State 0
      [ 0.0, 1.0, 1.0 ],

      [ 0.0, 0.0, 0.0 ] ],
        't':[ [ 0.0, 0.5, 0.5 ],
      [ 0.0, 0.0, 0.0 ],
      [ 0.0, 1.0, 1.0 ] ]
        }
    neg_infinity=float("-infinity")
    # 1e300000
```

Viterbiho algoritmus



databazy sekvencie nukleotidov