rií. Na horizontální přenos RAG–1 a RAG–2 ukazuje řada podobností s jinými transpozony: přepis DNA probíhá jako u jiných mobilních elementů, integrují se podobně jako viry HIV (původci onemocnění AIDS) a jejich integrázy štěpí DNA podobně jako integrázy HIV (tab. 1).

O horizontálním přenosu RAG je dnes přesvědčena většina vědců. Jako nejpravděpodobnější cesta přenosu se nabízí infekční proces. Je dokonce možné, že k tomuto přenosu byl již připraven terén, protože v r. 2005 byl objeven u kopinatců *(Cephalochordata)* homolog aktivátoru RAG–1. Samotný RAG–1 však v databázi genů kopinatce identifikován nebyl. Ani v genomu bezčelistných se RAG nenacházejí, a proto nemají imunoglobulinové receptory a netvoří ani protilátky.

Lze předpokládat, že lymfocyty prvních čelistnatých obratlovců byly dostatečně vhodným terénem pro přijetí genových kazet RAG. Podle současných znalostí mohly být zdrojem RAG ty druhy bakterií, které se dostaly do zárodečné lymfocytární linie, adaptovaly se na nitrobuněčný život a přenesly část svého chromozomu do chromozomu hostitelských buněk. Tyto bakterie mohly žít po dostatečně dlouhou dobu v buňkách jako komenzálové. Podobné případy jsou známy u členovců a kroužkovců, u nichž bakterie rodu *Wolbachia* určují chování nebo plodnost svých hostitelů. Někteří autoři dokonce prokazují, že RAG jsou velmi podobné invertázám a integračním faktorům prokaryot, které už byly identifikovány u bakterie *Escherichia coli* a salmonel.

RAG jsou pouze v lymfocytech

Tento fakt je dosud zahalen tajemstvím. Nevíme, proč RAG nejsou i v jiných buňkách, je ale jisté, že rekombinace genových segmentů u receptorů pro antigen je tak hlubokým zásahem do genomu, že musel být přísně omezen pouze na linii lymfocytů a musel být regulován zcela novými mechanismy oprav DNA. Jeho daní je obrovské množství chyb — lymfocytárních nádorových onemocnění. Sama inkorporace RAG je záhadná, protože tvorba lymfocytů je regulována kaskádou tzv. transkripčních faktorů z multigenových rodin, z nichž řada byla získána rovněž přenosem mobilních genových sekvencí. Problematika krvetvorby je však natolik složitá, že se omezíme na schema, které přináší jen nejzákladnější představu o účasti RAG a dalších transkripčních faktorů (obr. 3).

Otázky na závěr

Je nezvratně prokázáno, že horizontální přenos genetické informace může za určitých okolností hrát v evoluci závažnou úlohu, jejíž důsledky mohou být dalekosáhlé. Ukázali jsme si to na příkladu adaptivní imunity čelistnatců. V této souvislosti se přímo vnucují neodbytné otázky: Mohou se dnešní nitrobuněčné bakteriální parazité jako *Salmonella, Shigella, Francisella* nebo dokonce viry (viz podobnost RAG a HIV!) stát přenašeči nové genetické informace? Nebo to budou mobilní části DNA z jiných organismů, o jejichž působení vůbec nic nevíme? Jak dlouho bude trvat, než se evoluční inovace projeví? Bude to sto let, tisíciletí nebo déle? Mohou vůbec vzniknout nové vlastnosti, orgány a funkce u současných rostlin, živočichů nebo dokonce u lidí, podobně jak se to stalo s adaptivní imunitou v průběhu kambria?

Věnováno našemu učiteli Prof. Ctiradu Johnovi.

Genomika a bioinformatika

Jak se hledají geny

Jan Pačes

Genomika je v současnosti ve vědách o životě velmi frekventované slovo. Ne každý si však dokáže pod takovým slovem představit něco konkrétního. Ani krátké vysvětlení: „genomika je obor zabývající se analýzou genomů“ nám mnoho konkrétního neřekne. Termín bioinformatika pak může být pro laika úplně mystický. Opět krátce: bioinformatika se snaží udělat v počítači to, co dělá každá buňka našeho těla, každá bakterie, každý živý organismus dnes a denně: přečte si genetickou informaci uloženou v DNA, deoxyribonukleové kyselině, a tuto informaci analyzuje a používá. Ale jak se DNA analyzuje?

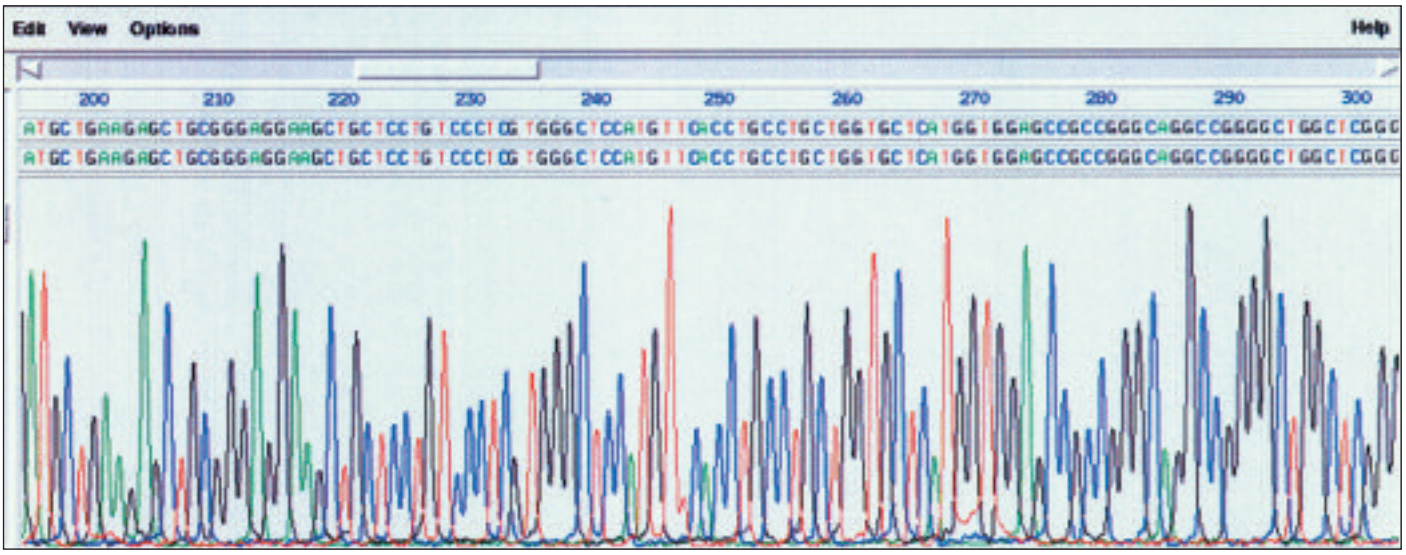
V tomto krátkém článku bych rád popsal, co vlastně vědec zabývající se genomikou a bioinformatikou ve skutečnosti dělá, když analyzuje genom.

Čtení DNA

Genomika začíná prací v laboratoři, kde se informace z molekuly DNA převádí do počítače. DNA je totiž právě ta jedinečná molekula, která je pro genomiku nejdůležitější, která nese informaci o tom, kým jsme a co se z nás stane.

Celá genetická abeceda má čtyři písmena: A, C, G a T, což jsou zkratky základních chemických komponent adeninu, cytosinu, guaninu a thyminu, které jsou v molekule DNA jako součást tzv. nukleotidů.

Obr. 1 Čtení DNA: jednotlivé různobarevné vlny jsou záznamem strojového čtení DNA. Zjednodušeně si proces čtení můžeme představit takto: každé písmenko genetické abecedy (nukleotid) je označeno jednou barvou (např. T červeně). Části DNA jsou protlačovány skrz kapiláru, na jejímž konci snímá digitální kamera barvu podle toho, které písmenko právě prochází kolem čidla. V horní části obrázku je vidět sled písmen, jak ho počítačový program odvodil ze spodního čtyřbarevného grafu (blíže v textu)



Takže onen první krok čtení DNA znamená vlastně určit pořadí těchto nukleotidů.

Přímo o metodě, jak určit sled nukleotidů, jen stručně: DNA studovaného organismu se vyčistí a enzymatickou (pomocí restriktáz) nebo fyzikální (sonikací — rozbití ultrazvukem) cestou se nastříhá na kousky potřebné velikosti — většinou v řádu tisíců až desetitisíců písmenek. Tento krok je nutný proto, že současná technologie umí při jednom čtení zaznamenat maximálně okolo tisíce písmen.

Tyto fragmenty DNA vložíme do bakterií k tomu uzpůsobených a vytvoříme tak „knihovnu". Ta sestává z mnoha bakterií, z nichž každá v sobě nese nějaký kus genetické informace z původního námi studovaného organismu. Pomůžeme si analogií s knihou: podobně bychom mohli vzít něčí memoáry, a to nejlépe hned v mnoha kopiích, a rozstříhat je po větách, odstavcích nebo zcela náhodně. Získali bychom tak knihovnu ústřižků, z nichž by opět šla získat informace obsažená v původní knize.

Jednotlivé bakterie pak z knihovny vybíráme, necháme je vyrobit mnoho kopií naší části DNA a tyto úseky čteme. Detaily vlastní fyzikálně-chemické metody vynechám, přidám však obrázek toho, jak takové čtení vypadá (obr. 1). Nejlepší současné přístroje umožňují číst zhruba 100 fragmentů DNA zároveň, z každého přečtou až 1 000 písmen a trvá jim to desítky minut. Abychom si byli jisti, že jsme celou DNA získali správně, je nutné přečíst každé písmenko v průměru 6–8×. I s tímto nadměrným počtem čtení to znamená, že přečíst genom bakterie je pro dobře vybavené laboratoře s dostatkem peněz záležitost několika týdnů. U složitějších organismů to jsou dnes pořád ještě desítky měsíců, ale technologie se vylepšuje obrovským tempem a není daleko doba, kdy i veliké a komplexní genomy budeme číst během několika týdnů a možná i dnů.

Po přečtení dostatečného množství úseků DNA z bakteriální knihovny nastane úmorné sestavování. Analogie s rozstříhanými memoáry nám opět pomůže s přiblížením: vybereme z naší knihovny jeden ústřižek a budeme hledat další, který na něj navazuje. Ty dva spojíme a hledáme další, který na ně navazuje a tak dále. Občas máme v ústřižku i nějakou informaci, která nám ho pomůže zařadit na konkrétní místo v knize. Např. číslo stránky nebo začátek kapitoly. To by v DNA mohlo odpovídat třeba genetickým značkám. S trochou štěstí budeme mít zanedlouho celou knihu (nebo genom) složenou.

Při sestavování genomu je situace složitější, než by byla při sestavování knihy o to, že máme pouze čtyři písmenka, takže se podobnosti hledají hůře. Také při čtení DNA děláme nutně chyby. Navíc se v genomech objevují stejné nebo velice podobné úseky mnohokrát na různých místech — asi jako kdyby se v knize několikrát opakovala stejná historka. Ale překonáme-li i tyto překážky, výsledek bude stát za to: v počítači budeme mít úplnou genetickou informaci studovaného organismu, tedy veškerou informaci, kterou si organismy předávají z generace na generaci a podle níž jsou vystavěny a žijí.

Hledání genů

Konečně máme tedy genom v počítači, buď celý v jednom kuse v případě bakterií nebo rozdělený do chromozomů u vyšších organismů. Můžeme zasednout k počítači a v klidu si v genomu zkoumaného organismu „počíst". Příběh je to pořádně rozvleklý, bakterie mají genom několik milionů písmenek dlouhý, vyšší organismy několik miliard. Třeba člověk má svoji genetickou informaci zapsánu ve více než třech miliardách písmenek. Ale co nám ten dlouhý sled písmen o bakterii, živočichu nebo rostlině říká? Zatím nic, teprve mu musíme porozumět.

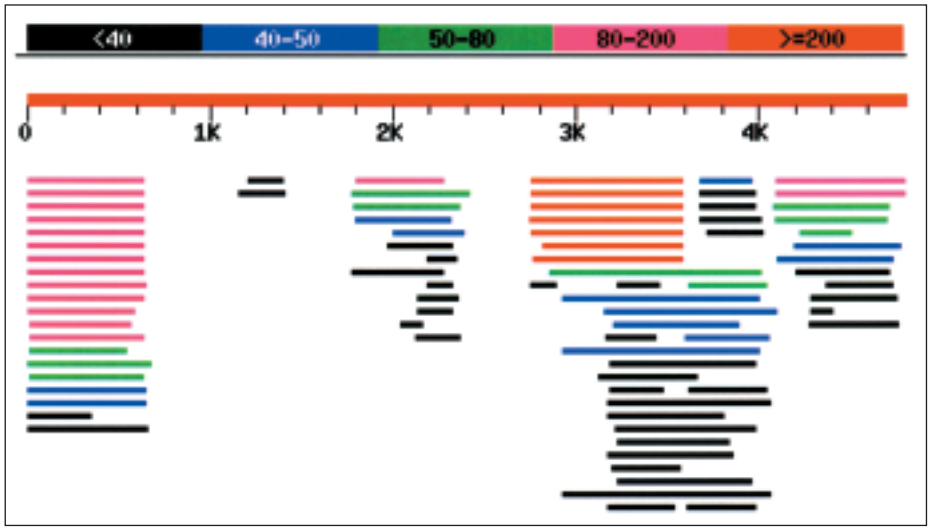
Pro pochopení toho, co je v DNA zakódováno, musíme nejdříve text rozdělit na menší části, a to tak, aby měly nějaký význam. Naše přirovnání DNA ke knize nám opět umožní analogii: v textu bychom hledali slova a věty, v DNA budeme hledat kodóny a geny.

Kodón se skládá ze tři písmenek a kóduje jednu aminokyselinu, která je základním stavebním kamenem proteinů. Několik (nebo spíše mnoho) kodónů v řadě za sebou dává dohromady gen. Gen představuje základní jednotku dědičné informace, při určitém zjednodušení bychom mohli říci, že gen je takový úsek DNA, ve kterém je zapsána informace o jednom proteinu. A proteiny jsou funkčními molekulami organismů. Zajišťují metabolismus, jsou stavebními prvky těla, tvoří podstatu imunitního systému atd. Naším prvním úkolem tedy bude najít v DNA geny. Jak se to dělá?

Hledání genů pomocí příbuznosti

Jedna z nejúčinnějších metod je založena na tom, že v přírodě velmi často platí, že protein vykonávající nějakou funkci v jednom organismu je podobný proteinu se stejnou nebo podobnou funkcí v jiném organismu. A protože už je mnoho objeveno a známo, bude nejlépe se podívat na databázi všech již objevených a popsaných proteinů a zejména genů, které o nich nesou informaci. Takové databáze existují dnes tři, jedna v Evropě, jedna v USA a třetí v Japonsku. Všechny obsahují víceméně stejné informace, podívejme se třeba na tu evropskou. Jmenuje se EMBL (European Molecular Biology Laboratory Database) a najdeme ji na internetu na adrese http://www.ebi.ac.uk/embl.

Porovnáme tedy naši DNA s databází. Použijeme k tomu některý ze speciálních programů vyvinutých právě pro porovnávání. Např. program jménem BLAST, který jsem použil na vyhledání úseků podobných něčemu známému v kousku DNA z bakterie *Rhodobacter capsulatus,* nám



Obr. 2 Schematické znázornění výsledků generovaných programem BLAST, který je určen pro hledání příbuzných sekvencí DNA nebo proteinů v databázi. Na obr. je vidět analýza DNA o délce asi 5 000 písmen — nukleotidů (jasně červená úsečka v horní části obr.). Barevné úsečky ve spodní části obrázku ukazují, ke kterým úsekům analyzované DNA existuje v databázi DNA podobná. Barvou je znázorněna míra této podobnosti (červená — nejvíce podobná, černá — nejméně podobná, číselně vyjádřeno v horní části obrázku)

ve výsledku ukáže takovýto obrázek (obr. 2). Z obrázku se dá vypozorovat, že pomocí podobností jsme v této DNA nalezli pravděpodobně čtyři geny.

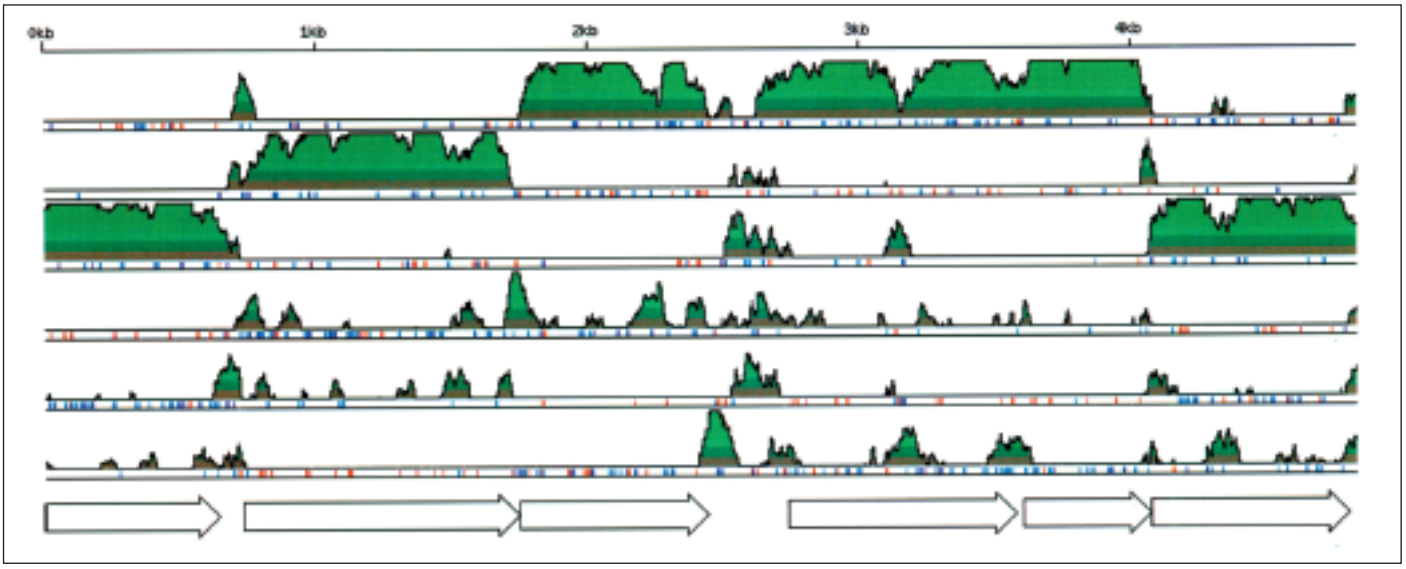
Hledání genů pomocí statistiky

Jak však najít ty geny, od kterých ještě žádný příbuzný nebo podobný gen z jiného organismu není znám? Geny úplně nové? Opět si vypomohu analogií knihy a kódu v DNA. Každý čtenář dobře ví, že jednotliví spisovatelé mají vlastní osobitý styl, používají různá slova s odlišnou frekvencí, mají oblíbená slovíčka, různé obraty a slovní spojení a používají třeba i různě dlouhé věty. Podle jejich způsobu psaní je často můžeme docela dobře identifikovat. A stejně je to i s DNA a jednotlivými organismy.

Protože kodóny jsou dlouhé tři nukleotidy a DNA se skládá ze čtyř různých druhů nukleotidů, existuje celkem 64 různých kodónů (4^3). Aminokyselin se ale v proteinech vyskytuje pouze 20. Proto je genetický kód (překlad DNA do proteinu) degene-

Tab. Preference kodónů. Kodón je tři písmena dlouhý úsek DNA, který je přeložen do jedné aminokyseliny — tedy ji kóduje. Jedna aminokyselina může být kódována více kodóny, v takovém případě odlišné organismy využívají jednotlivé kodóny různě často. V tab. je pro ilustraci uvedeno 6 kodónů pro aminokyselinu lysin a jejich využití bakteriemi Rhodobacter capsulatus a Escherichia coli

Kodón	*Rb. capsulatus* [% použití]	*E. coli* [% použití]
CTA	1	4
CTC	16	9
CTG	60	52
CTT	20	10
TTA	0	11
TTG	3	13



rovaný — jedna aminokyselina může být kódována více než jedním kodónem. Např. aminokyselinu lysin mají organismy zaznamenanou v DNA šesti různými způsoby. Jejich styl je dán výběrem, preferencí jednotlivých kodónů. Příklad toho, jak preference vypadají, uvádí tab.

Této vlastnosti s výhodou využijeme. Nejprve si ze známých genů našeho organismu odvodíme styl, preference pro jednotlivé kodóny. Pak vezmeme krátký úsek DNA a zjistíme, jaké všechny kodóny se v úseku vyskytují. Protože DNA je dvouvláknová molekula, můžeme v daném úseku kodóny získat šesti různými způsoby. Nejprve po trojicích v jednom směru a potom i na opačném vlákně (obr. 3). V tomto úseku DNA tedy může být šest genů. Zpravidla jen jeden z nich je reálný a skutečně v buňce vzniká. Abychom poznali, který to

T R M P L S L S M G M P I V H I H S C
N A N A S L S F D G Y A N C P H S L V L
E R E C L S L F R W V C Q L S T F T R V
GAACGCGAATGCCTCTCTCTCTTTCGATGGGTATGCCAATTGTCCACATTCACTCGTGT
CTTGCGCTTACGGAGAGAGAGAAAGCTACCCATACGGTTAACAGGTGTAAGTGAGCACA
F A F A E R E K S P Y A L Q G C E S T
R S H R E R K R H T H W N D V N V R T
V R I G R E R E I P I G I T W M * E H

Obr. 3 Překládání DNA do aminokyselin. Protože se DNA překládá po trojicích písmen (kodónech), lze každou část dvouvláknové DNA přeložit do aminokyselin šesti různými způsoby. Nejprve zleva doprava na horním vlákně: GAA CGC GAA..., AAC GCG AAT..., ACG CGA ATG... a potom na spodním vlákně zprava doleva: ACA CGA GTG..., CAC GAG TGA..., ACG AGT GAA. V horní a spodní části obrázku je znázorněno, jaký sled aminokyselin by teoreticky mělo šest možných genů

je, podíváme se, jak všechny odpovídají stylu používání kodónů ve zkoumaném organismu. Na obr. 4 je výstup z programu, který nám styl ukáže graficky. Testovaná DNA je ta samá, kterou jsme použili při hledání genů pomocí podobností. Z obrázku je patrné, že v tomto kousku DNA je genů šest.

Hledání genů pomocí entropie

Může nastat i situace, kdy je v DNA gen, který doposud není znám, takže ho nenajdeme pomocí příbuznosti, a navíc neznáme onen styl v používání jednotlivých kodónů, takže ani druhá metoda nám nepomůže. Pak nám zbývá poslední možnost. Tou je míra uspořádanosti „textu".

Tuto metodu si bioinformatici vypůjčili od kryptologů. Je založena na tom, že smysluplný text, který nese nějaké sdělení, je uspořádanější než text náhodný. Ukažme si to názorně. Mámě dva kousky textu, jeden s významem, druhý náhodný:

Chtělo by se mi zpívat a tančit.
Ghyd kmžruisól n shd ýžgst hutd.

Na první pohled bychom řekli, že více informace obsahuje horní řádek. Opak je však pravdou, větší informační obsah má věta v řádku dolním. Ověřte si moje tvrzení tak, že vypustíte jeden libovolný znak z horní a jeden ze spodní řádky. A nyní někoho požádejte, aby chybějící znak nahradil. V horním řádku se mu to ve většině případů povede bezchybně, pouze občas by mohla nastat nejasnost, např. chybějící „t" ve slově tančit by někdo mohl zaměnit za „j". Zato ve druhém řádku je správné nahrazení nemožné — část informace jsme vypuštěním jednoho znaku nenávratně ztratili.

Podobná situace jako v případě textu je i v DNA. Pokles informačního obsahu v kódujících oblastech lze využít k odlišení takových úseků DNA, které nesou sdělení, od těch, jež jsou díky mutacím náhodné. Tento způsob hledání genů je nejméně přesný, zejména proto, že v DNA se vyskytuje plno úseků, které sice žádný protein nekódují, ale přesto jsou nenáhodné. Např. regulační oblasti. Občas ale rádi vezmeme zavděk alespoň touto metodou.

Obr. 4 Šest grafů představuje šest možných způsobů čtení DNA. Výška zelených částí odpovídá tomu, jak jsou v těchto oblastech DNA jednotlivé kodóny využívány. Modré a červené značky pod grafem označují možné začátky a konce genů. Gen může začít u kterékoli modré a končit u nejbližší červené značky, v prvních třech grafech zleva doprava, v dalších třech zprava doleva. Všechny orig. J. Pačesa

Jak dál?

Z předcházejících ukázek by se mohlo zdát, že hledání genů je celkem snadné. V případě genomu bakterií tomu tak snad i je (dokud nenarazíme na nestandardní geny). Tam je úspěšnost hledání kolem 95 %. Horší situace je u vyšších organismů, které mají geny přerušeny negenovými oblastmi (exon–intronová struktura), jež jsou leckdy podstatně delší než vlastní gen. Tam ani kombinace všech zmíněných metod a několika jejich implementací nezaručí úspěšnost vyšší než 80–85 %. Musíme tedy pokračovat více biologickými experimentálními metodami než pouze metodami počítačovými. Např. budeme číst molekuly mRNA v buňce, protože mRNA vzniká přepsáním kódujících částí DNA.

Nalezení genů v DNA je pouze první krok. Stojí před námi otázka, jaké kódují proteiny a k čemu ty proteiny v buňce slouží, jakou mají funkci. Ale to už je povídání mimo tento článek.