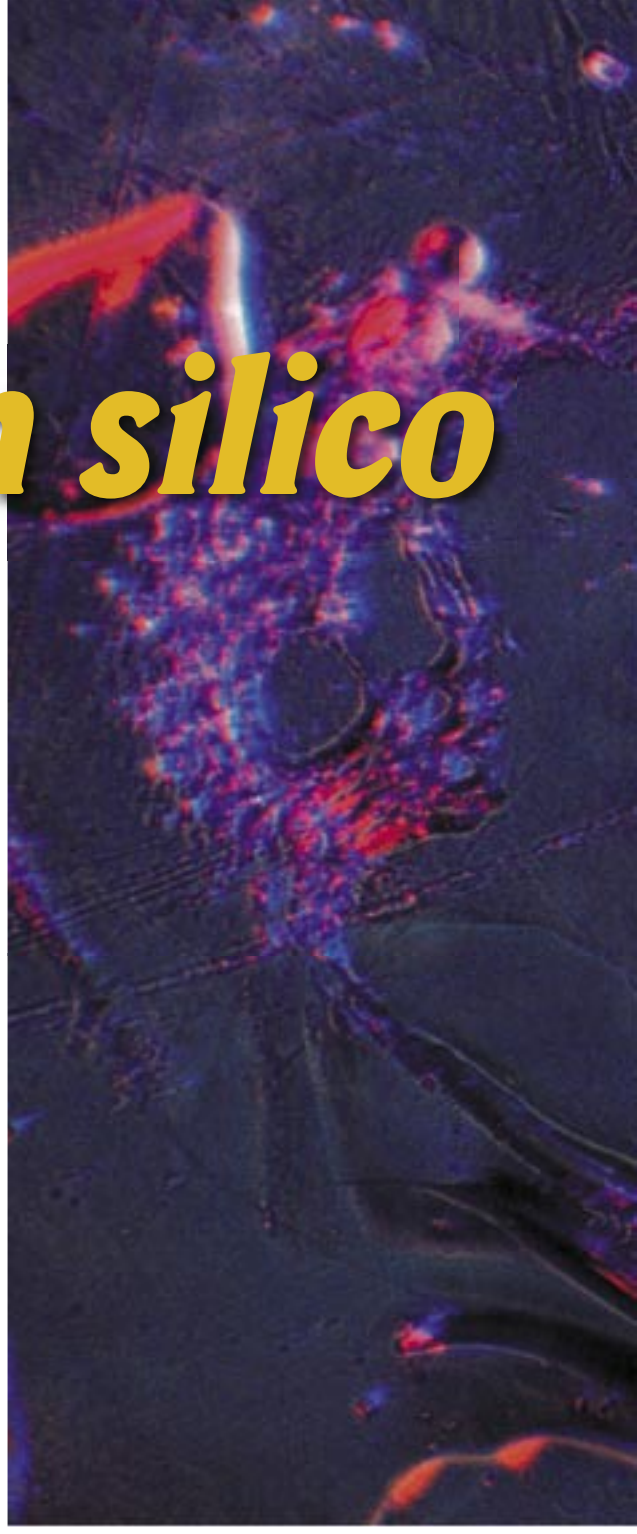
Ekologie genů z pohledu bioinformatiky[1]

Biologie *in silico*



RADKA STORCHOVÁ

PETR DIVINA

JIŘÍ VONDRÁŠEK

JAN PAČES

Tak jako se zoologové a botanici už po staletí zabývají vznikem, rozšířením, migracemi a vymíráním jednotlivých druhů rostlin a živočichů, mají nyní i molekulární biologové možnost studovat obdobné chování genů. Neustále přibývající počet sekvencovaných organizmů nám dnes dovoluje nahlédnout do života buněčného jádra i bez použití složitých mikroskopických či molekulárních technik. Protože většina dat je uložena ve veřejně přístupných databázích, potřebujeme k práci „pouze" výkonný počítač vybavený potřebnými programy a pár chytrých nápadů; někdy se mluví o studiu genomu *in silico* (tedy v počítači). V současné době se tímto přístupem zabývá již zcela samostatný obor, začínající se slibně rozvíjet i v České republice. Dostal název bioinformatika. Pracuje s daty, u nichž můžeme rozlišit tři kategorie:

- genom – nukleotidové sekvence genů a celých genomů;[2]
- transkriptom – informace o expresi genů[3] v jednotlivých tkáních a vývojových stadiích;
- proteom – informace o aminokyselinových sekvencích, struktuře a funkci proteinů.

Všechna data spolu samozřejmě úzce souvisejí a vzájemně se doplňují – bez sekvencí genomů bychom těžko určovali, které geny jsou ve studovaných tkáních exprimovány (vyjádřeny); data o expresi genů a sekvence proteinů nám zas pomáhají v genomu nacházet geny. V následujících odstavcích se pokusíme přiblížit práci dnešních bioinformatiků včetně výsledků, které nám jejich úsilí přineslo. Protože témat, která dnes bioinformatika řeší, je nepřeberné množství, zaměříme se hlavně na ta, na nichž se významně podílejí vědci z České republiky.

Uspořádání genů v genomu není náhodné

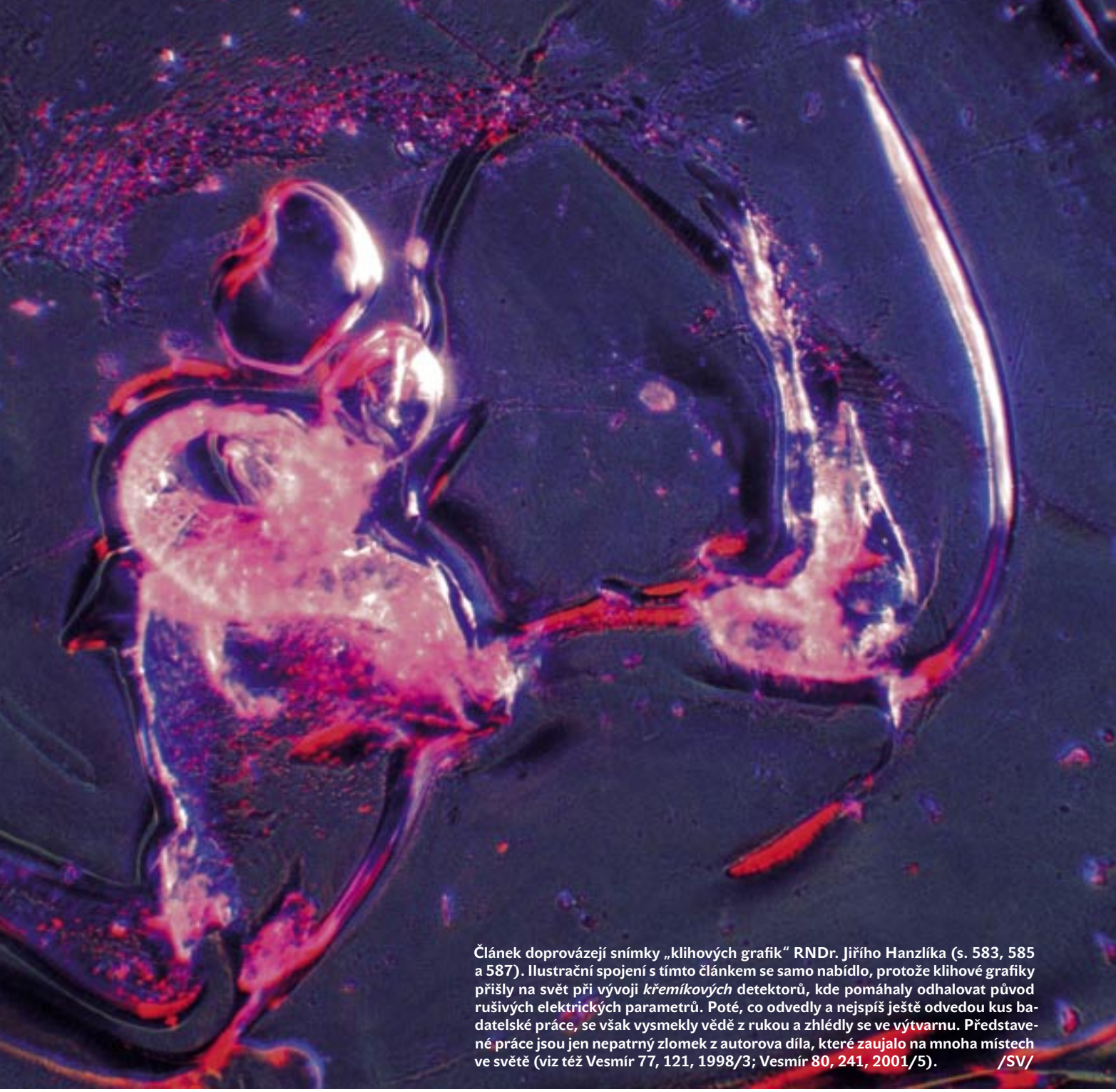
Po přečtení sekvence genomů prvních eukaryotických organizmů upoutalo zájem bioinformatiků rozmístění genů na chromozomech. Ještě nedávno se mělo za to, že rozmístění genů na chromozomech je v zásadě náhodné. Vládla zažitá představa o tzv. *trans-regulaci genové exprese* – to znamená, že transkripční faktory, které rozhodují o tom, jaké geny budou v daný okamžik v dané tkáni exprimovány, si v jádře vyhledají své cílové sekvence nezávisle na tom, v jakém místě genomu tyto sekvence leží. Při takové představě o regulaci genové exprese je tedy naprosto lhostejné, kde se geny nacházejí. Vědělo se sice o existenci genů, které se těmto pravidlům vymykají (např. Hox geny, jež u všech organizmů poslušně leží jeden za druhým a ještě ve stejném pořadí), ale ty byly považovány spíše za vzácnou výjimku. Výsledky bioinformatiků však tuto vžitou představu nabourávají. Ukazuje se totiž, že geny s podobnou expresí (ať už jsou to geny exprimované ve všech tkáních, nebo naopak tkáňově specifické geny) mají tendenci se v genomu shlukovat. O tom, proč tomu tak je, se dá zatím jen spekulovat – mluví se například o možné existenci jakýchsi chromozomálních smyček, které se rozbalují (a tím zpřístupňují transkripci) buď ve všech, nebo jen v určitých tkáních. Geny tedy mohou být umístěny jedině v takových smyčkách, které jim umožní, aby byly exprimovány ve správ-

Mgr. Radka Storchová (*1979) vystudovala biologii na Přírodovědecké fakultě UK v Praze. V Ústavu molekulární genetiky AV ČR se zabývá genetikou speciace u myší a evolucí pohlavních chromozomů.

Mgr. Petr Divina (*1975) vystudoval genetiku a molekulární biologii na Přírodovědecké fakultě MU v Brně. V Ústavu molekulární genetiky AV ČR se zabývá bioinformatickou analýzou genových expresních dat.

RNDr. Jiří Vondrášek, CSc., (*1963) vystudoval biofyziku na Matematicko-fyzikální fakultě UK v Praze. V Ústavu organické chemie a biochemie AV ČR se zabývá problematikou sbalování proteinů a návrhy strukturních databází.

Jan Pačes, Ph.D., (*1967) vystudoval biochemii na PřF UK v Praze. V Ústavu molekulární genetiky AV ČR se zabývá bioinformatickou analýzou endogenních retrovirových elementů.



Článek doprovázejí snímky „klihových grafik" RNDr. Jiřího Hanzlíka (s. 583, 585 a 587). Ilustrační spojení s tímto článkem se samo nabídlo, protože klihové grafiky přišly na svět při vývoji *křemíkových* detektorů, kde pomáhaly odhalovat původ rušivých elektrických parametrů. Poté, co odvedly a nejspíš ještě odvedou kus badatelské práce, se však vysmekly vědě z rukou a zhlédly se ve výtvarnu. Představené práce jsou jen nepatrný zlomek z autorova díla, které zaujalo na mnoha místech ve světě (viz též Vesmír 77, 121, 1998/3; Vesmír 80, 241, 2001/5). /SV/

ný okamžik tam, kde je to zapotřebí. Zatím tato teorie není ověřena, nicméně díky bioinformatice máme cenná data, která nám vůbec nějaké teorie o fungování buněčného jádra umožňují formulovat.

SeXY chromozomy

Kromě shlukování genů s podobnou expresí existuje ještě jedna skutečnost, která narušuje náhodné uspořádání genů v genomu – jsou jí pohlavní chromozomy (viz také Vesmír 84, 323, 2005/6). Dosud se předpokládalo, že soubor genů daného organizmu je náhodně rozházen mezi jednotlivé chromozomy. Bioinformatici však ukázali, že tento předpoklad neplatí pro chromozomy pohlavní, které svým genovým obsahem jednoznačně vybočují z řady. Pohlavní chromozomy si vysloužily svoje pojmenování díky tomu, že nesou geny, které jsou odpovědné za určení pohlaví. V současné době se však ukazuje, že si označení „pohlavní" zaslouží víc, než se dříve předpokládalo. Vzhledem ke genům, jež obsahují, to jsou totiž opravdu sexy chromozomy. Už delší dobu je známo, že chromozom Y, vyskytující se pouze v samčích buňkách, je obohacen o geny, které potřebují pouze samci (například geny odpovídající za vznik spermií). Překvapením však je, že i chromozom X, který se vyskytuje u obou pohlaví,[4] má výjimečný obsah genů. Ve srovnání s ostatními chromozomy nabízí více genů exprimovaných v ryze mužských orgánech (např. prostatě) nebo prekurzorech mužských pohlavních buněk. Zároveň však nese i více genů specifických pro orgány ženské (placentu a vaječníky). Je tedy maskulinizován a feminizován současně. Kromě toho

1) Ekologie je věda o uspořádáních a vztazích v živé přírodě.

2) Sekvence genu (genomu) je pořadí nukleotidů (A, T, G, C) ve vlákně DNA.

3) Exprese genu je přepis nukleotidové sekvence genu do nukleotidové sekvence molekuly mediátorové (informační) RNA, která popřípadě může být překládána do sekvence aminokyselin proteinu. Méně často se pro *expresi genu* užívá *vyjádření genu*; smůla je, že oba termíny v češtině evokují ještě další významy; v tomto textu se držíme termínů *exprese, exprimovat, exprimovaný.*

4) Samice savců nesou zpravidla kombinaci pohlavních chromozomů XX, samci XY.

Svítící DNA-čip. Mediátorová RNA byla izolována z HeLa buněk a hybridizována na DNA-čipu, který obsahoval vybraný soubor genů kódujících mitochondriální proteiny. Převzato z laboratoře Dr. Stanislava Kmocha, (Ústav dědičných metabolických poruch UK v Praze), snímek © A. Čížková.

je chromozom X obohacen o geny exprimované v orgánech, jejichž funkce se liší u obou pohlaví – například v mozku. Prozíraví evoluční biologové sice tento jev předpověděli už dávno, nicméně díky bioinformatice byl konečně prokázán.

Čilý dopravní ruch na chromozomu X

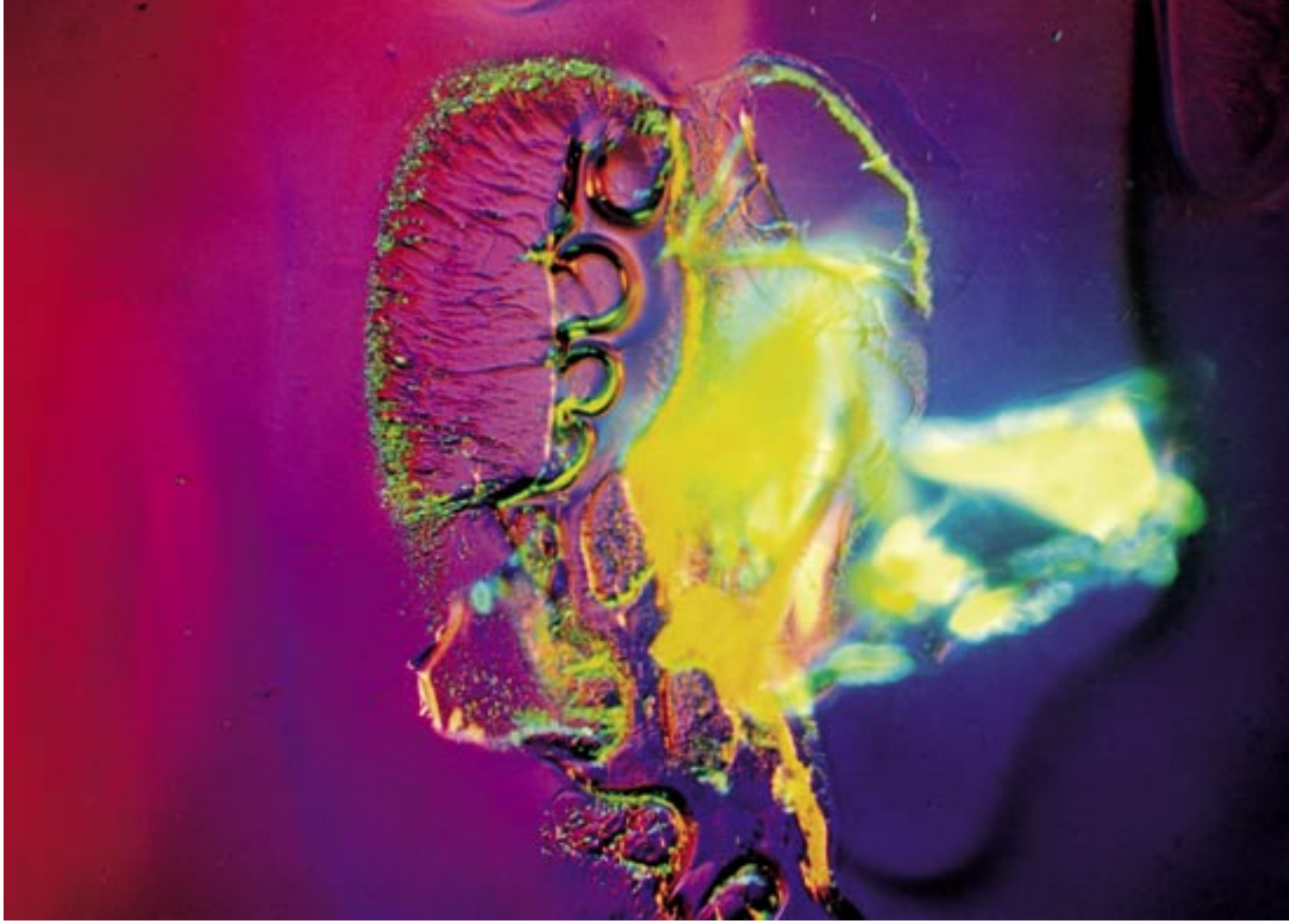
Geny na svých místech na chromozomech nezůstávají trvale, ale čas od času se mohou přesunout na jiné místo. Jedním ze způsobů cestování genů jsou *retropozice* (Vesmír 79, 273, 2000/5): geny se nejdříve přepíšou do molekuly mediátorové RNA, která se může v jádře volně pohybovat, a potom se zase zpětně přepíšou do molekuly DNA, která se může začlenit na úplně jiné místo v genomu. Protože jsou takové přesuny spojeny s určitými změnami v sekvenci, dá se poznat, která kopie genu je původní a která odvozená, čili můžeme spolehlivě určit, odkud a kam gen cestoval. Přesně tohle využil tým amerických bioinformatiků z Chicaga, který sledoval přesuny lidských a myších genů. Výsledky jejich studie ukazují, že k nejintenzivnějšímu genovému stěhování dochází mezi chromozomem X a ostatními chromozomy, a to v obou směrech. Geny se na chromozom X stěhují častěji než na kterýkoli jiný a zároveň tento chromozom nejčastěji opouštějí. Jaký k tomu mají důvod? Chromozom X má mezi ostatními chromozomy výjimečné postavení tím, že samci nesou pouze jednu jeho kopii, zatímco samice dvě. Tento rozdíl způsobuje, že chromozom X tráví dvě třetiny svého času v samicích a pouze jednu třetinu v samcích. Geny prospěšné pro samice tedy mohou mít zájem přesunout se na tento chromozom, a naopak některé geny prospívající samcům ho raději opouštějí. Jiné samčí geny zas mohou chromozom X vyhledávat, protože jenom na tomto chromozomu – který se u samců vyskytuje jen v jedné kopii – se projeví i recesivní mutace.[5] Dalším faktorem zvyšujícím migraci genů na chromozomu X je jeho neobvyklé chování při vzniku samčích pohlavních buněk. Chromozom X se v těchto buňkách zcela uzavře jakýmkoliv transkripčním faktorům a takto umlčen netečně spočívá na periferii buněčného jádra. Jestliže se tedy na tomto chromozomu objeví gen, který má být exprimován právě při vzniku samčích pohlavních buněk, musí se přestěhovat jinam. Zkráceně řečeno, protože se chromozom X chová jinak v samcích než v samicích, je pro určité geny, jejichž funkce se u obou pohlaví liší, atraktivním cílem a pro jiné zase zcela nevyhovujícím místem. To také vysvětluje jeho výjimečný genový obsah.

Třídění genetického harampádí

Abychom mohli analyzovat sekvence genomů vyšších organizmů, je často potřeba zbavit se nejdříve *repetitivních sekvencí*,[6] které například u lidí zabírají zhruba polovinu celého genomu (pro srovnání: samotné geny okupují jen asi 2% genomu). Většinou jde o tzv. *transpozony* neboli skákající elementy, což jsou „DNA paraziti“, kteří mohou škodit, když se náhodou v genomu začlení na nevhodné místo, na druhou stranu však činí genom dynamičtějším a umožňují jeho

5) Recesivní mutace se projeví pouze tehdy, není-li v buňce přítomna ani jedna dominantní verze daného genu.

6) Repetitivní sekvence nukleotidů jsou ty, které se v genomu mnohokrát opakují.



rychlejší vývoj. Pro usnadnění práce s genomem vyvinuli bioinformatici program zvaný *RepeatMasker*, který repetitivní sekvence dokáže rozpoznat a odstranit. Pomocí tohoto programu však pochopitelně můžeme udělat i přesně opačnou věc – vybrat repetitivní sekvence, které jinak odhazujeme jako nepotřebné harampádí, a začít je zkoumat, což vůbec není jednoduchý úkol. Většina těchto sekvencí totiž už dávno není celá. Postupem času se v nich nahromadily mutace, sekvence byly rozlámány a ztratily některé úseky. Pro srovnání si můžete představit skladiště starého harampádí, jaké má kdekdo doma na půdě. Tu se válí noha od stolu, blatníky od kola, provaz, o kus dál pneumatika, která se sem zakutálela, v osamělém šuplíku leží nějaké nářadí, pumpička a spouta dalších už jen těžko identifikovatelných předmětů.

Přesto se skupina českých bioinformatiků rozhodla tento složitý systém zkoumat. Nejdřív v něm však musela trochu udělat pořádek. Pro začátek si vybrali jeden typ transpozonů a zmapovali polohu všech jeho elementů (včetně neúplných) v lidském genomu. Jednotlivé fragmenty potom poskládali dohromady a rekonstruovali jejich původní podobu, což pro naše přirovnání znamená totéž jako vytvořit obrázek domácnosti před dvaceti lety podle harampádí z půdy.

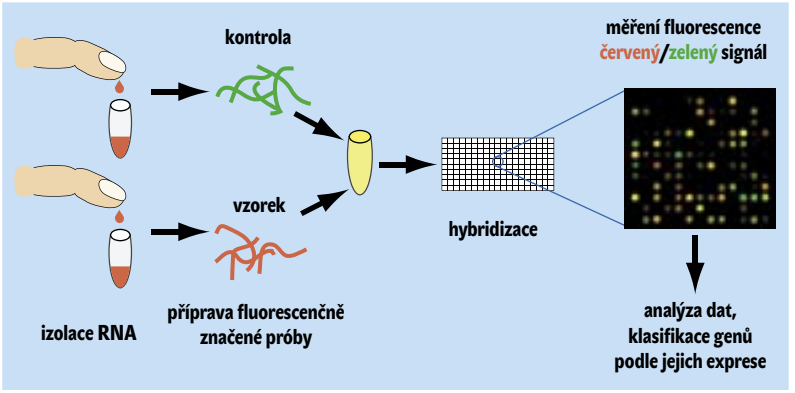
Dnes už jsou informace o transpozonech začleněny do celosvětově nejvíce používaného prohlížeče genomů *Ensembl*. Znalost jejich rozmístění v lidském genomu nám sice neřekne nic o nás samotných, protože transpozony skoro neovlivňují naše vlastnosti. Mohou však ledacos prozradit o struktuře chromozomů. Ví se totiž, že ani rozmístění transpozonů v genomu není náhodné, ale do značné míry kopíruje způsob uspořádání vlákna DNA v chromozomu.

Informace o genové expresi

Kromě samotné nukleotidové sekvence genů potřebujeme často informaci o tom, ve kterých tkáních jsou jednotlivé geny exprimovány. To ze sekvence pochopitelně nevyčteme, ale naštěstí už ani v tomto případě nemusíme brát do ruky pipetu a dělat experimenty, protože je někdo udělal za nás. V současné době se data o genové expresi získávají nejčastěji pomocí DNA-čipů (nebo-li mikroarrays) a metodou zvanou SAGE (sériová analýza genové exprese), které umožňují analyzovat expresi velkého množství genů zároveň. DNA-čipy jsou skleněné destičky, na jejichž povrchu jsou ukotveny kousíčky DNA. U celogenomových čipů jde o úseky všech známých genů daného organizmu.

Pokud chceme vědět, které z genů jsou exprimovány třeba v játrech, stačí z jater izolovat molekuly mediátorové RNA, označit je nějakou fluorescenční barvičkou a nechat hybridizovat na DNA-čipu. Molekuly mediátorové RNA se při tom navážou na odpovídající kousky DNA (spojí se spolu – hybridizují – komplementární úseky nukleových kyselin). Po odmytí nenavázané mediátorové RNA zjistíme podle intenzity fluorescence, která čipová DNA se spojila se zkoumanou mediátorovou RNA, což zároveň odhalí geny exprimované ve zkoumané tkáni. Pochopi-

více na www.vesmir.cz
www+



Postup při analýze genové exprese pomocí DNA-čipů. Na obrázku je příklad dvoukanálového systému, při němž se porovnává exprese genů ve dvou vzorcích – v analyzovaném a kontrolním. Molekuly mediátorové RNA z obou vzorků jsou označeny dvěma různými flourescenčními barvičkami (obvykle zelenou a červenou), smíchány a hybridizovány na čipu. Každá tečka na čipu představuje určitý gen. Poměr intenzit červeného a zeleného signálu jednotlivých teček odpovídá změně genové exprese mezi oběma vzorky. Svítí-li tečka více červeně, znamená to, že daný gen je exprimován více ve zkoumaném vzorku než ve vzorku kontrolním. Naopak, pokud svítí tečka více zeleně, dochází u našeho vzorku k poklesu exprese oproti kontrole. Na čipech je obvykle umístěno až několik desítek tisíc teček (u některých komerčních čipů i přes milion a odpovídají genům z celého lidského nebo myšího genomu). Pro vyhodnocování čipů se používají specializované statistické programy.

telně však dostaneme informaci pouze o těch genech, které jsme si předtím na sklíčko přichytili (viz obr. na s. 584 a 586).

Metoda SAGE je sice o trochu pracnější, zato však poskytuje informaci o expresi úplně všech genů, i těch dosud neznámých. Její princip spočívá v tom, že se z každé jednotlivé molekuly mediátorové RNA vystřihne jeden krátký úsek (14–21 nukleotidů), tyto úseky se potom pospojují a jako celek sekvencují. Úsek 14–21 nukleotidů je už dostatečně dlouhý na to, abychom poznali, z jakého genu pochází. SAGE neposkytuje informaci jen o tom, jaké geny jsou v daných tkáních aktivní. Díky tomu, že je z každé molekuly mediátorové RNA vystřižen právě jeden kousek, dostaneme informaci i o intenzitě genové exprese.

Data o genových expresích, sekvencích genů a genomů a informace o funkci genových produktů jsou uloženy ve vzájemně propojených veřejně přístupných databázích. Jedna z těchto databází je také spravována v České republice. Shromažďují se v ní veřejně dostupná data vytvořená metodou SAGE z myších tkání a buněčných linií. Díky tomu, že jsou data z různých laboratoří zpracována jednotně a shromážděna na jednom místě, lze je snadno prohledávat a analyzovat.

Soutěž o proteinovou strukturu

Sekvence genu ani informace o jeho expresi nám ještě nic neříkají o tom, jak daný genový produkt – protein – v buňce funguje. K objasnění funkce proteinu na molekulární úrovni totiž potřebujeme znát jeho trojrozměrnou strukturu. To je důležité například tehdy, chceme-li proti určitému proteinu navrhnout jako léčivo malou molekulu, která by nějak ovlivnila jeho funkci. Zatím bohužel nebyl objeven způsob jak z aminokyselinové sekvence odvodit prostorové uspořádání určitého proteinu. Tento problém, nazývaný také *sbalování proteinů*, je pro strukturální biology ohromnou výzvou skoro padesát let a je jisté, že rozluštění „sbalovacího kódu“ bude pro rozvoj biologie podobně přelomovou událostí, jako bylo rozluštění kódu genetického.[7]

Zatím však nezbývá než určovat struktury proteinů experimentálně. Nejčastějším postupem je vytvoření krystalické formy proteinu a vypočítání vzdáleností mezi jednotlivými atomy v krystalu podle toho, jak jím procházejí rentgenové paprsky. To je ovšem práce dost zdlouhavá a náročná. Proto se strukturální biologové nepřestávají povzbuzovat při vyvíjení programů pro predikci prostorového uspořádání proteinů a každé dva roky organizují mezinárodní soutěž o to, komu se podaří správně předpovědět strukturu proteinu, která v té době byla vyřešena experimentálně, avšak pro potřeby zmíněné soutěže zůstala v tajnosti.

V současné době je již vyřešeno více než 30 tisíc proteinových struktur. Každý protein přitom nemusí mít vždy jen jednu strukturu – například pro proteázu viru HIV, nejstudovanější protein vůbec (Vesmír 80, 332, 2001/6), známe už 250 různých trojrozměrných uspořádání. Proteáza HIV je nezbytným enzymem pro množení viru HIV, a je tedy jedním z cílů při hledání účinné léčby aidsu. Vytvoření léčiva, které by účinně zablokovalo funkci této proteázy, však komplikují mutace, jejichž vinou proteáza často mění tvar své vazebné dutiny, a uniká tak účinku navržených léčiv. Hledání univerzálního léčiva schopného blokovat všechny varianty enzymu by však mohla usnadnit speciální databáze, která obsahuje všechny známé trojrozměrné struktury proteázy HIV. Na jejím vybudování se podíleli také čeští bioinformatici.

Proč nemít strach z FOBIE

Bioinformatika je dnes obsáhlou vědou zabývající se sestavováním biologických databází, vytvářením nástrojů pro analýzu sekvencí DNA a proteinů, zpracováním dat o expresi genů, studiem uspořádání genů v genomech a jejich evolucí, stanovením struktury biologických makromolekul či modelováním metabolických drah. V nedávné době vzniklo v České republice sdružení FOBIA (Free & Open Bioinformatic Association), které bylo založeno jako sekce České společnosti pro biochemii a molekulární biologii. Sdružení FOBIA by mělo umožňovat kontakty a spolupráci mezi českými vědci, kteří se zajímají o různá odvětví bioinformatiky. Zájemci o činnost sdružení se mohou přihlásit na internetové adrese *http://fobia.img.cas.cz* nebo se mohou zúčastnit veřejně přístupných přednášek, které FOBIA pořádá obvykle jednou za měsíc.

K dalšímu čtení

J. Pačes et al.: HERVd: database of human endogenous retroviruses, Nucleic Acids Res. 30, 205, 2002
J. Vondrášek, A. Wlodawer: HIVdb: a database of the structures of human immunodeficiency virus protease, Proteins 49, 429, 2002
P. Divina, J. Forejt: The Mouse SAGE Site: database of public mouse SAGE libraries, Nucleic Acids Res. 32, D482, 2004
J. Emerson et al.: Extensive gene traffic on the mammalian X chromosome, Science 303, 537, 2004
L. D. Hurst et al.: The evolutionary dynamics of eukaryotic gene order, Nat. Rev. Genet. 5, 299, 2004
P. Divina et al.: Global Transcriptome Analysis of the C57BL6/J Mouse Testis by SAGE: Evidence For Nonrandom Gene Order, BMC Genomics 6, 29, 2005

7) Genetický kód je soubor pravidel, podle nichž je sekvence nukleotidů mediátorové RNA překládána do sekvence aminokyselin v proteinu.