BIOLOGICKÉ LISTY 68 (3): 239-246, 2003

Použití bioinformatiky v rostlinné biologii - analýza proteinových sekvencí

MARTIN POTOCKÝ

Ústav biochemie a mikrobiologie VŠCHT Praha, Technická 5, 166 28 Praha, (tel.: 221 951 685, e-mail: potockym@vscht.cz)

Úvod

Dokončení kompletní geonomové sekvence modelových rostlinných organismů jako *Arabidopsis thaliana* či rýže, a exponenciální nárůst probíhajících i dokončených sekvenačních projektů, přinášející ohromné množství primárních sekvenčních dat, vyžaduje nové postupy a prostředky k jejich zpracování. Zde se pokusím nastínit nejpoužívanější strategie analýzy proteinových sekvencí a možnosti jejich využití pro biologii rostlin. Text je členěn podle pravděpodobného schématu práce a zahrnuje identifikaci homologních proteinů, možnosti analýzy primární sekvence a predikci sekundární a terciární struktury proteinů. Ve výčtu příslušných programů se omezím pouze na software a služby dostupné přes internet, které jsou v akademické sféře volně distribuovatelné.

Databáze proteinových sekvencí

Od roku 1965, kdy Margaret Dayhoff publikovala poprvé slavný *Atlas of Protein Sequence and Structure* (první veřejná kompletní databáze, obsahující údaje o 65 proteinech, Dayhoff *et al.* 1965) se počet položek v proteinových databázích zvýšil na téměř 3 miliony, což kromě sekvenačního úsilí dokumentuje i nárůst redundance. Rovněž je třeba upozornit na fakt, že velká část proteinových sekvencí je výsledkem translace predikovaných kódujících oblastí genů, pro které neexistují žádné experimentální důkazy. Současný trend v proteinových databázích je tedy dvojí - na jedné straně kompletní databázové projekty odvozené z nukleotidových databází jako jsou GenBank\EMBL\DDBJ, které zároveň zahrnují výše zmíněné nevýhody, a nekompletní specializované, často manuálně redigované, databáze přinášející ověřené, dobře anotované sekvence na straně druhé. Přehled nejpoužívanějších zdrojů přináší tab. 1.

Vyhledávání sekvencí v databázích

Způsobů jak získat požadovaná data z proteinových databází je několik. Každá sekvence má svůj unikátní identifikátor, podle kterého ji můžeme jednoduše vyhledat. Další možností je hledání podle klíčových slov. Asi nejkompletnější databázová rozhraní Entrez (http://www.ncbi.nlm.nih.gov/entrez) a SRS-EBI (http://srs.ebi.ac.uk) nabízí řadu dalších vyhledávacích kriterií, jako je druh organismu, autor, rok původu a další.

Nejcitlivějším a nejkomplexnějším přístupem je vyhledávání orthologních sekvencí na základě homologie. Základním přístupem téměř všech metod je konstrukce tzv. „*pairwise alignments*“, tj. porovnání dvou sekvencí (z nichž jedna je vždy námi zadána), kvantifikace jejich vzájemné podobnosti a odhadnutí statistické významnosti výsledku. Tab. 2 uvádí nejpoužívanější algoritmy a jejich implementace.

Tab. 1. Přehled databází proteinových sekvencí

Název	URL odkaz	Poznámka
Entrez Proteins	http://www.ncbi.nlm.nih.gov/entrez	+ kompletní, integruje všechny ostatní databáze, provázána s literárními údaji, - redundantní, často špatná anotace
GenomeNet	http://www.genome.ad.jp/htbin/www_bfind?protein-today	+ kompletní, náhrada za Entrez - redundantní, uživatelsky nepříjemná
SWISS-PROT	http://www.expasy.org/sprot	+ ručně anotována, pouze ověřené sekvence, provázána s dalšími zdroji, nízká redundance - nekompletní
TrEMBL	http://srs.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-page+top	+ málo redundantní, předvoj a doplněk pro Swiss-Prot - aktualizována čtvrtletně
PIR	http://pir.georgetown.edu	+ řazení do rodin podle příbuznosti - nekompletní, často špatná anotace
Expasy Links	http://us.expasy.org/alinks.html	odkazy na velké množství specializovaných databází
Arabidopsis Gene Families	http://www.arabidopsis.org/info/genefamily/genefamily.html	odkazy na jednotlivé databáze genových rodin
Aramemnon	http://aramemnon.botanik.uni-koeln.de/	databáze membránových proteinů v *Arabidopsis*

Tab. 2. Zdroje pro vyhledávání proteinových sekvencí v databázích

Název	URL odkaz	Poznámka
Blast (NCBI)	http://www.ncbi.nlm.nih.gov/BLAST/	základní rozhraní, nejpoužívanější, uživatelsky nejpříjemnější
PSI-Blast (NCBI)	http://www.ncbi.nlm.nih.gov/BLAST/	detekce příbuzných sekvencí s nízkou podobností, iterativní, riziko falešně-pozitivních výsledků
Standalone BLAST (NCBI)	ftp://ftp.ncbi.nih.gov/blast/executables	program ke stažení na vlastní počítač, uživatelsky nepříjemný, pokročilé funkce (Windows, Linux)
DDBJ Homology Search System	http://www.ddbj.nig.ac.jp/E-mail/homology.html	soubor různých algoritmů (Blast, Psi-Blast, Fasta, Smith-Waterman), alternativa k NCBI
WU-Blast (EBI)	http://www2.ebi.ac.uk/blast2/	odlišný algoritmus od NCBI-Blast, někdy citlivější
FastA (EBI)	http://www2.ebi.ac.uk/fasta3/	vhodný pro dlouhé úseky s nízkou podobností
Smith-Waterman (EBI)	http://www.ebi.ac.uk/searches/blitz_input.html	citlivější než Blast a Fasta, pomalý
Propsearch	http://www.infobiosud.univ-montp1.fr/SERVEUR/PROPSEARCH/propsearch.html	hledání je založeno na aminokyselinovém složení a dalších fyzikálně-chemických vlastnostech proteinu
AA composition (ExPASY)	http://www.expasy.ch/tools/aacomp/aacomp0.html	hledání je založeno na aminokyselinovém složení, pI a molekulové hmotnosti

Vedle volby algoritmu je důležitým krokem volba tzv. substituční matrice, která definuje kritéria kvantifikace jednotlivých pozic v *alignment* (bonus za zachování, penalizace za záměnu). V dnešní době se nejčastěji používají matrice PAM (Dayhoff a Eck 1968) a Blosum (Henikoff a Henikoff 1992). Matrice PAM30 a Blosum80 se používají pro hle-

dání vysoce příbuzných sekvencí, použití matric PAM250 a Blosum45 je vhodné pro hledání vzdálených homologů s předpokládanou nízkou podobností. „Kompromisní" matrice Blosum62 se používá jako základní nastavení pro většinu vyhledávacích programů. Pro každou matrici pak existují experimentálně testované hodnoty tzv. „*gap parameters*" (penalizace za vytvoření a prodloužení mezery v *alignment*). Dalším prvkem ovlivňujícím citlivost hledání je hodnota parametru word (Blast) či K-tuple (FastA): pro citlivější (a delší) vyhledávání je možno zvolit nabízenou nižší hodnotu. Všechny používané vyhledávací programy také umožňují zapojení algoritmů, které rozpoznají a maskují sekvence s nízkou komplexitou (např. repetitivní či prolin bohaté oblasti); tyto sekvence mohou jinak dávat falešně pozitivní výsledky.

Významné zvýšení citlivosti při hledání vzdáleně příbuzných sekvencí přineslo zavedení tzv. pozičně-specifických matricí (PSSM - *position-specific weight matrix*). Čelním algoritmem využívajícím tento přístup je program PSI-BLAST (Altschul *et al.* 1997). PSI-BLAST využívá iterační přístup: po prvním kole hledání se z nalezených sekvencí vytvoří *multiple alignment* (viz níže) a generuje se PSSM (která přičítá vysoké bonusy za zachování pozic konzervovaných v alignmentu a obráceně). Takto vytvořená PSSM je použita k vyhodnocení druhého kola hledání, přidáním nově nalezených sekvencí se vytvoří nová PSSM a celý proces pokračuje po definovaný počet kol či do nasycení.

V případě, že výše popsanými metodami nemůžeme najít k zadané sekvenci žádné homologní sekvence, můžeme zkusit hledat podle aminokyselinového složení a dalších fyzikálně chemických vlastností proteinu (tab 2). Tento způsob může naznačit strukturní a funkční příbuznost mezi novou neznámou sekvencí a sekvencemi v databázích (Hobohm a Sander 1995).

Analýza proteinových rodin – *multiple alignment*

Zatímco metody *pairwise alignment* nacházejí použití především v prohledávání databází, pro simultánní analýzu většího množství homologních proteinů jsou kritické postupy tzv. „*multiple alignment*", tj. srovnání sady proteinů. Zatímco algoritmy *pairwise alignments* jsou podloženy robustní statistickou teorií a produkují tzv. optimální *alignment* (tj. s maximálním skóre), metody *multiple alignment* jsou pouze aproximační. V tab. 3 uvádím nejčastěji využívané programy využívající rozdílné algoritmy vhodné pro různě konzervované sekvence. Výstupem programu je pak buď globální (algoritmus se snaží srovnat všechny pozice v sekvencích) nebo lokální *alignmen*t (program vyhledává konzervované, často poměrně krátké úseky, mezi kterými jsou dlouhé úseky nesrovnaných sekvencí). Obecně platí, že pro srovnání dlouhých sekvencí s nízkou mírou podobnosti je lepší použít algoritmy vytvářející lokální *alignment*, zatímco metody globálních *alignments* se lépe hodí pro vysoce konzervované sekvence. Podoba výsledného *alignment* se může výrazně měnit s použitím různých substitučních matricí a *gap parameters*, podobně jako u *pairwise alignment*. V praxi je výhodné testovat několik algoritmů a výsledky (s kritickou analýzou) kombinovat.

Existuje rovněž řada softwarů umožňující rychlou manuální editaci *multiple alignments* a přípravu barevně kolorovaných a anotovaných *alignments* k publikaci; výčet nejčastěji používaných je uveden v tab 3.

Tab. 3. Programy pro vytváření a editaci *multiple alignments*

Název	URL odkaz	Poznámka
BCM Multiple Sequence Alignments	http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html	webové rozhraní nabízející několik algoritmů (ClustalW, MAP, PIMA, MSA, Block Maker, MEME)
ClustalW (EBI)	http://www.ebi.ac.uk/clustalw	nejpopulárnější program, globální *alignment*
T-Coffee	http://igs-server.cnrs-mrs.fr/Tcoffee/	nejlepší program na *alignment* vysoce konzervovaných sekvencí, umí i kombinovat *alignments* produkované jinými algoritmy
Dialign	http://bibiserv.techfak.uni-bielefeld.de/dialign/submission.html	nejlepší program na alignment málo konzervovaných sekvencí, lokální *alignment*
POA	http://www.bioinformatics.ucla.edu/poa/	nejrychlejší, vhodný pro velký počet sekvencí
ClustalX	http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html	verze ClustalX ke stažení a instalaci na vlastní počítač (Windows, Linux)
Macaw	http://www.bris.ac.uk/Depts/PathAndMicro/services/CGR/Common Files/MACAWinstructions.htm	program ke stažení, lokální *alignment*, vhodný pro málo konzervované sekvence, možnost ručního editování (Windows)
Boxshade	http://www.ch.embnet.org/software/BOX_form.html	server pro grafickou úpravu *alignment*s k publikaci
Chroma	http://www.lg.ndirect.co.uk/chroma/index.htm	program ke stažení, umožňuje anotaci a úpravu *alignments* k publikaci (Windows)
GeneDoc	http://www.psc.edu/biomed/genedoc/	program ke stažení, slouží k manuální editaci a grafické úpravě *alignments* (Windows)

Detekce proteinových domén a konzervovaných motivů v primární sekvenci proteinu

Identifikace strukturních domén či motivů v neznámé sekvenci je důležitý krok při vytváření hypotézy o biologické funkci zkoumaného proteinu. Realizace tohoto úkolu je však komplikována skutečností, že proteinové domény s podobnou prostorovou strukturou a funkcí často sdílejí jen velmi málo podobnosti na úrovni primární struktury a jejich detekce konvenčními metodami *pairwise* a *multiple alignments* proto není možná. Řešení přineslo zavedení pozičně-specifických matricí PSSM, které umožnilo vybudovat PSSM charakteristické pro konkrétní domény. Příslušné servery (tab 4.) potom obsahují (pravidelně obnovovanou) databázi jednotlivých PSSM, kterou je možno prohledávat zadanou sekvencí.

Dalším algoritmem umožňujícím detekci málo konservovaných domén se stalo využití tzv. „*hidden markov models*“ (HMM) v bioinformatice. HMM představuje dynamický statistický profil charakteristický pro určitou doménu či sekvenční motiv (vysvětlit princip HMM přesahuje možnosti tohoto článku, viz Hughey a Krogh 1996). Specializované servery (tab. 4.) opět disponují souborem prohledávatelných HMM profilů, výstup pak obsahuje statistické zhodnocení a grafické znázornění analýzy.

Existují i programy pracující opačným směrem, tj. v uživatelem zadaném souboru sekvencí vyhledávají konservované motivy, případně vytvářejí HMM profily. Dostupná je

rovněž řada programů predikujících potencionální místa posttranslačních modifikací jako jsou fosforylace, acylace apod. (nejznámější je Prosite, http://www.expasy.ch/prosite/), díky velmi vágně definovaným sekvenčním motivům je třeba výsledky predikce brát s rezervou.

Tab. 4. Zdroje pro detekci konzervovaných sekvenčních domén a motivů

Název	URL odkaz	Poznámka
SMART	http://smart.embl-heidelberg.de/	používá metodu HMM, manuálně editován, rychlý, přehledný, výborné grafické rozhraní
Pfam	http://www.sanger.ac.uk/Software/Pfam	používá metodu HMM, pomalejší ale komplexnější a někdy citlivější než SMART
InterPro (EBI)	http://www.ebi.ac.uk/interpro	integruje i data získaná pomocí SMART a Pfam
CDD (NCBI)	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	založeno na metodě PSSM, integruje i data získaná pomocí SMART a Pfam
Prosite (Expasy)	http://www.expasy.ch/prosite/	kromě domén předikuje i postranslační modifikace
Pratt (EBI)	http://www2.ebi.ac.uk/pratt/	vyhledá konzervované motivy ve skupině sekvencí
Pattinprot (NPS@)	http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_pattinprot.html	rozhraní k řadě programů

Analýza směrování a vnitrobuněčné lokalizace proteinů

Stejně jako detekce funkčních domén, je predikce intracelulární lokalizace proteinu nezbytnou součástí charakteristiky neznámé sekvence. Současné znalosti a prostředky dovolují poměrně spolehlivě predikovat tyto charakteristiky: signální peptidy, transmembránové domény a transitní peptidy směrující protein do mitochondrie nebo chloroplastu.

Standardem pro detekci signálních peptidů je program SignalP (http://www.cbs.dtu.dk/services/SignalP), který umožňuje predikci jak u eukaryot, tak u Gram-positivních i Gram-neativních bakterií s vysokou spolehlivostí.

Tab. 5 ukazuje některé metody používané pro predikci transmembránových segmentů. Ačkoli se jednotlivé algoritmy liší v detailech, základním principem většinou zůstává profil hydropathie podél sekvence proteinu (transmembránové domény jsou většinou α-helixy dlouhé 17 - 33 aminokyselin a obsahují hydrofobní a neutrální aminokyselinové zbytky). Sofistikovanější programy potom profil hydropathie kombinují s predikcí α-helikálních struktur. Autorova osobní zkušenost je taková, že kombinace 2 - 3 programů vede v 95% případů k správné predikci.

Podobně jako predikce signálních peptidů, je přítomnost transitních mitochondriálních/plastidových peptidů predikována programy založenými na bázi neuronových sítí, trénovaných na sadě proteinů s experimentálně ověřenou buněčnou lokalizací (viz tab. 6). Predikce mitochondriálních a plastidových signálních peptidů je však méně spolehlivá; podle autorových zkušeností je třeba kombinovat všechny dostupné algoritmy a i přesto se přesnost predikce pohybuje se jen asi okolo 80 %.

Tab. 5. Programy pro predikci transmembránových domén

Název	URL odkaz	Poznámka			
		Hranice TM	Graf hydrofobicity	Predikce α-helixů	Více sekvenci najednou
TMHMM	http://www.cbs.dtu.dk/services/TMHMM	ano	ano	ano	ano
TopPred2	http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html	ano	ano	ano	ano
PhDhtm	http://cubic.bioc.columbia.edu/pp/	ano	ne	ano	ne
DAS	http://www.sbc.su.se/~miklos/DAS/	ano	ano	ne	ne
TMpred	http://www.ch.embnet.org/software/TMPRED_form.html	ano	ano	ano	ne

Tab. 6. Programy pro predikci signálních peptidů a intracelulární lokalizace proteinů

Název	URL odkaz	Poznámka
iPSORT	http://hypothesiscreator.net/iPSORT/	predikce mitochondriálních, plastidových a signálních peptidů a místa štěpení
TargetP	http://www.cbs.dtu.dk/services/TargetP/	predikce mitochondriálních, plastidových a signálních peptidů a místa štěpení, možnost zadání více sekvencí, kvantifikace výstupu
ChloroP	http://www.cbs.dtu.dk/services/ChloroP/	predikce plastidových transitních peptidů
Predotar	http://www.inra.fr/predotar/	predikce mitochondriálních i plastidových TP, kvantitativní vyhodnocení
Mitoprot	http://mips.gsf.de/cgi-bin/proj/medgen/mitofilter	predikce mitochondriálních TP a místa štěpění, kvantitativní vyhodnocení

Predikce strukturních prvků v proteinech

Predikce strukturních prvků v proteinech často nemá přímou návaznost na predikci biologické funkce, ale je důležitým doplňkem analýzy primární sekvence. Jak bylo naznačeno výše, detekce oblastí s nízkou komplexitou je důležitým předpokladem pro hledání v databázích; predikce sekundárních struktur pak obvykle předchází modelování 3D struktury.

Detekce segmentů s nízkou komplexitou často signalizuje přítomnost domén neglobulárního charakteru, tzv. „*coiled-coil*“ domén. *Coiled-coil* domény mají superhelikální strukturu charakterizovanou přítomností periodicky se opakujících hydrofobních aminokyselin, často leucinů. Predikce založená na této periodicitě je základem programu COILS (http://www.ch.embnet.org/software/COILS_form.html).

Predikce sekundárních struktur sama o sobě dává málo informací o funkci proteinu či o jeho klasifikaci do proteinových rodin, je však důležitým doplňkem dalších metod. Evoluce proteinů často probíhá ve formě insercí neuspořádaných smyček, zatímco segmenty s charakteristickou sekundární strukturou jsou často konzervované. Spolehlivá

predikce těchto elementů pomáhá správnému srovnání vzdálených sekvencí a může odhalit konzervované oblasti, které by jinak unikly pozornosti. Detekce sekundárních struktur je rozvinuté odvětví bioinformatiky s řadou kompetujících algoritmů (tab. 7). Klasické metody jako Chou-Fasmanova nebo Garnierova zpracovávaly tehdy známé 3D struktury a počítaly pravděpodobnosti aminokyselinových zbytků tvořit nejpravděpodobnější sekundární strukturu pro daný úsek sekvence. Moderní metody (tab. 7) využívají neuronových sítí, které jsou trénované na databázích všech známých proteinových 3D struktur. Některé servery umožňují použít *multiple alignment* jako zadání či kombinují výsledky získané různými algoritmy, což významě zvyšuje pravděpodobnost správné predikce.

Tab. 7. Programy predikující sekundární struktury

Název	URL odkaz	Poznámka
Predict Protein	http://cubic.bioc.columbia.edu/predictprotein/	rozešle zadanou sekvenci dalším programům, které zpracují výsledky zvlášť
NPS@	http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html	rozešle zadanou sekvenci dalším programům a zpracuje konsensus
PHDsec	http://cubic.bioc.columbia.edu/predictprotein	asi nejpopulárnější program, část balíku Predict Protein
PSIpred	http://bioinf.cs.ucl.ac.uk/psipred/	se zadanou sekvencí provede tři iterace PSI-BLASTu a z *alignments* predikuje sekundární strukturu
Target99	http://www.cse.ucsc.edu/research/compbio/HMM-apps/	provede iterační hledání proti HMM profilům z proteinů se známými 3D strukturami
SSPAL, SSP, NNSSP	http://www.softberry.com/berry.phtml?topic=protein	jako zadání umožňují jednu sekvenci i *multiple alignment*
Jnet	http://www.compbio.dundee.ac.uk/Software/JNet/jnet.html	jako zadání umožňují *multiple alignment*, HMM profil a PSI-BLAST profil

Predikce a modelování 3D struktury proteinů

„*Sequence-structure threading*“, doslova „navlékání sekvence na strukturu“ zahrnuje rodinu informatických přístupů snažících se určit terciární strukturu zkoumaného proteinu na základě testování „vhodnosti“ různých typů experimentálně zjištěných 3D struktur. Základní myšlenka těchto přístupů tkví ve faktu, že struktura je mnohem více konzervována než sekvence. Obecně je principem většiny metod výpočet zbytkové kontaktní energie zadané sekvence „natažené“ na každou strukturu v databázi - struktura s nejmenší energií je potom prohlášena za nejpravděpodobnějšího kandidáta a dalším iteračním procesem se program snaží energii dále minimalizovat. Bylo navrženo několik statistických modelů, které testují pravděpodobnost, že navržená struktura odpovídá struktuře nativního proteinu. Dále je třeba podotknout, že výsledky získané kombinací „energetických“ metod predikce sekundárních struktur sekvenčních profilů jsou významně spolehlivější než data získaná pouze modelováním. Přehled některých z mnoha používaných metod je uveden v tab. 8.

Tab. 8. Programy pro predikci 3D struktury proteinů

Název	URL odkaz	Charakteristika
PDB	http://www.rcsb.org/pdb/	databáze experimentálně stanovených 3D struktur proteinů
SAM-T99	http://www.cse.ucsc.edu/research/compbio/HMM-apps/	provádí iterační hledání proti HMM databázi, z výsledků buduje nový HMM profil a pak prohledává databázi PDB..
InBGU	http://www.cs.bgu.ac.il/~bioinbgu/	porovnává sekvenční profily zadaného proteinu a proteinů s různými 3D strukturami, výsledná predikce je kombinací pěti různých metod
UCLA/DOE Fold Server	http://fold.doe-mbi.ucla.edu/	stejný postup jako InBGu, ale používá jinou knihovnu 3D struktur
GenThreader	http://bioinf.cs.ucl.ac.uk/psipred/	se zadanou sekvencí provede tři iterace PSI-BLASTu a s výsledným profilem buduje strukturu, nefunguje, pokud neexistují homology v databázi.
3D-PSSM	http://www.sbg.bio.ic.ac.uk/~3dpssm/	porovnává 1D a 3D profily spojené s predikcí sekundární struktury a solvatačního potenciálu

Tipy závěrem

Existuje mnoho formátů pro zpracovávání a uchovávání sekvencí. Autor doporučuje používat nejjednodušší a nejčastější z nich - FASTA formát. Ten je definován tak že začátek každé sekvenční položky začíná znakem '>', za nímž následuje jeden řádek informací o daném proteinu a od dalšího řádku vlastní sekvence. K převodu do dalších formátů lze s výhodou použít například program Readseq z nabídky serveru BCM Searchlauncher (http://searchlauncher.bcm.tmc.edu/seq-util/seq-util.html).

Je výhodné používat jako jeden z identifikátorů sekvence unikátní znak databází Genbank/EMBL/DDBJ.

Kromě zde uvedených programů existují akademické sféře volně dostupné programové balíky. Alternativou oblíbeného (nyní však komerčního balíku CGC package) může být soubor programů EMBOSS (http://www.emboss.org).

Literatura

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. 1997 - Nucleic Acids Res. **25**:3389.

Dayhoff, M.O., Eck, R.V. 1968. - In Dayhoff, M.O. Eck, R.V. (eds): Atlas of Protein Sequence and Structure **3**. S. 33. National Biomedical Research Foundation, Silver Spring, Maryland.

Dayhoff, M.O., Eck, R.V., Chang, M.A, Sochard, M.R. 1965. - Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Spring, Maryland.

Henikoff, S., Henikoff, J.G. 1992 - Proc. Natl. Acad. Sci. USA **89**:10915.

Hobohm, U., Sander, C. 1995. - J. Mol. Biol. **251**: 390.

Hughey, R., Krogh, A. 1996. - Comp. Appl. Biosci. **12**: 95.

Poděkování. Výzkum v laboratoři autora je podporován granty MŠMT LN00A081 a GAUK130/2002/B-BIO/Prf