evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

`

BIOCHEMICKÉ INFORMACE (2011/III)

Petr Skládal (Ústav biochemie PřF MU)

Výukový materiál - informační zdroje pro biochemii

- primární zdroje - publikované články, patenty
- sekundární zdroje - databáze obsahů časopisů nebo vydaných patentů (včetně klíčových slov a abstraktu)
- biochemické / biotechnologické databáze na internetu

stav odkazů aktualizován v březnu 2013

adresa aktuální verze tohoto souboru: http://biosensor.chemi.muni.cz/edu/bioinfo

Primární zdroje biochemických informací

Učebnice

tuzemské:

- Šípal Z: **Biochemie** (1992, SPN Praha) ... nalezl jsem v ISu
- Vodrážka Z: **Biochemie** (1996, 2002 Academia Praha)
- Racek J.: **Klinická biochemie** (1999, Galén Praha)
- Mikeš V.: **Základní pojmy z biochemie II** (CD ROM, MU Brno)

zahraniční:

klasické učebnice

- Voet D., Voet J.: **Biochemistry** (2011, Wiley; přeloženo - 1995 Victoria Publishing) na webu vydavatele je mnoho pomocných materiálů (náhled na Amazon)
- Stryer L.: **Biochemistry** (1995, Freeman) klasický text, další edice:
- Berg J.M., Stryer L., Tymoczko J.L.: **Biochemistry** (2007, Freeman) po registraci lze prohlížet na webu .
- D.E. Metzler: **Biochemistry. The Chemical Reactions of the Living Cells** (2001, Elsevier)
- Garrett R.H., Grisham C.M.: **Biochemistry** (2005, Thomson) nalezl jsem i přes google books .
- Nelson D.L., Coxx M.M.: **Lehninger Principles of Biochemistry** (2004, Freeman) po registraci lze prohlížet na webu .
- Devlin, T: **Textbook of Biochemistry with Clinical Correlations** (2001)
- Murray R.K., Granner D.K., Mayes P.A., Rodwell V.W.: **Harper's Illustrated Biochemistry** (2003, přeloženo)

zkrácené "rychlo"učebnice

- J. Koolman, K.H. Roehm: **Color Atlas of Biochemistry** (2005, Thieme)
- H.F. Gilbert: **Basic Concepts in Biochemistry: Studetn's Survival guide** (2000, McGraw Hill)
- J.T. Moore, R. Langley: **Biochemistry for Dummies** (2008, Wiley)

slovníky, encyklopedie

- A.D. Smith et al.: **Oxford Dictionary of Biochemistry and Molecular Biology** (2000, Oxford)
- J. Stenesh: **Dictionary of Biochemistry and Molecular Biology** (2000, Wiley) dostupné na webu Knovel
- : **Encyclopedia of Molecular Biology** (2005, 2010, Wiley) na MU dostupná i on-line verze

stačí zajít do knihkupectví, ev. Knihy.cz, Bohemia Starman, Amazon, ...

Ročenky, monografie

Methods in Enzymology (>390 svazků) - tématicky zaměřeno, nejen enzymologické, ale různé prověřené metodické postupy ze všech oblastí biochemie (proteiny, nukleové kyseliny, imunochemie, ...)

Annual Review of Biochemistry - každoročně, nejnovější objevy z posledního období

Archives of Biochemistry and Biophysics - obdobné

Časopisy

Přehledové a populárně-vědecké

Trends in Biochemical Sciences (TIBS) - vhodné pro doplnění poznatků při výuce, využíváno pro přípravu referátů na seminářích

Trends in Biotechnology

Nature případně **Science** (zaměřeny šířeji na přírodní vědy, ale velký podíl článků z oblasti biochemie). Přístup z počítačů PřF Nature resp. Science, vybrat odkaz "Přístup přes EZproxy (může se změnit...)"

Obecné

Biochemistry; Journal of Biological Chemistry (JBC); European Journal of Biochemistry (EJB); Proceedings of the National Academy of Sciences (PNAS); Biochemical Journal; Biochemical and Biophysical Research Communications (BBRC); Biochimica et Biophysica Acta (BBA) - existuje mnoho různých sekcí...

Speciální

Biotechnology and Bioengineering; Clinical Chemistry; Clinical Biochemistry; Enzyme and Microbial Technology; Analytical Biochemistry; Nucleic Acid Research; Molecular Immunology ... plus stovky dalších více či méně významných

Poznámky

- při publikování se názvy časopisů zkracují (standartní zkratky, např. J. Biol. Chem., Eur. J. Biochem., PNAS, ...)
- „kvalita“ časopisu je dána **impaktním faktorem** - je úměrný citovanosti prací publikovaných v daném časopise, základní informace lze najít na webové stránce příslušného časopisu
- pokud není časopis v knihovně nebo na internetu:
- je možné se stát předplatitelem (speciální nízké ceny pro individuální objednatele a studenty, často včetně on-line přístupu k několika posledním ročníkům ve formě PDF souborů (Acrobat Reader; jsou i rychlejší prohlížeče, např. FoxitReader)
- v zahraničí mají mnohem bohatší knihovny - využít stáží ke studiu literatury a pořizování kopií
- Internetový on-line přístup k časopisům (plné verse publikovaných článků) - rozsah a hloubka závisí na konkrétním

pracovišti (individuální smlouvy s poskytovateli)

Plné verze dostupných (buď zcela volné, nebo v rámci předplatného pro tištěnou verzi časopisu) článků publikovaných daným nakladatelstvím. Rozsah se liší rok od roku dle konkrétní uzavřené smlouvy (dostupnost finančních prostředků ...):

ScienceDirect - Elsevier

ACS - American Chemical Society

SpringerLink - Springer Verlag

Wiley - Wiley InterScience

Do jiných databází přístup na MU přes Ústřední knihovnu nebo přes El. informační zdroje MU, respektive knihovna chemických oborů v kampusu. V Brně - Moravská zemská knihovna, jiné knihovnické zdroje - WebLib, Státní technická knihovna, ...

Přístup z domu> na internetové zdroje přístupné (licencované) pouze pro fakultu - připojení přes proxy-server - návod, velmi pohodlně lze nastavit fakultní VPN připojení přes skript skript: **VPN-sci.muni.cz.pbk**, viz návod pro různé OS.

Příspěvky na konferencích, kongresech

- přednášky a plakátová sdělení (postery) dostupná pouze pro účastníky
- krátké souhrny přístupné ve formě sborníků abstrakt
- občas se objeví v rozšířenější formě jako speciální číslo nějakého časopisu
- nebo bývá vydáno i knižně

Patenty

Databáze patentových úřadů - často dostupné na internetu, některé i zdarma. Patenty jsou zejména v průmyslové oblasti důležitým zdrojem poznatků:

Úřad průmyslového vlastnictví - české CZ patenty (plus další chráněné materiály a vzory), EU celoevropské (lze i přes ESP@CEnet), WO celosvětové. Z národních patentových úřadů jsou nejdůležitější US americké, DE německé, JP japonské ...

Organizace

ČSBMB (Česká společnost pro biochemii a molekulární biologii)
ČSKB (Česká společnost pro klinickou biochemii)

FEBS (Federation of European Biochemical Societies) publikuje mimo jiné časopis FEBS Letters
EMBO (European Molecular Biology Organization)

ACS (American Chemical Society)

IUBMB (International Union of Biochemistry and Molecular Biology)
IUPAC (International Union of Pure and Applied Chemistry)

Sekundární zdroje

Web of Science (ISI Web of Knowledge) WOS

- V současné době asi hlavní vyhledávací nástroj používaný na PřF MU.
- součástí i Science Citation Index (SCI, Cited Reference Search) - sleduje, kde se citovaly publikované práce (jistý indikátor úrovně vědecké práce). Umožňuje netradiční literární rešerše: obvykle se hledá od současnosti směrem nazpět; SCI: najde se historická „průkopnická“ práce, zjistí se, které prameny ji citují

Chemical Abstracts CAS

nejobsáhlejší a klasický soubor informací z oblasti chemie: chemické a biochemické časopisy (včetně velmi exotických), patenty, knihy, sborníky konferencí; velmi kvalitní a také patřičně drahý informační zdroj

- **"papírová verze"**

dostupnost: vychází každý týden (kniha formátu A4 cca 1000 stran), 2 ročníky (volumes) za rok
organizace čísla:
textová část – články řazené dle sekcí, vždy název, zdroj, adresy autorů, abstrakt, ev. i vzorce, číslované v rámci ročníku - 1 až cca 400 000
indexová část - seznam autorů, klíčových slov, čísel patentů, sloučenin (formula index)
ke konci každého ročníku – indexy za celou řadu; občas vyjdou indexy za několik ročníků - **Collective Index** 12th: 1987-91; 13th: 1992-96

- knihovna Lachemy
- **CD ROM verze** - 1x měsíčně
- **online přístup** na Internetu (informace po roce 1970): buď přes webový prohlížeč (zatím nedostupné), nebo přes uživatelský interface - **SciFinder** (přístupový klient, existuje demo verze). K používání nyní dostupné na MU pro chemické ústavy v **knihovně kampusu na počítači číslo 35**

Current Contents CCC

(ISI, Philadelphia) Thomson Scientific)

- průběžné informace, vychází každý týden jako brožura, na disketách (vždy jedno číslo cca 5 MB, programy CC.EXE /DOS nebo CCWin.EXE /Windows), na CD-ROM (vždy uplynulý rok po současné datum), nebo on-line přístup; obsahy vycházejících časopisů, případně i s abstrakty

- členění na sekce:

Agriculture, Biology & Environmental Scieces; Social & Behavioral Sciences, Clinical Medicine, Ergineering, Technology & Applied Sciences,
Life Sciences - 1200 nebo 600 časopisů, Physical, Chemical & Earth Sciences

- prohlížení: podle oborů, klíčových slov, abecední seznamy, časopisy; výsledky: tisk žádanky o separát (dnes lépe přes e-mail autora), export do vlastní databáze separátů; ISI pošlou i reprint (platí se)
- vhodné pro každotýdenní sledování novinek, pro dlouhodobé rešerše – méně vhodné - mnoho manipulací s médii, pohodlná dostupnost: LIFE katedra biochemie (1993 - cca 2000, diskety, CD ROM), PCES - fakultní knihovna chemie (CD ROM OVID)

MedLine

obdobné jako Curr. Contents / Life Sciences, ale jiná forma; 2x ročně CD ROM (Lékařská fakulta MU), pohodlně (a

zdarma) dostupné on-line na PubMed (NCBI).

Scirus

- internetový vyhledávací nástroj specializovaný na vědecké informace - odfiltruje "nevědecké" zdroje
- "klasické" zdroje - časopisy, jiné databáze, webové stránky vědecky orientované (edu, org, ac.uk, gov, com)
- přístup přes **www.scirus.com** .

Stručně o bioinformatice

Bioinformatika používá informační systémy k analýze velkých biologických datových souborů - zejména sekvencí nukleových kyselin a bílkovin.

První úroveň může být definována jako návrh a použití metod pro sběr, organizování, třídění, uchovávání, zobrazování a analýzu biologických dat (genomy, transkriptosomy, proteomy, metabolomy, ...) či sekvencí (DNA, RNA, proteiny) a makromolekulárních struktur (případně i organizmů či ekologických systémů).

Další úrovní je odvozování znalostí o biochemických drahách, funci a interakcích genů (funkční genomika) a proteinů (proteomika) - biologická interpretace dat.

Pohled na funkci genů a proteinů probíhá prostřednictvím:

- Sekvenční analýza - studium sekvencí DNA a bílkovin, hledání spojení se strukturou, funkcí a kontrolními mechanismy
- Strukturní analýza - studium biologických struktur, hledání vazeb se sekvencí, funkcí a kontrolními mechanismy
- Funkční analýza - porozumění jak spojení sekvence a struktury vede k funkci

Obor bioinformatiky existuje na rozhraní biochemie, molekulární biologie, matematické biologie, klinické medicíny, sekvenční analýzy, databázových systémů a internetu.

Nejdůležitější odkazy:

European Bioinformatics Institute EBI
Bioinforatics Organization
The Institute for Genomic Research TIGR
Bioinformatics homepage Bioplanet
Bioinformatik.de
BioinfoMatix

Základní témata

Kódování aminokyselin

Tabulka ukazuje 20 aminokyselin, z nichž se sestávají proteiny a kodóny pro každou z nich

Ala	A	GCU, GCC, GCA, GCG	**Leu**	L	UUA, UUG, CUU, CUC, CUA, CUG
Arg	R	CGU, CGC, CGA, CGG, AGA, AGG	**Lys**	K	AAA, AAG
Asn	N	AAU, AAC	**Met**	M	AUG
Asp	D	GAU, GAC	**Phe**	F	UUU, UUC
Cys	C	UGU, UGC	**Pro**	P	CCU, CCC, CCA, CCG
		CAA, CAG			UCU, UCC, UCA, UCG, AGU,AGC

Gln	Q		Ser	S	
Glu	E	GAA, GAG	Thr	T	ACU, ACC, ACA, ACG
Gly	G	GGU, GGC, GGA, GGG	Trp	W	UGG
His	H	CAU, CAC	Tyr	Y	UAU, UAC
Ile	I	AUU, AUC, AUA	Val	V	GUU, GUC, GUA, GUG
Start		AUG, GUG	*Stop*		UAG, UGA, UAA

Genetický kód je degenerovaný, resp. redundantní, což znamená, že dva či více kodónů může kódovat jednu a tutéž aminokyselinu. Degenerované kodóny se obvykle liší ve své třetí pozici, viz kodóny GAA a GAG, které oba kódují glutamin. (tiché mutace)

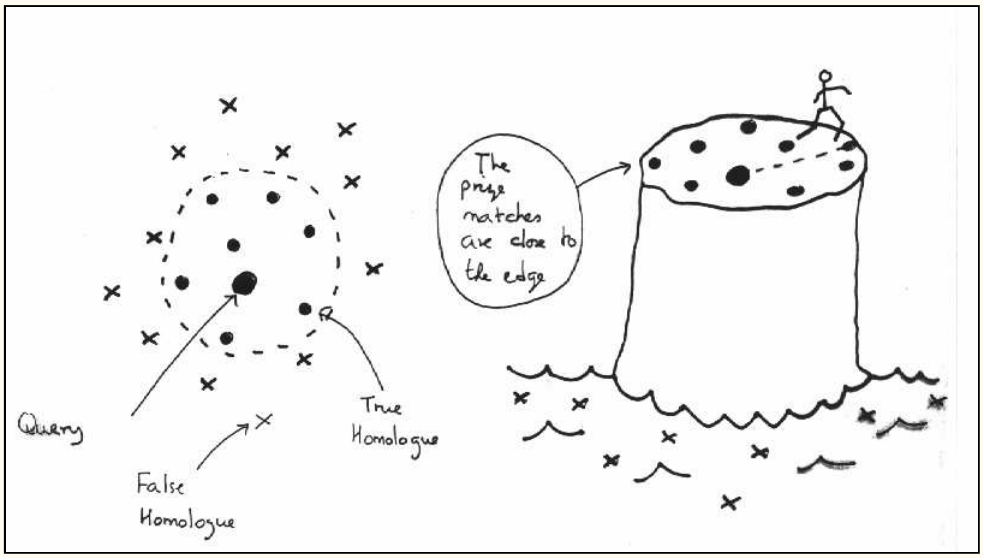
Porovnávání sekvencí (alignment)

cíl: pro dvě dané sekvence a hodnotící schéma nalézt optimální párování písmen:

RKVA--GMAKPNM
RKIAVAAASKPAV

účel:

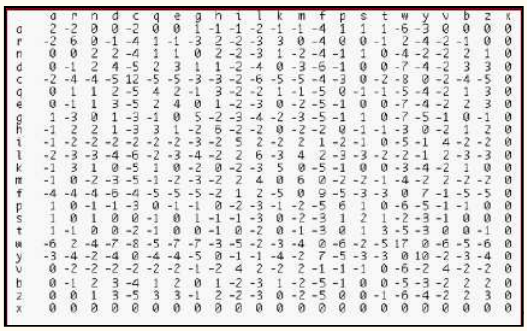
- existuje pouze málo sekvencí se známou strukturou a funkcí, u většiny sekvencí nejsou vlastnosti známé
- pokud jsou dvě sekvence podobné a u jedné z nich známe funkční / strukturní vlastnosti, poskytne srovnání další nové informace pro oba případy
- podobnost může sloužit jako důkaz homologie, avšak nezbytně nemusí znamenat homologii



Záleží na úhlu pohledu ... jak dopadne konecné hodnocení vyhledaných podobných sekvencí.

druhy:

- lokální / globální
- párové / multicetné



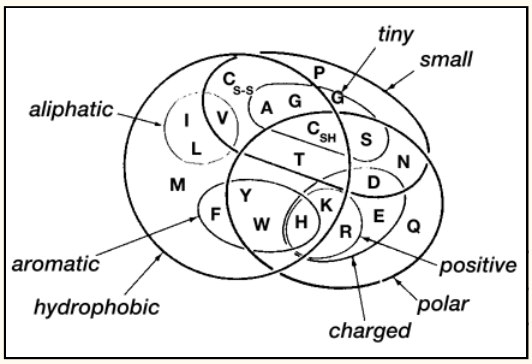
Score matrix

"měření" kvality překryvů - pomocí srovnávacích matic, vycházejících např. z vlastností aminokyselin (malé / velké, hydrofobní / polární, aromatické / alifatické, náboj, ...) či jiných elementů, "síla" podobnosti - procentuální identita

překryvové algoritmy

bodové matice

- dynamické programování (FASTA, BLAST, Psi-BLAST, Clustal)



Klastry (skupiny) podobných aminokyselin

Sekvenční motivy

motivy - krátké sekvenční úseky (subsekvence), které se vyskytují v mnoha sekvencích a mají určitý biologický význam:

motivy bílkovin často reprezentují strukturní rysy
DNA motivy poskytují signál pro vazbu bílkovin nebo vznik záhybů
v databázo PROSITE je kolekce více než tisíce motivů - manuálně vytvořený soubor spojený s různými proteinovými rodinami nebo funkcemi,
např. globin sequence signature (PDOC00933):

F-[LF]-x(5)-G-[PA]-x(4)-G-[KRA]-x-[LIVM]-x(3)-H

Hledání genů

cílem je identifikovat jednotlivé geny v rámci hrubé sekvence genomové DNA (vstupní informace) - přesné umístění elementů tvořících daný gen (exony, introny, jiné sekvenční anotace) ve studované sekvenci DNA

- relativně jednoduché u bakterií - DNA - mRNA - protein
- složité u vyšších organismů - DNA (exony a introny) - prekurzorová mRNA - RNA splicing (vyštěpení intronů) - mRNA - protein

Vyhledávací nástroje

BLAST

(The **Basic Local Alignment Search Tool**) Nástroj (program) BLAST vyhledává úseky s lokálními podobnostmi mezi srovnávanými sekvencemi. Zadanou sekvenci (aminokyseliny u peptidu, nukleotidy nukleové kyseliny) porovnává s údaji v sekvenčních databázích a počítá statistickou významnost nalezených výsledků ("matches"). Může sloužit k vyvozování funkčních a evolučních závislostí mezi sekvencemi, pomáhá nalézat členy genových rodin.

Hlavním přínosem je zrychlení vyhledávací procedury na rozdíl od kompletního překryvového porovnávání dvou sekvencí. Nejprve se mezi porovnávanými sekvencemi nalézají kratší podobnosti ("seeding") na základě vytvořených "words" a z nalezených souborů se pak vybírají ty "nejlepší" (viz wiki).

Základní BLAST

- **nucleotide blast:** prohledává nukleotidovou DB na výskyt nukleotidové sekvence, algoritmy: blastn, megablast, discontiguous megablast.
- **protein** blast: prohledává proteinovou DB na výskyt proteinové sekvence, algoritmy blastp, psi-blast, phi-blast.
- **blastx:** hledá v proteinové DB na základě přeložené nukleotidové sekvence.
- **blastn:** hledá v nukleotidové DB na základě přeložené proteinové sekvence
- **tblastx:** hledá v DB přeložených nukleotidových sekvencí pomocí dotazu na přeloženou nukleotidovou sekvenci

Specializovaný BLAST

Podle zadaných kriterií omezuje oblast a rozsah hledání, což poskytuje mnohem specifičtější a relevantnější výsledky.

FASTA

(FASTA = FAST-All, FAST-P, FAST-N) Implementuje Smith-Waterma prohledávací algoritmus - lokální porovnávání (viz wiki).

FASTA formát - textový způsob reprezentace oligonukleotidových nebo proteinových sekvencí pomocí jednopísmenných kódů:

;LCBO - Prolactin precursor - Bovine ; a sample sequence in FASTA format
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLSS
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*

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

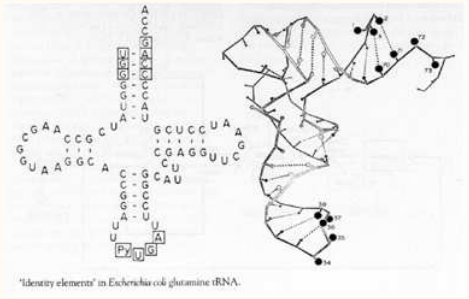
;pokusna sekvence oligonukleotidu
GTTCGGCGATGGCCGATGAGGTCGTCGCCGAGATTCGCGACAAGGGGGGCCGGGCGGTCGCCAACTACGACAGCG

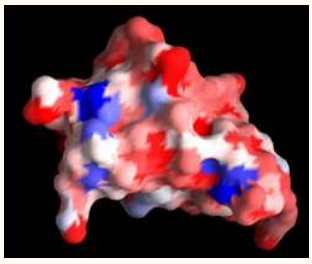
;pokusna sekvence peptidu TWDNGKPIRETSAADVPLAIDHFRYFASCIRAQEGGISEVDSETVAYH

Počítání s biologickými strukturami

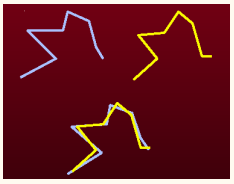
všeobecné úkoly:

- jak reprezentovat danou strukturu pro účel výpočtů
- jak porovnávat struktury
- jak sumarizovat strukturní rodiny



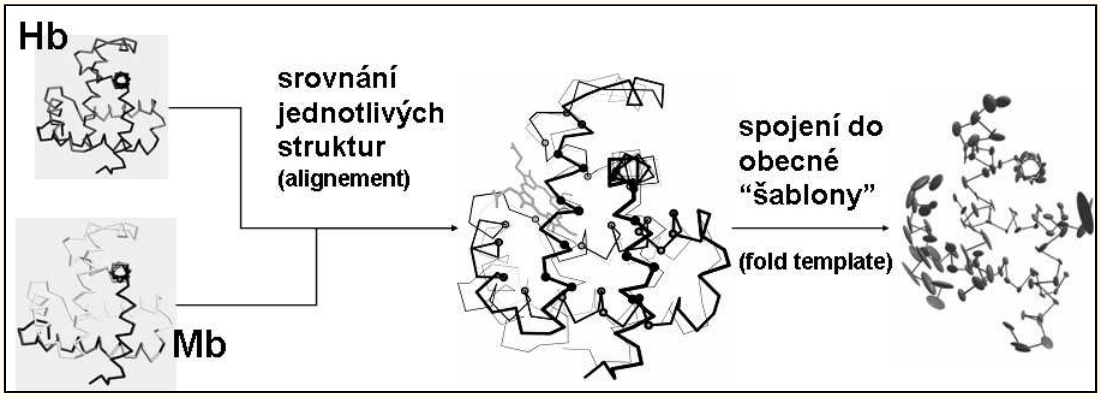


Možnosti reprezentace t-RNA



Prekrývání struktur

aplikace: porovnávání struktur - identifikovat pomocí překrývání struktur šablony různých ohybů (fold templates), budování knihoven strukturních elementů (fold libraries)



Porovnání struktur hemoglobinu a myoglobinu a nalezení spolecných strukturních elementu

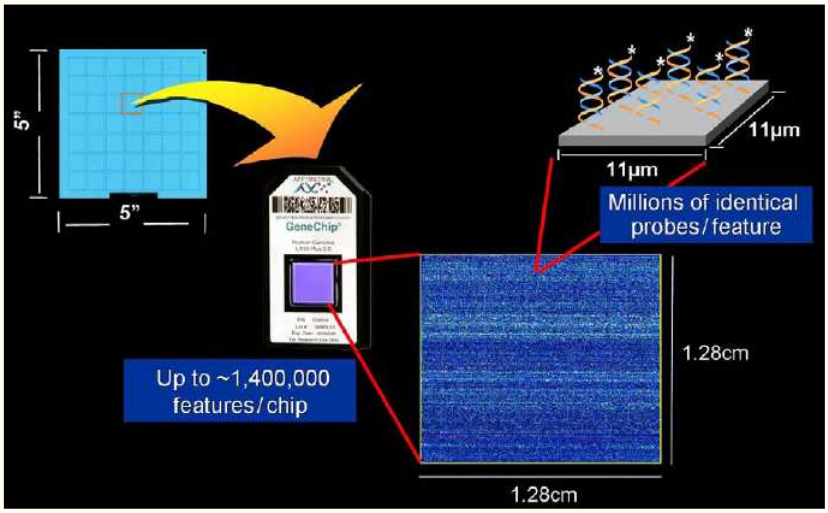
- porovnání struktur slouží jako "zlatý standard" pro porovnání sekvencí
- pro nehomologní proteiny je třeba identifikovat společné substrukturní elementy
- klasifikace bílkovin do klastrů na základě strukturní podobnosti (SCOP)
- predikce sekundární struktury RNA (program MFOLD)
- predikce sekundární struktury bílkovin (neuronové sítě)

Fylogenetické algoritmy

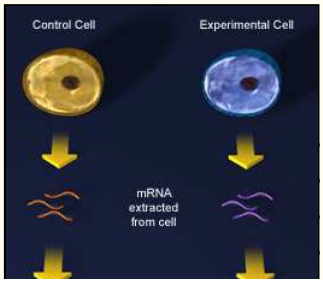
Proč vytvářet evoluční stromy:

- porozumět rodokmenům různých druhů
- vytvořit organizační princip pro taxonomické třídění druhů
- porozumět evoluci různých životních funkcí
- porozumět evolučním tlakům a omezením
- provádět mnohočetné překryvové srovnávání - u pokročilých metod probíhá současně s analýzou vytváření evolučních stromů;
- sekvenční porovnávání poskytuje kvantitativní údaje (scores), které mohou být považovány za nepřímo úměrné evoluční vzdálenosti srovnávaných druhů
- evoluční vzdálenosti pak slouží k tvorbě stromů, které poskytují multičetné překryvy prostřednictvím sdílených rodičů

Analýza dat z DNA biočipů



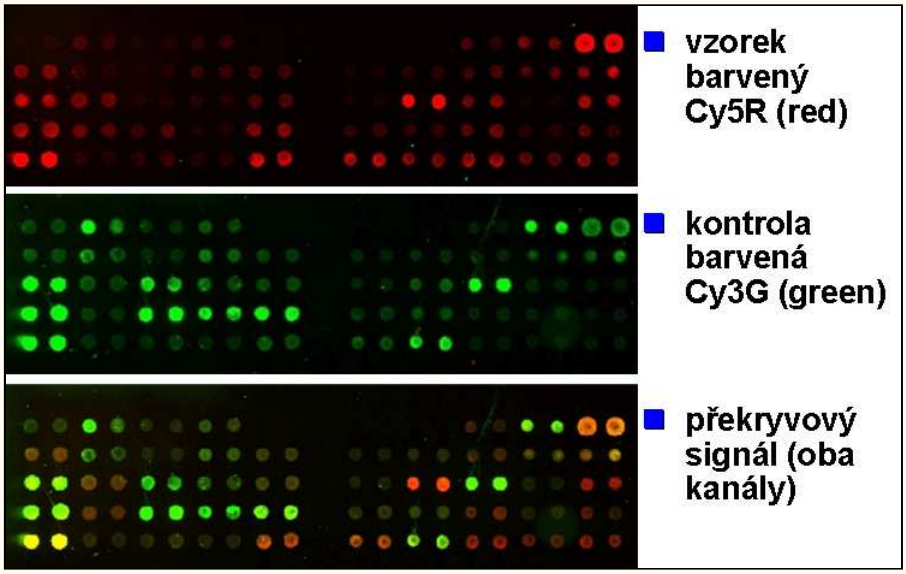
DNA čip (GeneChip, DNA microarray) je sensor nesoucí velký počet (100 až 10^6) oligonukleotidových prób o známých sekvencích; nechá se hybridizovat se vzorkem analyzované nukleové kyseliny vhodně označeným(případně současně s kontrolou - označenou jinou barvou) a dle zbarvení v místech prób se usuzuje na výskyt komplementární sekvence ve vzorku.



Obvykle se provádí tzv. expresní analýza - jak jsou které geny exprimovány:

v průběhu života populace (synchronizovaných) buněk
při reakci na externí podněty (léčiva, toxické látky, ...)
v případě patologických změn (např. rakovinné bujení)
prostřednictvím cDNA (complementary DNA - získá se z exprimované mRNA přepisem

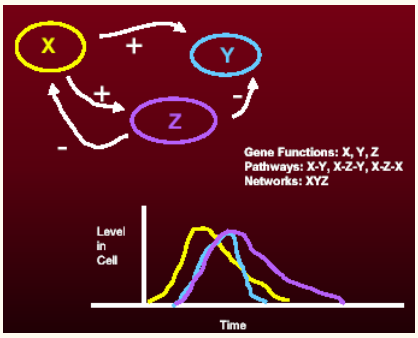
pomocí reverzní transkriptasy), která se transformuje do formy "knihovny" - cca 10^5 úseků 200-400 bp - EST (express sequence tags)



Výhodou je možnost sledovat současně mnoho různých genů. Sdružování genů do skupin (clustering):

- pokud jsou dva geny exprimovány stejným způsobem, mohou být funkčně příbuzné
- pokud má gen neznámou funkci, ale je v klastru s genem známé funkce, lze takto usuzovat na jeho funkci
- je možné vyvodit, jak se geny ovlivňují nebo kontrolují navzájem

Genové sítě



Ukázka genové síte ze trí genových elementu

(genetic networks) individuální geny mají funkci (např. konverze substrátů, vazba biomolekul) a soubory takových funkcí v průběhu sekvenace mohou vést k metabolickým drahám (produkt jedné transformace je substrátem pro druhou) a soubory metabolických drah pak vytváří genovou síť interakcí

Rekonstrukce genových regulačních sítí je náročný problém, pro N genů je možné exponenciální množství spojení, vzájemné interakci navíc nejsou jednoznačné (+/-) ale mění se kontinuálním způsobem. Počet možných interakcí genů se omezuje prostřednictvím znalostí o účasti v metabolických drahách a účasti v individuálních genových sítích.

Srovnávací genomika

Porovnává genomy ve velkém rozsahu za účelem porozumět biologické podstatě, extrahovat obecné principy platné pro skupiny genomů. Předpokládá se, že mnohé biologické sekvence, struktury a funkce jsou sdílené mezi organismy, kombinace genomů při analýze pak může vést k přesnějším výsledkům. Další úkoly:

- porovnávání velmi dlouhých sekvencí
- srovnávací přístupy k hledání genů a přiřazování jejich funkcí
- srovnávací přístupy při identifikaci klíčových regulačních oblastí

Proteomika

Proteom / proteomics - přípona -OMICS naznačuje v poslední době studium určitého jevu v komplexním pohledu na celý soubor, který ho zahrnuje

- proteomika - studium všech proteinů, které vzájemnou interakcí určují charakter buňky
- genomika - studium všech genů (chromozomální genom - genetická informace společná pro každou buňku organizmu)
- transkriptomika - studium všech mRNA transkriptů (exprimovaný genom - v dané buňce v daném vývojovém stadiu)
- metabolomika - studium všech metabolitů v buňce

Řešené problémy:

- které bílkoviny jsou genomem vytvářeny
- jaká je jejich struktura (3D)
- kde se nacházejí a jaká je jejich úloha
- s jakými dalšími bílkovinami interagují
- jsou v buňce posttranslačně modifikovány

Klíčové technologie a metody:

- stanovení prostorové struktury (3D - X-ray, NMR)
- dvojdimenzionální gelová elektroforéza posuzující všechny proteiny v buňce
- hmotnostní spektrometrie identifikující bílkoviny a jejich modifikace
- proteinové biočipy pro charakterizaci všech buněčných bílkovin pomocí protilátek nebo jiných rekogničních technik

Biologická ontologie

Pro efektivní komunikaci je zapotřebí společný jazyk a základní znalosti. Např. u metabolických drah jsou "jazykem" názvy produktů, enzymů a substrátů, "znalosti" pak zahrnují pojmy co to je reakce, jak se jí účastní enzymy a substráty, co jsou přijatelné složky dráhy

Geneová ontologie (www.geneontology.org) klasifikuje genové funkce, seznam tří hlavních typů funkcí: molekulární funkce, biologické procesy, buněčné komponenty

Dlouhodobé cíle

Matematický model fyziologie
- lze podat lék počítačovému modelu před tím, než ho podáme živým jedincům?

Návrh nových sloučenin pro lékařské a průmyslové využití
- lze navrhnout bílkovinu nebo nukleovou kyselinu se specifikovanou funkcí?

Vytváření nových biologických drah
- můžeme navrhnout metody pro vytváření a realizování nových metabolických schopností pro léčení nemocí?

Hledání nových poznatků (data mining)
- pomocí dotazů počítačovému programu zkoumat data v kontextu našich modelů a vytvářet tak nové znalosti?

Biochemické databáze

Slouží těmto hlavním účelům:

- hledání - je znám gen pro můj protein? - je známa mutace působící toto onemocnění?
- srovnávání - jsou známy sekvence podobné mé bílkovině? - jsou tyto dvě sekvence podobné (jak moc)?
- předpovídání – lze předpovědět aktivní místo tohoto enzymu? - lze zkonstruovat 3D model proteinu?

Odpovědi nemusí být nezbytně nalezeny pouze v jediné databázi - potřeba provádět kombinované hledání a integrovat nalezené výsledky - vzájemně kooperující databáze

---

NCBI National Center for Biotechnology Information

Národní centrum pro molekulárně-biologické informace v USA, existuje od roku 1988. NCBI vytváří veřejné databáze, zabývá se výzkumem v informatické biologii, vyvíjí programy pro analýzu genomu a šíří biomedicínské informace, vše za účelem lepšího pochopení molekulárních procesů ovlivňujících lidské zdraví a nemoci. Přehled tohoto zdroje informací je podán relativně detailně, účelem je demonstrovat široké možnosti a variabilitu dostupných informací. Struktura NCBI webu je na bázi typu požadovaných informací:

Chemicals & Bioassays ... chemikálie a biostanovení

- **Biosystems** ... DB sekupující literární odkazy, malé molekuly a sekvence podle biologických vazeb
- PubChem **BioAssay** ... data se vztahem k bioaktivitě a ke stanovením bioaktivity pro sloučeniny obsažené v DB PubChem Substance
- PubChem **Compound** ... unikátní validované chem. struktury malých molekul včetně odkazů na další DB
- PubChem**Substance** ... data o sloučeninách zadaná uživateli, včetně komentářů a odkazů na web "vkladatelů"

Data & Software ... programové vybavení

- ... nejrůznější typy programů pro stažení na lokální počítače (různé platformy) a přistupování k DB v rámci NCBI, např. BLAST (hledání překryvů sekvencí), Sequin (předávání informací), CN3D (prohlížení 3D struktur), odkazy do FTP depozitářů

DNA & RNA ... nukleové kyseliny

- **GenBank** ... databáze sekvencí oligonukleotidů s anotacemi, přístupné široké veřejnosti. Soubor sekvenčních genomových dat získaných ze sekvenačních projektů po celém světě (DDBJ, EMBL), každodenní aktualizace. Zadání sekvence vede k zobrazení jejího výskytu v řadě dalších typů DB v rámci NCBI. Základní zdroj bioinformatického výzkumu.
- **Nucleotide Database** ... kolekce sekvencí z různých zdrojů včetně RefSeq, GenBank, Third Party Annotation, PDB.
- **RefSeq** ... sekvence neopakujících se úseků genomové DNA, RNA transkriptů a odpovídajících proteinových sekvencí. Stabilní zdroj s odkazy na genom, identifikaci genů, hledání mutací a polymorfismů, expresní a srovnávací studie.
- **Trace Archive** ... data ze sekvenátorů a z různých sekvenačních projektů, hledání pomocí strukturovaných dotazů.
- **UniGene** ... DB transkriptů včetně informací o podobnostech proteinů a genetickém umístění.

Domains & Structures

- **Conserved Domain Database** (CDD) ... může být použit k identifikaci konzervativních domén v sekvencích bílkovin, zachovávaných v průběhu evoluce.
- **Structure (Molecular Modeling Database)** ... obsahuje makromolekulární 3D struktury odvozené z PDB, jakož i nástroje pro jejich prohlížení (program Cn3D)a porovnávání.
- Structure Přímý přístup ke strukturním nástrojům, nástin možností

Genes & Expression

- **Database of Genotypes and Phenotypes (dbGaP)** ... DB archivuje výsledky studií vztahů mezi genotypem a fenotypem, tj. genem a jeho projevy.

- **Gene** ... DB genů, s důrazem na plně osekvencované genomy.
- **Online Mendelian Inheritance in Man (OMIM)** ... lidské geny a genetické poruchy.

Genetics & Medicine

- **Database of Genotypes and Phenotypes (dbGaP)** ... vztahy genotypu a fenotypu s medicínskými aspekty.

Genomes & Maps

- **Database of Genomic Structural Variation (dbVar)** ... studie genomických změn, rozsáhlé inzerce, delece, translokace a inverze.
- **Genome** ... sekvence a mapy pro cca 1000 kompletně osekvenovaných organizmů, tak částečně sekvenovaných. Zahrnují mimo jiné: Bakterie - grafické representace kompletního bakteriálního genomu, zobrazení buď komplexní nebo detailní s odkazy na sekvenční data. Banánová muška (*Drosophila melanogaster*) - grafické znázornění všech chromosomů, možnost hledat cytogenetická i sekvenční data pro celý genom. Člověk - přehled dostupných zdrojů lidského genomu, včetně průběžných zpráv z Human Genome Project. Parasit malárie - data a informace se vztahem ke genetice a genomice malárie. Myš, krysa- soubor informací se vztahem k myším / krysím zdrojům, sekvence, mapování, klony, odkazy na různé kmeny a mutace. Nematoda - sekvenční data *Caenorhabditis elegans*. Genomy pro různé rostliny - grafické reprezentace chromosomů z různých genomů. Eukaryotické organely - přehled organel, popis referenčních sekvencí, odkazy na kompletně sekvenované, seřazeno taxonomicky a abecedně dle organismu. Další genomy - retroviry, kvasinky, plasmidy, viroidy.

Homology

- ... pohled na databáze z úhlu výskytu a studia homologií.

Literature ... Databáze literatury

PubMed - služba organizace National Library of Medicine, která poskytuje přístup k více než 12 mil. citací z databáze MEDLINE a z dalších časopisů; včetně odkazů na kompletní články, pokud jsou volně dostupné.

PubMed Central - digitální archiv časopisových informací z oblasti věd o životě, jehož prostřednictvím NCBI zachovává volný přístup k elektronické literatuře.

Bookshelf - elektronická knihovna příruček a učebnic konvertovaných do elektronické podoby.

OMIM - katalog lidských genů a genetických poruch.

PROW (Protein Reviews on the Web) - mezinárodní zdroj informací o lidských bílkovinách. Systém tvořený PROW Guides, což jsou autoritativní a strukturované přehledy o proteinech and proteinových rodinách, členěno na cca 20 standardtních kategorií informací (abstrakt, biochemická funkce, ligandy, odkazy, aj.).

Proteins

- **Protein Database** ... proteinové sekvence z různých zdrojů - GenPept, RefSeq, Swiss-Prot, PIR, PRF, a PDB

Sequence Analysis

- **Nástroje pro analýzu sekvencí (porovnávání sekvencí - BLAST), návrhy primerů pro PCR metody (Primer-Blast).**

Taxonomy

- **Obsahuje názvy a fylogenetické vazby pro organismy (přes 160 tis.), které v NCBI mají nějaké popsané biomolekuly. Pro daný organismus podává souhrn odkazů v různých DB.**
-

Training & Tutorials

- Návody, manuály, příklady, referenční příručky, FAQs, ...

-

Variation

- Databáze a přehledy pozměněných variant normálních sekvencí, struktur, ...
-

Nástroje pro vyhledávání a předávání dat

Hledání na bázi textových údajů: Entrez.
LinkOut - registrační služba pro tvorbu odkazů z článků, časopisů nebo biologických dat Entrezu na externí webové stránky.

Cubby - umožňuje uživatelům Entrez ukládat a aktualizovat vyhledávání a zobrazování výsledků.

Citation Matcher - umožňuje nalézt PubMed ID nebo MEDLINE UID článků z databáze PubMed na základě bibliografických informací.

Vyhledávání podobných sekvencí: BLAST Home Page (Basic Local Alignment Search Tool) - programy, přehledy, nápovědy, dokumentace a FAQ. BLink - zobrazuje výsledky hledání pomocí BLAST pro sekvence bílkovin z Entrez Protein databáze. Network BLAST - TCP/IP klient-server verse Entrez. Přímé spojení s NCBI databázemi přes internet. BLAST je k dispozici i pro lokální použití.

Taxonomické vyhledávání: Taxonomy Browser - nástroj pro hledání v NCBI taxonomických databázích. Taxonomy BLAST - seskupuje výsledky BLASTu na základě zdrojových organismů. TaxTable - shrnuje BLAST taxonomická data a zobrazuje vzájemnou příbuznost organizmů pomocí barevně kódovaných grafů. ProtTable - poskytuje souhrn oblastí genomu kódujících bílkoviny. TaxPlot - poskytuje různé pohledy na genomové podobnosti.

Předávání nalezených sekvencí: Sequin - nástroj pro předávání dat, obsahuje modul ORF Finder, zobrazovač / editor překrývajících se úseků. BankIt - WWW předávací nástroj pro jednoduché sekvence.

Nástroje pro 3D zobrazování a porovnávání

Srovnávací analýza makromolekul a 3-dimensionálních struktur:

Cn3D - překryvové porovnávání pro 3-dimensionální struktury a sekvence.

Conserved Domain Architecture Retrieval Tool - zobrazuje funkční domény tvořící bílkovinu a podává přehled bílkovin s podobnou doménovou architekturou.

VAST Search - služba pro vyhledávání strukturních podobností, porovnává nově zjištěné 3D koordináty struktury bílkovin s obdobnými údaji z MMDB/PDB databáze.

Threading - algoritmus pro rozpoznávání struktury bílkovin (protein folding).

Nástroje pro sekvenční analýzy

COGs (Clusters of Orthologous Groups) - systém genových rodin z kompletních genomů.

COGnitor - program k porovnávání uživatelských sekvencí s COGs databází za účelem identifikace orthologních skupin, ke kterým náleží.

GEO (Gene Expression Omnibus) - zdroj dat genové exprese s dostupnými zdroji z různých organismů i umělých zdrojů.

HomoloGene - porovnává nukleotidové sekvence mezi páry organismů pro nalezení putujících podobností (putative orthologs).

Conserved Domain Database - souhrn sekvenčních překryvů a profilů reprezentujících bílkovinné domény zachované v průběhu evoluce.

LocusLink - poskytuje jednoduchý dotazovací systém pro práci s názvy genů, genovými místy a LocusID čísly.

MGC (Mammalian Gene Collection) - zdroje komplementárních cDNA sekvencí plné délky.

Clone Registry - databáze účastnických center sekvenujících lidský a myší genom, pro vzájemnou informovanost o zpracovávaných úsecích.

Trace Archive - podobné jako CloneRegistry, uchovává sekvenční data v hrubém primárním stavu.

ORF Finder - grafický analytický nástroj pro hledání otevřených čtecích rámců dané minimální velikosti v uživatelem zadané či databázové sekvenci.

VecScreen - nástroj pro identifikaci úseků nukleotidových sekvencí, které mohou pocházet z vektoru, linkeru apod, před zařazením do databází.

e-PCR - pro porovnávání sekvencí se zmapovanými označenými úseky.

Mapy

Přístup k různým genetickým a fyzikálním mapám.

MapViewer - poskytuje integrující pohledy na chromosomální mapy, překrývající se úseky.

ModelMaker - umožňuje zkonstruovat MRNA sekvence z genomových dat, vybírá introny na bázi překryvů mRNA a EST, edituje vzniklé kombinace, testuje otevřené čtecí rámce a ukládá data; dostupný i v rámci MapVieweru jako mm odkazy.

OMIM Gene Map - cytogenetické umístění genů uvedených v literatuře a určených různými mapovacími metodami.

OMIM Morbid Map - abecední seznam nemocí a odpovídajících umístění na genetických mapách.

Human-Mouse Homology Maps - tabulka porovnávající homologní úseky DNA.

Další genetické mapy: krysa, zebrafish, moskyt, nematoda, Drosophila.

GeneMap'99 - fyzikální mapa více než 35 tis. markerů lidských genů, zkonstruováno organizací International Radiation Hybrid Mapping Consortium.

Výzkum zhoubného bujení

Řada projektů ve spoluprácí s National Cancer Institute (NCI) zaměřených na výzkum zhoubného bujení.
SKY/CGH (Spectral Karyotyping SKY and Comparative Genomic Hybridization CGH Database) - uložení veřejně předaných SKY a CGH údajů.
CCAP (Cancer Chromosome Aberration Project) - definice a detailní charakteristiky vybraných chromosomálních změn přiřazených k maligním transformacím.
CGAP (Cancer Genome Anatomy Project) - interdisciplinární program k identifikaci lidských genů exprimovaných při různých stavech rakovinného bujení.
Mitelman Database of Chromosome Aberrations in Cancer - genomová mapa chromozomálních zlomů při lidské rakovině.

SAGE Analysis - diferenciální exprese SAGE tagů v rakovinných knihovnách.
SAGEmap (Serial Analysis of Gene Expression) - experimentální technika kvantitativního rozsahu genové exprese.

Lidský genom

Kdokoliv s počítačem a internetovým připojením se může podílet na výzkumu lidského genomu.

DBGET

Spadá pod japonský server GenomeNet, DDBJ (DNA Data Bank of Japan). Je to jednoduchý databázový vyhledávací systém pro molekulárně-biologická data, databáze je považována za souborový systém, kde každé položce (charakterizované unikátním identifikátorem) odpovídá jeden či více souborů. Rozsah typů souborů zahrnuje jak textové, tak grafické formáty. Tak je možné přistupovat k nejrůznějším databázím po celém světě stejným způsobem: **dbname:identifier**. Genové katalogy systému KEGG jsou zpracovávány podobným způsobem: **organism:gene**. Databáze obsahují mimo jiné křížové odkazy, takže vytváří vlastní webovou strukturu dat a odkazů na data; DBGET obsahuje tuto strukturu uvnitř LinkDB databáze. DBGET má tři základní příkazy (nebo mody pro webovou verzi):

bget provádí stažení databázových položek specifikovaných kombinací **dbname:identifier**.

bfind je používán pro hledání pomocí klíčových slov.

blink pak provádí stahování podobných položek z dalších databází.

ExPASy Expert Protein Analysis System

Server organizace Swiss Institute of Bioinformatics (SIB) zaměřený na proteomiku - analýza sekvencí a struktur bílkovin, 2-D PAGE. Obhospodařuje následující databáze:

SWISS-PROT a TrEMBL - informace o bílkovinách

PROSITE - proteinové rodiny a domény. Databáze napomáhá spolehlivě odhalit, ke které proteinové rodině (pokud vůbec) náleží nově nalezená sekvence aminokyselin. V současnosti obsahuje a podrobně popisuje více než tisíc různých domén.

SWISS-2DPAGE - dvoudimenzionální elektroforéza v polyakrylamidovém gelu

ENZYME - enzymová nomenklatura

SWISS-3DIMAGE - prostorové modely bílkovin a jiných biomakromolekul

SWISS-MODEL Repository - automaticky generované modely bílkovin

CD40Lbase - CD40 ligandové defekty

SeqAnalRef - bibliografické reference zaměřené na sekvenační analýzu

K dispozici jsou také nejrůznější programové nástroje:

Zaměřené na proteomiku a analýzu sekvencí: proteomika [PeptIdent, PeptideMass, ...], exprese DNA -> Protein [Translate], hledání podobností [BLAST], hledání profilů [ScanProsite], post-translační modifikace a topologické předpovědi primární struktury [ProtParam, pI/MW, ProtScale], návrh sekundární a terciární struktury [SWISS-MODEL, Swiss-PdbViewer], překrývání sekvencí [T-COFFEE, SIM], biologická textová analýza

Melanie 3 - software pro 2-D PAGE vyhodnocování

Roche Applied Science's Biochemical Pathways - komplexní schematický grafický pohled na metabolické dráhy

SRS

Server spravovaný organizací European Biotechnological Institute (Velká Británie), funguje jako vstup pro ~ 80 bází sekvencí, metabolických drah, transkripčních faktorů, mutací aj. Přístup do EMBL (European Molecular Biology

Laboratory) - primární data sekvence proteinů. Napojení na další databáze:

- nukleotidy (EMBL Nucleotides)
- proteiny - Uniprot
- literatura - Medline
- genom - Entrez gene
- mutace - OMIM
- metabolické dráhy - Kegg

PDB Protein Data Bank

Zahrnuje struktury biomakromolekul. Poskytuje nástroje a zdroje pro studium 3D struktury a její vztah k sekvenci, funkci a onemocněním. Pro 3D zobrazení je třeba nainstalovat vhodný software, který dokáže vstupní PDB data převést do grafického znázornění:

- KiNG
- Jmol
- WebMol
- QuickPDB
- Protein Workshop

případně lze nainstalovat doplnění internetového prohlížeče:

- Rasmol
- Swiss PDB Viewer

různé možnosti přístupu do databází: přes identifikátor PDB_ID, přes web, FTP, db dotazy (query).

Formáty dat: PDB, PDBML/XML, mmCIF (macromolecular crystalography information file)

PIR různé informace o proteinech

PIR (protein information resources)

PIRSF (structural families) - klasifikace proteinových struktur

iProClass - znalostní databáze informací o proteinech

iProLink - literatura, odkazy

UniProt unverzální proteinově-orientované zdroje

---

spojení databází Swiss-Prot, TrEMBL a PIR v jeden centralizovaný systém

---

OWL neredundantní sekvence

Další zdroje odkazů

Katedra biochemie PřF MU, kde je i biochemický server

Harvard University, Dept. Mol.Cell. Biol (MCB) poskytuje řadu velmi cenných odkazů z oblasti biologie, biochemie, biodatabáze či výukové zdroje.