Bioinformatika

Charakteristika, aplikácie, metódy

Autori: Arpád Nagy (kapitoly 2 – 3)
Marek Serafín (kapitoly 4 – 6)

E-mail: adinis@adinis.sk

Bratislava Január 2010

1. Úvod

Množstvo a charakter údajov v biovedách si vo svete vynútilo v týchto vedných disciplínach používanie a aplikáciu štatistických, matematických a machine learning metód i algoritmov. Táto štúdia má za cieľ krátko informovať o aktuálnom stave a rozvoji metód bioinformatiky, najmä z pohľadu aplikačného ako aj matematicko-informatického. Snahou je priblížiť a ukázať širšej vedeckej obci pracujúcej v biovedách na Slovensku možné aplikácie bioinformatických riešení, vrátane stručného prehľadu používaného matematicko-štatistického aparátu, metód a algoritmov, ktoré je možné úspešne a efektívne využívať vo výskume aj praxi.

Mnohé z týchto algoritmov sú obvykle vo všeobecnejšej a jednoduchšej forme súčasťou vyhodnocovacích procesov zabudovaných v moderných analytických prístrojoch. Individuálnosť výskumných prác i projektov a potreba jedinečných riešení vo výskume, však vyžadujú adaptívny rozvoj a aplikáciu štatistických, matematických a machine learning metód na hľadanie, spracovanie a vyhodnotenie údajov priamo z experimentov, či databáz. Tu bude dôležitá vzájomná spolupráca medzi pôvodcom/prijímateľom výsledkov (bio-výskumník) a spracovateľom údajov (štatistik, matematik, tvorca softwaru). Pri postupnej adaptívnej implementácii bioinformatických algoritmov budú výsledky práce výskumných pracovníkov presnejšie a úplnejšie. Zásadne sa môže zmeniť aj obsah poznania možností používaných nástrojov, čo môže viesť aj k novým metódam skúmania a novým výsledkom.

V tomto dokumente je postupne opísaná charakteristika bioinformatiky a hlavné oblasti jej aplikácie. Následne je uvedený matematicko-informatický aparát slúžiaci k definovaniu a riešeniu úloh bioinformatiky, ďalej nasleduje popis všeobecných prístupov k riešeniu úloh, za využitia výpočtovej techniky (ide o metodiku machine learning) a popis konkrétnych softvérových nástrojov používaných v tejto oblasti.

Záverečná časť je venovaná štúdiu bioinformatiky na najpoprednejších univerzitách sveta a Európy a z voľne dostupných materiálov mapuje situáciu pri výučbe na druhom a treťom stupni vysokoškolského štúdia so zameraním na matematicko-informatické prístupy v bioinformatike. V dostupných prípadoch uvádza aj možné smery aplikácie pre bioinformatiku, respektíve výpočtovú biológiu (viaceré smery sú etablované, niektoré môžu pôsobiť inšpiratívne).

2. Stručná charakteristika bioinformatiky

2.1 Definície bioinformatiky

Termín *bioinformatika* zaviedla v r. 1979 Paulien Hogeweg pre štúdium informačných procesov v biotických systémoch. V súčasnosti je bioinformatika definovaná rôznymi spôsobmi, ale v princípe je vždy považovaná za kombináciu biologických vied a informatiky, spolu s ďalšími prispievajúcimi disciplínami. Často sa diskutuje, či bioinformatika a počítačová biológia (computational biology, doslovne biológia zahrňujúca počítanie) predstavujú to isté, alebo sú odlišné. V širšom zmysle je bioinformatika ponímaná ako využívanie počítačov na spracovanie akýchkoľvek s biológiou súvisiacich informácií, napr. DNA sekvencií, ale aj RTG snímkov prsníka. Preto aj oblasti ako medicínske zobrazovacie techniky, analýza obrazov a iné, by mohli byť považované za súčasť bioinformatiky. Ale v

praxi sa používa aj oveľa užšia definícia, v ktorej bioinformatika je synonymom pre výpočtovú molekulárnu biológiu, t.j. akékoľvek využitie počítačov na charakterizáciu molekulárnych komponentov živých organizmov (Bioinformatics Organization, Inc.[1]). Pre ilustráciu rozmanitosti definícií bioinformatiky uvádzame niektoré z nich:

- (Molekulárna) Bioinformatika je konceptualizácia biológie z hľadiska molekúl (v zmysle fyzikálnej chémie) a následná aplikácia techník informatiky (odvodených z disciplín ako aplikovaná matematika, počítačové spracovanie informácií a štatistika) s cieľom pochopiť a usporiadať informácie veľkého rozsahu spojené s týmito molekulami (Gerstein Bioinformatics Group - Yale University [2]).
- Bioinformatika je oblasť vedy, v ktorej sa spájajú biológia, informatika a informačné technológie, aby vytvorili jednu disciplínu. Konečným cieľom odboru je umožniť objavovanie nových biologických poznatkov, ako aj vytvorenie globálnej perspektívy, z ktorej možno odvodiť zjednocujúce princípy v biológii (National Center for Biotechnology Information – NCBI [3]).
- Bioinformatika je interdisciplinárna vedná oblasť na rozhraní biologických a počítačových[4] vied. Konečným cieľom bioinformatiky je odhaliť bohatstvo biologickej informácie ukrytej vo veľkom množstve dát a získať jasnejšiu predstavu o základoch, na ktorých je založená biológia organizmov. Získané nové poznatky môžu mať vážny dosah na rôzne oblasti, ako je ľudské zdravie, poľnohospodárstvo, životného prostredie, energetika a biotechnológie (The European Bioinformatics Institute – EBI [5])
- Výskum, vývoj, alebo aplikácia výpočtových nástrojov a prístupov s cieľom rozšíriť využívanie biologických, lekárskych, behaviorálnych alebo zdravotných údajov, vrátane výskumu, vývoja a aplikácie nástrojov a prístupov na získavanie, ukladanie, organizovanie, archivovanie, analyzovanie a vizualizovanie týchto údajov (National Institute of Health – NIH [6]).

US Environmental Protection Agency[7] definuje (z metodického hľadiska) 3 hlavné úlohy bioinformatiky nasledovne:

1) vývoj nových algoritmov a štatistických nástrojov na posúdenie vzťahov medzi údajmi vo veľkých súboroch dát,
2) analýza a interpretácia rôznych typov dát, vrátane nukleotidových a aminokyselinových sekvencií, proteínových domén a proteínových štruktúr,
3) vývoj a implementácia nástrojov, ktoré umožňujú pohotový prístup a efektívny manažment rôznych typov informácií.

2.2 Bioinformatika a výpočtová biológia

Bioinformatika aj výpočtová biológia úzko súvisia s vedami o živote.

[1] http://www.bioinformatics.org/wiki/Bioinformatics
[2] http://www.gersteinlab.org/courses/452/09-spring/pdf/cbb752-mg-spr09-bioinfo-intro.pdf
[3] http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html
[4] Počítačová veda môže byť definovaná ako štúdium teoretických základov informácie a spracovania informácie a praktických techník na ich implementáciu a aplikáciu v počítačových systémoch.
[5] http://www.ebi.ac.uk/2can/bioinformatics/bioinf_what_1.html
[6] http://www.bisti.nih.gov/docs/CompuBioDef.pdf
[7] http://www.epa.gov/comptox/glossary.html

- Bioinformatika uplatňuje princípy informačných vied a technológií, aby sa veľké, rôznorodé a komplexné údaje vied o živote stali zrozumiteľnejšími a užitočnejšími. Bioinformatika zahŕňa najmä manažment biologických databáz, získavanie dát z databáz, modelovanie dát, ako aj IT-nástroje na ich vizualizáciu.
- Počítačová biológia využíva matematické a výpočtové prístupy k riešeniu teoretických aj experimentálnych otázok v biológii. Snaží sa o riešenie biologických problémov počítačovými nástrojmi ako sú modelovanie, algoritmy, heuristika a iné.

Hoci bioinformatika a počítačová biológia sú v princípe odlišné, existuje aj výrazný prekryv na ich rozhraní. Výsledky získané metódami počítačovej biológie sú často súčasťou bioinformatických riešení. Napr. doplňovanie chýbajúcich experimentálnych bodov v expresných profiloch s využitím Expectation Maximization algoritmu je využitím počítačovej biológie. Ak sa však vytvorí databáza, pomocou ktorej sa budú ukladať a sprístupňovať doplnené expresné profily, môžeme hovoriť o bioinformatickom riešení. V mnohých prípadoch sa stáva, že bioinformatika a počítačová biológia sa voľne zamieňajú a hranica medzi nimi je často nepostrehnuteľná.

2.3 Pre- a post-genomická bioinformatika

Pre-genomická bioinformatika je charakterizovaná najmä využívaním výpočtovej techniky na ukladanie, získavanie, analýzu a predikciu zloženia a štruktúry biomolekúl. V centre klasickej bioinformatiky boli najmä nukleové kyseliny a proteíny a jej ťažisko bolo orientované na analýzu sekvencií.

Najväčší úspech bioinformatických metód, *Project ľudského genómu*[8], je prakticky kompletný. V dôsledku tejto skutočnosti sa v post-genómovej ére povaha a priority bioinformatického výskumu a jeho aplikácií zmenili. To ovplyvňuje bioinformatiku niekoľkými spôsobmi:

- Teraz, keď sú k dispozícii rôzne úplné genómy, môžu sa hľadať rozdiely a podobnosti medzi všetkými génmi z viacerých druhov. Z týchto štúdií je možné vyvodiť špecifické závery o jednotlivých druhoch ako aj všeobecné závery o evolúcii. Tento typ vedy je často označovaný pojmom *komparatívna genomika*.
- Existujú technológie navrhnuté na meranie relatívneho počtu kópií genetickej informácie (úrovne génovej expresie) v rôznych štádiách vývoja, choroby, alebo v rôznych tkanivách. Používanie takých technológií, ako sú *DNA microarray* a ďalších príbuzných technológií bude stále významnejšie.
- Význam ďalších priamejších metód identifikácie génových fukcií a asociácií vo veľkom rozsahu (napr. metódy *yeast two-hybrid*) bude rásť a spolu s nimi aj súvisiaca oblasť bioinformatiky zameraná na *funkčnú genomiku*.
- Bude všeobecný posun (obzvlášť sekvenčnej analýzy) od génov samotných ku génovým produktom. To povedie k:
    - k pokusom katalogizovať činnosť a charakterizovať interakcie medzi všetkými génovými produktami (u ľudí): *proteomika*,
    - k pokusom o kryštalografickú analýzu a/alebo predpovedanie štruktúry všetkých proteínov (u ľudí): *štrukturna genomika*.

[8] International Human Genome Sequencing Consortium, Nature 2001, *409*, 860-921. Venter et al., Science 2001, *291*, 1304-1351.

- Využívanie metód *výskumnej* alebo *medicínskej informatiky*, manažment všetkých biomedicínskych experimentálnych dát spojených s určitými molekulami alebo pacientami - od hmotnostnej spektroskopie až po klinické *in vitro* testy vedľajších účinkov - sa bude posúvať od ľudí pracujúcich vo výskume liekov a nemocničných informačných technológiach do hlavného prúdu bunkovej a molekulárnej biológie a bude migrovať z komerčnej a klinickej do akademickej sféry.

Stojí za zmienku, že všetky vyššie uvedené post-genomické oblasti výskumu závisia na zavedených pre-genomických technikách sekvenčnej analýzy.

3. Aplikačné oblasti bioinformatiky

Aplikačné oblasti bioinformatiky a počítačovej biológie je možné charakterizovať rôznymi spôsobmi. Na ilustráciu uvádzame 3 rôzne pohľady. Z hľadiska využívania metód na úrovni génu, genómu alebo ich produktov je možné aplikáciu bioinformatických nástrojov rozdeliť do 3 úrovní [9].

1) Analýza sekvencie jedného génu (proteínu). Napríklad:
   - Podobnosť s inými známymi génmi
   - Fylogenetické stromy, evolučné vzťahy
   - Identifikácia dobre definovaných domén v sekvencii
   - Charakteristické vlastnosti sekvencie (fyzikálne vlastnosti, väzbové miesta, modifikačné miesta)
   - Predikcia subcelulárnej lokalizácie
   - Predikcia sekundárnej a terciárnej štruktúry
2) Analýza kompletných genómov. Napríklad:
   - Ktoré rodiny génov sú prítomné, ktoré chýbajú
   - Umiestnenie génov na chromozómoch, korelácia s funkciou alebo evolúciou
   - Expanzia/duplikácia rodín génov
   - Prítomnosť či neprítomnosť biochemických dráh
   - Identifikácia "chýbajúcich" enzýmov
   - Významné udalosti v evolúcii organizmov
3) Analýza génov a genómov z hľadiska funkčných údajov. Napríklad:
   - Analýza expresie; microarray dáta, merania koncentrácie mRNA
   - Proteomika; merania koncentrácie proteínov, kovalentné modifikácie
   - Porovnanie a analýza biochemických dráh
   - Delécia alebo mutantné genotypy versus fenotypy
   - Identifikácia základných génov, alebo génov zapojených do špecifických procesov

Odborný časopis Bioinformatics, podľa impact faktoru druhý medzi časopismi zameranými na matematickú a počítačovú biológiu, publikuje odborné články v nasledovných oblastiach bioinformatiky:

1) Sekvenčná analýza
2) Analýza genómu
3) Fylogenetika
4) Expresia génu

[9] http://www.bioinfo.se/kurser/swell/per/bioinfo-general.html

5) Systémová biológia
6) Genetika a populačná biológia
7) Štruktúrna bioinformatika
8) Vyhľadávanie a spracovanie údajov a textov
9) Databázy a ontológie

Bessant et al.[10] v knihe venovanej vytváraniu praktických bioinformatických riešení v programovacích prostrediach Perl, R a MySQL uvádzajú 7 aplikačných oblasti bioinformatiky:

1) Analýza sekvencií
2) Analýza microarray
3) Proteomika
4) Metabolomika
5) Systémová biológia
6) Štruktúrna biológia
7) HĺbkovÁ analýza literatúry (Literature Mining - LM)

V nasledujúcich podkapitolách sú v krátkosti charakterizované jednotlivé oblasti spolu s metódami a algoritmami používanými v bioinformatike a počítačovej biológii.

3.1 Analýza sekvencií

Sekvenčná analýza v molekulárnej biológii a bioinformatike predstavuje automatizované, počítačové skúmanie charakteristických fragmentov, napríklad vlákien DNA alebo proteínov. Do analýzy sekvencií je možné zahrnúť napr.:

- Porovnanie sekvencií s cieľom nájsť podobnosti a odlišnosti v porovnaných sekvenciách (zarovnanie sekvencií).
- Identifikáciu génových štruktúr, čítacích rámcov, distribúciu intrónov a exónov a regulačných prvkov.
- Zistenie a porovnanie bodových mutácií alebo jednonukleotidových polymorfizmov (SNP) v organizme za účelom získania genetických markerov.
- Odhalenie evolučnej a genetickej diverzity organizmov, vytváranie fylogenetických stromov.
- Anotácia funkcie génov.

Používané metódy[11]

- Dynamické programovanie (algoritmy: Needleman-Wunch pre globálne zoradenie sekvencií, Smith-Waterman pre lokálne zoradenie).
- Heuristické metódy na vyhľadávanie sekvenčných similarít v databázach veľkého rozsahu (nástroje BLAST a FASTA).
- Analýza sekvencií pomocou bodovej matice.
- Párové zoradenie s využitím skrytých Markovových modelov.
- Zoradenie viacerých sekvencií s využitím metód multidimenzionálneho dynamického programovania, progresívne zoradenie pomocou modifikovaného algoritmu Feng-Doolittle, ktorý je základom programu ClustalW. Metódy na báze profilových skrytých

[10] Bessant et al. Building Bioinformatics Solutions. Oxford University Press. Oxford 2009.
[11] Durbin et al. Biological sequence analysis. Ninth printing. Cambridge University Press, Cambridge 2004.

Markovových modelov a s využitím špecializovaného EM algoritmu a simulovaného žíhania, ktoré patria k základným nástrojom v oblasti strojového učenia.

- Zostavovanie fylogenetických stromov:
  - s využitím vzdialenostných matíc a klastrovacích algoritmov UPGMA a Neighbour-joining alebo stromov na báze najväčšej úspornosti (parsimony),
  - s využitím Bayesovských pravdepodobnostných modelov, metóda Maximal likelihood.
- Hľadanie génov v genóme s použitím Markovových reťazcov a skrytých Markovových modelov.

3.2 Analýza microarray

Expresia mnohých génov môže byť určená meraním hladiny mRNA pomocou viacerých techník, vrátane microarray, EST, SAGE, MPSS[12], alebo rôznych aplikácií multiplexnej in-situ hybridizácie. Všetky tieto metódy sú veľmi citlivé na pozadie a na skreslenie pri meraní. Preto je dôležitou úlohou počítačovej biológie vývoj štatistických nástrojov pre oddelenie signálu od šumu pri štúdiach génovej expresie vo veľkom rozsahu. Využívanie techniky microarray je v súčasnosti na vzostupe a používa sa na mapovanie génov, na zisťovanienie mechanizmu pôsobenia terapeutík, v diagnostike klinických ochorení ako aj v ďalších prípadoch.

Používané metódy[13]

- Spracovanie obrazu zo skenera microarray pomocou počítačových algoritmov na identifikáciu spotov, stanovenie intenzity spotov a pozadia.
- Normalizácia dát pomocou pomerovej štatistiky alebo regresnej techniky.
- Analýza dát:
  - Hierarchické klastrovanie, K-priemer klastrovanie
  - Využívanie neurónových sietí, Self Organised Maps
  - Analýza hlavných komponentov (Principal Component Analysis)
  - Informované účenie, Support Vector Machines
  - Metódy využívajúce Bayesovské siete
- Databázové riešenia na uchovávanie expresných dát a súvisiacich informácií o vzorke, experimentálnych podmienkach a pod. (tzv. metadát).

3.3 Proteomika

Proteomika je zameraná na kvantifikáciu expresných úrovní kompletnej sady proteínov (proteomu) v bunke v danom okamihu. Proteomický výskum bol pôvodne zameraný na 2-D gelové elektroforézy pre separáciu a identifikáciu proteínov, ťažisko proteomiky sa teraz zameriáva na postupy, ktoré charakterizujú funkciu veľkých súborov proteínov. Proteomika sa často používa ako synonymum pre funkčnú genomiku[14]. Na analýzu expresie proteínov sa používajú proteínové microarray a najmä hmotnostná spektrometria (MS). Vyhodnocovanie MS dát je zamerané na porovnávanie veľkého objemu experimentálnych hmotnostných dát s hmotnosťami predikovanými na základe sekvencií z proteínových databáz a komplikované

[12] EST - expressed cDNA sequence tag, SAGE - serial analysis of gene expression, MPSS - massively parallel signature sequencing

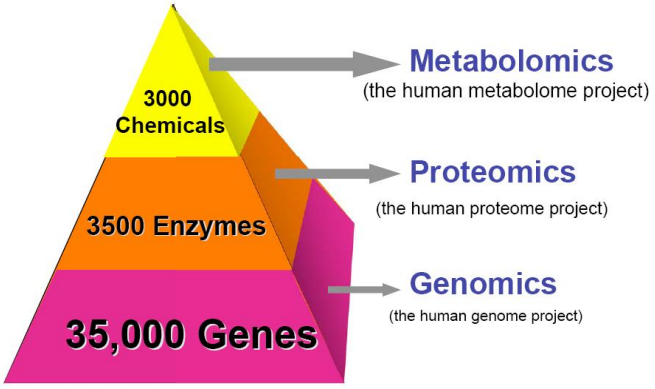
[13] Whitehead Institute for Biomedical Research: http://jura.wi.mit.edu/bio/education/bioinfo2002/lecture10-color.pdf

[14] http://www.pasteur.fr/recherche/unites/Binfs/definition/bioinformatics_definition.html

štatistické analýzy vzoriek, v ktorých je detekovaných veľa neúplných peptidov z každej bielkoviny prítomnej vo vzorke.

Používané metódy

- Obrazová analýza a kvantifikácia 2-D gélov.
- Štatistické metódy podobné s metódami na analýzu expresie génov.
- Porovnanie experimentálnych a teoretických peptidových fragmentov pomocou vyhodnocovacích funkcií (scoring functions), SEQUEST, Mascot alebo novšej pravdepodobnostnej funkcie odvodenej pomocou skrytých Markovových modelov[15].
- Vývoj nových algoritmov pomocou teórie grafov, heuristickej optimalizácie MCMC (Markov chain Monte Carlo) alebo skrytých Markovových modelov, najmä pre de-novo sekvenčné analýzy.
- Vývoj experimentálnych a výpočtových metód na kvantitatívne stanovenie proteínov pomocou MS.



Obr. 1 Pyramída života [16]

3.4 Metabolomika

Zaoberá sa kvantitatívnym meraním metabolických profilov modelových organizmov, aby charakterizovala ich fenotyp alebo fenotypovú odozvu na genetické alebo nutričné perturbácie [17]. Pod metabolitom sa rozumie každá v organizme detekovateľná organická molekula s hmotnosťou < 1500 Da. Metabolomika využíva experimentálne metódy UPLC, HPLC, CE/mikrofluidiku, LC-MS, FT-MS, QqQ-MS, NMR, RTG kryštalografiu, GC-MS, LIF detekciu a umožňuje:

- Vytvárať metabolické „podpisy“
- Monitorovať/merať toky metabolitov
- Monitorovať kinetiku enzýmov/metabolických dráh
- Posúdiť/identifikovať fenotypy
- Monitorovať interakcie génu s okolím
- Sledovať účinky spôsobené toxínmi/perturbantami
- Monitorovať dôsledky zablokovania činnosti génov

[15] Colinge et al. Introduction to Computational Proteomics. Computational Biology 2007, *3*, 1151-60.
[16] David Wishart: http://incob.apbionet.org/incob07/presentation/wishart.pdf
[17] http://incob.apbionet.org/incob07/presentation/wishart.pdf

- Identifikovať funkcie neznámych génov

Výzvy pre bioinformatiku
- Metabolomika produkuje veľmi rozmanité dátové typy.
- Veľmi málo metabolomických dát je dostupných elektronicky, existujú len 2 databázy metabolických dráh KEGG a BioCyc a nová databáza Human Metabolome Database[18].
- Hĺbková analýza dát z existujúcich databáz – BioSpider, PolySearch.
- Problematika manažovania a sprístupňovania dát produkovaných na rôznych miestach – aplikácia LIMS (Laboratory Information Management System).
- Spracovanie surových analytických dát, využívanie štatistických metód.
- Metabolomické štandardy a ontológia.
- Integrácia údajov a matematické modelovanie metabolických sietí v rámci systémovej biológie.

Používané metódy
- Metabolomické dáta sú analyzované štatistickými metódami a metódami strojového učenia[19].
    - algoritmy pre učenie sa bez učiteľa: hierarchické klastrovanie, PCA (Principal Component Analysis) a SOM (Self-Organizing Maps),
    - algoritmy využívajúce informované účenie: ANOVA, PLS (Partial Least Squares) a DFA (Discriminant Function Analysis).

Poznámka: Špecializovaná bioinformatika pre glykomiku je charakterizovaná v práci [20].

3.5 Systémová biológia

Je jednou z najvýznamnejších novo vznikajúcich interdisciplinárnych oblastí vedy. Prepojením genomiky, proteomiky, bunkovej biológie, lekárstva, molekulárnej biológie a genetiky, s matematikou, bioinformatikou, strojárstvom a výpočtovými metódami, umožňuje objavovanie doposiaľ neznámych princípov fungovania živých buniek. Zároveň vytvára testovatelné a prediktívne modely zložitých bunkových dráh a možno aj celej bunky, ktoré sú užitočné pre účinný experimentálny dizajn a bioinžinierstvo a pre dizajn liekov a liečebných postupov založených na sieťach[21]. V modernej systémovej biológii sa riešia napr. nasledovné problémy[22]:

- Analýza architektúry a dynamiky celulárnych sietí.
- Počítačové modelovanie dráh bunkovej signalizácie.
- Charakteristika bunkových sieťových štruktúr, ktoré ich odlišuje od náhodne generovaných sietí.
- Vzťahy štruktúry sietí a biologických funkcií.
- Otázky zachovania a evolučného vývoja sieťových štruktúr.
- Sú špecifické topologické vzory preferované v určitom čase alebo za určitých podmienok a pod.

---

[18] http://www.hmdb.ca

[19] Shulaev V. Metabolomics technology and bioinformatics. Briefings in Bioinformatics 2006, *7*, 128 - 139.

[20] Aoki-Kinoshita K.F. An Introduction to Bioinformatics for Glycomics Research. Computational Biology 2008, *4*, 1-7.

[21] Federation of European Biochemical Societies: http://www.febssysbio.net/flyer_sysbio2009.pdf

[22] Qi J. et al. Modularity and Dynamics of Cellular Networks. Computational Biology 2006, 2, 1502-10.

- Dizajn databáz na ukladanie heterogénnych dát a vývoj nových metód na analýzu a vizualizáciu údajov.

Používané metódy
- Bayesianské siete, orientované pravdepodobnostné grafické modely.
- Klastrovacie algoritmy s využitím skrytých Markovových modelov na hľadanie modulov.
- Klastrovanie s využitím topológie siete.
- Integrovanie sústav obyčajných a parciálnych diferenciálnych rovníc na modelovanie dynamiky sietí.

3.6 Štruktúrna biológia

Štruktúrna biológia sa venuje štúdiu fyzikálnej architektúry biologických molekúl, najmä proteínov.Výskum je typicky zameraný na vzťah medzi štruktúrou a funkciou, štruktúrnou podobnosťou medzi proteínmi, sumuláciou interakcie medzi proteínmi a inými molekulami a vzťahmi medzi sekvenciami proteínov a ich štruktúrou. Experimentálne sa štruktúra určuje najmä pomocou RTG difrakcie alebo NMR. Cieľom bioinformatiky je predikcia štruktúry použitím počítačových metód a ukladanie štruktúr do databáz. ako napr. Protein Data Bank (PDB)[23].

Používané metódy
- Predpovede štruktúry a funkcie využitím skrytých Markovových modelov, viacvrstvových perceptronov, rozhodovacích stromov.
- Predpovedanie sekundárnej štruktúry na základe znalosti primárnej
    - Metódy Chou-Fasman/GOR
    - Metóda najbližšieho suseda (Nearest Neighbor)
    - Metódy strojového učenia - neurónové siete, Support Vector Machines
- Predpovedanie terciárnej štruktúry
    - Sekvenčná homológia – použitie homológnej sekvencie ako templátu
    - Threading – hľadanie štruktúr, ktoré majú podobné konfigurácie zvinutia bez zjavnej sekvenčnej podobnosti

3.7 Hĺbková analýza literatúry

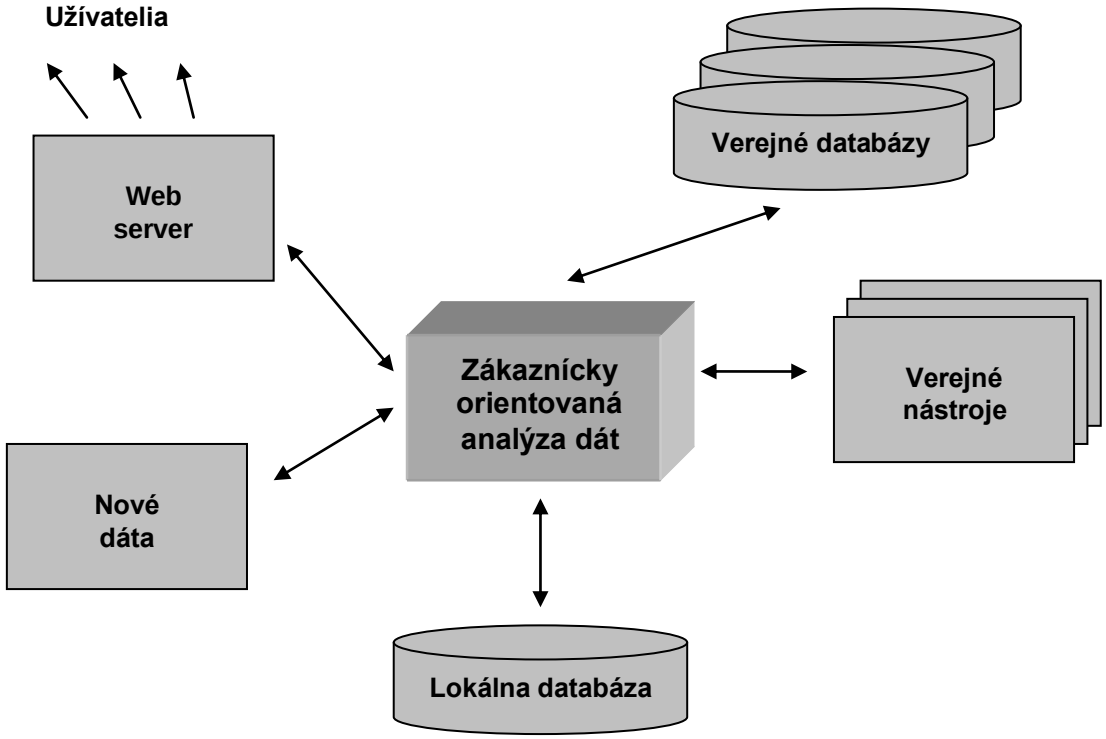
Predstavuje proces uplatňovania techník analýzy údajov (data mining) na databázy publikovanej literatúry. Cieľom LM je pomôcť užívateľom extrahovať oveľa ľahšie ako tradičné vyhľadávače skryté znalosti existujúce v extrémne veľkých sytémoch, v ktorých je literatúra uložená. Idea hĺbkovej analýzy literatúry vychádza z predpokladu, že uživateľ je v „objavnom režime“ a hľadá zaujímavé súvislosti medzi zdanlivo nesúrodými poznatkami, ktoré môžu pomôcť pri riešení zadanej úlohy.

Používané metódy[24]
- NLP (natural language processing)
- Používanie kontrolovanej terminológie (napr. nomenklatúra génov HUGO, klasifikácia chorôb ICD, génova ontológia, kompilácia proteínových synoným z viacerých terminológií BioThesaurus)
- Využívanie UMLS (Unified Medical Language System)

[23] http://www.pdb.org
[24] Rodriguez-Esteban R., Biomedical Text Mining and Its Applications, Computational Biology 2009, *5*, 1-5.

3.8 Vytváranie bioinformatických riešení

Priklad ilustrujúci komponenty systému na vytváranie bioinformatických riešení je na Obr. 2. Vo všeobecnosti v jednotlivých aplikáciach nie sú vyžadované všetky komponenty, napr. je možné analyzovať nové dáta získané lokálne bez prístupu do externej databázy. Na druhej strane nie je potrebné mať vlastné dáta, ale je možné analyzovať verejne dostupné data vlastnými, alebo verejnými bioinformatickými nástrojmi.



Obr. 2 Generické bioinformatické riešenie s typickými komponentami, ktoré môžu byť použité [25]

4. Matematicko-informatický aparát bioinformatiky

4.1 Štandardné matematicko-informatické metódy

Pri rozvoji Bioinformatiky sa z hľadiska matematicko-informatického štandardne (podľa [1]) používajú tieto prístupy:

Matematika

- vektory, matice
- integrácia, diferenciálne rovnice
- numerický diferenčný a integrálny výpočet
- polynomická interpolácia
- metóda najmenších štvorcov
- iteratívne metódy na riešenie systémov lineárnych rovníc
- riešenie nelineárnych rovníc

[25] Bessant et al. Building Bioinformatics Solutions. Oxford University Press. Oxford 2009.

- formálne gramatiky a Chomského hierarchia
- permutácie, kombinatorika a rekurentné vzťahy
- teória grafov: tranzitívny uzáver, najkratšia cesta, farbenie grafov, planárne grafy, vetviaci sa strom (spanning tree)

Štatistika
- popis dát: priemer, medián, modus, kvantily, štandardná (smerodajná) odchýlka, rozptyl, koeficient rozptylu
- lineárna regresia a korelácia
- teória pravdepodobnosti: axiómy pravdepodobnosti, podmienená pravdepodobnosť, nezávislé udalosti, Bayesov vzorec
- náhodné premenné
- diskrétne a spojité rozdelenia, Chi-kvadrát, Studentov t-test, Fisher-Snedecorovo F rozdelenie
- teória odhadu, interval spoľahlivosti
- limitné vety

Databázy so zameraním na biologické dáta
- dátová nezávislosť, fyzická organizácia dát: hašovanie, B-stromy, indexovacie súbory
- relačná algebra a normálové formy
- relačné dotazovacie jazyky a optimalizácia dotazov
- úvod do Data mining: klasifikácia, klastrovanie dát, dátové sklady

Machine learning
- Skryté Markovove modely (HMMs)
- Support vector machines (SVM)
- Neurónové siete: učenie perceptrónu, Hebbovské učenie, algoritmus Backpropagation, Asociatívne učenie, Súťaživé učenie, Rekurentné siete, Hopfieldova sieť, Stochastické siete
- Genetické algoritmy
- Fuzzy množiny
- Pravdepodobnostné grafové modely

Pre bioinformatiku v zásade každý biologický , alebo informatický predmet môže byť považovaný za jej predmet. Na popredných univerzitách, sa vyučuje v rôznej forme, najčastejšie je jej smerom možné vysokoškolské štúdium tretieho stupňa (PhD.).

4.2 Techniky všeobecného prístupu

V tejto časti sú uvedené všeobecné prístupy k riešeniu problémov v bioinformatike [2, 3].

Dynamické programovanie
Vo svojej podstate ide o optimalizačnú techniku, keď sa problém dá rekurzívne rozdeliť na dva podproblémy menšej veľkosti tak, že zloženie podproblémov poskytne riešenie pre nadproblém. Je to jeden z kľúčových algoritmických prístupov pri sekvenčnej analýze a pri vyhodnocovaní sekvenčnej pravdepodobnosti. Príkladom použitia dynamického programovania sú algoritmy: Needleman-Wunch, Smith-Waterman a Viterbi algoritmus.

Metóda klesajúceho gradientu (gradient descending method)
Nachádza svoje uplatnenie konkrétne pri neurónových sieťach v podobe *algoritmu backpropagation* a pri skrytých Markovových modeloch (HMMs) v podobe *procedúry forward-backward*.

Metóda EM
Expectation Maximization (EM) sa používa pri rôznych aplikáciách a pri sekvenčnej analýze. V skrytých Markovových modeloch sa implementuje v podobe *Baum-Welch* algoritmu. Je užitočný v modeloch a situáciách so skrytými premennými - bežným príkladom použitia je situácia, keď dáta chýbajú, alebo sú nepozorovateľné.

Metódy MCMC
Markov chain Monte Carlo (MCMC) je dôležitá trieda stochastických metód. Sú založené na tom, že najprv sa konštruuje Markovov reťazec (Markov chain) pomocou ktorého postupne získame požadovanú distribúciu – ako stav reťazca po mnohých krokoch. Dva hlavné MCMC algoritmy sú: *Gibbs sampling* a algoritmus *Metropolis*.

Simulované žíhanie (simulated annealing)
Je to optimalizačný algoritmus, hľadajúci dobrú aproximáciu globálneho minima danej funkcie vo veľkom prehľadávanom priestore. V každom kroku sa aktuálne riešenie nahrádza „blízkym“ riešením, ktorého pravdepodobnosť závisí od rozdielu medzi korešpondujúcimi funkčnými hodnotami a tiež od globálneho parametru T (teplota), ktorý je postupne počas procesu znižovaný. Často osobitné uplatnenie nadobúda, keď je prehľadávaný priestor diskrétny.

Evolučné a genetické algoritmy
Evolučné algoritmy sú optimalizačné algoritmy, ktoré si berú inšpiráciu z terajšieho pohľadu na evolúciu a jej vnútorné mechanizmy. Spoločným znakom je, generovanie náhodných porúch a zmien, za prítomnosti funkcie, ktorá dokáže ohodnotiť kvalitu zmien (mutácií) a následne sa dajú oddeliť zbytočné od užitočných. Genetické algoritmy sú podtriedou evolučných. Ide o prehľadávaciu techniku, určenú k nachádzaniu riešení optimalizačných a prehľadávacích problémov.

4.3 Vizualizácia dát

Mnohorozmerné dáta musia byť pre vizualizáciu špeciálne upravené, často sa využívajú rôzne metódy, ktoré najprv viacrozmerné dáta premietnu do 2D / 3D priestoru podľa osi kde majú dáta najväčší rozptyl. Na takéto spracovanie dát a redukciu dimenzionality sa napr. používa

- PCA
- Faktorová analýza

Dáta je potom možné zobraziť pomocou 2D / 3D grafov. Balíky ako napr. Matlab, Mathematica poskytujú mnoho spôsobov vizualizácie a rôzne typy grafov v interaktívnom prostredí, kde používateľ môže kontrolovať uhol pohľadu, veľkosť detailu atď.

Vizualizácia dát v 3D prostredí môže byť skvalitnená použitím nových 3D efektov a pokročilých grafických technológií s použitím moderných grafických kariet (GPU), ktoré slúžia napr. na vizualizáciu nových, graficky náročných 3D hier. Vizualizácia dát tak môže byť vytvorená s rôznym typom priehľadných, priesvitných objektov, ktoré môžu prezentovať dátové body. Moderné grafické karty sú naviac schopné renderovať obrovské množstvo dát s

veľkou rýchlosťou, takže je možné dáta vizualizovať v reálnom čase interaktívnym spôsobom. Podobné grafické technológie sa momentálne používajú na vizualizáciu 3D snímkov ľudských orgánov v medicíne [4], fMRI[26] snímkov ľudského mozgu [5] atď.

4.4 Konkrétne prístupy a aplikácie

Vo všeobecnosti sa úlohy delia na:

1. Sekvenčne orientované problémy
2. Dátovo zamerané problémy

Popis konkrétnych postupov a techník pri riešení problémov. Táto časť je podobne ako predchádzajúca inšpirovaná knihou [2].

Neurónové siete
Možnosťou je viacero architektúr, napríklad dopredná, či rekurentná. Štandardne sa používajú v spojení s algoritmom *backpropagation*, ale občas sa môže použiť aj prístup EM, či simulované žíhanie.

Použitie neurónových sietí:
- dokážu zachytiť koreláciu medzi jednotlivými časťami sekvencie (proteínov atď.)
- v oblasti grafickej vizualizácie: snaha o predpovedanie druhotnej proteínovej štruktúry
- predpovedanie signálnych peptidov a ich štiepnych miest
- aplikácie pre nukleotidové sekvencie DNA a RNA
- štruktúra a pôvod genetického kódu
- vyhľadávanie eukaryotických génov
- sekvenčná analýza pozorovaním poradia, v akom sa učí neurónová sieť
- predpovedanie „intron splice sites“

Skryté Markovove modely
Učenie môže prebiehať rôznymi spôsobmi, pomocou *Baum-Welch* algoritmu, prístupom klesajúceho gradientu, či *Viterbi* algoritmom. Po natrénovaní sa môže použiť na data-mining, roznorodú klasifikáciu, štruktúrnu analýzu a vyhľadávanie vzorov. Porovnaj s [3].

Použitie skrytých Markovových modelov:
- hľadanie génov
- určovanie tried sekvencií
- modelovanie druhotných štruktúrnych prvkov
- modelovanie vzťahu medzi párom DNA sekvencií
- objavovanie slabých vzorov v DNA sekvenciách

Pravdepodobnostné grafové modely, Bayesovské siete (Bayesian networks)
Mnoho rôznych problémov sa dá popísať pomocou závislostí medzi premennými. Pravdepodobnostné grafové modely [6, 7], umožňujú vyjadrenie takýchto závislostí pomocou grafov, ktoré vyjadria závislosti v pravdepodobnostnej distribúcii pre daný model. Takéto grafové modely sa dajú výhodne použiť v prípadoch zložitých modelov a vzťahov, ak je tiež prítomná neistota meraní, pozorovaní, atď. Pomocou grafových modelov vieme robiť inferenciu, čiže odvodenie stavov neznámych veličín alebo iných udalostí.

[26] functional Magnetic Resonance Imaging

Pre každú aplikáciu grafových modelov sa potrebujeme najprv naučiť parametre daného modelu, pretože grafový model vyjadruje len všeobecný charakter danej pravdepodobnostnej distribúcie. Učenie prebieha na pozorovaných dátach, na učenie sa používa metóda najväčšej zhody (maximum likelihood), alebo EM algoritmus, ak niektoré veličiny v modeli chýbajú. Špeciálnym typom grafových modelov sú napr. Skryté Markovove modely, neurónové siete, rôzne klastrovacie algoritmy atď.

Použitie [8]
- modelovanie proteínov
- gene expression, microarray analýza
- proteín-proteín interakcia (priradenie génových funkcií a propagácií v *protein interaction network)*
- analýza sekvencií DNA (mutácie, podobnosť, vývoj sekvencií)
- genetika (analýza fylogenetických sietí)

Support vector machines (SVMs)
Je množina príbuzných metód učenia s učiteľom využívaných na klasifikáciu a regresiu. Ak máme trénovaciu množinu, ktorej každý prvok patrí do jednej z dvoch kategórií, tak na princípe SVM [9] môžme model najprv natrénovať a následne použiť na predikciu príslušnosti nových prvkov. SVM si môžeme predstaviť ako model, ktorý reprezentuje príklady ako body v priestore a tie sú namapované tak, že oddelené kategórie sú rozdelené s jasnou čo najširšou medzerou. Nové príklady sú následne namapované do tohto priestoru a háda sa ich kategorické zaradenie podľa toho, na ktorú stranu medzere sa dostanú.

Použitie:
- klasifikácia génov a proteínov
- predikcia druhotnej proteínovej štruktúry
- regresia

Stochastické gramatiky
Vo všeobecnosti mnoho problémov výpočtovej molekulárnej biológie sa dá vyjadriť v rámci teórie formálnych jazykov. Všeobecnou úlohou je metódami strojového učenia zostrojiť k dátam korešpondujúcu gramatiku. Aplikuje sa pri nich EM prístup, ale môžu sa použiť podobne ako pri Skrytých Markovových modeloch aj iné postupy. Podrobnejšie tiež v [3].

Použitie
- modelovanie druhotnej štruktúry RNA
- štúdium molekúl RNA

4.5 Ďalšie špecializované prístupy

- Hybridná architektúra: Neurónová sieť/Skrytý Markovov model
- Dvojsmerná rekurentná neurónová sieť
- Pravdepodobnostné modelovanie evolúcie pomocou fylogenetických stromov (phylogenetic tree)
- Pravdepodobnostné modelovanie Microarrays
- Vytváranie a analýza klastrov (s učiteľom/bez učiteľa): napríklad pomocou EM

4.6 Softvérové nástroje

Na spracovanie biologických dát sa podľa [10] používajú tieto nástroje (ťažisko je na voľne šíriteľných MySql, Perl a R):

- Biologické databázy: slúžia na uloženie informácií podobného charakteru, uľahčujú ich zhromažďovanie, výber a umožňujú efektívne narábanie. Kľúčovým nástrojom sú relačné databázy a prístup k nim pomocou MySQL. Ako alternatíva sa ponúka PostgreSQL a komerčné nástroje Oracle a Microsoft Access.
- Automatizovanie procesov pomocou programovania: niektoré zložitejšie a často opakované úkony sa vykonávajú pomocou rôznych programovacích jazykov, ktoré uľahčujú prácu a niekedy sú pri predspracovaní aj samotnom spracovaní údajov nevyhnutné. Konkrétne je opísaná syntax a práca pomocou jazyka Perl. Alternatívou sú ďalšie skryptovacie jazyky ako Python, či Ruby. Pokiaľ ide o komplexnejšie, alebo výpočtovo náročnejšie aplikácie vhodným nástrojom sú interpretované programovacie jazyky ako Java, C/C++, či C#.
- Ďalšou kľúčovou časťou je numerická dátová analýza: matematické a štatistické spracovanie dát s následnou vizualizáciou. Osobitná pozornosť je venovaná softvérovému nástroju R. Výhodou oproti softvérovým balíkom ako napríklad Microsoft Excel, GenStat a Statistica je, že tie napriek výborným nástrojom na analýzu a vizualizáciu, nie sú až tak flexibilné na pridávanie nových funkcionalít, spolupráca s ďalšími nástrojmi nie je priamočiara a tieto nástroje priamo predpokladajú prítomnosť užívateľa, ktorý vykonáva osobne analýzy cez grafické ovládacie prostredie. Na druhej strane bežné programovacie jazyky neobsahujú zabudované neelementárne matematické funkcie a tie je nutné naprogramovať skoro od základov. A práve medzeru medzi týmito prístupmi sa snaží vypĺňať R. Ďalšími možnosťami sú komerčné programy S-Plus, Matlab a voľne šíriteľný Octave.

Nové aktuality pozri na [11].

Weka – Waikato Environment for Knowledge Analysis, je populárny voľne šíriteľný nástroj machine learning, vyvinutý na univerzite Waikato. Obsahuje [12, 13] sadu vizualizačných nástrojov a algoritmov slúžiacich k dátovej analýze a predpovedaniu, pričom poskytuje grafický interface. Medzi výhody ďalej patrí vysoká prenosnosť, pretože je plne implementovaná v prostredí Java a obsahuje techniky na predspracovanie dát a modelovanie. Nevýhodou je, že nie je schopná multirelačného dátového miningu a tiež že vyžaduje, aby dátové body boli popísané pevným počtom atribútov. Implementované schémy obsahujú:

- rozhodovacie stromy
- učenie sa pravidiel
- support vector machines
- lokálne váženú regresiu
- klastrovacie metódy
- učenie sa asociačných pravidiel

Mathematica – je výpočtový softvér používaný pri vedeckých a matematických výpočtoch [14]. Je spoplatnený. Obsahuje knižnicu s elementárnymi a špecializovanými matematickými funkciami, nástroje na 2D i 3D vizualizáciu a vlastný programovací jazyk podporujúci procedurálne, funkcionálne a objektové programovanie. Ďalej nástroje pre vizualizovanie, analyzovanie grafov a data mining. Nevýhodou je, že nie vhodná na vytváranie komplexnejších a rozsiahlejších programov.

Matlab – je numerické výpočtové prostredie a štvrtogeneračný programovací jazyk. Umožňuje manipuláciu s maticami, vykreslovanie funkcií a dát, implementáciu algoritmov a vytváranie používateľských rozhraní [15]. Jeho súčasťou môžu byť aj tieto rozšírenia [16]:

- Statistics toolbox: obsahuje funkcie a interaktívne nástroje na modelovanie dát, analýzu historických trendov, simulovanie systémov, vizualizáciu (vrátane mnohorozmerných nelineárnych modelov) a vývoj štatistických algoritmov
- Bioinformatics toolbox: prístup ku genomickým a proteomickým dátovým formátom, technikám analýzy špecializované vizualizácie pre genomické a proteomické sekvencie amicroarray analýzu
- Neural Network toolbox: má uplatnenie pri identifikácii nelineárnych systémov a rozpoznávaní vzorov
- A tiež mnohé ďalšie toolboxy často vytvárane užívateľmi a následne voľne zdielané

Okrem obrovského prínosu a nesporných výhod, medzi nevýhody môžeme radiť [15], že je tiež spoplatnený. Ďalšou nevýhodou je, že používa zátvorky na indexovanie poľa aj volanie funkcie, čo vyžaduje pri čítaní veľkú pozornosť. Občas sa stane, že Matlab interpretuje kód inak, než zamýšľal užívateľ a tiež kód napísaný pre špecifickú verziu, nemusí spolupracovať s inou.

Záver
V tejto časti sme uviedli a popísali rôzne softvérové nástroje, ktoré môžu slúžiť k spracovaniu biologických dát. Každý nástroj má svoje špecifické určenie, svoje silné stránky a svoje nevýhody. Často ide o všeobecné nástroje, ktoré nie sú úzko špecializované na danú oblasť. Z toho vzniká potreba a oprávnenosť vývoja a nasadenia úzko špecializovaného softvérového nástroja uspôsobeného priamo na mieru.

4.7 Výskumné oblasti bioinformatiky

Medzi hlavné smery bioinformaticky orientovaného základného výskumu napríklad aj podľa [17] patrí:

- sekvenčná analýza
- meranie biodivezity
- výpočtová evolučná biológia
- analýza proteínovej expresie
- analýza génovej expresie
- predpovedanie proteínovej štruktúry
- analýza mutácií rakoviny
- komparatívna genomika
- „protein – protein docking“
- vývoj vhodného softvéru a nástrojov

5. Študijné programy na vybraných univerzitách

5.1 Výber univerzít

Výber prehľadov univerzít bol robený s prihliadnutím na vysoké umiestnenie v dostupných rebríčkoch univerzít a vzhľadom na dobrú voľnú dostupnosť k ich študijným programom. Konkrétne ide o tieto univerzity: Stanford university, Princeton university, Cornell university,

University of Oxford, University College London, Imperial College London, ETH Zürich, University of Copenhagen. Ich postavenie na svete je všeobecne dobre známe a môže sa overiť napríklad v rebríčkoch [18], [19] a [20]:

Konkrétne podľa vyhodnotenia univerzít s ohľadom na Informatiku (Počítačovú vedu) rebríček GRE [18] umiestňuje Stanford university na prvé miesto, Princeton university, Cornell university na piate a Princeton university na šieste miesto (medzi univerzitami v USA). Zároveň sa Stanford university zaraďuje na prvé miesto aj v oblasti biologických vied.

Podľa vyhodnotenia [20] sa v rámci Európy University of Oxford umiestnila na druhom mieste, Imperial College London na treťom, University College London na štvrtom, ETH Zürich na siedmom a University of Copenhagen na dvanástom mieste.

5.2 Stanford University

Na Stanford University sa podľa [21] vyučuje Biomedicínska informatika a Biomedicínske výpočty. Ak sa chce niekto zamerať na smer bioinformatiky, na prvom stupni štúdia sa má orientovať na predmety:

- počítačové programovanie
- databázy
- pravdepodobnosť a štatistika
- molekulárna biológia a fyziológia

Následne na druhom stupni štúdia, si môže vyberať z predmetov označených ako bioinformatické. Ako bioinformatické predmety sa na Stanford university prednášajú nasledovné predmety:

- Representations and Algorithms for Computational Molecular Biology: kurz programovania a základov algoritmov
- Translational bioinformatics: pokrýva použitie bioinformatiky v translačnej medicíne
- Computer Applications in Molecular Biology: predmet pre biológov, ktorý vovádza do problematiky bioinformatiky
- Genomics: predmet pre biológov, programovanie v Perl a analýza genómových dát
- Protein Architecture, Dynamics and Structure Prediction: základné koncepty molekulárnej štruktúry a výpočtov
- Computational Genomics: dôležité algoritmy pre výskum genómu a komparatívna genomika
- Algorithms for structure and function in biology: algoritmy na modelovanie v molekulárnej biológii
- Algorithms in Biology: detailná štúdia nových algoritmov v bioinformatike
- Computational methods for analysis and reconstruction of biological networks: algoritmy a dátové štruktúry slúžiace na analýzu a rekonštrukciu biologických sietí
- Computational Systems Biology: úvod do výpočtov biologických systémov
- Biomedical Informatics: úvod do informatiky s jej aplikáciou v biológii a klinike

Situácia je taká, že ľudia študujú PhD. buď informatiku (computer science), alebo biológiu (či podobný odbor), následne v závislosti od toho či pracovali na prieniku týchto dvoch disciplín, sú kvalifikovaný na samostatnú vedeckú činnosť v bioinformatike, alebo sa potrebujú ešte dovzdelať v oblasti čo im chýba. Potom je samozrejme kategória ľudí, čo priamo študovali bioinformatiku a majú priamo relevantné skúsenosti s touto oblasťou.

5.3 Cornell University

Podľa [22] sa tu dajú študovať nasledovné predmety zamerané na biometriu a štatistiku:

- Biological Statistics I: vizualizácia dát, ohodnotenie parametrov, vzorkovanie, testovanie hypotéz, normálna a iné pravdepodobnostné distribúcie
- Biological Statistics II: lineárna regresia, inferencia, zovšeobecnené lineárne modely, ANOVA (jedno a viac faktorová analýza variácie), nelineárne modelovanie
- Multivariate Analysis: multivariantná normálna distribúcia, multivariantná regresia, analýza hlavných komponentov, korelácia, klastrovanie a diskriminačná analýza
- Statistical Genomics: dôležité pravdepodobnostné distribúcie, pravdepodobnosť a Bayesovská inferencia, použitie štatistických metód v „linkageanalysis“, „Quantitative Trait Locus mapping“, „analysis of pedigrees“, molekulárna populačná genetika a genomika, fylogenetická inferencia
- Statistical Methods I: deskriptívna štatistika, testovanie hypotéz, inferencia, porovnávanie dvoch populácií, porovnanania medzi populačným priemerom, korelácia, regresná analýza
- Statistical Methods II: spôsob zberu dát, metóda najmenších štvorcov, viacnásobná regresia, výber vplyvných bodov, variancia a kovariancia
- Statistical Methods III: logická regresia, log-lineárny model, tabuľky pre stratifikáciu, párová analýza, ordinálne dáta

A ďalšie predmety zamerané na strojové učenie:

- Machine Learning: rozhodovacie stromy, support vector machines, Bayesovské učenie s parametrami, Skryté Markovove modely, učenie bez učiteľa, zosilnenie učenia
- Advanced Artificial Intelligence: reprezentácia znalostí, stochastické chápanie, prehľadávacie procedúry, vymedzovacie problémy
- Advanced Topics in Machine Learning: informované učenie (zovšeobecňovacie a diskriminatívne), neinformované učenie (k-ty najbližší sused, klastrovanie, redukcia dimenzií) a robotické učenie (Kalmanov filter, zosilnenie učenia)

5.4 Princeton university

Na Princeton University sa na oddelení Kvantitatívnej a výpočtovej biológie dá študovať PhD. (Ph.D.). Z matematicko-informatického pohľadu sa podľa [23] ponúkajú tieto predmety:

- Foundations of Machine Learning: matematické základy strojového učenia, teoretické základy strojového učenia, návrh a analýza učiacich sa algoritmov. Učenie sa z náhodných príkladov, ale aj reálnych dát (napr. v oblasti investovania), učí sa ako zlepšiť výkon slabých algoritmov, učenie s dotazovaním a support vector machines.
- Probability and Stochastic Systems: úvod do teórie pravdepodobnosti. Náhodné premenné, nezávislosť. Brownov pohyb, Markovove reťazce a Markovove procesy, Poissonov proces. Ďalej stochastické modely pre frontu, komunikačné systémy, náhodné signály a spoľahlivosť.
- Optimization under Uncertainty: kvantitatívne prístupy pre hľadanie optimálnych rozhodnutí pri neurčitosti a zložitosti. Používajú sa rozhodovacie stromy, simulácia Monte Carlo a stochastické programovanie. Predpovedanie a plánovanie systémov.
- Regression and Applied Time Series: metóda najmenších štvorcov a robustnosť, neparametrické techniky (splajny, neurónové siete). Časové série: modely stavového priestoru, trendy, klinické modely

- Statistical Design of Experiments: hlavné metódy štatistiky aplikované v inžinierskych a fyzikálnych vedách. Konštrukcia empirických modelov a ich aplikácia na predpovedanie a robenie rozhodnutí pod neistotou
- Introduction to Monte Carlo Simulation: simulácia a priamy výpočet pri analýze stochastických modelov a interpretovaní reálnych fenoménov. Generovanie diskrétnych a spojitých náhodných premenných, stochastické usporiadanie a štatistická analýza simulovaných dát, validačné techniky, nestacionárne Markovove reťazce a Markov chain Monte Carlo metódy.
- Dynamic Programming: úvod do stochastického dynamického programovania, diskrétne a spojito stavové dynamické programy, konečné a nekonečné horizonty, stacionárne a dynamické dáta

5.5 University of Copenhagen

Podľa [24] sa na univerzite v Copenhagen (Dánsko) študuje Bioinformatika ako druhý stupeň vysokoškolského štúdia. Z matematicko-informatického pohľadu sú dôležité tieto predmety:

- Machine Learning: ako jeho súčasť sa vyučuje Bayesovská inferencia, neurónové siete, support vector machines, prístup Monte Carlo
- Statistics for bioinformaticians: využitie štatistiky špeciálne v bioinformatike a v problematike molekulárnej a evolučnej biológie. Pokrýva pravdepodobnostné miery na diskrétnych a spojitých vzorkových priestoroch, binomické, multinomické a normálne rozdelenie, Gumbelovo rozdelenie. Popisné štatistické metódy: empirické miery, tabuľky, priemerná hodnota, štandardná odchýlka. Ďalej histogramy, box-ploty a QQ-ploty. Simulácia, štatistický program R, parametrizované modely a ohodnotenie maximálnej pravdepodobnosti, predpovedanie a klasifikácia. Evolučné modely, spoľahlivostné množiny a bootstrapping.

5.6 University of Oxford

Podľa [25] sa na univerzite v Oxforde ponúkajú na oddelení Bioinformatiky a systémovej biológie nasledovné moduly:

- Perl Programming for Bioinformatics: programovanie v Perle v prostredí Linuxu
- Introduction to Molecular Biology
- Statistics for Biosciences: deskriptívna štatistika, vizualizácia dát, pravdepodobnosť, podmienená pravdepodobnosť, Bayesova veta, pravdepodobnostné stromy, rozdelenia a náhodné premenné, testovanie hypotéz a vytváranie testov, ANOVA, F-test, lineárne modely a regresia, kategorická dátová analýza, kontingenčné tabuľky, Kruskall-Wallisov test, Wilcoxonov test, Chi-kvadrát test
- The Power of Bioinformatics in Modern Research: databázy, sekvenčné analýzy, analýza microarrays, vyhľadávanie génov, na „alignment“ založené sekvenčné metódy
- Algorithm Design and Complexity
- Microarray Bioinformatics
- Statistical Data Mining: data mining, vizualizácia dát a rozoznávanie vzorov, viacrozmerové škálovanie, Kohonenove samoorganizujúce sa mapy, klastrová analýza, lineárne a nelineárne (neurónové siete) klasifikačné metódy, expertné systémy
- Structural Bioinformatics: vizualizovanie makromolekulových štruktúr (röntgenová kryštalografia, NMR spektroskopia, elektrónový mikroskop), interpretovanie štrukturálnych dát a predpovedanie proteínovej štruktúry

- Systems Biology
- Database Management Systems: UML, dotazový jazyk SQL, použitie XML
- Bioethics
- High Throughput Experimental Techniques: rôznorodé objavovanie pri DNA sekvenovaní, identifikácia SNP, genotypové dáta, expresné profily a funkčná genomika
- Molecular Evolution and Comparative Genomics
- The Experimental Interface: návšteva európskych konferencií o bioinformatike
- Advanced Functional Genomics: pokročilá analýza microarray dát použitím štatistického data miningu

5.7 ETH Zürich

Na ETH Zürich sa podľa [26] vyučuje na druhom stupni vysokoškolského štúdia a zároveň sa dá získať titul z Výpočtovej biológie a Bioinformatiky. Vyučujú sa aj tieto predmety:

- Computational Statistics: viacnásobná regresia, neparametrické metódy pre regresiu a klasifikáciu (splajny, klasifikačné stromy, aditívne modely, neurónové siete a ďalšie prístupy). Interpretovanie problémov, spoľahlivé predpovedanie.
- Introduction to Database Systems: modelovanie dát (UML diagramy), relačný dátový model, normálové formy, SQL, integrita databáz, bezpečnosť, transakcie a dátové sklady
- Grafy a algoritmy: blokový rozklad a kompozícia grafov, párovanie bipartitných grafov, Hamiltonovské cykly, planárne grafy, farbenie grafov, extremálna teória grafov
- Probabilistic Modeling in Molecular Evolution: Bayesovská inferencia, ohodnotenie maximálnej pravdepodobnosti, Markovove modely, fylogenetická rekonštrukcia, modelovanie heterogénnej evolúcie, algoritmy MCMC, simulovanie evolúcie
- Computational Biology: matematické modely evolúcie, zoradenie sekvencií proteínov a DNA, významy tohto zoradenia, vytváranie fylogenetických stromov, predpovedanie druhotnej štruktúry, molekulárna dynamika
- Computational Systems Biology: prístup teórie grafov na odhaľovanie sieťovej organizácie, pravdepodobnostné (Bayesovské) siete, štruktúrna sieťová analýza založená na reakčnej „stoichiometries“, kvalitatívne metódy na dynamické modelovanie a simuláciu, modelovanie využívajúce diferenciálne rovnice a stochastické simulačné metódy
- Evolutionary Dynamics: evolučná teória hier, stochastické modely pre konečné populácie, evolučná teória grafov, evolučná dynamika rakoviny, rýchlosť adaptívnej evolúcie
- Introduction to Machine Learning: Bayesovská rozhodovacia teória a ohodnotenie maximálnej pravdepodobnosti, cross validácia, bootstrap a test Jackknife, testovanie hypotéz, klasifikačné techniky (perceptrón, support vector machines), vyhodnotenie hustoty, učenie bez učiteľa (Skryté Markovove modely), posilnené učenie, redukčné techniky
- Statistical Methods for the Analysis of Gene Expression Data: preprocesorové spracovanie dát, skúmanie dát (klastrovanie, analýza hlavných komponentov), testovanie hypotéz, klasifikácia (najbližší sused, support vector machines) a funkčná analýza

5.8 Imperial College London

Ponúka sa druhostupňové štúdium v Bioinformatike a teoretickej systémovej biológii [27]. Študenti si musia osvojiť aj nasledovné vedomosti:

- študenti si v rámci genomiky osvojujú fyzikálne mapovanie, projekt ľudského genómu, sekvencovanie, proteínové a DNA databázy
- pravdepodobnostnú teóriu, Bayesovské a frekvenčné metódy, popisnú štatistiku veľkých dátových množín, axiómy pravdepodobnosti a jej interpretovanie, diskrétne a spojité náhodné premenné, infereciu
- programovacie jazyky, dizajn programov a relačné databázy
- sekvenčná analýza DNA, algoritmy na zoskupovanie DNA, proteínová a DNA homológia a jej využitie, identifikácia a zobrazenie proteínových rodín, fylogenetická analýza proteínových sekvencií a konzervovanie zbytkov
- funkčná genomika: eukaryotická a mikrobiálna genetika, génové funkcie, signálové siete, transkripčné metódy, metódy a analýza proteínovej štruktúry
- štatistická genetika: genetická epidemológia, segregačná analýza „and path analysis“, parametrická a neparametrická „linkage analysis“, QTL analýza, „linkage disequilibrium analyses“, fylogenéza a „cladistics“

5.9 University College London

V ponuke je predmet IS in Bioinformatics s nasledovným obsahom:

biologické databázy (CATH databáza proteínovej štruktúry), identifikácia génov (neurónové siete a skryté Markovove modely), metódy na odvodzovanie vzťahov medzi génmi a proteínmi (dynamické programovanie, hierarchické klastrovanie, skryté Markovove modely), metódy na predpovedanie sekundárnej a terciárnej proteínovej štruktúry (neurónové siete, support vector machines, genetické algoritmy, stochastická optimalizácia), metódy na testovanie génovej expresie a microarray (testovanie hypotéz, klastrovanie, support virtual machines), genómová analýza využívajúca inteligentných softvérových agentov a matematické modelovanie biologických systémov

5.10 Záver

Uviedli sme všeobecný aj konkrétny matematicko-informatický aparát, ktorý už nachádza, respektíve môže nájsť ďalšie uplatnenie vo výskume v oblasti bioinformatiky, respektíve výpočtovej biológie. Ťažiskom je práve nasadenie výpočtovej techniky na jednotlivé biologické úlohy za využitia metód machine learning, pričom pri rozsiahlejších projektoch je poukázané na vhodnosť vývoja špecializovaného softvérového nástroja. Prehľad stavu výučby na niekoľkých najvýznamnejších svetových univerzitách potvrdzuje aktuálnosť, uplatnenie a význam týchto metód pre popredné výskumné pracoviská na svete.

6. Použité zdroje v kapitolách 4 a 5

[1] http://www.bhu.ac.in/mmv/bioinfo.pdf

[2] Baldi P., Brunak S. Bioinformatics – The Machine learning approach. The MIT Press, 2001.

[3] Durbin R., Eddy S., Krogh A., Mitchison G. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press, 1998.

[4] Kuttera O., Shamsb R., Navab N.Visualization and GPU-accelerated simulation of medical ultrasound from CT images. In Computer Methods and Programs in Biomedicine. 94 (3), pp. 250-266, 2009.

[5] Rößler F., Tejada E., Fangmeier T., Ertl T., Knauff M. GPU-based Multi-Volume Rendering for the Visualization of Functional Brain Images. In Proceedings of SimVis 2006, pp. 305-318, 2006.

[6] Bishop, C. Pattern Analysis and Machine Learning. Springer-Verlag, New York, 2007.

[7] Jordan M. I. (Ed.). Learning in Graphical Models, Cambridge, MA: MIT Press. 1999.

[8] http://genomics10.bu.edu/bioinformatics/kasif/bayes-net.html

[9] http://en.wikipedia.org/wiki/Support_vector_machine

[10] Bessant C., Shadforth I., Oakley D. Building bioinformatics solutions with Perl, R and MySQL. Oxford University Press, 2009.

[11] www.bixsolutions.net

[12] http://packages.debian.org/sk/squeeze/weka

[13] http://en.wikipedia.org/wiki/Weka_(machine_learning)

[14] http://en.wikipedia.org/wiki/Mathematica

[15] http://en.wikipedia.org/wiki/MATLAB

[16] http://www.mathworks.com

[17] http://wapedia.mobi/en/Bioinformatics

[18] http://www.greguide.com/comps.html

[19] http://www.topuniversities.com/university-rankings/world-university-rankings/2008/regional-rankings/europe

[20] http://latestuniversityranking.blogspot.com/2008/05/graduate-school-rankings-2009-computer.html

[21] http://www.helix.stanford.edu/peoplepaltman/bioinformatics.html

[22] http://courses.cuinfo.cornell.edu/CoScourses.php?college=ALS&dept=Biometry+and+Statistics

[23] http://www.princeton.edu/qcbgrad/courses/

[24] http://www.binf.ku.dk/programmes/Master/

[25] http://bioinfomsc.stats.ox.ac.uk/courses/index.html

[26] http://www.cbb.ethz.ch/

[27] http://www3.imperial.ac.uk/lifesciences/postgraduate/courselist/bioinformatics

[28] Fielding A. H. Cluster and Classification Techniques for the Biosciences. Cambridge University Press, 2007.

Príloha 1

Táto príloha čerpá z [28], ozrejmuje niektoré dôležité pojmy a zameriava sa na použitie klastrovacích a klasifikačných techník v bioinformatike.

Vysvetlenie niektorých pojmov

Klasifikácia
Je umiestňovanie objektov do preddefinovaných skupín. Tento proces sa často nazýva rozoznávanie vzorov (pattern recognition) a je časťou strojového učenia.

Klastrovanie
Je podobné ako klasifikácia, obe techniky triedia objekty do skupín, alebo tried. Hlavný rozdiel je, že pri klastrovej analýze triedy nie sú preddefinované.

Učenie
Učenie z pohľadu machine learning pozná dva prístupy: učenie s učiteľom (informované učenie, supervised learning) a učenie bez učiteľa (unsupervised learning).

Učenie s učiteľom: toto učenie je iteratívnym procesom, pričom pri každom iterovanom kroku sa stále viac znižuje chyba - používa sa pri klasifikácii. Učí sa na trénovacích príkladoch, o ktorých už dopredu vieme ich zaradenie do tried a tak môžeme porovnávať a následne zlepšovať spôsob klasifikácie jednotlivých objektov.

Učenie bez učiteľa: nie je prítomný učiteľ a klasifikátor si sám vytvára triedy - skupiny klastrov. Najväčším prínosom tohto prístupu je pre prostredia s nedostatkom znalostí, lebo sa môže použiť ako prieskumná analýza. Často je potrebné následné dospracovanie, či vyhodnotenie používateľom. V tomto zmysle ide skôr o generovanie a nie testovanie hypotéz.

Machine learning
Zahŕňa veľké množstvo myšlienok a metód z matematiky, či štatistiky a využíva algoritmy, ktoré sa dokážu učiť z dát. Učenie je zväčša založené na hľadaní vzťahov medzi znakmi a tiež na označenie tried. Z prístupu logiky sa používa indukcia a dedukcia.

Indukčný prístup môže byť napríklad použitý na generovanie klasifikačných pravidiel z príkladov pomocou vytváranie príčinno-závislostných vzťahov. Indukčný prítup je pri strojovom učení najbežnejší.

Pri dedukcii sa skôr priamo pracuje s chápaním, využívajúc existujúce znalosti. Môže sa použiť na predpovedanie následkov udalostí, alebo na odhalenie nutných podmienok k nastaniu pozorovanej udalosti. Tieto prístupy sa špeciálne osobitne uplatňujú pri vytváraní expertných systémov.

Vyhodnotenie pomocou klastrových a klasifikačných techník

Prieskumná dátová analýza
Pred hlbšou dátovou analýzou je dobré skúmať integritu a zistiť niečo o distribúcii dát. To je cieľom prvotnej dátovej analýzy, medzi ktorej kľúčové prvky môžeme radiť:

- získanie vhľadu do dát
- odhalenie základnej štruktúry
- získanie dôležitých premenných
- vyhľadanie dátových anomálii

- preverenie základných predpokladov
- vyvinutie jednoduchého modelu a optimálne navolenie zložiek

Základom je získať jednoduché číselné, alebo ešte lepšie grafické zhrnutia dát. Začína sa od analýzy jednej premennej a často sa pristúpi aj k multivariačnej analýze. Zložitejšie techniky pomocou lineárneho i nelineárneho mapovania redukujú dimenziu dát.

Metódy klastrovej analýzy
Všeobecne pri klastrovaní vznikajú skupiny, pričom ich členovia sú si podobnejší navzájom, než k členom v iných skupinách.

Rozdeľovacie metódy
Používajú sa na problém rozdelenia *n* špeciálnych prípadov, opísaných *p* premennými na malý počet *(k)* diskrétnych tried. Bežne sa používajú tieto algoritmy:

- k-priemer
- k-medián, PAM (partition around medoids)
- SOM (samoorganizujúce sa mapy)
- Mixture modely – zväčša sú založené na algoritme Expectation Maximization

Hierarchické metódy
Netreba pri prvotnej špecifikácii špecifikovať počet tried *(k)* a ďalej umožňuje ľahkú vizualizáciu. Sú použiteľné na vytváranie veľkých taxonómií (samozrejme podľa očakávaní časť klastrov je biologicky irelevantných, ale za použitia indexu spoľahlivosti môže byť až 78% prípadov klastrov pomenovaných s veľkou mierou spoľahlivosti).

Metódy klasifikácie
Medzi hlavné metódy patria:

- Naivný Bayesovský klasifikátor
- Diskriminačná analýza
- Logická regresia
- GAM (zovšeobecnené aditívne modely)
- Rozhodovacie stromy
- Support vector machines
- Neurónové siete
- Genetické algoritmy

Príloha 2

Prezentácia softvéru na „počítanie“ expresných profilov, vyvinutého firmou ADINIS s.r.o. Softvér umožňuje porovnávať expresné profily, ktoré boli získané pri odlišných experimentálnych podmienkach. Ak niektoré body v expresných profiloch neboli namerané, softvér ich dopočíta pomocou EM algoritmu. Pomocou programu je možné identifikovať páry expresných profilov, ktoré sú odlišné.

Počítanie expresných profilov

G.L.C.

ADINIS s.r.o.

Dopĺňanie chýbajúcich hodnôt :

Použitý je Gaussian Mixture Model (GMM), ktorého parametre sú trénované pomocou Expectation Maximization (EM) algoritmu.

- http://www.autonlab.org/tutorials/gmm.html
- http://en.wikipedia.org/wiki/Mixture_model

Chýbajúce hodnoty je potom možné doplniť z natrénovanej pravdepodobnostnej distribúcie.

Štatistické testy

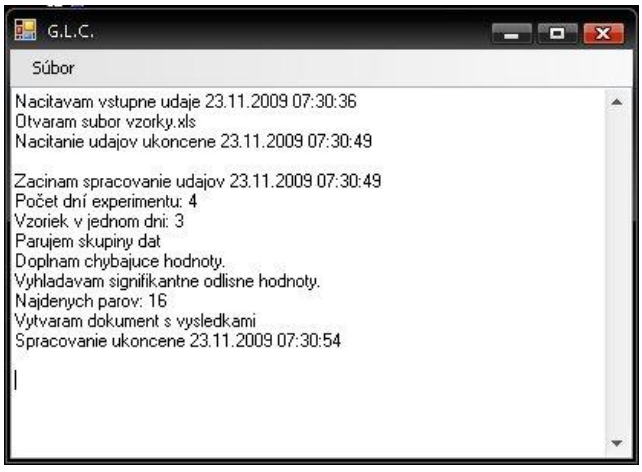
Použitý je Grubbs test na detekciu outlierov

- http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm
- http://en.wikipedia.org/wiki/Grubbs%27_test_for_outliers

Test je ovládaný s parametrom štatistická významnosť, ktorého hodnota môže byť 0.4, 0.25, 0.20, 0.15, 0.1, 0.05, 0.025, 0.01, 0.005, 0.001.
Hodnota 0.25 prislúcha nájdeniu takých párov o ktorých môžme povedať s 75% istotou že sú výchylky, atď.
Treba si uvedomiť, že reálne dáta sú väčšinou veľmi netypické a nesprávajú sa podľa vzorových prípadov, tak test napr. nemusí objaviť žiadne výchylky.

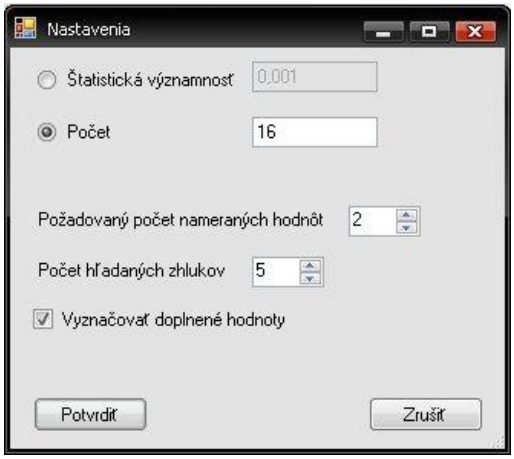
Druhý test je zoradenie párov podľa hodnoty ich vzdialenosti, kde používateľ môže voliť počet N najviac vzdialených párov, ktoré sú označené ako výchylky.

Hlavné okno programu



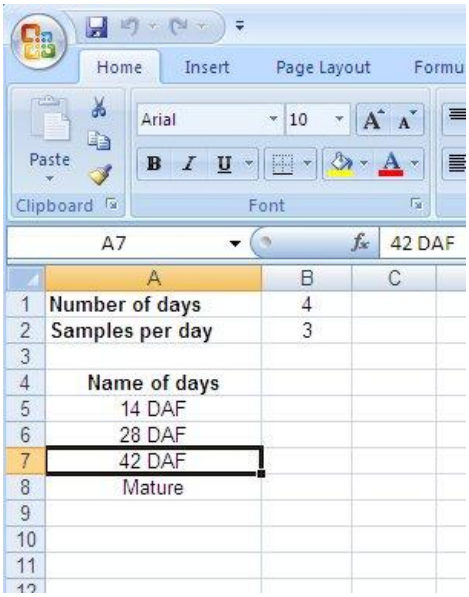
Spracovanie vstupných údajov z excelovského súboru.

Možnosti nastavenia programu

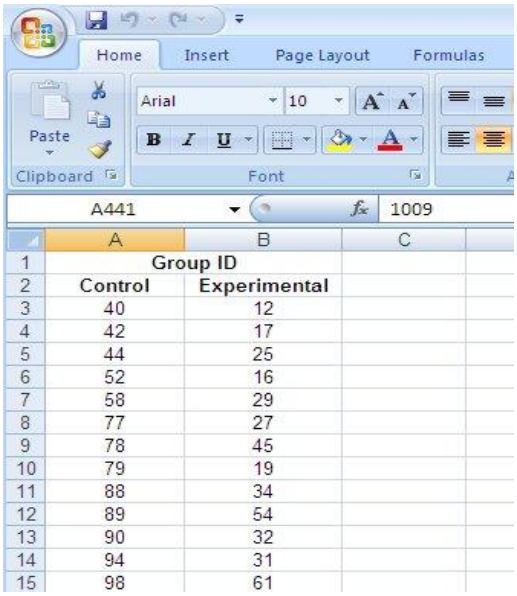


Obrazovka s nastaveniami programu.

Vstupný súbor vo forme Excelovského dokumentu

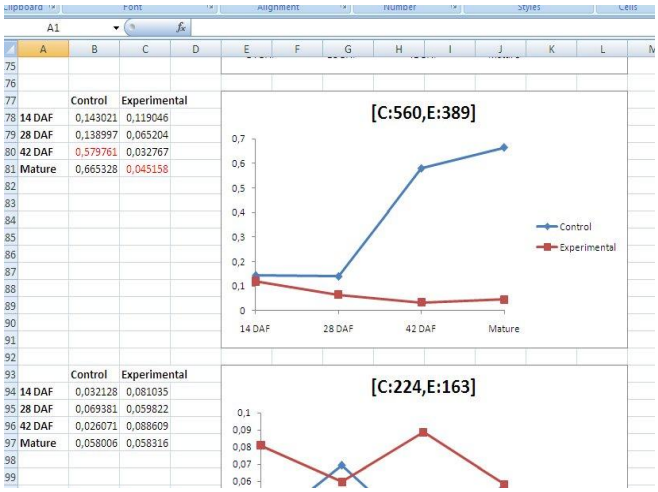


	A	B	C
1	Number of days	4	
2	Samples per day	3	
3			
4	Name of days		
5	14 DAF		
6	28 DAF		
7	42 DAF		
8	Mature		
9			
10			
11			

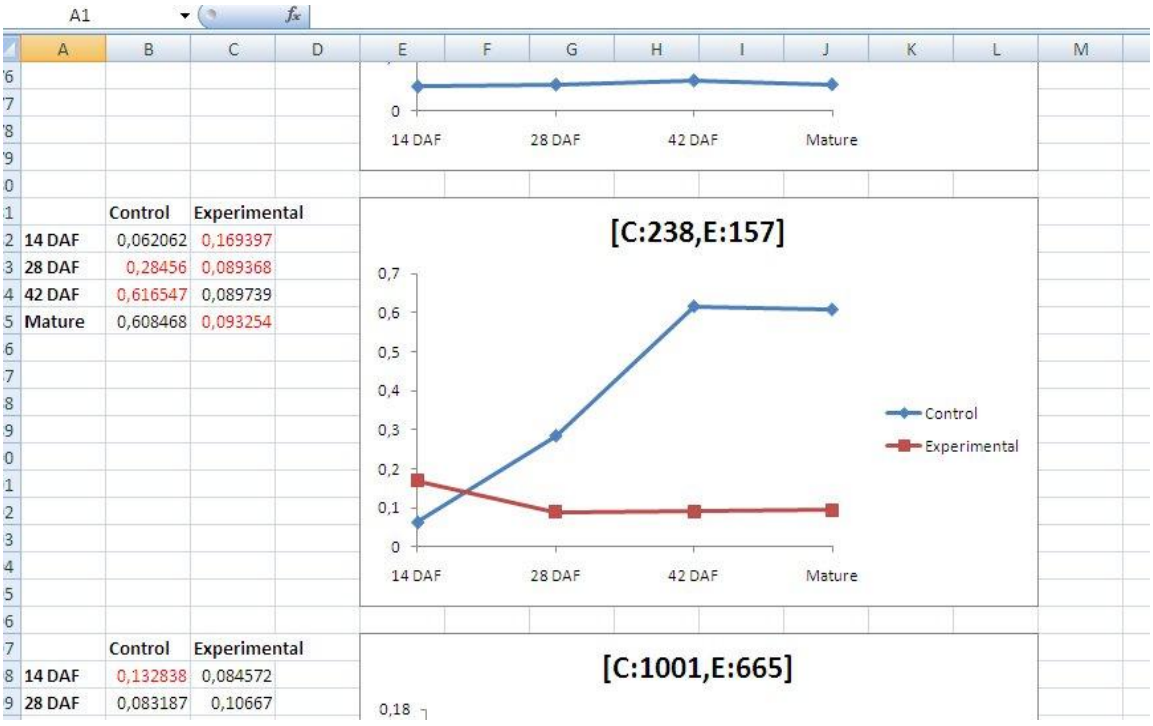


	A	B	C
1	Group ID		
2	Control	Experimental	
3	40	12	
4	42	17	
5	44	25	
6	52	16	
7	58	29	
8	77	27	
9	78	45	
10	79	19	
11	88	34	
12	89	54	
13	90	32	
14	94	31	
15	98	61	

Generovanie výstupu do Excelu



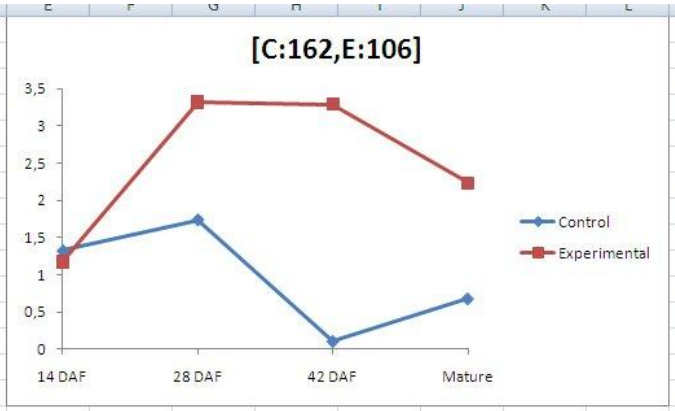
Výstup z programu ide priamo do Excelu vo forme číselných údajov a grafu. Hodnoty zvýraznené červenou farbou sú dopočítané pomocou EM algoritmu.



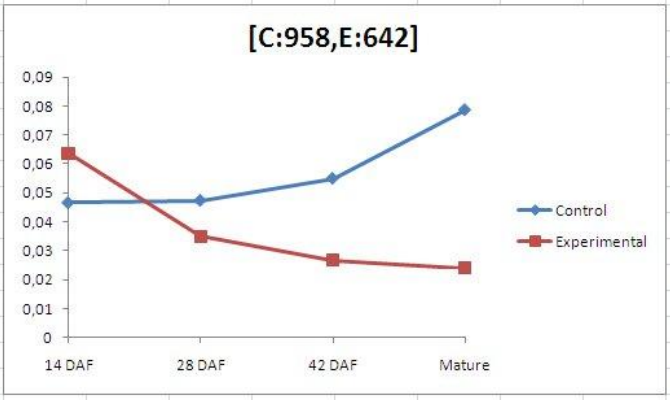
	Control	Experimental
14 DAF	0,062062	0,169397
28 DAF	0,28456	0,089368
42 DAF	0,616547	0,089739
Mature	0,608468	0,093254

	Control	Experimental
14 DAF	0,132838	0,084572
28 DAF	0,083187	0,10667

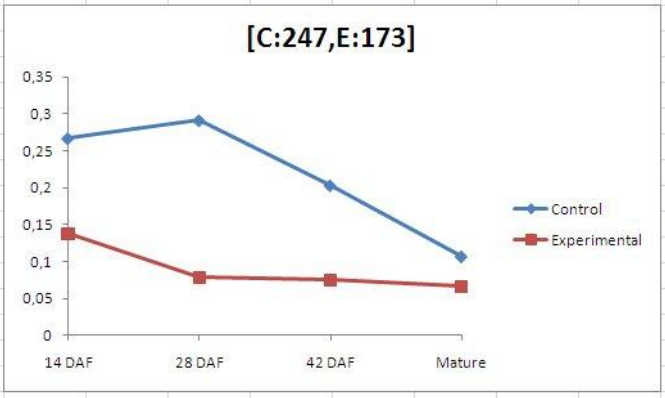
	Control	Experimental
14 DAF	1,321077	1,174816
28 DAF	1,737114	3,316404
42 DAF	0,107934	3,29036
Mature	0,678398	2,230373



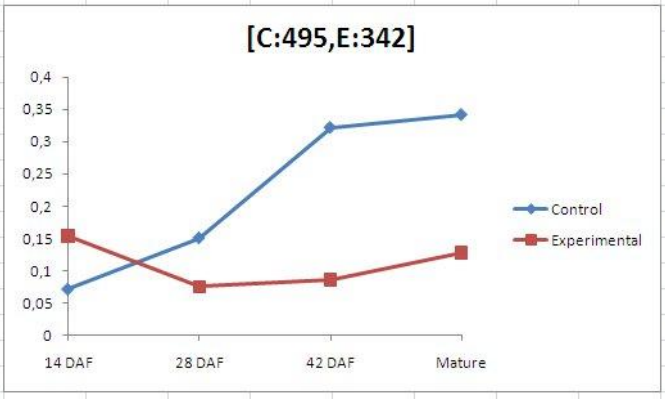
	Control	Experimental
14 DAF	0,046569	0,063621
28 DAF	0,047385	0,034889
42 DAF	0,054955	0,02675
Mature	0,078582	0,024039



	Control	Experimental
14 DAF	0,266588	0,137762
28 DAF	0,290876	0,078541
42 DAF	0,202898	0,074738
Mature	0,106655	0,066232

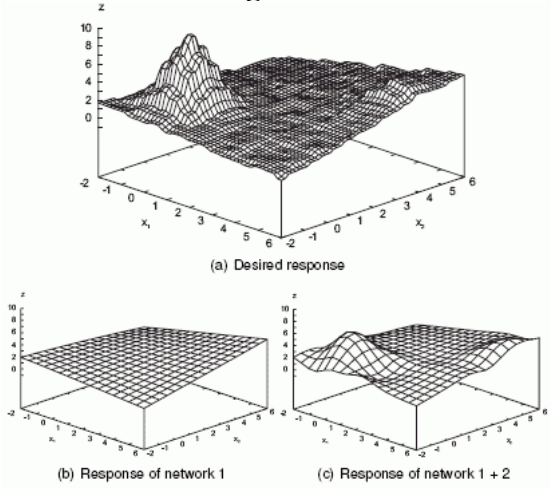


	Control	Experimental
14 DAF	0,071205	0,154044
28 DAF	0,150868	0,075656
42 DAF	0,32171	0,086203
Mature	0,341676	0,127866



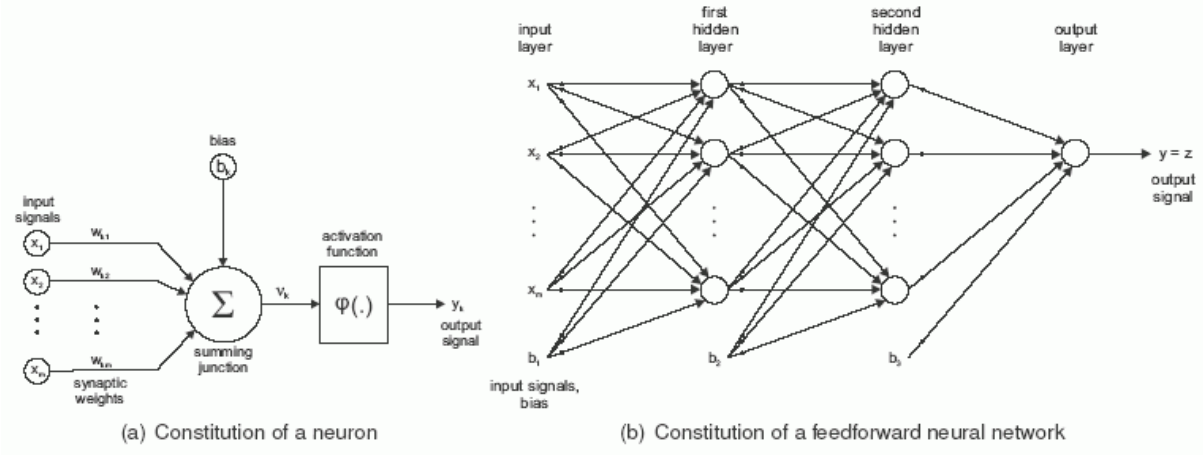
Obrázková príloha

Neurónová sieť – grafické znázornenie učenia sa



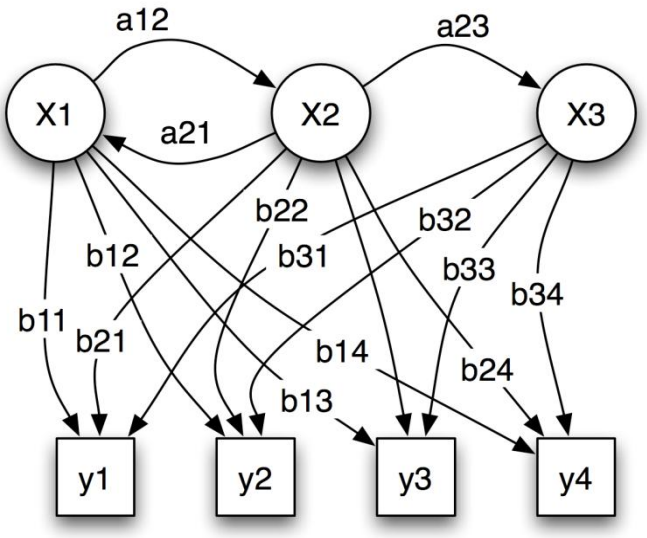
Znázornenie postupne: (a) požadovanej odozvy – citlivosti na vstupy neurónovej siete, (b) nastavenie odozvy jednej neurónovej siete, (c) odozva vhodného spojenia dvoch neurónových sietí

Architektúra neurónu a neurónovej siete



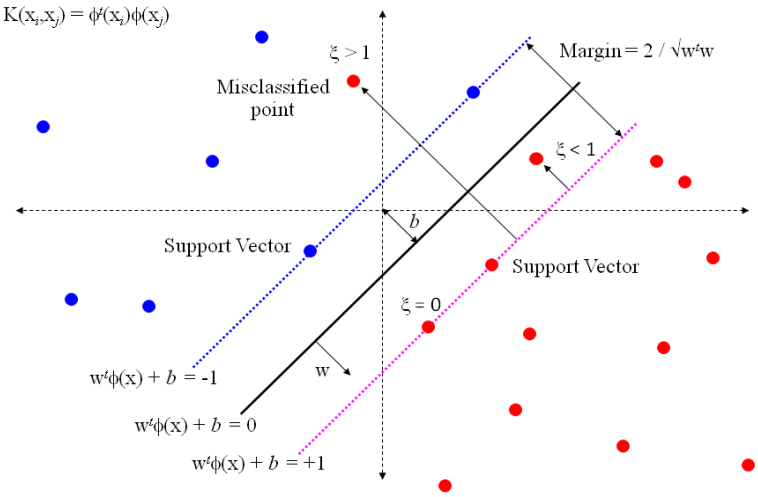
Znázornenie: (a) stavby umelého neurónu – vstupy sú ohodnotené váhami a ak ich súčet prekročí prah, pomocou vhodne zvolenej aktivačnej funkcie získame výstupný signál,
(b) stavba doprednej umelej neurónovej siete – signál sa šíri dopredne z každého neurónu nižšej vrstvy, na každý neurón vrstvy nasledujúcej

Architektúra skrytého Markovového modelu (HMM)



HMM je štatistický model, kde je modelovaný systém považovaný za Markovov proces s neznámymi parametrami a výzvou je nájsť skryté parametre z pozorovaných parametrov. Na obrázku sú znázornené pravdepodobnostné parametre HMM, kde *x* označuje stavy, *y* sú možné pozorovania, *a* sú pravdepodobnosti presúvania sa medzi stavmi a *b* sú pravdepodobnosti prechodu na výstup.

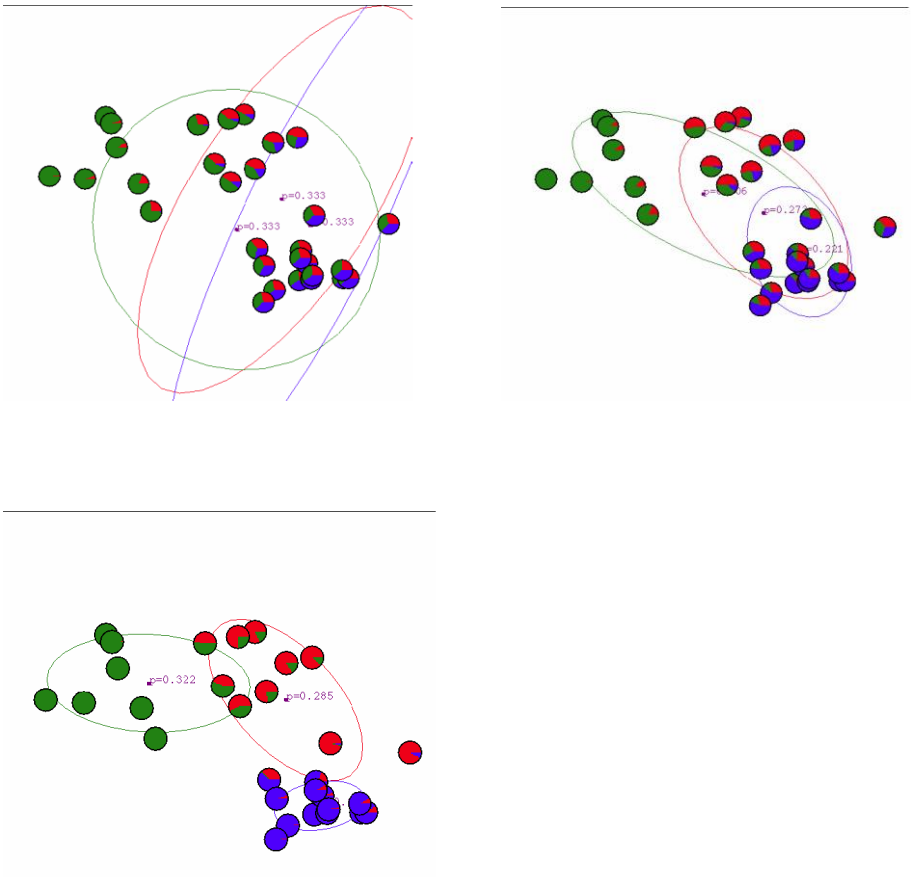
Support vector machine (SVM)



Cieľom SVM algoritmu je oddeliť dve triedy označených dátových bodov v p-rozmernom Euklidovskom priestore.

Mnoho dichotomických experimentálnych dát je nemožné priamo oddeliť pomocou hyperroviny a SVM algoritmus najprv dáta pretransformuje do iného, vysoko rozmerného priestoru, kde je možné pomocou inej hyperroviny dve triedy bodov rozlíšiť jednoduchšie. Príklady bodov na okrajoch separujúcej hyperroviny sa nazývajú podporné vektory (support vectors), pretože oddeľujúca hyperrovina je definovaná práve pomocou týchto podporných vektorov.

Expectation maximization – príklad postupného klasifikovania



Každý bod predstavuje pravdepodobnosť ohodnotenia daného parametra - môže patriť do jedného z troch klastrov (vyznačených ako modrý, zelený a červený) s istou pravdepodobnosťou vyznačenou koláčovým grafom. Pri procese učenia sa postupne zvyšuje pravdepodobnosť príslušnosti zaradenie do spoločného klastra pre príbuzné body.

Zdroj obrázkov: internet